Understanding Security Behavior of Real Users: Analysis of a Phishing Study

Mingqing Kang
Johns Hopkins University
mkang31@jhu.edu

Matthew Shonman
Cybersecurity and Infrastructure Security Agency
mshonma1@alumni.jh.edu

Anshul Subramanya
Johns Hopkins University
anshul.subramanya@gmail.com

Haoruo Zhang
Johns Hopkins University
harol@jhu.edu

Xiangyang Li
Johns Hopkins University
xyli@jhu.edu

Anton Dahbura
Johns Hopkins University
antondahbura@jhu.edu

Abstract

This paper presents a set of statistical analyses on an empirical study of phishing email sorting by real online users. Participants were assigned to multitasking and/or incentive conditions in unattended web-based tasks that are the most realistic in any comparable study to date. Our three stages of analyses included logistic regression models to identify individual phishing "cues" contributing to successful classifications, statistical significance tests assessing the links between participants' training experience and self-assessments of success to their actual performance, significance tests searching for significant demographic factors influencing task completion performance, and lastly k-means clustering based on a range of performance measures and utilizing participants' demographic attributes. In particular, the results indicate that multitasking and incentives create complex dynamics while demographic traits and cybersecurity training can be informative predictors of user security behavior. These findings strongly support the benefits of security training and education and advocate for customized and differentiated interventions to increase users' success of correctly identifying phishing emails.

1. Introduction

Most studies to understand users' security behavior have been conducted in controlled lab environments where participants might vary their behaviors at the presence of distracting or even intrusive factors. Apart from measuring broad performance indicators of task success based on predetermined criteria (e.g., "right" or "not very secure"), many of these studies relied on self-reported feedback from participants. Partly due to such constraints, it is hard to comprehensively examine the relationships between users' operations, performance, and demographic characteristics.

The empirical study presented in this paper tasked participants with sorting 40 emails as legitimate or phishing within a set time duration. It also introduced a multitasking requirement and a monetary incentive. Participants were recruited through the Amazon Mechanical Turk human subject pool to conduct web-based tasks remotely. Data capture techniques administered through the webmail platform Roundcube (https://roundcube.net) and the survey platform Qualtrics (https://www.qualtrics.com) collect fine details of user operations, e.g., mouse movements and clicks, continuously.

This work aims to make two contributions. First, the empirical study used real emails in a realistic task environment on computer users in an unattended fashion to capture security behaviors of high fidelity. The analyses lend insight into patterns of participants' performance on phishing detection tasks. It also showed that a user's demographic background, e.g., education, training, and age, bear correlation with performance indicators of security risk and operating time in such email sorting tasks. Second, our methodology combines a range of statistical tests with clustering analysis to explore a broad degree of useful knowledge. The analyses range from traditional significance tests, to regression modeling using phishing cues, and to k-means clustering that simultaneously considers multiple performance measures. These methods are able to identify unique subpopulations among participants that exhibited differing behaviors. The results further revealed the complexity of different users interacting with email elements and how a range of internal and external factors impact their security decisions.

2. Background

2.1. Cybersecurity user studies

Several studies have examined how users respond

to phishing emails such as the impact of using mobile devices. In one prominent effort, Vishwanath et al. developed the Suspicion, Cognition, and Automaticity Model (SCAM) to describe the cognitive, preconscious, and automatic processes contributing to phishing success [8]. The model specifically highlights interrelations between factors such as individuals' preexisting cyber-risk beliefs, suspicion, and multiple information processing modes. Two empirical tests supported SCAM predictions, finding, for instance, that greater awareness of online cyber risks led university students to process suspected emails more thoroughly, while students with less concern for such risks tended to judge emails according to simple decision rules.

In a study of similar phishing email sorting tasks [4], the authors illustrated the effectiveness of using the double system lens model, a judgment analysis technique with linear regression, to understand both how users synthesized phishing cues to make a judgment and how effective those cues were in the environment in which the judgments were made. For each participant model, the cue weights represented how the user used each cue. For the environment model, the cue weights represented how diagnostic that cue was for identifying a phishing email. Although not generalizable beyond the emails analyzed in the research, this effort identified lack of signer details, lack of branding/logos, and the presence of suspicious links as the most relevant cues in classifying an email.

Extensive research in computer security, and more generally in psychology, has utilized unattended web-based questionnaires, camera recording, and other participant self-reported data. For example, Bianchi et al. [2] used Amazon Mechanical Turk to study Android users' ability to resist GUI confusion attacks, which utilize social engineering principles similar to those in phishing attempts. They developed an emulated Android GUI remotely accessible via a web browser.

In [9], the authors performed a series of significance tests on the data of the present empirical study, finding that multitasking worsened participants' sorting accuracy and that, in general, differences between the conditions affected participants' ability to sort phishing emails, but not legitimate emails. Incentive alone, by contrast, made no difference in either multitasking or no-multitasking cases. Multitasking and incentive showed opposite effects on email processing time: multitasking reduced users' email processing time, while the incentive increased this value. However, spending more time on individual emails did not guarantee better sorting accuracy.

2.2. Analysis of user demographics

A study by Sheng et al. [5] is one of the most relevant studies on phishing susceptibility with a focus on demographic analysis. Through a large-scale roleplay survey using simulated emails, the authors found that gender and age can predict risks of falling for phishing attempts, and that educational material helped to reduce it. The results were consistent with several other studies including that by Kumaraguru et al. [3].

A number of previous studies explored the role of demographic factors of users in other cybersecurity applications. For example, Akhawe and Felt [1] studied web users' tendencies to dismiss or respond to browser security warnings. They found behavioral differences between early adopters of new browser updates and users who waited for default browser releases, attributing some of these distinctions to varying levels of technical ability among individuals. By contrast, Sunshine et al. studied user responses to Secure Socket Layer (SSL) warnings and found little effect from technical expertise [7]. In a second experiment comparing the effects of various real and designed SSL notifications, the researchers additionally found no significant differences from gender.

3. Methodology

3.1. A user study of email sorting

The research team progressed through our university's IRB review and approval process. 177 participants from the United States progressed through this study in late 2017. The analysis focused on the 146 participants who sorted all 40 emails in the given time.

Experimental Design - Participants functioned as a personal assistant directed to classify emails into either a "keep" or "suspicious" folder. As shown in Table 1, participants were randomly assigned to one of four experimental conditions. Multitasking participants answered 20 sets of questions through Qualtrics while sorting the emails, as shown in Figure 1. Each question set was presented for a maximum of two minutes; participants could manually advance to the next 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. The right side of the screen showed a countdown timer instead of the multitasking questions.

Table 1. Experimental condition and the number of participants (in parentheses)

	Incentivized	Non-incentivized
Multitasking	1 (35)	3 (34)
No-multitasking	2 (42)	4 (35)

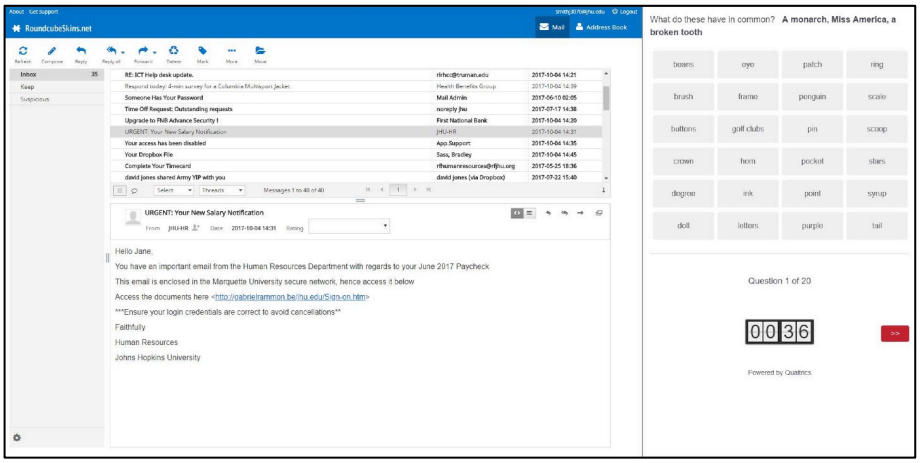

Figure 1. Multitasking condition where a participant classifies emails and answers questions concurrently

Incentivized participants could earn additional monetary compensation based on the number of correctly sorted emails. For those participants in the multitasking with incentive condition, earning extra money also depended on the number of multitasking questions answered accurately. Participants were eligible for the incentive if they correctly sorted 30/40 emails and correctly answered 15/20 multitasking questions. More information is available at http://behavior.isi.jhu.edu/.

Email Design and Phishing Cues - All 40 emails were created from real emails with personally identifiable information modified. Twenty (20) phishing emails were derived from a semi-random sample of emails in Cornell University's "Phish Bowl" database (https://it.cornell.edu/phish-bowl). The other 20 legitimate emails were derived from emails received by the research team.

We analyzed a series of 12 information or phishing cues, contained within the emails, implying whether those emails are legitimate or phishing. Crucially, legitimate emails may contain suspicious cues, such as misspellings or an absent greeting, while phishing emails may contain non-suspicious cues to seem legitimate. However, phishing emails on average contained more suspicious cues than legitimate emails, providing a path to accurate classification. All cues and definitions are available at http://behavior.isi.jhu.edu/ with some examples given below.

- *URL Hyperlinking*: are displayed hyperlinks mismatched with the underlying links?
- *Spelling and Grammar Errors*: does the text contain spelling or grammar mistakes?
- *Use of Threatening Language*: does the email threaten a negative consequence if instructions unfollowed?

3.2. Data collection

Log Files and Information Extraction - Each log file represented a single user, and every line in the file was an event. Each event record included timestamps, the user identification, the operation taken, and additional information relevant to the operation. These operations consisted of common user interactions with a webpage, including mouse clicking, hovering, scrolling, and moving over objects on the Roundcube and Qualtrics user interfaces, such as menus, buttons, links, email attachments, questions, and answer choices. Moreover, specific events were triggered when a user's mouse moved between the Roundcube and Qualtrics windows.

The log file data also enabled determination of metrics such as total email processing time. For the no-multitasking conditions, this value was defined as the time interval between a user opening (clicking on) an email and classifying (moving) that email. For the multitasking conditions, processing time excluded any period when the mouse cursor was in the Qualtrics window.

Demographics and Self-Reported Information - We collected self-reported demographics and other information on participants' experiences to aid in interpreting the experimental results. Items considered in our analysis included age, education level, and experience with network or cybersecurity courses/certificates. We also utilized participants' self-rated confidence in each email classification decision (1: not confident at all, to 10: extremely confident) and their own estimates of email sorting accuracy.

Behavioral Performance Measures - Six performance measures were directly extracted from log files for each participant (Table 2). The first pair is the

Table 2. Six user performance measures

Performance Measure	Definition
Processing time (phishing)	Measured in seconds, range of value is approximately (0,1050)
Processing time (legitimate)	Measured in seconds, range of value is approximately (0, 900)
Average rating (phishing)	Range [0,10]
Average rating (legitimate)	Range [0,10]
False negative rate (FNR)	Range [0,1]
False positive rate (FPR)	Range [0,1]

processing times for phishing and legitimate emails respectively, measured as the total time spent on all the emails in each category. The second pair is the average confidence ratings for phishing and legitimate emails respectively. The third pair is false negative rate (FNR) and false positive rate (FPR), which are the error rates for phishing and legitimate email classifications respectively.

3.3. Data analysis methodology

This work aimed to further understand our detailed empirical study results, focusing on identifying emerging subpopulations that are different in their security behaviors. As shown in Figure 2, we employed a three-stage approach toward this goal.

Cue and User Confidence Analyses - The first-stage analyses looked further into participants' performance in different condition groups. Principal component analysis (PCA) was performed to identify correlations between different email cues, ultimately reducing the full cue set to a smaller number of non-correlated cues. Next, for each participant, a logistic classifier was trained using that participant's email classification results in order to weight the significance of each cue for a given user's decisions. One-way ANOVAs, Tukey's test, and Welch's *t*-test were performed to identify whether experimental condition influenced the use of each cue. Finally, correlation tests were conducted between individuals' self-estimated and actual sorting accuracy.

Demographic Partitioning Analyses - We then concentrated on analyses that partitioned the participants according to their demographic factors of age, education, and cybersecurity training. ANOVAs were performed within each condition group to assess demographic differences in participants in terms of how they valued the various cues in an email. Additional statistical analysis was performed to determine the impact of past training on sorting accuracy. Correlations between estimated and actual performance were again examined, this time distinguishing between trained and untrained participants.

Clustering Analysis - Some of the above analyses identified statistically significant differences on individual performance measures. However, we had difficulty clearly categorizing and understanding the entire participant population through these tests. This may suggest that the subpopulations did not prominently vary along any single performance measure. After trying several clustering methods, we utilized *k*-means clustering to simultaneously consider all six performance measures.

4. Analysis and findings

4.1. Phishing cue and decision confidence analyses

We first performed PCA by evaluating correlations between the appearance of unique cues among the 40 emails. Significant correlations (Pearson correlation coefficient $p < 0.05$) were present between multiple sets of cues, and so we narrowed our scope to a subset of eight non-correlated cues from the original twelve: *Suspicious Sender Display Name, URL Hyperlinking, Poor Overall Design, Generic Greeting, Use of Time Pressure, Use of Emotional Appeals, Too Good to be True Offers,* and *Request for Personal Information*. Further analysis used only these non-correlated cues.

Next, we trained a logistic model using the email classification results for each participant. For an email classified as "phishing," y is '1'; for emails classified as "legitimate," y is '0.' Each of these emails has its own set of indicators, X, corresponding to whether each of the eight cues selected above is present. Thus, the

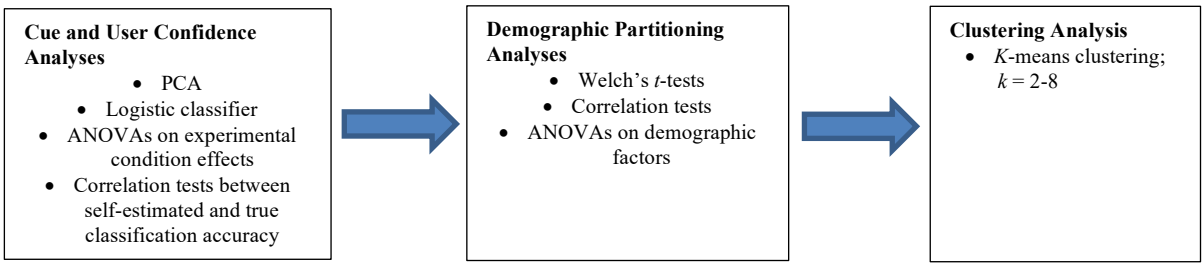

Figure 2. An analysis approach combining three stages

Table 3. Average phishing cue weights of logistic models for experimental conditions

Experimental Condition	Suspicious Sender	URL Hyperlinking	Poor Overall Design	Generic Greeting	Time Pressure	Emotional Appeal	Too Good to be True	Request for Personal Information
1. Incentivized Multitasking	0.49	-0.32	1.02	-0.04	0.36	-0.53	0.15	0.30
2. Incentivized No-multitasking	0.53	-0.18	1.00	-0.16	0.45	-0.39	0.12	0.28
3. Non-incentivized Multitasking	0.51	-0.25	1.01	-0.11	0.39	-0.35	0.18	0.34
4. Non-incentivized No-multitasking	0.59	-0.09	0.90	-0.13	0.39	-0.42	0.32	0.25

Table 4. Confidence ratings for experimental conditions

Condition	Rating	*T*-test (Between Two Conditions)
1. Incentivized Multitasking	M=7.682, SD=0.430	
2. Incentivized No-multitasking	M=7.757, SD=0.652	
3. Non-incentivized Multitasking	M=7.493, SD=0.544	3,4: t=-2.269, p=0.026
4. Non-incentivized No-multitasking	M=7.783, SD=0.596	3,4: t=-2.269, p=0.026

Table 5. Correlation between self-estimated and actual numbers of correctly sorted emails

Condition	Correlation Coefficient	*T*-test
1. Incentivized Multitasking	0.253	p=0.142
2. Incentivized No-multitasking	0.272	p=0.082
3. Non-incentivized Multitasking	0.467	p=0.005
4. Non-incentivized No-multitasking	0.336	p=0.048

independent variables per user constituted a 40 x 8 binary matrix, and our dependent variables composed a 40 x 1 binary matrix. Our analysis assumed that every user processed each cue in an email.

In this logistic model, the weights indicate each cue's importance to the participant's classification of an email. These weights were then averaged over each condition group as shown in Table 3. For example, *Poor Email Design* was highly characteristic of a phishing email in all four groups, while *Too Good to be True Offers* and *Generic Greetings* were less indicative.

Eight separate one-way ANOVAs were performed, one for each non-correlated cue, to compare the effect of experiment condition on the use of the cue. Significant differences at the $p < 0.05$ level occurred for only two cues: *URL Hyperlinking* ($p = 0.029$) and *Too Good to be True Offers* ($p = 0.031$). However, performing Tukey's test for these two cues failed to demonstrate any significant difference in weights between these condition groups.

Next, multiple *t*-tests revealed a difference in confidence ratings between conditions 3 and 4 (Table 4). When the incentive was absent, the single-task participants reported slightly higher confidence ratings than those assigned the secondary task. However, applying Bonferroni corrections for multiple comparisons revealed no significant difference in performance between experimental groups, indicating that the former group's sorting accuracy did not improve despite spending more time per email and their higher confidence ratings. This may show a challenge that participants faced in asserting their decisions, even when they focused only on email sorting.

Lastly, we calculated correlation coefficients between participants' self-estimated and actual numbers of correctly sorted emails (Table 5). Significant correlations were identified for conditions 3 and 4. We found no significant correlation for the incentive participants (conditions 1 and 2); counterintuitively, this result may represent the incentive having pressured participants into *overthinking* their email sorting actions.

4.2. Demographic analyses based on partitioning participants

Previous analyses had alerted us to the complicating effect imposed by multitasking on participants' behaviors. We first focused demographic comparisons of the no-multitasking participant groups (2 and 4). For age-based tests, participants were split into two groups (age ranges 20-38 and 39-56 for condition 2; 22-41 and 42-61 for condition 4) and then three groups (age ranges 20-33, 34-47, 48-61 for condition 2; 22-34, 35-48, 49-61 for condition 4), with divisions based on keeping group sizes equal. Welch's *t*-tests were taken comparing all of these condition groups. Similar partitions were taken for participants' highest education achieved (one of seven values ranging from no high school to doctorate) as well. However, none of these tests found significant results.

Similarly, we performed ANOVAs to analyze the role that demographic factors played on participants' perception of different email cues, including:

Table 6. Demographic analysis on phishing cue weights of logistic models

Experimental Condition	Demographic	Email Cue	Population 1	Population 1 Cue Weight	Population 2	Population 2 Cue Weight
1. Incentivized Multitasking	Student Status	Generic Greeting	Student	0.323	Non-Student	0.101
	Age	Poor Overall Design	Ages 30-39	1.276	Ages 60-70	0.377
			Ages 50-59	1.189	Ages 60-70	0.377
2. Incentivized No-multitasking	Highest Education	Generic Greeting	Some College	-0.318	Bachelor's Degree Holders	-0.021
			Bachelor's Degree	-0.021	Master's Degree Holders	-0.502
3. Non-incentivized Multitasking	Student Status	Generic Greeting	Student	0.661	Non-Student	0.133
		Request for Personal Information	Student	-0.477	Non-Student	0.367
	Highest Education	URL Hyperlinking	High School	0.569	Two-Year Associate	0.094
4. Non-incentivized No-multitasking	Student Status	Use of Emotional Appeal	Student	0.064	Non-Student	0.458

Table 7. Impact of cybersecurity training experience on sorting accuracy

Group	Accuracy	*T*-test (Between Two Groups)
1. Trained Multitasking	M=0.704, SD=0.121	1,2: t=-3.027, p=0.022
2. Trained No-multitasking	M=0.864, SD=0.048	1,2: t=-3.027, p=0.022 2,3: t=4.093, p=0.002 2,4: t=5.360, p=0.0000734
3. Untrained Multitasking	M=0.779, SD=0.093	2,3: t=4.093, p=0.002 3,4: t=2.025, p=0.045
4. Untrained No-multitasking	M=0.741, SD=0.113	2,4: t=5.360, p=0.0000734 3,4: t=2.025, p=0.045

- Student Status (Yes/No)
- Highest Education Achieved (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)
- Cybersecurity Training Experience (Yes/No)
- Age (20-29, 30-39, 40-49, ...)

Performing Tukey's test on the results found some differences among participant subpopulations, as in Table 6. For example, being a student or not was a significant differentiator in several cases including the usage of *Generic Greeting* in decision making. In another example, *Poor Overall Design* affected several age groups differently.

Furthermore, 13 participants previously completed a network engineering or cybersecurity course/certificate. For *t*-tests exploring this factor, we divided the participants into four groups based on the presence or absence of past cybersecurity training and whether those participants had faced a multitasking requirement. We chose to not divide these groups by the incentive condition, as it did not seem to affect performance.

Table 8. Correlation between self-estimated and actual numbers of correctly sorted emails based on cybersecurity training experience

Group	Correlation Coefficient	*T*-test
Trained	0.552	p=0.050
Untrained	0.322	p=0.0001

Table 7 shows the trained no-multitasking group achieved the highest email sorting accuracy. Cybersecurity training did improve the classification accuracy quite significantly when email sorting was the only task. However, the trained multitasking group did not perform better than the untrained multitasking. This shows again the challenge presented by multitasking, which complicated the effectiveness of training. Using a Bonferroni correction for multiple comparisons yielded significant differences between group 2-group 3 and group 2-group 4.

Lastly, we expanded statistical tests on the correlation coefficients between participants' self-estimated and true numbers of correctly sorted emails by looking at their cybersecurity training experience. As shown in Table 8, the 13 participants with cybersecurity knowledge achieved a higher correlation coefficient, seeming to suggest that they possessed better self-awareness of their performance than did other subjects. This conclusion would be further supported with a larger sample of such participants.

The complexity of the above statistical tests to find subpopulations among participants and the required computational cost are obvious. Any single performance measure may not be discriminating enough to efficiently find heterogenous participant subsets. The small sample sizes resulting from demographic partitioning present an additional challenge. These observations strongly influenced our decision to pursue clustering analysis, described below.

4.3. Clustering analysis

As the monetary incentive showed no significant effect on participants' performance, participants were sorted based on the multitasking condition and placed into either the multitasking or no-multitasking group. We applied the *k*-means algorithm using the six performance measures on these two groups respectively (Table 2). Furthermore, we used two different methods to normalize values of the six performance measures; i.e., the standard L2-norm function, also known as the Euclidean norm, as well as the MinMaxScaler function that normalizes the minimum and maximum bounds to 0 and 1. These two methods generated compatible results.

We experimented with different *k* values, from two to eight, and the most informative findings emerged for a division of *three* distinct subpopulations:

- An "*overachiever*" cluster with strong overall performance, shown by blue circles in plots;
- A "*conservative*" cluster featuring lower FNR and higher FPR, shown by green triangles;
- A "*naive*" cluster featuring lower FPR and higher FNR, shown by red squares.

These three clusters not only had clear partitions with regards to FNR, FPR, and other behavior measures,

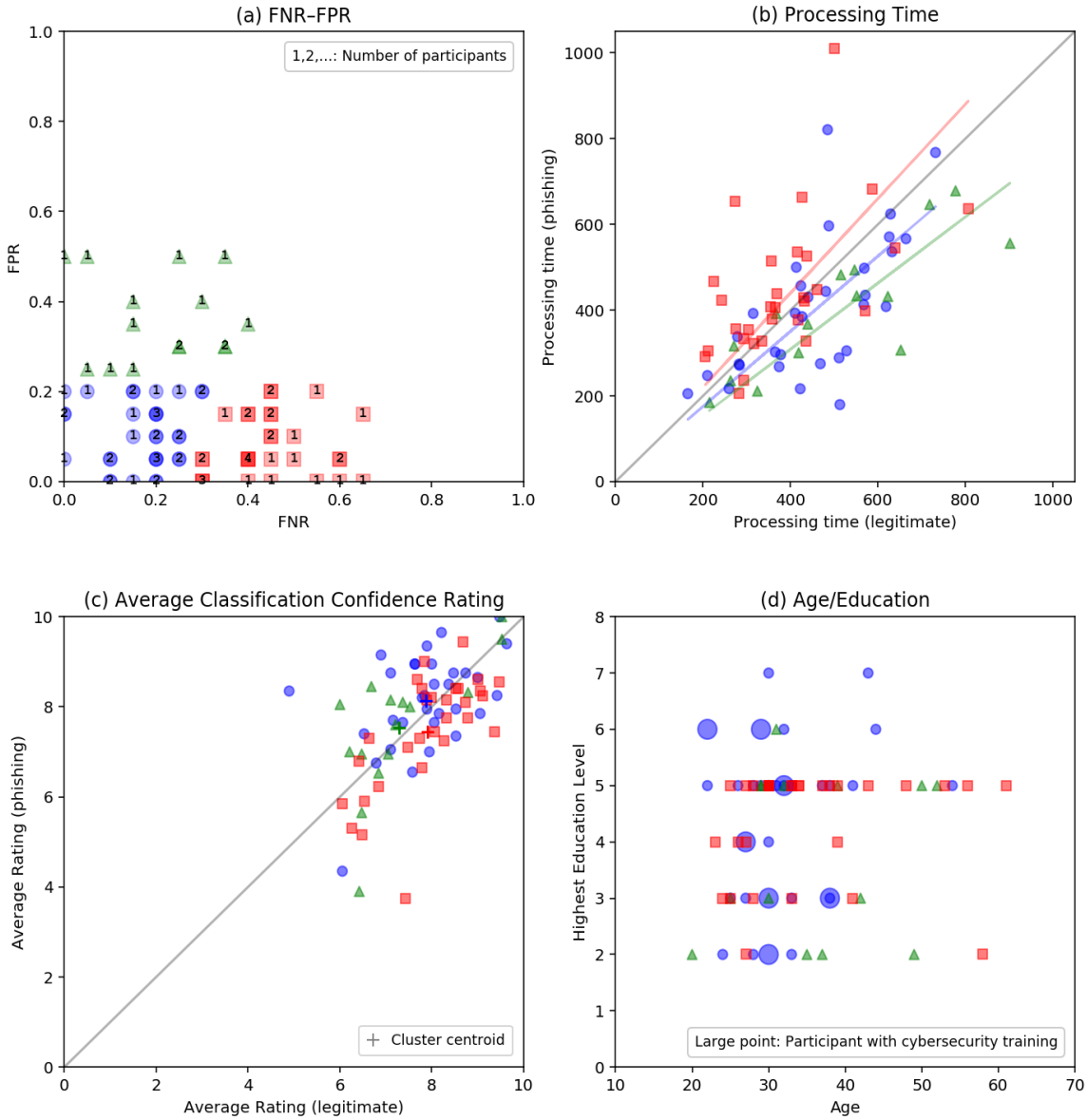

Figure 3. Clustering of participants in the no-multitasking condition using L2-norm normalization

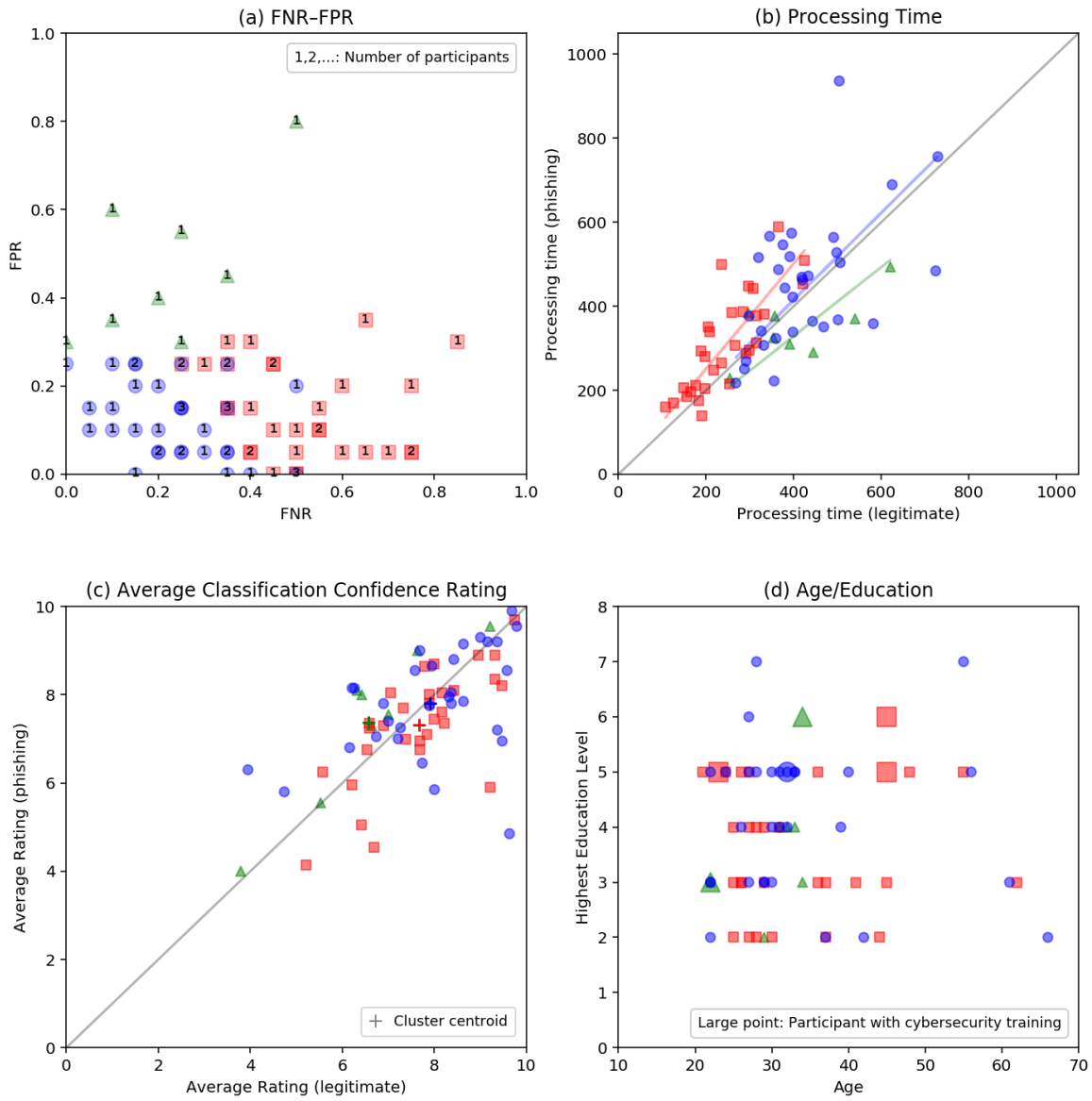

Figure 4. Clustering of participants in the multitasking condition using L2-norm normalization

but also possessed interesting patterns in the demographics of their member participants.

The clustering results are shown below as a set of two-dimensional scatter plots (Figures 3-4). Each figure has four subplots, of which (a)-(c) are the performance measures used in clustering. In each (a) subplot, displaying FNR and FPR, a numeric label denotes the number of overlapping points, i.e., participants with the same FNR and FPR values. Each (b) subplot shows the two processing times of phishing emails vs. legitimate emails, also featuring linear regression lines fit on each cluster. Each (c) subplot, on participants' average decision confidence ratings for phishing emails vs. legitimate emails, displays the three cluster centroids by cross markers (+). Finally, the (d) subplots show three pieces of important demographic information associated with the clusters, namely age, education level, and cybersecurity training, revealing a few patterns of note.

Figure 3 depicts the clustering results for the no-multitasking participants using the L2-norm normalization method. As shown in Figure 3(a), naive-cluster participants demonstrated comparatively high FNR, signifying that they were less successful in detecting phishing emails. Not coincidentally, as shown in Figure 3(b), 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 (and the monetary incentive may have led some to overthink and misclassify legitimate emails as phishing).

The overachiever cluster includes those participants with both low FNR and FPR. These participants also had the highest confidence level among the three clusters. The corresponding linear regression line in Figure 3(b) 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 thoroughly, e.g., checking more phishing cues, before confidently moving it to the "keep" folder. However, they only needed to find enough evidence of suspicion to correctly classify a phishing email. This seems to support a similar strategy used in the simulation study of no-multitasking users as reported in [6].

Intuitively, higher confidence ratings would be associated with better performance. As shown in Figure 3(c), confidence ratings of different clusters in general 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 3(b) also features overlap between clusters on email processing time.) These observations, consistent with findings in our previous reports, highlight the difficulty of relying on one or two criteria to characterize security behaviors, and the necessity of a comprehensive approach such as clustering.

Figure 3(d) highlights several interesting demographic trends on the roles of cybersecurity training experience, advanced education, and age 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. This seems to suggest that academic study or training can effectively improve a person's security behavior. Additionally, only one of the participants older than 45 is in the overachiever cluster. This may suggest a negative effect of aging on phishing classification, representing a widespread challenge to confront in societies with aging populations. We note that none of the above-45 participants possessed a graduate degree or had cybersecurity training, complicating our ability to determine whether age, lack of education/training, or a combination influenced their comparatively poorer performance. Further effort is highly desirable to study this phenomenon.

The MinMaxScaler normalization method resulted in clusters similar to, but seemingly less distinct than, the L2-norm function. The average confidence ratings apparently played a more significant role in forming these clusters. While participants reporting the highest confidence ratings mostly fell in the overachiever cluster, the three clusters are mixed with regards to FNR and FPR. Demographically, these clusters exhibit almost identical patterns to those in Figure 3(d).

Figure 4 shows the clustering results for the multitasking participants. Performance of the multitasking participants was generally poorer than single-tasking participants, as previously reported. Consequently, the data points are further away from the best performance, i.e., (0,0) in Figure 4(a) and (10,10) in Figure 4(c). As shown in Figure 4(b), these multitasking participants used less time overall to process emails.

The results indicate that multitasking significantly impacted the patterns among the three clusters. This is shown by the increased cluster scattering in Figures 4(a) and 4(c) as compared to corresponding subplots in Figure 3. Still, some broader trends remain present: naive-cluster participants spent more time processing phishing emails, while conservative-cluster participants spent more time on legitimate emails. Interestingly, the corresponding linear regression line in Figure 4(b) indicates that multitasking participants in the overachiever cluster now spent relatively less time on legitimate emails than phishing emails. Such behavior contrasts with that of the no-multitasking participants in this cluster, as seen in Figure 3(b), and might signal a shift in the strategy described previously.

Lastly, Figure 4(d) does not exhibit the demographic patterns as seen for no-multitasking participants. Multitasking participants with previous security training exhibited no increase in performance, showing multitasking to be a negative performance equalizer. (We note that a high number of multitasking participants older than 45 fell in the naive cluster, indicating that the secondary task might have more greatly impacted their ability to detect phishing emails.)

5. Further discussion

Interpretations - The complexity of human security behaviors makes sophisticated analysis regimes necessary for revealing insightful patterns. Initially, understanding the roles of incentives and demographics on user performance proved challenging. Through the combined efforts of the statistical tests and clustering analysis described in this work, as well as the analyses in several previously-published reports, we have developed reasonable conclusions as to the influence of these factors.

Multitasking has a significant negative effect on participants' capability to identify phishing emails and greatly changes patterns shown for participants only occupied by the email sorting task. This is unsurprising, considering the overhead caused by frequently switching between the two tasks. Surprisingly, monetary incentives are ineffective at improving critical decision making during phishing recognition. These findings would demand appropriate adjustment of strategies for security behavior interventions.

More importantly, demographic and background traits, including education level, experience of cybersecurity training, age, and knowledge of phishing cues, represent useful and reliable predictors for a participant's security behaviors. Such findings echo other studies including that by Sheng et al. [5]. These results offer further support for the role of education and security training in improving critical skills in security tasks. They are especially informative for developing individualized and customized anti-phishing strategies.

Limitations - The first challenge comes from the constraints of an unattended empirical study that relied on maintaining participants' attention. We had to consider factors including the session time, Internet connection, types of web browser, etc., in experiment design and execution. We could only accommodate 40 emails that likely do not accurately reflect the real-world ratio of legitimate to phishing emails (although the 50/50 split is consistent with previous phishing research). The phishing cues may not be fully representative of all possible cases. Moreover, arguably a larger number of participants per condition could have enabled more conclusive findings.

The second challenge lies in information availability and quality. Certain information modalities are not available from unattended experiments. For example, we were not able to capture eye gaze movement so could not conclusively determine whether participants focused on a certain element in an email. While we disabled all unused functions on the Roundcube interface, the nature of such a remotely-conducted, Internet-based user study means that noise might still be introduced if participants did not follow instructions closely. However, such variation is an inherent risk for any user study conducted in similar real-world settings.

6. Conclusion

This work is unique in characterizing empirical data of phishing decision-making, gathered on real users in a real-world scenario, and through the range of clustering and other analyses. The presented findings touched only a part of the rich information available in the complete dataset of our phishing study, publicly available at http://behavior.isi.jhu.edu/. Ongoing and future efforts aim to examine additional information in these data.

7. Acknowledgement

This work was supported by the National Science Foundation through Award 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 Cybersecurity and Infrastructure Security Agency or the Department of Homeland Security.

8. References

[1] D. Akhawe and A. P. Felt, "Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness", Proceedings of the 22nd USENIX Security Symposium, 2013, pp. 257–272.

[2] A. Bianchi, J. Corbetta, L. Invernizzi, Y. Fratantonio, C. Kruegel, and G. Vigna, "What the App is That? Deception and Countermeasures in the Android User Interface", IEEE Symposium on Security and Privacy, 2015, pp. 931-948.

[3] P. Kumaraguru, S. Sheng, A. Acquisti, L.F. Cranor, and J. Hong, Teaching Johnny not to fall for phish, Technical Report, Carnegie Mellon University, 2005.

[4] K.A. Molinaro and M.L. Bolton, "Evaluating the applicability of the double system lens model to the analysis of phishing email judgments", Computers & Security, 2018, vol. 77, pp. 128-137.

[5] S. Sheng, M. Holbrook, P. Kumaraguru, L.F. Cranor, and J. Downs, "Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2010, pp. 373-382.

[6] M. Shonman, X. Li, H. Zhang, and A. Dahbura, "Simulating Phishing Email Processing with Instance-Based Learning and Cognitive Chunk Activation", Brain Informatics (BI 2018), Lecture Notes in Computer Science, Springer, 2018, vol. 11309.

[7] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L.F. Cranor, "Crying Wolf: An Empirical Study of SSL Warning Effectiveness", 18th USENIX Security Symposium, USENIX Association, 2009, pp. 399–416.

[8] A. Vishwanath, B. Harrison, and Y.J. Ng, "Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility", Communication Research, 2016, vol. 45(8), pp. 1146-1166.

[9] H. Zhang, S. Singh, X. Li, A. Dahbura, and M. Xie, "Multitasking and Monetary Incentive in a Realistic Phishing Study", Proceedings of British Human Interaction Conference (British HCI-2018), Belfast, Northern Ireland, 2018.