Using a Computational Cognitive Model to Understand Phishing Classification Decisions

Matthew Shonman
U.S. Department of Homeland Security
mshonma1@alumni.jh.edu

Xiaoyu Shi
Johns Hopkins University
Information Security Institute
xshi23@alumni.jh.edu

Mingqing Kang
Johns Hopkins University
Information Security Institute
mkang31@jhu.edu

Zuo Wang
Johns Hopkins University
Information Security Institute
mwang2@atlassian.com

Xiangyang Li
Johns Hopkins University
Information Security Institute
xyli@jhu.edu

Anton Dahbura
Johns Hopkins University
Information Security Institute
antondahbura@jhu.edu

Numerous studies of human user behaviours in cybersecurity tasks have used traditional research methods, such as self-reported surveys or empirical experiments, to identify relationships between various factors of interest and user security performance. This work takes a different approach, applying computational cognitive modelling to research the decision-making of cybersecurity users. The model described here relies on cognitive memory chunk activation to analytically simulate the decision-making process of a user classifying legitimate and phishing emails. Suspicious-seeming cues in each email are processed by examining similar, past classifications in long-term memory. We manipulate five parameters (Suspicion Threshold; Maximum Cues Processed; Weight of Similarity; Flawed Perception Level; Legitimate-to-Phishing Email Ratio in long-term memory) to examine their effects on accuracy, email processing time and decision confidence. Furthermore, we have conducted an empirical, unattended study of US participants performing the same task. Analyses on the empirical study data and simulation output, especially clustering analysis, show that these two research approaches complement each other for more insightful understanding of this phishing detection task. The analyses also demonstrate several limitations of this computational model that cannot easily capture certain user types and phishing detection strategies, calling for a more dynamic and sophisticated model construction.

Keywords: Phishing, Security behaviour, Cognitive model, Simulation, User study

1. INTRODUCTION

The modern digital society has heralded a growing need to overcome persistent information security challenges, particularly those facing human users. Social engineering and phishing remain among the most serious dangers to all computer users (Vergelis et al. 2019), with a significant portion of data breaches and security incidents stemming from the theft of digital credentials after email users click on a malicious phishing link that appears legitimate (Verizon 2021).

Researchers have sought to identify the various environmental and user-specific factors contributing to such threats, utilizing empirical experiments and self-reporting by both everyday users and security experts. For example, several empirical studies identify user personality traits and informational cues in legitimate and suspicious emails to quantify their impact on user performance (Molinaro and Bolton 2018; Veksler and Buchler 2016; Vishwanath et al. 2016).

Computational cognitive modeling may reveal additional insight into the mental challenge of identifying phishing emails. This practice "imputes computational processes… onto cognitive functions," producing algorithmic and analytic descriptions of specific psychological mechanisms that can be simulated through a computational and accrual model (Sun 2008). Such models may be contrasted against "product theories," which describe the mental inputs and outputs that produce behaviors but do not represent internal psychological processes. Cognitive architectures, such as ACT-R and Soar (Anderson 1996; Laird 2012), provide frameworks to develop specific cognitive models. In addition to predicting potential issues such as errors in decision making or delays in reaching a task goal, computational cognitive modeling sheds light into plausible causes, based on emerging cognitive conditions, to provide guidance toward an effective remedy.

This paper builds upon our team's two recent efforts: a simulation study using computational cognitive modeling to examine cybersecurity decision-making

(Shonman et al. 2018) and a recent empirical study of users classifying emails as legitimate or phishing (Zhang et al. 2018). The current work offers two contributions to this ongoing investigation:

- We refine our original ACT-R-based cognitive model of phishing detection (Shonman et al. 2018) by adding two model parameters to the original three, which lend greater complexity to our representations of both user perception of a suspect email and a user's past experience with phishing and legitimate emails. We have conducted new simulations to observe this model's behavior on a range of input values.
- We compare the simulation results with data from the previous empirical study to assess our model's validity. Multiple simulation results align with existing insights and best practices for phishing email identification, pointing to the utility of this modeling method.

Our code and collected user data are available at https://behavior.isi.jhu.edu.

2. LITERATURE REVIEW

Veksler et al. (2018) explored potential uses of cognitive modelling 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. Veksler and Buchler (2016) presented 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 understanding and responding to cyber attackers.

The computational cognitive model described by Dutt et al. (2013) strongly influenced our work. They use instance-based learning theory to simulate the behaviour of a security analyst in determining whether a series of network events constitutes a cyberattack. The model represents situation information as a series of attributes denoting 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 past similar experiences, stored as individual "chunks" in memory. Per the ACT-R architecture, chunks are scored based on similarity, retrieval recency, and other factors, and the chunk scored highest is used to classify the event under consideration. For each event sequence, a counter increments for each new event judged as a threat. When the counter surpasses a set threshold, the entire sequence is classified as a cyberattack.

Our model of email sorting (in Shonman et al. 2018 and in Section 3.2 below) adapts this work to describe step-by-step (cue-by-cue) processing of a suspicious email, adding additional parameters to introduce more complexity to each email judgment.

Independently, Cranford et al. (2019) have also applied ACT-R-based cognitive modelling to the study of phishing detection. Their model differs in several aspects from ours: it processes an entire email in one step, as opposed to our model which judges individual phishing cues; it uses an ACT-R blending mechanism to calculate a "consensus value" from many similar memories, while ours retrieves the single memory fragment judged most similar to the current cue; and it uses semantic closeness between emails to compute the similarity of each memory to the current email, while ours calculates similarity as the difference between coded attributes within the current email cue and respective attributes within each memory.

3. METHODOLOGY

3.1. Phishing User Study Design and Execution

The research team went through our university's Institutional Review Board approval protocol. 177 participants from the United States were recruited through Amazon Mechanical Turk in late 2017.

3.1.1. User Study Design and Execution

Participants functioned as a personal assistant directed to classify 40 emails into either a "keep" or "suspicious" (phishing) folder. Emails appeared in a random order for each participant. After classifying each email, participants rated their confidence in that classification decision on a scale from 1-10.

Zhang et al. (2018) details the study's various experimental conditions (single-tasking vs. multitasking and incentive vs. no-incentive), which we will not discuss here. Our analysis focused on the 77 participants in single-task conditions who sorted all 40 emails within the 30-minute time limit.

Participants viewed emails in the Roundcube webmail system (https://roundcube.net) with a countdown timer displayed on the screen. Through several pilot studies that tested the protocol and parameters, we judged that the time pressure was sufficient to keep participants engaged during the full study period and to help reduce potential bias introduced by having informed them to look for phishing emails.

3.1.2. Phishing Cue and Email Design

All 40 emails were created from real emails with personally identifiable information modified. The 20 phishing emails stem from a semi-random sample of emails in Cornell University's "Phish Bowl" database (https://it.cornell.edu/phish-bowl). The 20 legitimate emails were derived from emails received by the research team and consisted of promotions, notifications (like "Final Reminder for Warranty Activation") and requests for information.

***Table 1**: Phishing Cue Definitions (Shonman et al. 2018).*

Cue Type	Cue Definition
No Branding/ Logos	*Does the email lack company branding and/or logos?*
Overall Design	*Does the overall email quality appear poor?*
Suspicious Sender Name	*Does the sender display name appear suspicious?*
Subject	*Does the subject line direct the receiver to take an action?*
Lack of Sender 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 URL text suggest a webpage different from the true link?* • *Does the URL 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	*Does the email present a too-good-to-be-true offer?*
Personal Information	*Does the email request personal information?*

We analysed 12 phishing cues defined in Molinaro and Bolton (2018) that imply whether an email is legitimate or phishing, such as whether the email requested personal information. (The simulation uses the modified 13-cue set in Table 1.) Legitimate emails may contain individual suspicious cues, such as misspellings or an absent greeting, while phishing emails may contain non-suspicious cues and thus seem legitimate. However, phishing emails on average contained more suspicious cues than did legitimate emails, providing a path to accurate classification. For example, Suspicious Sender Name appeared in 12/20 phishing emails but only 4/20 legitimate emails, and Lack of Sender Details was present in 15/20 phishing emails but only 3/20 legitimate emails.

All emails were manually coded by the research team to identify the cues present.

3.1.3. Data Collection and Performance Measures

Our analysis considered individuals' self-reported age, education level and experience with network or cybersecurity courses/certificates. We also utilized participants' self-rated confidence in each email classification decision (1: no confidence, to 10: extremely confident). Six performance measures were recorded for each participant:

i. *False negative rate [FNR]:* error rate for phishing email classifications (range [0,1]);
ii. *False positive rate [FPR]:* error rate for legitimate email classifications (range [0,1]);
iii. *Total processing time for all legitimate emails* (range approximately (0,900) seconds);
iv. *Total processing time for all phishing emails* (range approximately (0,900) seconds);
v. *Average confidence rating for legitimate email classifications* (range [0,10]);
vi. *Average confidence rating for phishing email classifications* (range [0,10]);

Measure (v) was averaged across the 20 legitimate emails, with measure (vi) averaged across phishing emails.

3.2. Phishing Detection Model and Simulation

3.2.1. Model Design

Our model (Figure 1) represents the cognitive process of an individual determining whether a series of emails are phishing or legitimate, drawing upon the single-task scenario in the empirical phishing study. In the model, the "user" classifies an email by evaluating the email's individual cues as "threat" or "non-threat." Each cue is classified by comparison to individual "chunks" in the user's simulated long-term memory (the mode of memory that retains information indefinitely, as opposed to short-term memory that holds "active" information for less than one minute). Chunks represent previously encountered cues for which the email nature (phishing/legitimate) is known.

3.2.2. Model Parameters

This study investigated how changes in experimental parameters influence the model's phishing classification performance. The model used five parameters. Parameters (i)-(iii) were examined in the previous simulation report (Shonman et al. 2018). Parameters (iv) and (v) are new to this study.

i. *Suspicion Threshold*: This term denotes the number of suspicious cues classified before the user marks an email as phishing. Values were whole numbers from 2 to 6, always less than the Maximum Cues Processed parameter value.
ii. *Maximum Cues Processed*: If the suspicion threshold is not crossed, this term denotes the highest number of cues per email that a user evaluates before deciding. Values were whole numbers from 7 to 12. The lower

```
Start of the simulation
For each of 100 users
   Populate the long-term memory with all the cue chunks derived from the 40 emails according to Legitimate-Phishing (L-P) Ratio(a);
   For each email of the 40 emails randomly ordered
      Reset Number of Processed Cues to 0;
      Reset Suspicion Level to 0;
      For every cue in the email processed either in linear order or simultaneously according to the cue type (b)
         If Number of Processed Cues <= Maximum Cues Processed AND Suspicion Level < Suspicion Threshold
            Update the activation values of all the related cue chunks in long-term memory (c);
            If the utility of the cue chunk being retrieved with the highest activation value is 1
               Increase the Suspicion Level by 1.
            Endif
         Endif
      End
      If the Suspicion Level = Suspicion Threshold
         This email is classified as phishing.
      else
         This email is classified as normal.
      Endif
   End
End
End of the simulation
```

Figure 1*: A computational model based on cognitive chunk activation revising and instance-based learning (a: see 3.2.3 Cue Chunks in Long-term Memory; b: see 3.2.4 Cue Processing and Email Classification; c: see 3.2.5 Cognitive Chunk Activation).*

bound ensured that at least one cue beyond the first six (fixed-ordered) cues would be classified; the upper bound was selected because one email only had 12 cues.

iii. <u>*Weight of Similarity*</u>: This parameter corresponds to P_l in Equation 3 (Section 3.2.5), which weights the similarity term used in the memory chunk activation equation (Equation 1). Values were whole numbers from 1 to 7. This variation allowed us to examine the similarity term's interaction with the base-level learning and noise terms. In comparing a chunk in memory to the cue currently under consideration, greater weights translate to a higher value of the similarity term within the activation equation.

iv. <u>*Flawed Perception Level*</u>: This value determines the probability that the "user" correctly classifies any information cue in the current email being processed. Higher values raise the likelihood that the user will code the information cue wrongly. The term varied from 0.0 to 0.5, incrementing by 0.1.

v. <u>*Legitimate-to-Phishing Email Ratio [L-P Ratio] in long-term memory*</u>: This value represents the experience of previously seen legitimate and phishing emails for a user. The term started with 20 legitimate and 20 phishing emails (1:1). Since this ratio for real-world emails is estimated at roughly 3000:1 (Symantec 2018), we input the ratios 10:1, 100:1, 3000:1 and 5000:1 to observe the parameter's potential influence.

3.2.3. Cue Chunks in Long-Term Memory

The simulated long-term memory was populated with chunks derived from the 40 empirical study emails. (Note that each email contained 0-13 hyperlinks, all encoded as distinct chunks.) In this way, these emails are the source of cue chunks, i.e. past knowledge facts, that are associated with legitimate and phishing emails. Considering that legitimate emails outnumber phishing emails in real-world experiences, we duplicated the cue chunks associated with the 20 legitimate emails multiple times for different L-P Ratio settings.

Memory chunks contain the following components:

- <u>*Cue type*</u>: One of the 13 different cue categories (Table 1).
- <u>*Attribute score(s)*</u>: Each is 0 if the question is answered "No," and 1 otherwise (Table 1).
- <u>*Utility*</u>: Value is 0 if the email associated with this past cue was normal; 1 for phishing.

At the beginning of each simulation run, all long-term memory chunks were entered simultaneously.

3.2.4. Cue Processing and Email Classification

The model processes an email one cue at a time. During classification of a cue, every cue chunk of that same cue type in the long-term memory receives an updated "activation" score according to a formula described below in 3.2.5. If the cue chunk with the highest score belonged to a phishing email, the current cue is classified as "threat." The model maintains a counter starting at zero 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 the Suspicion Threshold level.

Not all cues in each email were processed. One model parameter sets the maximum number of cues that can be classified per email, separate from the Suspicion Threshold. When this number is reached, the email is classified as normal if the Suspicion Threshold has not been crossed.

Information cues were visited in a manner combining fixed steps and random elements. Expert input and a pilot study suggested 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 seven cues, their order is not fixed, and the model treats these cues as processed simultaneously by the user. All memory chunks corresponding to these seven cue types are likewise pooled together; the memory chunk being activated determines which cue is processed next.

Flawed Perception Level sets the probability that the simulated user flips the attribute score of a cue upon reading it (equivalent to misinterpreting a suspicious cue as benign or vice versa).

3.2.5. Cognitive Chunk Activation

In the ACT-R cognitive architecture, declarative knowledge (i.e. facts and events) is stored as discrete "chunks" in long-term memory (Anderson 1996). Information chunks relevant to a present situation are selected according to an activation value calculation, simplified in Dutt et al. (2013) as:

$$A_i = B_i + Sim_i + \varepsilon_i \qquad (1)$$

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 information cue. ε_i is a random noise term to model imperfection in human cognition. The components of this equation are detailed below, drawn from Dutt et al. (2013).

For the *i*th memory chunk:

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

{1,…,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.

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

P_l is a weight term which we varied as one model parameter (i.e. Section 3.2.2). M_{li} represents the raw similarity score comparing the *l*th information attribute with the present situation. M_{li} was scored as 0 if the *l*th attribute value in a memory chunk matched that of the current cue under consideration, or -1 if the two values were unequal.

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

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.

3.2.6. Simulation Output

The simulation was run 100 times for each combination of parameter settings. Corresponding to the performance measures from the empirical study (Section 3.1.3), we defined six performance measures based on the simulation output:

- *False negative rate [FNR]:* equals # false negatives [FN] / (# FN + # true positives);
- *False positive rate [FPR]:* equals # false positives [FP] / (# FP + # true negatives);
- *Average processing time, negative [TN]:* average time spent assessing a legitimate email, measured as number of cues processed;
- *Average processing time, positive [TP]:* average time spent assessing a phishing email, measured as number of cues processed;
- *Confidence rating, negative [CRN]:* equals 1 – (suspicion_counter of a legitimate email) / (number of cues_checked);
- *Confidence rating, positive [CRP]:* equals 1 – (suspicion_counter of a phishing email) / (number of cues_checked).

Note that a higher CRN or CRP means a higher confidence. We conjecture that simulated users who check more cues will make a more informed decision, translating to a greater confidence rating. As with the empirical study metrics, CRN and CRP were distinguished based on the true classification of each email.

4. RESULTS

4.1. Clustering Analysis of Empirical Data

Our initial significance tests (Zhang et al. (2018)) failed to clearly characterize the participants. This suggests that the subpopulations we sought did not prominently vary along individual performance measures. Therefore, we utilized k-means clustering to examine their interactions by simultaneously considering all six performance measures (as in Section 3.1.3). We normalized the minimum and maximum bounds of all performance measures to 0 and 1. After experimenting with k values from two to eight, the most informative findings emerged for a division of three distinct subpopulations:

- An "*overachiever*" cluster with strong overall performance (*n* = 34);
- A "*conservative*" cluster featuring lower FNR and higher FPR (better at identifying phishing than legitimate emails) (*n* = 16);

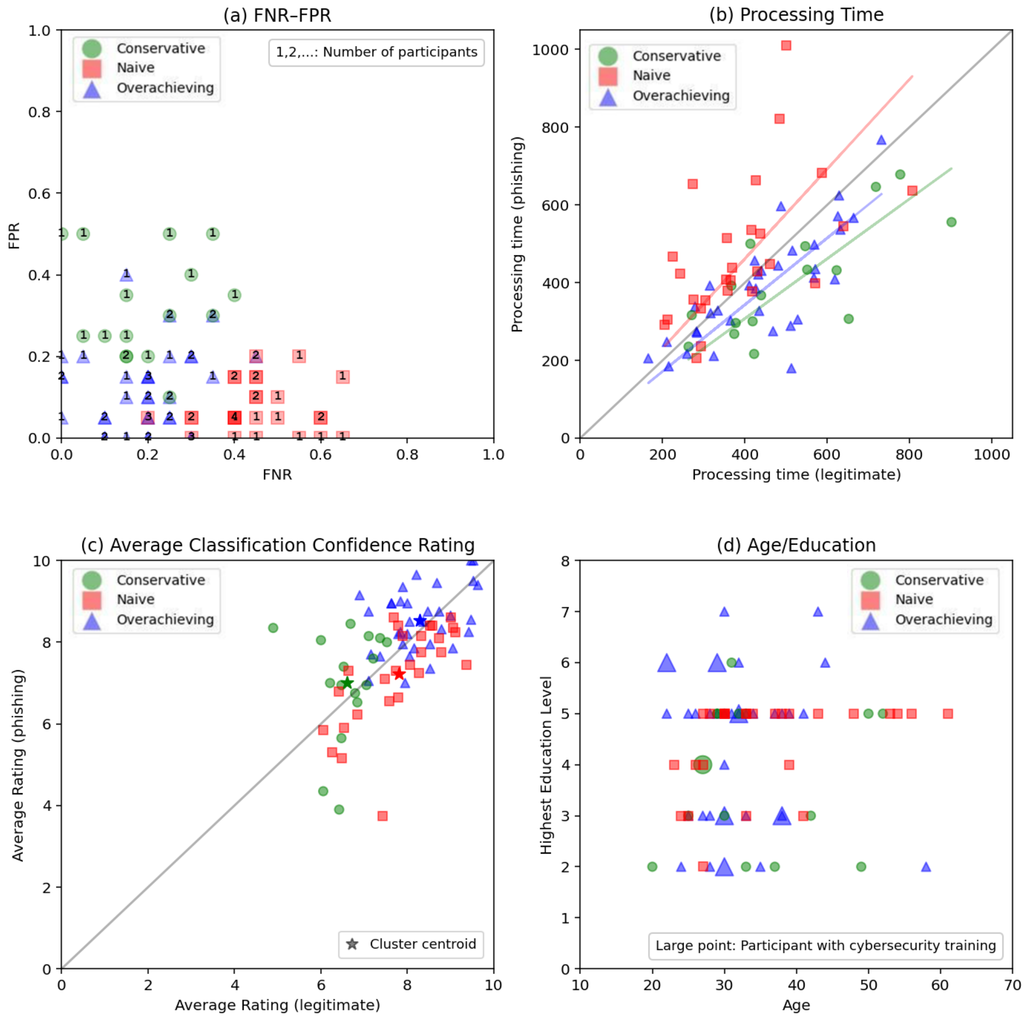

Figure 2*: Clustering of participants in the no-multitasking condition where performance measures were normalized into three user types. Education levels in subplot d were coded as follows: 1: No High School Diploma; 2: High School; 3: Some College; 4: Two-Year Associate; 5: Four-Year Bachelor; 6: Master's Degree; 7: Doctorate Degree.*

- A "*naive*" cluster featuring lower FPR and higher FNR (more accurate at identifying legitimate than phishing emails) (*n* = 27).

Figure 2 shows the clustering results as a set of two-dimensional scatter plots. In Figure 2a, displaying FNR and FPR, a numeric label denotes the number of overlapping points, i.e. participants with the same FNR and FPR values. Figure 2b compares processing times for phishing and legitimate emails, also fitting linear regression lines on each cluster. Figure 2c shows participants' average decision confidence ratings for phishing and legitimate emails. Finally, Figure 2d shows participants' age, education level and cybersecurity training along with their cluster.

As shown in Figure 2a, naive-cluster participants demonstrated comparatively high FNR, signifying less success in detecting phishing emails. Not coincidentally, as per Figure 2b, these participants also spent more time classifying phishing emails than legitimate ones. Similarly, conservative-cluster participants exhibited relatively high FPR: they experienced more difficulty classifying legitimate emails despite spending more time on these emails.

The overachiever cluster mostly includes participants with both low FNR and FPR. These participants also reported the highest confidence level among the three clusters. The corresponding linear regression line in Figure 2b indicates that these participants showed an overall slight tendency to spend less time on phishing emails. One potential explanation is that they had to examine a legitimate email more thoroughly, for example, by checking more phishing cues, before confidently moving it to the "keep" folder. However, they only needed to find "enough" suspicious evidence to correctly classify a phishing email. This seems to support a similar strategy used in the simulation study of single-task users as reported in Shonman et al. (2018).

Intuitively, higher confidence ratings would be associated with better task performance. As shown in Figure 2c, confidence ratings of different clusters generally reflected their relative success at detecting phishing, legitimate or both types of emails. However, points from different clusters are interspersed: some conservative-cluster participants were less confident on legitimate emails, and some naive-cluster participants expressed higher confidence on phishing emails. (Similarly, Figure 2b also features overlap between clusters on email processing time.) These observations, consistent with findings in our previous reports, highlight the difficulty of relying on just one or two performance criteria to characterize security behaviours, and the necessity of a comprehensive approach such as clustering.

Figure 2d highlights the potential influence of cybersecurity training experience and advanced education on phishing classification. All participants with cybersecurity training, across all education levels, lie in the overachiever cluster, as do all but one individual possessing master's or doctoral degrees. No participants older than 45 possessed a graduate degree or had cybersecurity training, and only one individual in that age group is in the overachiever cluster. These observations seem to support previous research, including Gavett et al. (2017), holding that academic study or training can effectively improve a person's security behaviour. Given the findings of Gavett et al. and Lin et al. (2019) that aging did not show direct impacts on phishing success, we conclude that the over-45 population's performance is likely better explained by their lack of training and advanced education rather than directly by their age. Additional research, with a study population including older participants who possess greater training and education, may further clarify the roles of these factors.

4.2. Clustering Analysis of Simulation Data

To obtain insights into the model's efficacy, similar clustering analysis methods as used on the empirical study data were applied to the simulation results. The k-means clustering analysis used all six performance measures (as in Section 3.2.6), each normalized on a [0, 1] scale, to identify distinct user groups (represented by particular combinations of input parameters in the simulation).

This analysis pursued two goals. First, we questioned whether our model could accurately represent the types of users apparent from the empirical study. We therefore "hunted" for user categories corresponding to the three clusters identified in the empirical study data. Second, we hoped that analysing the parameter settings associated with individual clusters would help identify the factors related to these users' unique performance characteristics. This is similar to the demographic analysis in the empirical study.

One notable difference emerged when comparing the clustering analyses of simulation and empirical results. Each instance in the simulation represents a distinct combination of model parameter values, corresponding to one "type" of user. These parameter values are varied continuously across given ranges. However, each instance in the empirical study corresponds to a single real individual, many of whom may exhibit similar traits that do not vary continuously across a spectrum. As a result, a certain empirical user type may correspond to multiple individuals in data, resulting in more apparent patterns represented by clusters. Clusters may be dense with specific user types, with sparse or no instances of other types in between.

Figure 3 shows the results for k=3, the same number of clusters identified in the empirical study results, for simulations with an L-P Ratio of 3000:1. Figure 3a displays sorting accuracy (FNR and FPR). Clustering analysis on these data produces less distinct clusters than those from the empirical data. Specifically, simulated overachiever-cluster users do not exhibit a significantly stronger performance in either FNR or FPR, compared to real users from this cluster in the empirical study.

Figure 3b highlights email processing time. Unlike the empirical results, the simulation data demonstrate clear differences between clusters for this metric; average time is longest for the overachiever cluster and shortest for the conservative cluster. According to the respective linear regression lines, all simulated users generally spend more time on legitimate emails than on phishing emails. This is reasonable given that they must examine a legitimate email more thoroughly, for example, by traversing more cues, before asserting the email to be legitimate. However, correctly classifying a phishing email only requires that enough suspicious cues are found to cross the suspicion threshold level, so simulated users understandably take less time to identify such emails.

As in Figure 3c, simulated overachiever-cluster users possess the highest confidence scores, then naive-cluster users, with the lowest scores among conservative-cluster users. Compared to conservative-cluster users, naive-cluster users examine more information cues and exhibit higher confidence scores while still tending to misclassify more phishing emails as legitimate. This behaviour suggests that greater confidence does not necessarily indicate more accurate classification decisions in the model, potentially due to simulated users simply lacking appropriate evidence to judge certain emails. By contrast, confidence ratings in the empirical study data show less distinction between the three clusters, as in Figure 2c.

Figure 3d presents box plots to summarize the model parameter settings within each cluster and

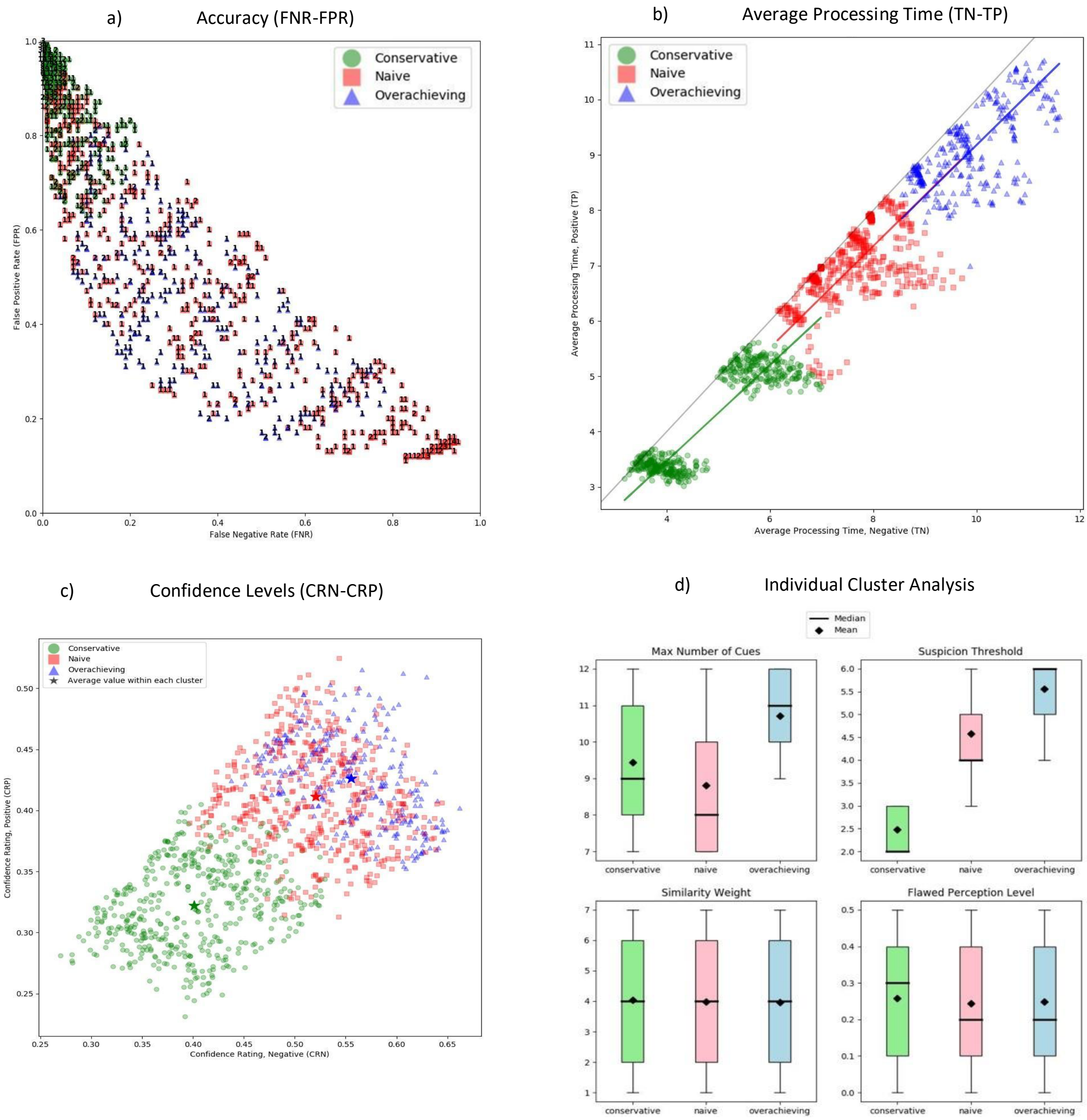

***Figure 3**: Clustering of simulation results using three clusters (L-P Ratio = 3000:1).*

highlight relevant trends. Overachiever-cluster users tend to have the highest upper bounds on number of cues examined, corresponding to the fact that these users spend more time on emails and show higher confidence in their classifications. Overachiever-cluster users also exhibit higher Suspicion Thresholds, followed by naive-cluster users. A higher Suspicion Threshold allows more cues to be examined before an email is classified, decreasing the risk that a few suspicious cues encountered initially will skew the classification decision. Flawed Perception Level lacks an obvious distinction among the three clusters. However, comparing the median and mean values, conservative-cluster users skew toward a higher value, with the other clusters skewing opposite.

Arguably, the clustering analysis does not effectively capture overachiever-cluster users, as overachiever and naive clusters are closely intertwined in several plots. This suggests that real users who perform well may employ more sophisticated phishing detection strategies. Their suspicion level may be based on correlating multiple cues together, instead of examining each cue in isolation as the model does. Moreover, real-life “email suspicion level” is presumably more dynamic than our model’s version, with the ability to decrease as well as rise.

Results for other L-P Ratios are very similar.

5. DISCUSSION

5.1. Interpretations and Suggestions

Insights from this simulation study align with successful real-world phishing identification strategies, point to further avenues for improving the model, and highlight ways in which specific combinations of input parameters can produce simulation runs aligning to salient characteristics of user populations from the empirical study.

Two observations from the simulation align with the standard recommendation that real users thoroughly examine all emails to detect phishing indicators (per Parsons et al. (2019), Vishwanath et al (2016), etc.).

- Simulated overachiever-cluster "users," the best-performing group, featured significantly higher settings for the Maximum Cues Processed parameter than did simulated users in the other two clusters. In other words, this group's greater success rate was associated with being most likely to inspect many aspects of an email for suspicious evidence.
- Simulated users from all groups tended to spend more time judging legitimate emails than phishing emails (Figure 3b). To identify a phishing email, users need only spot a few suspicious cues; however, they must traverse almost all cues to verify an email as legitimate. This behaviour is largely consistent with the empirical study results, for which the majority of participants spent more time on legitimate emails (Figure 2b).

The model found no effect on phishing detection from varying L-P Ratio (i.e. degree of past exposure to phishing emails stored in long-term memory). This might imply that such experience does not guarantee higher detection accuracy in the future, but previous empirical studies disagree: Gavett et al. (2017) found that previous phishing knowledge did influence older adults' susceptibility to phishing emails, and Singh et al (2019) concluded that higher ratios of phishing emails in training sets led to greater detection by users. An alternate explanation is that other parameters (Maximum Cues Processed and Suspicion Threshold) outweigh L-P ratio in determining model performance. Future model iterations might explore more complex ways to represent the effect of past phishing exposure.

In addition, some characteristics of simulated users using these model parameters also correspond to the demographic background of real users. For example, cybersecurity training and education are strong predictors of user performance, as shown in the empirical study. In the simulation, higher values for the Maximum Cues Processed and Suspicion Threshold parameters, which can represent greater education, positively correlate with user performance (as in Figure 3). A higher Flawed Perception Level in the simulation could potentially represent increased age, or a combination of age and lack of cybersecurity training. As previously stated, these factors were difficult to separate in the empirical study; more research would illuminate which demographic factors can be associated with this parameter.

5.2. Limitations and Future Work

While the model echoes overall trends of the three identified user types, discrepancies exist between simulated and real users. Notably, simulated "overachievers" perform more poorly than real participants in that cluster. Additionally, the average Suspicion Threshold of simulated overachievers is higher than that of simulated naive users, even though the latter group, which tend to classify inordinately many emails as legitimate, should intuitively exhibit the highest values for this parameter.

As noted in Section 4.2, one distinction between the empirical and simulation data is that the empirical results are not evenly distributed across the range of potential user "types." For example, multiple individuals might come from similar educational backgrounds, be of similar ages, and exhibit similar performance on the study task. By contrast, instances in the model are evenly spread across the full spectrum of parameter combinations. This distinction made direct quantitative comparisons between the two data sets inherently difficult, leading us to pursue qualitative comparisons in this paper. Future research might look for the subset of simulation values that match observed "types" of empirical study participants, thus permitting accurate quantitative analyses between the two datasets.

Real users presumably weight various cues differently and use additional strategies beyond cue identification to spot phishing emails. For example, poorly formatted emails might still be trusted if sent from a known source, and real-world "suspicion threshold" may shift dynamically based on factors like situational urgency. Additional research could explore how and when humans adjust their mental equivalents to the model's parameters.

Finally, this simulation uses the same set of 20 legitimate and 20 phishing emails for both testing and long-term memory construction. The lack of a "training" step in modelling and the similarity between training and testing data might sway the simulation results. Future work should utilize more representative email datasets.

6. CONCLUSION

Effectively combatting phishing threats will require further understanding of the boundaries and limitations of human cognition and security-related decision making. Computational cognitive modelling offers a promising approach to complement empirical user studies and tackle emerging hard problems in this field. This study set to identify how closely the initial model could reinforce existing real-world phishing detection strategies and the extent to which user subgroups observed in the empirical study could be replicated using parameter settings in the simulation. Future work can build upon these modelling strategies, utilizing more dynamic and sophisticated mechanisms to fully represent and capture the mental complexities that result as humans attempt to identify phishing threats.

ACKNOWLEDGMENTS

This work is supported under the National Science Foundation Award No. 1544493. We would like to thank Nathan Bos and Kylie Molinaro from Johns Hopkins University Applied Physics Laboratory for their help with the user study and data analysis. The views expressed in this paper do not necessarily reflect those of the United States Department of Homeland Security.

REFERENCES

Anderson, J.R. (1996) ACT: A Simple Theory of Complex Cognition. *American Psychologist*, 51, 4 (April 1996). 355-365.

Cranford, E.A., Lebiere, C., Rajivan, P., Aggarwal, P., and Gonzalez, C. (2019) Modeling Cognitive Dynamics in End-User Response to Phishing Emails. *Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modelling*, 35-40.

Dutt, V., Ahn, Y., and Gonzalez, C. (2013) Cyber Situation Awareness: Modeling Detection of Cyber Attacks With Instance-Based Learning Theory. *Human Factors* 55, 3 (June 2013). 605-618. https://doi.org/10.1177/0018720812464045

Gavett, B.E., Zhao, R., John, S.E., Bussell, C.A., Roberts, J.R., and Yue, C. (2017) Phishing Suspiciousness in Older and Younger Adults: The Role of Executive Functioning. *PLoS ONE* 12, 2 (February 2017). https://doi.org/10.1371/journal.pone.0171620

Laird, J.E. (2012) The Soar Cognitive Architecture. The MIT Press, Cambridge, MA.

Lin, T., Capecci, D.E., Ellis, D.M., Rocha, H.A., Dommaraju, S., Oliveira, D.S., and Ebner, N.C. (2019) Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM Transactions on Computer-Human Interaction* 26, 5, Article 32 (July 2019). https://doi.org/10.1145/3336141

Molinaro, K.A. and Bolton, M.L. (2018) Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. *Computers & Security*, 77 (August 2018). 128-137. https://doi.org/10.1016/j.cose.2018.03.012

Parsons, K., Butavicius, M., Delfabbro, P., and Lillie, M. (2019) Predicting Susceptibility to Social Influence in Phishing Emails. *International Journal of Human-Computer Studies* 128 (August 2019). 17-26. https://doi.org/10.1177/0018720816665025

Shonman, M., X. Li, H. Zhang, and A. Dahbura (2018) Simulating Phishing Email Processing with Instance-Based Learning and Cognitive Chunk Activation. Brain Informatics (BI 2018) (December 2018). Lecture Notes in Computer Science, 11309. 468-478. https://doi.org/10.1007/978-3-030-05587-5_44

Singh, K., Aggarwal, P., Rajivan, P., and Gonzalez, C. (2019) Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails. *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting*, 453-457. https://doi.org/10.1177/1071181319631355

Sun, R. (2008) Introduction to Computational Cognitive Modeling. In The Cambridge Handbook of Computational Psychology. Cambridge University Press, pages 3-19.

Symantec (2018) Internet Security Threat Report, vol. 23. Symantec Corporation. symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf (retrieved 1 March 2022).

Veksler, V.D. and Buchler, N. (2016) Know Your Enemy: Applying Cognitive Modeling in Security Domain. Presented at the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, PA, August 2016.

Veksler, V.D., Buchler, N., Hoffman, B.E., Cassenti, D.N., Sample, C., and Sugrim, S. (2018) Simulations in Cyber-Security: A Review of Cognitive Modeling of Network Attackers, Defenders, and Users. *Frontiers in Psychology* 9, Article 691 (May 2018). https://doi.org/10.3389/fpsyg.2018.00691

Vergelis, M., Shcherbakova, T., and Sidorina, T. (2019) Spam and phishing in Q1. Securelist. https://securelist.com/spam-and-phishing-in-q1-2019/90795 (retrieved 1 March 2022).

Verizon (2021) 2021 Data Breach Investigations Report. https://www.verizon.com/business/resources/reports/2021/2021-data-breach-investigations-report.pdf (retrieved 20 March 2022).

Vishwanath, A., Harrison, B., and Ng, Y.J. (2016) Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research* 45, 8 (December 2018). 1146-1166. https://doi.org/10.1177/0093650215627483

Zhang, H., S. Singh, X Li, A Dahbura, and M. Xie (2018) Multitasking and Monetary Incentive in a Realistic Phishing Study. *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI)*. https://doi.org/10.14236/ewic/HCI2018.115