CyberPhishing: A Game-based Platform for Phishing Awareness Testing

Matthew L. Hale
University of Tulsa
mlhale@unomaha.edu

Rose F. Gamble
University of Tulsa
gamble@utulsa.edu

Philip Gamble
University of Tulsa
philip-gamble@utulsa.edu

Abstract
Phishing attacks sap billions of dollars annually from unsuspecting individuals while compromising individual privacy. Companies and privacy advocates seek ways to better educate the populace against such attacks. Current approaches examining phishing include test-based techniques that ask subjects to classify content as phishing or not and in-the-wild techniques that directly observe subject behavior through distribution of faked phishing attacks. Both approaches have issues. Test-based techniques produce less reliable data since subjects may adjust their behavior with the expectation of seeing phishing stimuli, while in-the-wild studies can put subjects at risk through lack of consent or exposure of data. This paper examines a third approach that seeks to incorporate game-based learning techniques to combine the realism of in-the-wild approaches with the training features of testing approaches. We propose a three phase experiment to test our approach on our CyberPhishing simulation platform, and present the results of phase one.

1. Introduction
Phishing, including spear phishing and whale phishing, has become a serious problem prompting researchers and practitioners to find the most effective way to limit its impact. In 2009, it was documented that approximately 156 million phishing emails are sent out world-wide every day [1]. Even with increased research in filtering and classification techniques [2-4], about 16 million make it through email provider filters. Worse, approximately 8 million are opened and, in the end, 80,000 people are baited into giving away their personal information or allowing malicious activity that obtains financial and intellectual property information of organizations [5]. Social networking sites, such as Facebook, twitter, and renren (in China), increase the problem because the malicious content may appear to be coming from a “friend.” Websites that embed legitimate logos give the user a false sense of security when clicking on what appears to be a legitimate link. In 2012, the total annual cost of phishing attacks was estimated to be around 1.5 billion dollars [6].

Besides filtering and enhanced detection techniques, researchers also study the kinds of alerts users might take notice of in email and on websites [7], classes on phishing attacks [8], embedded training initiatives [9, 10], and game play focusing on a specific content type [11, 12] to address the human aspect of phishing and what allows it to work [2, 13, 14]. Training remains a necessary component [15], but the training must be such that users become aware of the multitude of possibilities for malicious activity given online content so that they can adapt as these activities become more sophisticated.

Game-based learning has been advocated for a number of areas where it is difficult to engage, motivate, and train users [16]. For game improvement to occur, learning goals, learning support features in the game, learner engagement, and learning outcome assessment must be well understood. For phishing awareness and training, two game features appear appropriate to affect learning: authenticity and repetition. Authentic simulation games can immerse learners in an environment of real world scenarios for the purpose of training [17]. Repetition, as training, can be used in games to enhance memory, such as with suspicious content that may similarly appear in an email, post, or browser, while at the same time immersing the learner in an engaging activity [18].

Data accuracy is an important consideration for phishing studies to prevent the subject expectancy effect [19], where the subject expects a certain stimuli (i.e., phishing attempts), adjusts his or her behavior accordingly (i.e., becomes generally more suspicious of all content), and thus, skews data away from what might be seen in real world settings. Research shows that data accuracy and the subject expectancy effect in phishing studies is directly tied to the realism of the phishing content and to user unawareness, i.e. not being aware that they will be examining phishing content as part of the study [20].

Test style phishing studies, such as [11, 19], ask subjects to rate the trustworthiness of various phishing content using screenshots [12] or out of context snippets of phishing content. These studies
lack realism, as subjects tend not to be immersed in the subject matter, and lack in unawareness as subjects know they are looking at phishy content.

In contrast to test-style phishing research are so-called in the wild studies such as in [21] and [22]. In these studies subjects were recruited without their knowledge and sent modified, but real, phishing stimuli via real world communication mediums, such as email and social networking. In such studies, the phishing attack stimuli pointed to fake, malicious URLs hosted by the study researchers. While such measures may have a good deal of realism and subject unawareness, they present ethical and legal challenges since they may require spoofing of legitimate sites, such as eBay (as in [22]), or central authentication pages (as in [21]). In such studies, when real users are duped by phishing attempts and then made aware via post-study postmortem debriefing they may feel a sense of victimization or anger [20] at the fact that their real world credentials were exposed to fake login servers accessible by the researchers. This type of training is adverse to subjects, can cause snags in the IRB process, and should be avoided if possible.

In this paper, we develop a two-tiered classification of phishing content features and a three phase methodology for examining how our simulation platform, CyberPhishing, can be used to present authentic content, which allows for realism and unawareness in a protected game environment that limits or eliminates subject risk exposure. Once a user participates, their responses can be studied and used to customize game content towards game-based awareness training. Our content classification consists of degradations specifying negative trust cues that can evoke suspicion in phishing content and sophistications that make phishing more difficult to detect. We outline the phases of experimentation and report the results of a Phase 1 alpha test.

2. Background

2.1. Phishing Tactics

Hong [7] overviews the effects of modern phishing attacks. Such effects include corporate information loss and national security secrets. This work shows that phishing attacks are increasingly more invasive and sophisticated, e.g. spear phishing with personalized content. Hong highlights phishing countermeasures including (1) filtering, black listing sites, and taking down sites, (2) user interface assistance, e.g. better and more appropriate warnings, and (3) proactive user training towards the improved recognizing and avoiding of attacks.

Many studies focus on classifier algorithms as a first line of defense for improved detection of malicious email or websites. For instance, Dhamija, et al., [5] investigate user awareness of website design. They found that attacks were more successful using various forms of visual deception that rely on a lack of user knowledge of security indicators. Their work included 22 participants examining 20 legitimate and fraudulent websites and found that often people miss trust cues, such as padlocks, and suspicious cues, such as an incorrect web address in the address bar. Findings in [4] suggest that these cues may be missed because of positive site features that elicit trust, such as brand logos or mass notifications by a recognized bank relying on a moderate chance that the user has an account.

2.2. Conducting Phishing Studies

Work by Jakobsson, et al. [20] examines the various legal, ethical, and technical implications that must be considered when developing a phishing study. Specifically the work highlights the fine balance between data accuracy, ethics, and legality that guide the experiment design, IRB process, and any ramifications that may result from the study. They developed a test-based phishing experiment to demonstrate the relevance of these factors with regard to user suspicion. The experiment consisted of three phases. In phase 1, subjects were given a test with 5 phishing emails as screenshots. The researchers varied the number of phishing emails in the test, hypothesizing that user suspicion would be independent of the number of actually suspicious emails in the sample. After the test, in phase 2, the subjects were presented with educational content regarding phishing. Finally, in step 3, the users were re-tested with 5 different content pieces. The results showed that ability to detect phishing content did not improve with education, although user suspicion did increase – meaning users were more suspicious of all emails. Thus, the subject expectancy effect, i.e. where subjects expect a certain type of stimuli and change their behavior accordingly must be handled in phishing awareness training research. Hence, fear of phishing attacks may increase via training with high expectancy effects, but not the ability to distinguish phishing attacks from non-phishing attacks.

3. Cyberphishing Platform

In order to bring the level of realism provided by naturalistic online environments, yet maintain a controlled environment that limits subject exposure and provides for training opportunities, we created a
A web simulation platform called CyberPhishing [23]. Coded as a client-side web application, the platform is a multi-facetted tool that provides simulation and data tracking capabilities across three mediums, i.e. email, web browsing, and social media. All mediums are conducted as part of a game scenario and mediated by the simulation modules shown in the game architecture in Figure 1. An economy engine reacts to user actions and drives the game scenario forward, while usage data is collected and analyzed.

Figure 1. Game architecture [23]

In addition to the user-facing simulation modules and UI, the CyberPhishing platform provides a rich Admin interface that allows researchers to build content dynamically with a live-preview client-side interface and a customizable experiment generator.

The core idea behind the application is that realism must pervade user experience in order to prevent the subject-expectancy effect, extract meaningful data from users for phishing research, and direct training. Through capture and analysis, the application will be able to compare user data against other phishing data and tailor later training exercises to deficiencies in phishing awareness.

To provide realism, the application features three primary interfaces that mirror real world equivalents. A story dashboard, shown in Figure 2, serves as a landing page. It describes the story arc using both video and text and lists user tasks as they appear in the role play scenario. Each blue link provides a basic task description that involves a user decision. Clicking a link takes users to one of the three interfaces based on the content medium. Each interface provides simulation and data tracking features for a) presenting the content realistically, and b) capturing all relevant usage and action information from users. All interfaces include a 1-10 scale for how trustworthy an item is and two trust actions that either perform an action in the game (i.e., trust the content) or ignores/blocks/rejects the action (i.e., distrusts the content). These are visualized in each of the interfaces shown below as blue and red buttons.

Figure 2. Story dashboard

3.1. Email Simulation

The first simulation interface, shown in Figure 3, presents email content to users, that includes malicious phishing email and legitimate emails. Each email is HTML-enabled allowing researchers to construct any sort of content that may be seen in the real world. Content is much more than simple screenshots as is often seen in test-based games. It may include fake links, pictures with fake links (seen by mouse hovering), attachments, and javascript simulation capabilities for interactions that mirror what is seen in standard inbox interfaces, such as Gmail. Importantly, all clickable links or pop-ups do allow users to leave the game domain (i.e. users cannot be directed to actually malicious sites). Items are presented in an inbox format as shown in Figure 3 and, when clicked, are shown in full as in Figure 4.

Figure 3. Email inbox simulation interface
The content item in Figure 4 is a spoofed google docs email that tells the user that a co-worker has shared a document form with them. The link in the email appears legitimate, but actually points to a false URL (as shown by the platform in the lower left consistent with what would be seen in modern browsers on mouseover). We believe the interface’s user immersion and richness will provide results closely mirroring real-world environments, while preventing subject exposure. This hypothesis is under investigation as discussed in Section 4.

3.2. Web Browser Simulation

To investigate user actions that occur while browsing, we built a simulated web browser with capabilities to a) show the https “lock” icon, b) show certificate information for simulated sites c) show URLs of simulated sites and d) provide all of the capabilities present in the email interface (i.e., mouseover link inspection, and HTML and javascript enabled content). To use screenshots of existing web content, researchers upload an image and then overlay hotspot areas that provide user interaction capabilities, such as mouseover links or javascript popups. As an example, Figure 5 shows a malicious spoof of a Red Cross donation site inside of the browser interface (top navigation bar). Each of the red arrows point to malicious links and the URL is shown as redx.com instead of redcross.com.

3.3. Social Media Simulation

The social media interface simulates a twitter-like interface. It allows for link spoofing and for users to mouseover links. In addition to these basic features, researches can set a list of business partners that correspond to “friends” or “followers” and provide a referential basis for users to potentially trust other users (i.e., they have partners in common). The interface also displays contextual clues, such as the username, as well as displaying the name of the user as is typical of twitter. For instance, Figure 6 shows several content items displayed in the social media interface, called My bVerse (for Business universe) in the role play scenario.
3.4. Content design Factors

To develop content, we created an initial baseline set of features that we believe make content more or less difficult to correctly identify as phishing. For factors that make phishing more difficult to identify, we create a list of *sophistications*. For factors that make phishing content more easily identifiable, we create a list of *degradations*. In general, we hypothesize that the more sophistications a content item has, the harder it is to detect, while the converse is true for degradations. This hypothesis is being reviewed as part of the study methodology discussed in the next section. Each sophistication is listed using an identifier S* where * is the number of the item. Similarly for degradations, D* is used. The initial lists below guide the content design process for the alpha test.

**Sophistications**

S1. Legitimate Logos  
S2. Duplicated look and feel of legitimate sites  
S3. Contextual or personal information given in content  
S4. Legitimate links where malicious content can be hidden  
S5. Sense of trust  
S6. Mimic intercepted communication  
S7. Formal grammar/style in writing

**Degradations**

D1. Suspicious URL identifying content sender  
D2. Suspicious links and Pop-ups within content  
D3. Multiple instances of poor spelling and/or grammar  
D4. Use of certain greetings and catchy phrases  
D5. Unnecessary warnings  
D6. People posing as friends or acquaintances  
D7. The use of a full name of receiver rather than a nickname  
D8. References to products not commonly available  
D9. Information or Item prices too good to be true  
D10. Missing security designators, e.g. https, padlock, certificate  
D11. Direct request to input personal data  
D12. Picture-in-picture  
D13. Survey requests with links  
D14. Appeals to an emotion, e.g., urgency and greed  
D15. Suspicious attachments to email  
D16. There should be links but they are missing

4. Study Methodology

The following sections discuss the specific methodology used for each phase of the overall experiment that seeks to distill out which trust cues elicit suspicion and which are overlooked in demographic groups to ultimately train users toward more awareness of phishing attempts. Phase 1 of the experiment focuses on alpha testing the CyberPhishing simulation platform to ensure user experience and interface factors match the levels expected for effective content simulation and presentation. Phase 2 involves a larger scale beta test of the CyberPhishing application that will target all demographic groups to develop a baseline assessment profile and assess the efficacy of game-based content as part of the overall experiment. Phase 3 will introduce advanced, customized training capabilities toward full game-based learning, in which CyberPhishing presents subjects with specific feedback based on content they misclassify and may adapt to provide additional challenges or may provide repetitive content where deficiencies are manifested. In addition, the effect of the training capabilities will be examined to determine if subjects simply become more fearful, or actually better able to detect suspicious content.

4.1. Phase 1 (alpha Test) Methodology

Currently, the CyberPhishing platform is in its first user ready state. Phase 1 provides a baseline for content development practices and removes bugs and issues with the platform to ensure a quality user experience (UX).

Accounting for these factors, our Phase 1 methodology involved the authors developing a baseline set of legitimate and malicious content across the three mediums (email, websites, and social networks). Developed content spans the content classification features so that a baseline understanding of how the different factors affect subject trust decisions can be understood before moving on to full scale content generation for the beta test. The second step in Phase 1 involved gathering a few subjects from each demographic group and allowing them to play through the developed content, knowing it was an alpha test. Subjects were anonymous and identified only by their self-made username and their session ID. Before consenting to testing, subjects were asked to self-report basic demographic information including their 1) age, 2) gender, and 3) education level.

Upon starting game play, subjects were given basic instructions for using the simulation platform and a primer for the role play scenario that exposed them to relative story elements. Subjects were then presented with 9 content items of each type (27 total across the mediums) across 3 *game days*, i.e., delineated time periods in the game. Each content
item allowed two specific actions to be performed by a subject, as shown by the blue and red bars in Figures 4-6.

For experimentation, the CyberPhishing platform tracks the “trust/don't trust” decisions made by the user. Later versions will enforce the trustworthiness rating to understand trust at a finer granularity, but this feature was not present in the alpha test. The Phase 1 analysis involved examining subject trust decisions en bloc across demographic groups to baseline the content development process. User experience and usability information were captured to determine if the platform was usable, such as if subjects could view and rate all content in the system. At the end of game play, a comment box allowed for optional subject feedback on their positive or negative experience with the application. The results of Phase 1 testing are presented later in Section 5.

4.2. Phase 2 (Beta test) Methodology

Development tasks for Phase 2 include doubling the amount of content using more game days, refining existing content based on alpha test feedback, and introducing repetitive training elements and feedback mechanisms to the simulation game platform. This will provide a diverse set of phishing and legitimate content for experimentation.

The experiment design will shift from focusing on UX and UI assessment towards content and training mechanism assessment. For Phase 2, we will recruit approximately 60 test subjects and attempt equal coverage across demographic groups. The Phase 2 experiment design will subdivide beta test subjects into three groups. Two of the three groups will be control groups and will, together, constitute about half (25% each) of the subject pool. Group 1 (control) will receive all phishing content examples during gameplay. Conversely, (control) Group 2 will receive all legitimate content during gameplay. The remaining subjects will be divided in half. Group 3 (25% of overall subject pool) will receive a mix of content types ranging from highly phishy to pristine. The last group, Group 4, will be divided into two equal sub groups. Group 4-1 will be given a standard phishing test that consists of a mix of phishing and non phishing content in the form of screenshots while Group 4-2 will be given all pristine content.

This design will allow us to determine the effect, if any, that the game play platform and/or role playing story has on whether or not users trust content. Our hypothesis is that subjects using our platform will exhibit a degree of subject-expectancy bias, as measured by the trust assessments in Groups 1 and 2 compared to those in Group 3, but that this degree of subject-expectancy bias will be significantly less than that found in Group 4.

The next part of the experiment design in Phase 2 will have subjects utilize the developed training mechanisms. Group 1 will be given specific training feedback that targets each of the phishing content items in their content pool. Group 2 will be given generic phishing training, since all of their initial content was pristine. Group 3 will be given specific training, consistent with that of Group 1 for items they incorrectly trusted during the first round of game play. Group 4-2 will be given generic training consistent with that of Group 2 while Group 4-1 will receive training consistent with that of Groups 1 and 3 (i.e. specific feedback regarding each item incorrectly identified as phishing content).

At this point a second round of content will be given to subjects and they will complete game play (in Groups 1, 2, and 3) or testing (in Groups 4-1 and 4-2). The overall detection rate of phishing content will be compared between the groups and between rounds 1 and 2. This information will inform us as to the effect, if any, of different training regimes on suspicion levels and ability to detect phishing content in both game play and traditional test-based methods.

Our first hypothesis is that we will see generally higher suspicion levels across the board due to increased subject-expectancy bias. Our second hypothesis is that we will see significantly better ability to detect phishing content for subjects given customized training based on their previous content over those that are given general phishing training, such as that provided by various government agencies or online tips. Our third hypothesis is that subjects playing the game will have better training outcomes than those that are simply taking a test. This last hypothesis is based on game-based learning research that highlights the value of game play as a medium for learning.

Overall the results of both content assessment and training mechanism assessment will inform the creation and adaptation of the platform for full scale testing in Phase 3. More specifically, these results will provide a baseline understanding of the factors involved in game based phishing research and in game-based learning for this training domain.

4.3. Phase 3 (Production) Methodology

Phase 3 will open the game to the public. This will mean a much larger subject pool across demographic groups. The first part of Phase 3 will be to gauge, using the previously developed and tested content and training mechanisms, the suspicion and awareness levels of different demographic groups on
a larger scale. The second part of Phase 3 will examine the efficacy of training and feedback mechanisms towards increasing awareness of phishing lures and cues across demographic groups.

The last part of Phase 3 will determine the effect of gamification towards improving training outcomes, i.e. making users better able to detect phishing content. For this part, Phase 3 will introduce gaming features including a score, badges, and titles into the game that awarded when subjects based on how well subjects detect phishing in the game. To examine the effect of gamification, subjects will be auto-divided into three equal groups. The first group will receive gamification rewards equitably based on how well they are actually performing. The second group will receive gamification rewards below how well they are actually performing, while the third group will receive no gamification rewards.

Our hypothesis, with regard to the first two goals, is that we will see varying degrees of suspicion and phishing awareness by demographic groups. We believe it likely that higher education will correlate positively with higher awareness and suspicion levels. We do not have a hypothesis regarding age, gender, or region, so we will posit the default hypothesis (i.e. no correlation with phishing awareness and suspicion levels). We also do not have a hypothesis as to the effect, if any, of age, gender, or education on training efficacy.

For the last part of Phase 3, we hypothesize that gamification will positively correlate with improved training outcomes. We also believe that gamification will improve training outcomes better if it is more accurately attuned to subject actions in the game. This hypothesis fits with operant conditioning psychological models that suggest behavior can be strengthened or weakened using reinforcement or punishment - in the case of the application, positive reinforcement upon successful classification of content as phishing or non-phishing.

5. Alpha Test Results

Following the methodology for Phase 1, this section presents the data collected during the alpha test of the CyberPhishing platform. The alpha test period allowed anonymous testing during a one week window following email recruitment, with 14 alpha testers registered on the site. Of those, 10 completed all content items while 13 completed all content in day 1. Though unit tests and usability testing were conducted, these results are not included below. Section 5.1 reviews the demographic breakdown of self-reporting alpha test subjects. Sections 5.2-5.4 examine the baseline content analysis for each content type and discuss factors of interest that arose from the alpha test.

5.1. Demographics

Phase 1 collected demographic information including self-reported gender, age, and education level. Of the 14 registered users, 8 were male and 6 were female. Figure 7 decomposes gender and age in a population pyramid format. The mean subject age was 28.3, meaning that the propensity to be familiar with phishing and the internet in general is likely higher than it might be in general.

![Figure 7. Population pyramid of testers vs age and gender](image)

These age results fit with expectations given that recruitment included emails to university students. Similarly, with this form of subject we would expect higher levels of education than seen in a random sample of the populace. These expectations fit with the data as shown in Figure 8, where 4 subjects identified as having 2 or fewer years of college, 3 identified as having bachelor’s degrees, and 5 identified as having a graduate degree. The other 2 listed themselves as high school or below.

![Figure 8. Decomposition of alpha test Subjects by education level](image)

The results in the following Sections should be interpreted in light of the low average age and
relatively high level of education. Given the small size of the subject pool, covariant analysis would be meaningless. Thus, per content analysis is presented en bloc as the number of users that trusted or did not trust each content item by type.

5.2. Email Content Analysis

The first examined content type was email phishing attempts. For email content, we constructed 7 malicious content items and 2 legitimate ones during Phase 1 preparation. These items were presented using the CyberPhishing simulated email inbox Figure 3. Table 1 presents the raw data captured for email content during the Alpha test. The item identifies the email content item. Items 1-3 were presented on Day 1 of the game scenario, items 4-6 on Day 2, and finally items 7-9 on Day 3. The trust row indicates the number of test subjects that performed the action on the content item. The don’t row indicates the number of subjects that selected not to perform the action. The # of sophistications indicates the number of these features applied to the content item, while # of degradations is the number of degraded features applied to the item. Finally the class row indicates the classified item type, with M for malicious and L for legitimate. The same format is used for referencing web and social content items. The green highlighted items indicate that subjects correctly identified legitimate content, while the light orange highlight indicates that subjects correctly identified phishing content.

Table 1. Data for email content assessments

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>trust</td>
<td>6</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>don't</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td>3</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td># of Sophistications</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td># of Degradations</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Class</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

For email, the data shows that in every case the majority of subjects correctly identified the content item. As a better way to visualize the trust decision breakdown, the data from Table 1 is shown as a stacked bar graph in Figure 9. The green bar on the left indicates the % of subjects that trusted the content item, while the light orange bar indicates the % of subjects that did not trust the content. Each Section for social and web content will present an equivalently formatted graph for content inspection.

In two instances (items 1 and 9) there seemed to be a level of sophistication that tripped up some of the subjects. A closer look at these two items showed they share one of the same sophistications, specifically S7 (formal grammar/style in writing) from the classification. Another interesting aspect of the content data was for items 5, 6, and 7, where 100% of subjects correctly identified the content as phishing. All of these content items shared degradations D11 (request for personal data) and D14 (appeals to emotion, urgency, or greed), indicating that these degradations may be readily identifiable.

5.3. Social Content Analysis

Using the same analysis techniques, we examined the social content presented to subjects using the social media interface shown in Figure 6. Table 2 displays the raw data results while Figure 10 presents the visual display of trust dynamics. With social media content, items 1, 2, 7, and 9 (all legitimate items) caused the most ambiguity. These items had no degradations, but did indicate one or more sophistications. Items 1, 7, and 9 shared S1 (legitimate logos), while items 2, 7, and 9 shared S5 (sense of trust indicated by common connection). It is unclear if the hesitation to trust these items stemmed from the nature of the role play scenario, a subject-expectancy effect, or the mode of communication.

Table 2. Data for social content assessments

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>trust</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>don't</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td># of Sophistications</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td># of Degradations</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Class</td>
<td>L</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>M</td>
</tr>
</tbody>
</table>

Figure 9. Trust dynamics in alpha test email content

Figure 10. Trust dynamics for social content
We hypothesize that the tendency to doubt the content stemmed from the latter effect, given that all of the legitimate posts involved legitimate requests for business transactions with established entities. This hypothesis is augmented by anecdotal commentary provided by a subject who questions the use of business transactions in social media. On the malicious content side, we found that an overwhelming majority correctly classified the content. Although the degradations were different across the items, they all had 3 or more degradations. We believe that the confluence of these multiple trust cues helped subjects correctly identify the content. In Phase 2, we will shape content to provide a finer granularity of coverage across classification elements to distill out at what level of degradation make it difficult for subjects to correctly identify content.

Item 8 was of particular interest in the results, because it was the only item that was classified incorrectly across all of the content in the alpha test. The content item was a web 2.0 format the simulate grassroots disaster recovery websites often seen after large scale regional disaster events domestically or world-wide. The item asked subjects to donate money, but had no degradations. We hypothesize that sites design for seeking donations are inherently more phishy than other sites. This factor will be examined further in an investigation of false positives in subsequent experiments.
play activity, what expectations are associated with the various mediums having phishing content, where confusion arose, and how feedback is needed.

Worked well! There was not a space between some of the links to 'tasks' at the bottom of the page for websites.

There is a typo in the intro video where it says kidapped instead of kidnapped. Also I'm not sure if that is a way IRS would reach out to people, however I do not know myself.

Some of the things that we were supposed to rate the trustworthiness of were ambiguous, like the fact that some people were making business deals over Twitter. Were we supposed to just accept that, or was it supposed to seem sketchy on its own? For some of the websites it was obvious that there had been a copy-and-paste to apply the websites to this situation, and the way that some of the text didn't match the rest of the site was jarring. Otherwise it was all good.

The application was well done and seemed to work very well. I noticed a few issues with the content pages seeming to change slight after loading - but I assume it was just executing a script in the background. I think the role play scenario could be clarified a little. There seem to be a lot of random association with a timber business, but maybe that is intentional?

After spending the time to go through everything, I'm a little disappointed that there isn't a score like you see on facebook stuff.

Figure 12. Sample comments from phase 1

6. Conclusion

In this work we overviewed our phishing simulation platform, CyberPhishing, and detailed a three part study to investigate the usefulness of role playing games towards understanding trust cues in phishing content and improving user awareness through training. Our results for Phase 1 highlighted various factors in the content design process that will be helpful when moving forward with the subsequent phases of the experimentation and user training.

7. References

23. Hale, M, and R. Gamble, Toward Increasing Awareness of Suspicious Content through Game Play, in 10th World Congress on Services, 2014.