Real-World Phishing Experiments: A Case Study

Markus Jakobsson
School of Informatics
Indiana University
Bloomington, IN 47408
markus@indiana.edu

Jacob Ratkiewicz
Dept. of Computer Science
Indiana University
Bloomington, IN 47408
jpr@indiana.edu

ABSTRACT

We describe a means for constructing phishing experiments which achieve the mutually competitive goals of being *ethical* and *accurate*. We present an implementation of these experiments based on the user interface of a popular online auction site, and the results gained from performing these experiments on several hundred subjects. In particular, we find that cousin domains (such as `ebay.secure-name.com`) are more effective (from a phisher's perspective) than IP addresses; approximately 11% of users will yield their credentials to a cousin domain, compared to approximately 7% for an IP address. Portions of this work appeared in a paper presented at WWW '06; these are marked, and a discussion of new material is given at the end of the introduction.

1. INTRODUCTION

Despite the increasing prevalence and use of anti-phishing measures – which include browser toolbars, spam filters, and user education – phishing continues to be a very real threat to Internet commerce. While it is of importance to understand what makes phishing attacks successful, there is to date very little work done in this area. Dominating the efforts are surveys, as those performed by the Gartner Group in 2004 [10]; these studies put a cost of phishing attacks around $2.4 billion per year in the US alone, and report that around 5% of adult American Internet users are successfully targeted by phishing attacks each year. (Here, a successful phishing attack is one which persuades a user to release sensitive personal or financial information, such as login credentials or credit card numbers). However, we believe that this is a lower bound: the statistics may severely underestimate the real costs and number of victims, both due to the stigma associated with being tricked (causing people to under-report such events), and due to the fact that many victims may not be aware yet of the fact that they were successfully targeted. It is even conceivable that this estimate is an upper bound on the true success rate of phishing attacks, as some users may not understand what enables a phisher to gain access to their confidential information (e.g. they may believe that a phisher can compromise their identity simply by sending them a phishing email).

We believe that the best (indeed, the only) way to measure the success rate of phishing attacks is to perform experiments on real user populations. The main drawback of this is clearly that the experiments have to be ethical, i.e., not harm the participants. Unless particular care is taken, this restriction may make the experiment sufficiently different from reality that its findings do not properly represent reality or give appropriate predictive power.

We believe it is important not only to assess the danger of *existing* types of phishing attacks but also of *not yet* existing types – e.g., various types of context-aware [7] attacks. We are of the opinion that one can only assess the risk of attacks that do not yet exist in the wild by performing experiments. Moreover, we do not think it is possible to argue about the exact benefits of various countermeasures without actually performing studies of them. This, again, comes down to the need to be able to perform experiments. These need to be *ethical* as well as *accurate* – a very difficult balance to strike, as deviating from an actual attack that one wishes to study in order not to abuse the subjects may introduce a bias in the measurements. Further complicating the above dilemma, the participants in the studies need to be unaware of the existence of the study, or at the very least, of their own participation – at least until the study has completed. Otherwise, they might be placed at heightened awareness and respond differently than they would normally, which would also bias the experiment.

In this paper, we describe an ethical experiment to measure the success rate of one particular type of attack. Namely, we design and perform an experiment to determine the success rates of a particular type of "content injection" attack. (A content injection attack is one which works by inserting malicious content in an otherwise-innocuous communication that appears to come from a trusted sender). As a vehicle for performing our study we use the popular online auction site eBay. We base our study on the current eBay user interface, but want to emphasize that the results generalize to many other types of online transactions. Features of the eBay communication system make possible our ethical construction (as will be discussed later); this construction is orthogonal with the success rate of the actual attack. Our work is therefore contributing both to the area of designing phishing experiments, and the more pragmatic area of assessing the usability of anti-phishing measures deployed and used by our experimental subjects.

Related Work. As previously mentioned, current work to determine the success rate of phishing attacks is predominately in the form of surveys. Mailfrontier [2] released in March '05 a report claiming (among other things) that people identified phishing emails correctly 83% of the time, and legitimate emails 52% of the time. Their conclusion is that

when in doubt, people assume an email is fake. We believe that this conclusion is wrong – their study only shows that when users *k*now they are are being tested on their ability to identify a phishing email, they are suspicious.

Another technique for assessing the success rate of a phishing attack is by monitoring of ongoing attacks, for instance, by monitoring honeypots. The recent efforts by The Honeypot Project [3] suggest that this approach may be very promising; however, it comes at a price: either the administrators of the honeypot elect to not interfere with an attack in progress (which may put them in a morally difficult situation, as more users may be victimized by their refusal to act) *or* they opt to protect users, thereby risking detection of the honeypot effort by the phisher, and in turn *affecting* the phenomenon they try to measure – again, causing a lower estimate of the real numbers.

We are aware of only a few studies which evaluate phishing success through experiments. The first study, by Garfinkel and Miller [6] indicates the (high) degree to which users are willing to ignore the presence or absence of the SSL lock icon when making a security-related decision; and how the name and context of the sender of an email in many cases matter more (to a recipient determining its validity) than the email address of the sender. While not immediately quantified in the context of phishing attacks, this gives indications that the current user interface may not communicate phishy behavior well to users.

Another paper by Wu, Garfinkel and Miller [13] discusses the (lack of) impact that commercial "security toolbars" have on users' security-related decisions. In particular, they find that users are spoofed by false websites 34% of the time, even when a security toolbar is installed. Dhamija et al. similarly find [4] that users frequently ignore browser-based security cues in making security decisions, leading to incorrect classification of potential spoof sites 40% of the time.

An experimental study of relevance is that performed by Jagatic et al.[12], in which a social network was used for extracting information about social relationships, after which users were sent email appearing to come from a close friend of theirs. This study showed that more than 80% of recipients followed a URL pointer that they believed a friend sent them, and over 70% of the recipients continued to enter credentials at the corresponding site. This is a strong indication of the relevance and effectiveness of context in phishing attacks. However, the study also showed that 15% of the users in a control group entered valid credentials on the site they were pointed to by an unknown (and fictitious) person within the same domain as themselves. This can be interpreted in two ways: either the similarity in domain of the apparent sender gave these user confidence that the site would be safe to visit, or the numbers by Gartner are severe underestimates of reality.

Organization. The next few sections (§ 2, § 4) introduce in detail some phishing attacks that may take place in the context of user-to-user communication. In particular, we describe several scenarios involving the specific phishing attacks that we would like to study. These sections constitute largely a review of material first published in [8]. In this paper we further discuss the anti-phishing measures that are commonly applied by eBay users. We then describe our experiment in § 5, and argue that it is both ethical and safe to perform, and simulates a real phishing attack; this discussion is also largely similar to that in [8].

Finally, we outline the implementation of the experiment in section § 6. We discuss findings in § 7, including several interesting implications. One is that the greeting that eBay includes in each message, meant to certify that it is genuine, seems to be largely ignored by users. It also seems that attacks employing a look-alike domain for the malicious site do better than those which simply use an IP address. This is especially interesting when the look-alike domain can stand up to scrutiny no better than the IP; the user seems to derive a feeling of reassurance from a quick glance at the look-alike domain. We find that this type of attack has an approximate 11% success rate. This section largely contains material not found in [8].

2. REVIEW: USER-TO-USER PHISHING ON EBAY

We contrast a user-to-user phishing attempt with an attempt that is purported to come from an authority figure, such as eBay itself. Before discussing what types of user-to-user phishing attacks are possible, it is useful to describe the facilities eBay provides for its users to communicate.

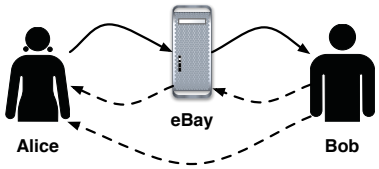
2.1 eBay User-to-User Communication

eBay enables user-to-user communication through an internal messaging system similar to internet email. Messages are delivered both to the recipient's email account, and their eBay message box (similar to an email inbox, but can only receive messages from other eBay users). The sender of a message has the option to reveal her email address. If she does, the recipient (Bob in Figure 1) may simply press 'Reply' in his email client to reply to the message (though doing this will reveal his email address as well). He may also reply through eBay's internal message system, which does not reveal his email address unless he chooses. See Figure 1 for an illustration of this scenario.

In messages sent to a user's email account by the eBay message system, a 'Reply Now' button is included. When the user clicks this button, they are taken to their eBay messages to compose a reply (they must first log in to eBay). The associated reply is sent through the eBay message system rather than regular email, and thus need not contain the user's email address when it is being composed. Rather, eBay acts as a message forwarding proxy between the two communicating users, enabling each user to conceal their internet email address if they choose. An interesting artifact of this feature is that the reply to a message need not come from its original recipient; the recipient may forward it to a third party, who may then click the link in the message, log in, and answer. That is, a message sent through eBay to an email account contains what is essentially a `reply-to` address encoded in its 'Reply Now' button – and eBay does not check that the answer to a question comes from its original recipient. This feature will be important in the design of our experiment.

2.2 Abusing User-to-User Communication

A user-to-user phishing attempt would typically contain some type of deceptive question or request, as well as a malicious link that the user is encouraged to click. Since eBay does not publish the actual internet email addresses of its users on their profiles, it is in general non-trivial to determine the email address of a given eBay user. Typically, a



(a) Communication path

(b) Features of email

Figure 1: Normal use of the eBay message system. In (a), Alice sends a message to Bob through eBay. If she chooses to reveal her email address, Bob has the option to respond directly to Alice through email (without involving eBay); in either case, he can also respond through the eBay message system. If Bob responds through email, he reveals his email address; he also has the option to reveal it while responding through eBay. The option to reveal one's email address implies that this system could be exploited to harvest email addresses, as will be discussed later. Figure (b) illustrates the features of an email in a normal-use scenario. This will be contrasted to the features of various types of attacks (and attack simulations) which will arise later.

phisher wishing to attack a eBay user would do one of the following:

1. Determine a particular eBay user's email address, and send a message to that user's internet email. Disguise the message to make it appear as though it was sent through the eBay internal message. This is a *spoofing* attack. (Spoofing in this context refers to the act of sending an email message whose sender has been falsified).

2. Spam users with messages appearing to have been sent via the eBay message system, without regard to what their eBay user identity is[1]. This may also use spoofing, but does not require knowledge of pairs of eBay user names and email addresses; the resulting message will therefore not be of the proper form in that the user name cannot be included. Since eBay tells its users that the presence of their username in a message is evidence that the message is genuine, this may make users more likely to reject the message.

[1]We note that an attacker may still choose only to target actual eBay users if he first manages to extract information from the users' browser caches indicating that they use eBay. See [11] for details on such browser attacks.

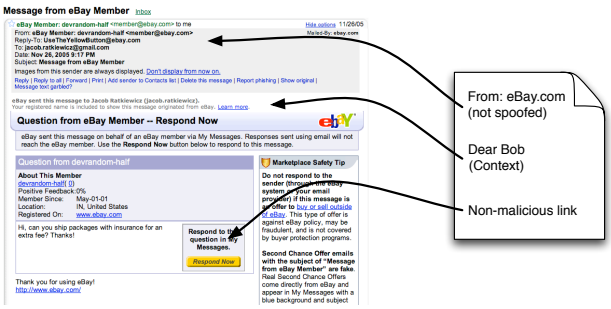


Figure 2: A message forwarded to an email account from the eBay message system. Note the headers, greeting (including real name and username), and "Reply Now" button.

3. CURRENT COUNTERMEASURES

The countermeasures a eBay user might be expected to use are summarized below. Some are certainly more prevalent than others (while we might imagine that almost all internet users make use of some sort of spam filter, it is likely that a smaller fraction use some dedicated anti-phishing tool, such as a browser toolbar).

Spam Filter. Spam filters can be very effective in identifying spoofed messages, or messages likely to be phishing attacks (perhaps as determined by keywords).

Browser Toolbar. eBay has made available a browser toolbar that provides direct access to many of its website's features. It also has a built-in database of known spoof sites, and can identify new ones by some simple heuristics. A visual warning is displayed to a user if they are visiting a site judged to be potentially malicious. As noted earlier, Wu et al. [13] find that this countermeasure is not likely to have significant impact.

User Education. eBay has made a concerted effort to educate its users on the dangers of phishing attacks, especially through spoofed messages. In particular, eBay includes in each user-to-user communication the registered eBay username of the recipient, as well as their real-life first and last name. It does this because this information would (supposedly) be difficult for a phisher to discover, and hence it claims that the presence of this information is evidence that the message is genuine.

4. REVIEW: PHISHING SCENARIOS

In considering the phishing attempts we discuss, it is useful to contrast them with the normal use scenario:

- *Normal use* – Alice sends a message to Bob through eBay, Bob answers. If Bob does not reply directly through email, he must supply his credentials to eBay in order to answer. This situation occurs regularly in typical eBay use, and corresponds to Figure 1). Important for later is the fact that when a user logs in to

answer a question through the eBay message system, he is reminded of the original text of the question by the eBay web site.

The following are some scenarios that involve the eBay messaging interface. In each, a user (or phisher) Alice asks another user (or potential victim) Bob a question. In order to answer Alice's question, Bob must click a link in the email sent by Alice; if Bob clicks a link in an email that is actually a phishing attack, his identity may be compromised.

- *Attack 1: Context-aware spoofing attack* – Alice spoofs a message to Bob, bypassing eBay. If Bob chooses to respond by clicking the link in the message (which Alice has crafted to look exactly like the link in a genuine message), he must supply his credentials to a potentially malicious site. Alice controls the contents of the email, and thus may choose to have the link direct Bob to her own website, which may harvest Bob's credentials. Alice includes contextual information in her comment to make Bob more likely to respond. This corresponds to Figure 3(a).
- *Attack 2: Contextless spoofing attack* – This is a spoofing attack in which Alice makes certain mistakes - perhaps including incorrect context information, or no information at all. This corresponds to an attack in which a phisher does not attempt to determine associations between eBay user names and email addresses, but simply sends spoofed emails to addresses he has access to, hoping the corresponding users will respond. The degree to which Bob is less likely to click a link in a message that is part of this attack (with respect to a message in the context-aware attack above) measures the impact that contextual information has on Bob, which is an important variable we wish to measure. This corresponds to Figure 3(b).

(a) Attack 1 – Includes context information

(b) Attack 2 – Incorrect, or missing, context information

Figure 3: Two possible spoofing attacks. Attacks currently in the wild, at the time of writing, are closer to (b), as they do not often contain contextual information.

5. EXPERIMENT DESIGN

In our study we wished to determine the success rates of Attacks 1 and 2 as described in the previous section, but we cannot ethically or legally perform either – indeed, performing one of these attacks would make us vulnerable to lawsuits from eBay, and rightly so. Thus one of our goals must be to develop an experiment whose success rate is strongly correlated with the success rate of Attacks 1 and 2, but which we can perform without risk to our subjects (or ourselves). The ethical principles underlying the design and evaluation of this experiment are described in the concurrent submission [5] by Finn and Jakobsson.

To this end we must carefully consider the features of the attacks above that make them different from a normal, innocuous message (from the recipient's point of view):

1. Spoofing is used (and hence, a message constituting one of these attacks may be caught by a spam filter).
2. An attack message contains a malicious link rather than a link to `eBay.com`.

More carefully restated, our goals are as follows: we wish to create an experiment in which we send a message with both of the above characteristics to our experimental subjects. This message must thus look exactly like a phishing attack, and must ask for the type of information that a phishing attack would (login credentials). We want to make sure that while we have a way of knowing that the credentials are correct, we never have access to them. We believe that a well-constructed phishing experiment will not give researchers access to credentials, because this makes it possible to prove to subjects after the fact that their identities were not compromised.

Let us consider how we may simulate each of the features in a phishing attack – spoofing and a malicious link – while maintaining the ability to tell if the recipient was willing to enter his or her credentials.

5.1 Experimenting with Spoofing

The difficulty in simulating this feature is not the spoofing itself, as spoofing is not necessarily unethical. When one visits a site such as a newspaper's online portal, there is often a link named something similar to 'Share this story with a friend'. This link allows a user to send a link to the story to some email address. If one chooses to do this, the site often employs 'benevolent' spoofing so that the friend will receive a message from the original user which contains the link. (One might intuit that the newspaper considers someone more likely to click a link that comes from their friend – much as in a context-aware phishing attack!) Our challenge, then, is not to avoid spoofing – but to make it possible for us to receive a response from our subject even though she does not have our real return address. Fortunately, eBay's message system includes a feature that makes this possible.

Recall that when one eBay user sends a message to another, the reply to that question need not come from the

original recipient. That is, if some eBay user Alice sends a message to another user Cindy, Cindy may choose to forward the email containing the question to a third party, Bob. Bob may click the 'Respond Now' button in the body of the email, log in to eBay, and compose a response; the response will be delivered to the original sender, Alice. Note that this feature is not a security flaw as it is of no use to a phisher; in our case, however, it makes replies possible even when spoofing is used.

Using this feature, consider the situation shown in Figure 4. Suppose that the researcher controls the nodes designated Alice and Cindy. The experiment – which we call *Experiment 1* – proceeds as follows:

1. Alice composes a message using the eBay question interface. She writes it as though it is addressed to Bob, including context information about Bob, but sends it instead to the other node under our control (Cindy).

2. Cindy receives the question and forwards to Bob, hiding the fact that she has handled it (e.g., through spoofing). The apparent sender of the message is still `member@ebay.com`.

   Note that at this point, Cindy also has the option of making other changes to the body of the email. This fact will be important in duplicating the other feature of a phishing attack – the malicious link. For now, assume that Cindy leaves the message text untouched except for changing recipient information in the text of the message (to make it appear as though it was always addressed to Bob).

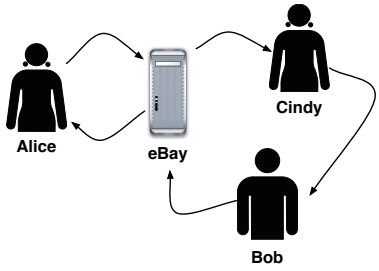
3. If Bob chooses to respond, the response will come to Alice.

We measure the success rate of this experiment by considering the ratio of responses received to messages we send. Notice that our experiment sends messages using spoofing, making them just as likely to be caught in a spam filter as a message that is a component of a spoofing attack (such as the attacks described above). However, our message does not contain a malicious link (Figure 5(a)) – thus it simulates only one of the features of a real phishing attack.

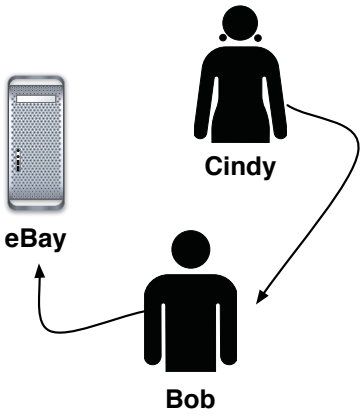
It's important to note that spam filters may attempt to detect spoofed or malicious messages in many different ways. For the purposes of our experiments we make the simplifying assumption that the decision (whether or not the message is spam) is made without regard to any of the links in the message; however, in practice this may not be the case. We make this assumption to allow us to measure the impact that a (seemingly) malicious link has on the user's likelihood to respond.

Note that in order to respond, Bob must click the 'Respond Now' button in our email and enter his credentials. Simply pressing "reply" in his email client will compose a message to `UseTheYellowButton@ebay.com`, which is the `reply-to` address eBay uses to remind people not to try to reply to anonymized messages.

Note that Experiment 1 is just a convoluted simulation of the normal use scenario, with the exception of the spoofed originating address (Figure 1). If Bob is careful, he will be suspicious of the message in Experiment 1 because he will see that it has been spoofed. However, the message will be



(a) Our experiment's communication flow



(b) C spoofs a return address when sending to B, so B should perceive the message as a spoofing attack.

Figure 4: Experimental setup for Experiments 1 and 2. Nodes A and C are experimenters; node B is the subject. A sends a message to C through eBay in the normal way; C spoofs it to B. The involvement of node C is hidden, making node B perceive the situation as the spoofing attack in (b); but if B answers anyway, the response will come to A.

completely legitimate and in all other ways indistinguishable from a normal message. Bob may simply delete the message at this point, but if he clicks the 'Respond Now' button in the message, he will be taken directly to eBay. It is possible he will then choose to answer, despite his initial suspicion. Thus Experiment 1 gives us an upper bound on the percentage of users who would click a link in a message in a context-aware attack. This is the percentage of users who either do not notice the message is spoofed, or overcome their initial suspicion when they see that the link is not malicious.

To measure the effect of the context information in Experiment 1, we construct a second experiment by removing it. We call this Experiment 2; it is analogous to the non-context-aware attack (Figure 5(b)). In this experiment, we omit the eBay username and registered real-life name of the recipient, Bob. Thus, the number of responses in this experiment is an upper bound on the number of users who would

(a) Experiment 1 - Spoofed originating address, but real link

(b) Experiment 2 - Spoofed originating address, real link, but poorly written message text

Figure 5: Spoofed messages without malicious links. These messages have a strong chance of being caught in a spam filter, but may appear innocuous even to a careful human user.

be victimized by a non-context-aware phishing attack.

5.2 Experimenting with a Malicious Link

Here, our challenge is to simulate a malicious link in the email – but in such a way that the following are true:

1. The site linked from the malicious link asks for the user's authentication information
2. We have a way of knowing if the user actually entered their authentication information, but,
3. The entering of this authentication information does not compromise the user's identity in any way – in particular, we must never have the chance to view or record it.

Recall that Cindy in Experiment 1 had the chance to modify the message before spoofing it to Bob. Suppose that she takes advantage of this chance in the following way: instead of the link to eBay (attached to the 'Respond Now' button) that would allow Bob to answer the original question, Cindy inserts a link that still leads to eBay but *appears* not to. One way that Cindy may do this is to replace `signin.ebay.com` in the link with the IP address of the server that `signin.ebay.com` refers to (for more information on why this works, see [1]). Another way to do this is to create a domain that maps to `signin.ebay.com` by another name. We did this by registering a domain, `static-address.com`, and defining `ebay.static-address.com` to be an alias for `signin.ebay.com`.

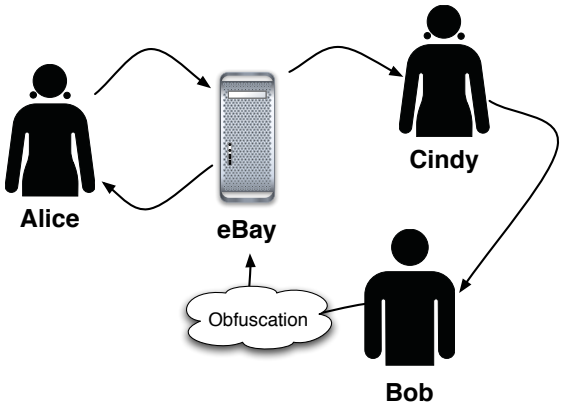


Figure 6: Communication flow for experiments 3 and 4. Node C uses spoofing to make the message to B appear to come from `member@ebay.com`, and obfuscates the link to `contact.ebay.com` to make it appear malicious. B should perceive the communication as a phishing attack.

Either of these links then fulfill the three requirements above – not only do they certainly appear untrustworthy, but they request that the user log in to eBay. We can tell if the user actually did, for we will get a response to our question if they do – but since the credentials really are submitted directly to eBay, the user's identity is safe.

5.3 Simulating a Real Attack

Combining the two techniques above, then, lets us simulate a real phishing attack. The experiment performing this simulation would proceed as follows:

1. Alice composes a message as in Experiment 1.
2. Cindy receives the question and forwards to Bob, hiding the fact that she has handled it (e.g., through spoofing). Before forwarding the message, Cindy replaces the 'Respond Now' link with the simulated malicious link.
3. If Bob chooses to respond, the response will come to Alice.

Call this experiment *Experiment 3* if it includes a simulated malicious link with an IP address, or *Experiment 4* if the "malicious" link is a sub-domain alias. See Figure 6 the setup of this experiment, and Figure 7 for a summary of the features of this experiment email. Note that Alice's participation is necessary in order to open a channel for a response through eBay; if Bob chooses to answer the (spoofed) question coming from Alice, he will actually be led to eBay. Alice creating the question initially allows this to succeed.

Note that the message that Bob receives in this experiment is principally no different (in appearance) than the common message Bob would receive as part of a spoofing attack; it has a false sender and a (seemingly) malicious link. Thus, it is almost certain that Bob will react to the

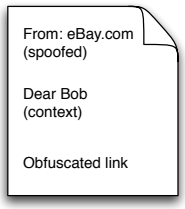


Figure 7: Experiments 3 and 4 – Context aware attack simulations with a spoofed originating address and simulated malicious link.

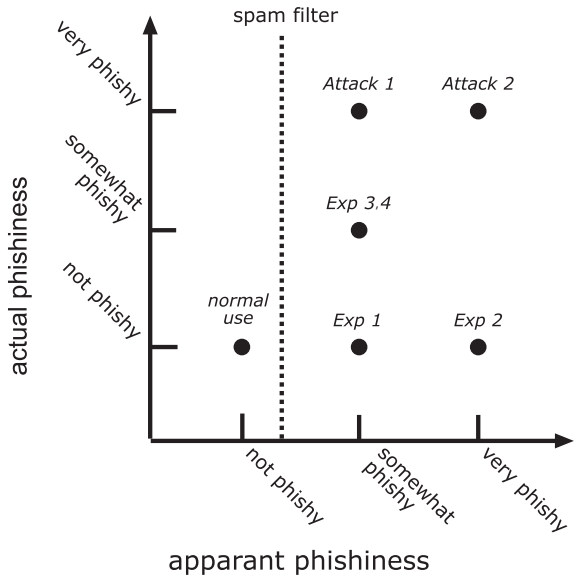


Figure 8: Our four experiments, contrasted with the phishing attacks they model, and the normal use scenarios which they imitate. Attacks that appear "somewhat phishy" are those that can be recognized by close scrutiny of the source of the message, but will look legitimate to casual investigation. "Very phishy" appearing attacks will be rejected by all but the most careless users. In the context of actual phishiness, "somewhat phishy" messages with deceptive (but not malicious) links, and "very phishy" messages are those which attempt to cause a user to compromise his identity. Any message to the right of the "spam filter" line may potentially be discarded by a spam filter.

message exactly as he would if the message really was a spoofing attack.

5.4 Experiment Design Analysis

In summary, we have constructed experiments that mirror real phishing attacks, but do so in a safe and ethical manner. The emails in our experiments are indistinguishable from the emails in the corresponding attacks.

That is, if in Experiment 3 we receive (through Alice) an answer from Bob, we know that Bob has entered his credentials to a site he had no reason to trust – so we can consider the probability that we receive a response from Bob to be strongly indicative of the probability Bob would have compromised his credentials had he received a real phishing attack. Refer to Figure 8; our goal is to have each experiment model a real attack's *apparent* phishiness (that is, to a user, and to automated anti-phishing methods), while not actually being a phishing attempt.

In the above, we use the term *indistinguishable* in a different manner than what is traditionally done in computer security; we mean indistinguishable to a human user of the software used to communicate and display the associated information. While this makes the argument difficult to prove in a formal way, we can still make the argument that the claim holds, using assumptions on what humans can distinguish. Thus, we see that experiment 1 (normal use, but spoofed) is indistinguishable from experiment 3 (obfuscated link and spoofed) for any user who does not scrutinize the URLs. This is analogous to how – in the eyes of the same user – an actual message from eBay (which is simulated in experiment 1) cannot be distinguished from a phishing email with a malicious link (simulated by experiment 3). However, and as noted, we have that messages of both Experiments 1 and 3 suffer the risk of not being delivered to their intended recipients due to spam filtering. This is not affecting the comparison between experiment 1 (resp. 3) and real use (resp. phishing attack).

More in detail, the following argument holds:

1. A real and valid message from eBay cannot be distinguished from a delivered attack message, unless the recipient scrutinizes the path or the URL (which typical users do not know how to do.)

2. A delivered attack message cannot be distinguished from an experiment 3 message, under the assumption that a naïve recipient will not scrutinize path or URLs, and that a suspicious recipient will not accept an obfuscated link with a different probability than he will accept a malicious (and possibly also obfuscated) link.

5.5 What gives it away?

The following is a summary of the features of a phishing attack (or one of our experiments) that might tip off a victim (or subject) of their true nature.

- *Spam designation* - A phishing attack might be flagged as spam and still shown in a user's Inbox, moved to a 'junk mail' folder, or deleted entirely. Instances of our experiments employ spoofing just as phishing attacks typically do, making them equally likely to be marked as spam.

- *Fraudulent links* - A phishing attack is typically focused on persuading a user that they should trust a fraudulent link. These are typically disguised to look like genuine links in one of several ways:

    1. A phisher might simply disguise the link in the message so that it looks legitimate before being clicked. This is the easiest to do, but when a user clicks a link of this type the true URL of the site visited will appear in her address bar. Often, a link of this type is simply an IP address, making it easy to tell that the link is not what it claims to be. This corresponds to the obfuscation by IP address in our experiments.

2. A somewhat more sophisticated method that a phisher might employ is to register a domain that looks similar to the one that he wishes to imitate. This can be done in a variety of different ways, ranging from the insidious (Unicode homograph attacks) to the easily detectable (as in our experiments, where we use ebay.static-address.com to stand in for signin.ebay.com).

   The intuition here from a phisher's perspective is that the presence of the legitimate site's name in the URL, even if it is not in the right place, might lull a user who glances at it into a false sense of security.

We expect that links of the latter type might be harder for a user to detect as fraudulent than links of the former type; this is suggested by our results, but not to a statistically significant degree. It is certainly intuitive, however, that a cleverly constructed look-alike URL is more convincing than the string of numbers that makes up an IP address.

6. METHODOLOGY

6.1 Identity Linkage

The first step in our overall meta-experiment was establishing a link between eBay users' account names and their real email addresses. To gather this information, we sent 93 eBay users a message through the eBay interface. We selected the users by performing searches for the keywords 'baby clothes' and 'ipod' and gathering unique usernames from the auctions that were given in response.[2]

We chose not to anonymize ourselves, thus allowing these users to reply using their email client if they chose. A previous experiment by Jakobsson [7] had suggested that approximately 50% of users so contacted would reply from their email client rather than through eBay, thus revealing their email address. In our experiment, 44 of the 93 users (47%) did so, and we recorded their email addresses and usernames.

We also performed Google searches with several queries limited to `cgi.ebay.com`, which is where eBay stores its auction listings. We designed these queries to find pages likely to include email addresses.[3]

We automated the process of performing these queries and scanning the returned pages for email addresses and eBay usernames; by this means we collected 237 more email and username pairs. It's important to note that we cannot have complete confidence in the validity of these pairs without performing the collection by hand. We chose to do the collection automatically to simulate a phisher performing a large-scale attack.

6.2 Experimental Email

Our goal was to try each experiment on each user, rather than splitting the users into four groups and using each user as a subject only once. This gives us more statistical significance, under the assumption that each trial is independent – that is, the user will not become 'smarter', or better able to identify a phishing attack, after the first messages. We believe this assumption is a fair one because users are exposed to many phishing attacks during normal internet use. If receiving a phishing attack modifies a user's susceptibility to later attempts, the user's probability to respond to our experiments has already been modified by the unknown number of phishing attacks he or she has already seen, and we can hope for no greater accuracy.

In order that the experimental messages appear disjoint from each other, we used several different accounts to send them over the course of several days. We created 4 different questions to be used in different rounds of experiments, as follows:

1. `Hi! How soon after payment do you ship? Thanks!`
2. `Hi, can you ship packages with insurance for an extra fee? Thanks.`
3. `HI CAN YOU DO OVERNIGHT SHIPPING??`
4. `Hi - could I still get delivery before Christmas to a US address? Thanks!!` (sent a few weeks before Christmas '05).

eBay places a limit on the number of messages that any given account may send in one day; this limit is determined by several factors, including the age of the account and the number of feedback items the account has received.

Because of this, we only created one message for each experiment. We sent this message first to another account we owned, modified it to include an obfuscated link or other necessary information, and then forwarded it (using spoofing) to the experimental subjects.

As discussed earlier, a real phisher would not be effectively hampered by this limitation on the number of potential messages. They might use accounts which they have already taken over to send out messages; every account they took over would increase their attack potential. They might also spam attacks to many email addresses, without including a eBay username at all.

Note that, technically, we also had the ability to send each user an arbitrary message because we had the chance to modify the text of the message along with the link, before spoofing it to the subject. We chose not to do this because after the subject clicks the 'Respond Now' link, if indeed they do so, they will be reminded of the question that was originally attached to the question ID in the link. This would no longer be the question in the email, had it been changed. This could have caused suspicion and was an artifact of our ethical construction, so we chose to write more vague messages instead. The most important piece of context information in the message was the user's eBay username. This was chosen because eBay claims that the presence of a user's registered name in an email certifies that the email is genuine. Thus, we argue that the fact our construction necessitated a vague question does not overly affect the results.

Debriefing Issues. Significantly, we elected *not* to debrief our participants. While this at first glance may seem unethical (in light of the tenants of informed consent), the Human Subjects Committee allowed this for two reasons:

- The subjects had been placed at no risk and results were anonymous, and

[2]Most data collection was done in the Perl programming language, using the `WWW::Mechanize` package [9].

[3]These queries were "`@ site:cgi.ebay.com`", "`@ ipod site:cgi.ebay.com`", and "`@ "baby clothes" site:cgi.ebay.com`"

Experiment	Response Rate
No name, good link	19% ± 5%
Good name, good link	15% ± 4%
Good name, "evil" IP link	7% ± 3%
Good name, "evil" Subdomain link	11% ± 3%

Figure 9: Results from our experiments. It's interesting to note that the presence or absence of a greeting makes no significant difference in the user's acceptance of the message. The intervals given are for 95% confidence.

- Debriefing would arguably do more harm than good – explaining the experiment to a sufficient level of detail to convince the layman that he was safe (and that he could trust eBay) was deemed too risky. At worst, the committee (and the authors) feared that a debriefing would result in users who feared their identity had been stolen (though it had not), as well as damage to eBay's brand. For a more detailed discussion of these difficult issues, see [5].

7. ANALYSIS

The results of our experiments are summarized in Figure 9, with intervals given for 95% confidence. 'Response rate' refers to the percentage of users in each experiment who logged in to eBay and composed a reply to us; thus, it is indicative of the number of accounts we could have compromised, had we been actually performing a phishing attack.

These results indicate that the absence of the greeting text at the top of each message has little to no effect on the user's chance to trust the contents of the message. This finding is significant, because eBay states that the presence of a user's registered name in a message addressed to them signifies that the message is genuine. It seems that users ignore this text, and therefore its inclusion has no benefit; identity linkage grants no improvement in the success rate of an attack.

However, we observe a significant drop in the number of users who will follow a link that is designed to look malicious. Note that the success rate for the attack simulated by a subdomain link is significantly higher than that predicted by Gartner. Further, Gartner's survey was an estimation on the number of adult Americans who will be victimized by at least one of the (many) phishing attacks they receive over the course of a year. Our study finds that a single attack may have a success rate as high as $11 \pm 3\%$ realized in only 24 hours.

Cousin domains are dangerous.. The fact that users are more likely to follow a textual link (i.e. a cousin domain) than an IP link was one we found very interesting. We feel that the registration of cousin domains to perform phishing attacks will become increasingly common, both due to their beneficial psychological aspects and the increasing filtering of IP-links by spam filters. Of course organizations cannot possibly register all potential cousin domains for their brands, but we feel that they should attempt to register common names, especially in response to emerging threats. As an example of this, one of the authors of this paper registered `wamu-rewards.com` after the "Chase Rewards" phishing attack began circulating.[4] Not only had Washington Mutual not registered this name, there was also no mechanism in place to prevent it from being registered. We feel this situation should be remedied.

A better means to identify legitimate messages. eBay's attempt to perform a limited amount of "mutual authentication" with its email recipients by including their registered names in an email is well-intentioned, but we feel that user inattention renders it effectively useless. A similar scheme is used by banks which include the last four digits of a user's credit card in emails; we feel this scheme also is insecure, for two reasons. Firstly, the last four digits of a credit card number are often printed on receipts (and shown on web pages e.g. during the Amazon.com ordering process). Secondly, the digits might conceivably be replaced by the *first* four digits (easy to guess), with an appropriate textual change and without the user noticing.

We feel that a personalized image, rather than a user's registered eBay name, would better identify a legitimate message. Our reason for this is twofold – a personalized image would be harder for a phisher to guess (or determine) than the user's registered name, and the absence of this personalized image would be harder for the recipient to overlook. Indeed, this solution is marketed by PassMark Inc., but for websites into which credentials must be entered rather than emails. It's worth noting that mutual authentication on the eBay login page is not sufficient; a spoofed email sent to a user might persuade them to visit a site performing a man-in-the-middle attack. This site might be able to present the mutual authentication information to the user after gaining their username, simply by giving the username to the real site. Thus, authenticating emails to users, e.g. by some form of mutual authentication, remains important.

8. FUTURE WORK

An interesting extension of this work would be a study to determine the effectiveness of a PassMark-like system for email, as we have suggested. This study could easily be conducted by a service provider, such as eBay, but could also be conducted by independent researchers – so long as they have access to an MTA for a number of experimental subjects.

1. Given a group of users, perform an initial experiment analogous to ours, in which a greeting is left out of an otherwise legitimate message. Determine the likelihood with which a user will fail to notice that the greeting is gone.

2. Train the group of users to expect a personalized image in each of their messages. While this can be done by any researcher with access to a user's email inbox (to insert the personalized image in messages from eBay), it would be far easier for an agency such as eBay itself to perform.

3. Once the each user has become accustomed to seeing the personalized image in each of their messages,

[4]This was a phishing attack that attempted to get user's credentials in return for a promised reward for banking with Chase. It did not actually use a domain like `chase-rewards.com`, but used an IP link; perhaps it would have been even more successful otherwise.

perform a second round of experiments in which the image is left out. Determine the likelihood with which the user will fail to notice its absence.

Of course, we expect that the likelihood in 3 will be significantly smaller than that in 1; however, a study revealing the extent of the difference in these would be very intereseting.

9. REFERENCES

[1] DNSRD - the DNS Resource Directory. `http://www.dns.net/dnsrd/`.

[2] Mailfrontier Phishing IQ Test. `http://survey.mailfrontier.com/survey/quiztest.html`.

[3] Know your enemy : Phishing. behind the scenes of phishing attacks. `http://www.honeynet.org/papers/phishing/`, 2005.

[4] Dhamija, R., Tygar, J. D., and Hearst, M. Why phishing works. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), ACM Press, pp. 581–590.

[5] Finn, P., and Jakobsson, M. Designing and conducting phishing attacks. In *IEEE S&T Special Issue on Usability and Security, to appear.* (2006).

[6] Garfinkel, S., and Miller, R. Johnny 2: A user test of key continuity management with S/MIME and Outlook Express. Symposium on Usable Privacy and Security.

[7] Jakobsson, M. Modeling and preventing phishing attacks. In *Financial Cryptography* (2005).

[8] Jakobsson, M., and Ratkiewicz, J. Designing ethical phishing experiments: A study of (ROT13) rOnl query features. WWW2006, Edinburgh, UK.

[9] Lester, A. WWW::Mechanize - handy web browsing in a Perl object. `http://search.cpan.org/~petdance/WWW-Mechanize-1.16/lib/WWW/Mechanize.p%m`, 2005.

[10] Litan, A. Phishing attack victims likely targets for identity theft. *FT-22-8873, Gartner Research* (2004).

[11] M. Jakobsson, T. Jagatic, S. S. Phishing for clues. `www.browser-recon.info`.

[12] T. Jagatic, N. Johnson, M. J., and Menczer, F. Social phishing. 2006.

[13] Wu, M., Miller, R. C., and Garfinkel, S. L. Do security toolbars actually prevent phishing attacks? In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), ACM Press, pp. 601–610.