Designing and Conducting Phishing Experiments

Peter Finn Markus Jakobsson
Dept. of Psychology School of Informatics
Indiana University
Bloomington, IN 47406

Abstract

We describe ethical and procedural aspects of setting up and conducting phishing experiments, drawing on experience gained from being involved in the design and execution of a sequence of phishing experiments (second author), and from being involved in the review of such experiments at the Institutional Review Board (IRB) level (first author). We describe the roles of consent, deception, debriefing, risks and privacy, and how related issues place IRBs in a new situation. We also discuss user reactions to phishing experiments, and possible ways to limit the perceived harm to the subjects.

Keywords: ethical, experiment, deception, human subjects, phishing, risks

1 Introduction

While fraud has been part of human society for as long as we know, the automated type of fraud that is known as phishing is a relatively recent phenomenon. It is becoming clear to society that phishing is a problem of quite catastrophic dimensions. Namely, phishing is not limited to the most common attack in which targets are sent spoofed (and often poorly spelt) messages imploring them to divulge private information. Instead, and as recently documented both in academic and criminal aspects, phishing is a multi-faceted techno-social problem for which there is no known single silver bullet. As a result of these insights, an increasing number of researchers and practitioners are attempting to quantify risks and degrees of vulnerabilities in order to understand where to focus protective measures.

There are three principal approaches in which people try to quantify the problem of phishing. The first, and most commonly quoted approach uses some form of survey, whether based on reports filed, polls relating to recent losses, and polls relating to recent corruptions of systems and credentials. A drawback of this approach is that it is likely to underestimate the damages. This is since many victims may be unaware of being attacked, or unwilling to disclose falling for them. However, it also is possible that surveys overestimate the risks, given the limited understanding among the public of what exactly constitutes

phishing. For example, a person who finds that his or her credit card bill contains charges for purchases he or she has not authorized, may think this is due to phishing or identity theft, whereas it may instead simply be fraud. (The distinction is often made in terms of whether the aggressor is able to initiate any new form of request, which should only be possible to the legal owner of the account; fraudulent use of credit cards numbers is typically not counted herein.). This makes surveys somewhat untrustworthy; moreover, they only allow the researcher to understand the risks of existing attacks in the context of existing countermeasures; no new attacks or countermeasures can be assessed. A related approach, with similar drawbacks, is to monitor honeypot activity. This poses the additional ethical problem to the researcher of whether to stop attacks in spite of the fact that this may alert the attacker, thereby tainting the results of the study.

A second approach is to perform "closed-lab" experiments. This approach also covers common tests, such as "Phishing IQ tests". While this approach allows the evaluation of attacks and countermeasures that are not yet in use, it has the significant drawback of alerting the subjects that they are being part of a study. This may significantly affect their response, typically causing an underestimate of the real risks. At the heart of the problem, we see that the *knowledge of the existence* of the study biases the likely outcome of the study.

A third approach is to perform experiments that mimic real phishing attacks, thereby measuring the actual success rates simply by making sure that the study cannot be distinguished (by the subjects) from reality. This poses a thorny ethical issue to the researcher. Clearly, if the study is identical to reality, then the study constitutes an actual phishing attempt. On the other hand, if the study is too dissimilar to reality, then the measurements are likely to be influenced by the likely difference in user response between the experiment and the situation it aims to study. Phishing experiments have the benefit of being able to measure the danger of attacks that do not yet occur in the wild, and of the success rates of countermeasures that may not yet be commonly deployed.

Our contribution

When academic researchers plan phishing studies, they are faced with the reality that such studies must not only be conducted in an ethical manner, but they also must be reviewed and approved by their Institutional Review Board (IRB). This requirement can be daunting. To begin with, we see that the phishing researcher typically would have to request a waiver of aspects of the informed consent process and request the use of deception when performing an experiment. This is in order to be able to ensure the validity of the study by the use of experimental, fake, phishing attacks that the subject/victim can not distinguish from *real* phishing attacks. The ethical issues relating to waiving aspects of informed consent are controversial and there is little consensus among IRB members and ethicists. Such issues are particularly controversial in the domain of online research, especially phishing research, which is relatively new to IRBs and ethicists in general.

This paper provides an overview of the review process used by IRBs, an outline of the section of the federal regulations, 45 CFR 46 [5], 116(d)(1-4), that provide the circumstances where aspects of the informed consent process can be waived. Moreover, it contains a discussion of the controversial ethical issues inherent in phishing studies that request a waiver of aspects of the informed consent requirement. Finally, this paper outlines the process of designing and analyzing phishing experiments in an ethical manner, and in accordance with the principles and regulations guiding IRBs.

We argue that phishing research, and the request for waiver of aspects of informed consent, involve a collection of new issues. A *first* key purpose of this paper is to outline the unique ethical issues and IRB approval issues raised by phishing research. While deception and the complete waiver of informed consent are a necessity in some types of studies on human subjects, it is usually avoided to the extent it is possible, and is typically only allowed by IRBs when the expected benefits of the study outweigh the anticipated risks, and the study meets certain conditions outlined in the federal regulations governing human subjects research. Here, the risks considered include any potential psychological harm that may be associated with being deceived. A *second* key component of this paper relates to debriefing. Namely, phishing experiments pose a rather unique and complex situation that questions the ethics of using debriefing as a means of harm reduction. We discuss the unique factors that might contribute to online debriefing in phishing studies causing more damage than good, in contrast to how debriefing is normally used to *avoid or rectify* damages. Thus, we reason that using debriefing in phishing research may be in immediate conflict with the standard IRB best practices. In this paper, we describe these issues, discuss the ethical principles and controversies regarding waiver of aspects of informed consent, and illustrate the concepts using recently performed phishing experiments, and their associated IRB reviews. We also illustrate the technical issues associated with how to mimic phishing reality without extracting identifying information, and, in fact, without *being able to* extract such information. The latter may be significant both as it comes to convincing IRBs of the ethicality of a study design, and to provide evidence to law enforcement that no abuse took place.

Remark about authors: Peter Finn is the chair of the IRB at Indiana University at Bloomington, which has processed several applications for human subjects approval for phishing experiments. Markus Jakobsson is pursuing research on this topic, and is the principal investigator or the faculty advisor for several studies involving phishing, many of which have been reviewed by the IRB.

2 Ethics and Regulation

2.1 The IRB and Research Ethics

The IRB has the mandate to review, approve or disapprove, and oversee all research conducted with data collected from human subjects to ensure that the research is conducted in compliance with the code of federal regulations, 45 CFR 46 [5], and in a manner that is consistent with the three ethical principles outlined in the Belmont Report [2]. The federal regulations codify the Health and Human Services policy for the protection of human subjects. The code covers the requirements for IRB structure, function, and review procedures, institutional responsibilities and the review requirements for researchers. It also covers the requirements for informed consent and altering informed consent, the analysis of risk, and special protections for vulnerable populations, such as pregnant women, fetuses/neonates, prisoners, and children. The Belmont Report's three ethical principles are (i) respect for persons, (ii) beneficence, and (iii) justice. *Respect for persons* means that individuals are treated as autonomous agents, capable of self-determination. Practically applied, respect for persons means that participants are allowed to freely consent to participate and should be fully informed of the nature of participation. *Beneficence* refers to the obligation that researchers secure the wellbeing of participants and requires that any risks associated with participation are out-weighed by the benefits of the study, and that researchers are diligent in removing, or appropriately managing, the risks of participation. Finally, *justice* refers to the principle that the benefits and risks of the research be fairly distributed across the general population. No subset of the population should gain most of the benefits, while another bears most of the burdens.

Both the use of deception – which is a necessity in many phishing research studies – and a complete waiver of informed consent – which is necessary in naturalistic studies of phishing – clearly challenge the principle of respect for persons. While these procedures are ethically controversial, the federal code allows for the use of deception and waiver of consent under certain circumstances that are outlined in the next section. It is important to point out that IRBs typically ask researchers to look for alternative experimental designs that do not require deception, and require cogent explanations justifying the use of deception, prior to approving such research. In fact, there are clear precedents and accepted justifications for the necessity of using deception in some types of studies [7]. Acceptable justifications usually include the following points: (i) the experiment involves no more than minimal risk (as defined below) and does not violate the rights and welfare of the individual, (ii) the study could not be conducted without the use of deception, and (iii) the knowledge obtained from the study has important value. Deception is typically used in social psychological and sociological studies. It is also used in some kinds of pharmacological studies that use placebos, where it is deemed a necessity to retain validity in studies of spontaneous behavior that cannot easily be elicited in a laboratory setting. Another example where it is used is in studies of the effects of drugs (such as

alcohol), or interventions where the expectations of subjects can contaminate their responses. For instance, the use of deception and waiver of informed consent have provided valuable information on the influence of social context in determining whether bystanders will help others in cases of emergencies or victims of crime [15]. The use of deception as placebo manipulations in studies on the effects of alcohol on psychological processes, such as pain or discomfort [6] or addictive processes [17], have provided valuable information about the role of pharmacological and psychological processes in the development of addiction. In studies that involve deception, the subject typically first consents to participate, but the researchers withold key information relating to the deception and then debrief the subject after the experiment has been completed. Debriefing, which is discussed at greater length below, involves explaining the nature and purpose of the deception and attempting to allieviate any discomfort the subject might experience upon learning that he/she was deceived. A critical facet of the process of debriefing is the personal, face-to-face, contact between researcher and subject, where the researcher can engage the subject immediately in a discussion and respond to the subject's concerns to ensure that subject understands all aspects of the study and can actively question the researcher about his/her concerns and possible misconceptions about the study. This facet of the debriefing process is impossible to duplicate in an online context.

It also is not uncommon for naturalistic observation studies to request and be granted by the IRB a waiver of the informed consent requirement. These are typically unobtrusive studies of people's public behavior, when it is either impossible to obtain consent from everyone observed, or when consent would potentially change the behavior of those being observed. The IRB typically will consider allowing for a waiver of informed consent when the researcher does not interact with the subjects in any fashion (i.e., is unobtrusive) and the behavior is clearly occurring in a public situation, where it is clearly observable to anyone. A naturalistic phishing study would not fall into such a category, because the researchers are interacting with the subject by sending fake phishing attacks, and, as such, an IRB would typically not approve such a study without some manner of debriefing the subject and only if it was clear that the study did not pose any risk to the subject. However, we note below that debriefing in naturalistic phishing studies sometimes raises more concerns than it addresses, and increases – rather than decreases – the potential for harm. The conditions under which a waiver for any aspect of the consent process is allowable is outlined in the next section.

2.2 IRB review process and phishing studies

From a historical perspective, online research in general is relatively new to IRBs and has presented a number of challenges to the IRB review process. Online research raises issues for IRBs, such as how to conduct and document informed consent; what is public data; how to deal with minors masquerading as adults; and how to protect the confidentiality of data collected online. Phishing research is entirely new to IRBs and presents unique ethical and legal challenges

to both researchers and IRBs, often in addition to those challenges posed by its common online nature. As far as we know, our IRB at Indiana University at Bloomington is the only IRB with experience reviewing and approving phishing studies. Hopefully, this paper will serve to highlight key ethical and IRB issues in phishing studies and to stimulate further debate on how best to address these issues. Our experience shows that there is not unanimity among IRB members or ethicists, as to the best way to handle the ethical challenges of phishing research. Clearly, more will be learned from experience, discussion, and debate among ethicists and IRBs. Thus, some of the perspectives presented in this paper are the perspectives of the authors and do not necessarily represent the perspective of the Indiana University, its IRB, or any other IRB. It is our hope, though, that there will be many commonalities between IRBs – especially given that different IRBs share the common goal of aiming to provide society with a screening and monitoring of research efforts. The use of deception, waiver of consent, and the nature of the risk in phishing research are the key issues that will be the focus of IRB review. As noted above, the federal regulations 45 CFR 46 [5], 116(d) provide conditions under which the IRB can alter the elements of informed consent, which allows for the use of deception, or waiver of the informed consent requirement entirely. Because it often is impossible to do valid phishing experiments without altering the informed consent process, phishing researchers must request that the IRB approve the use of deception or waiver of informed consent. Typically, IRBs will not allow for the alteration, or waiving, of the informed consent process unless the study has clear benefits. Phishing research has many potential benefits, given the catastrophic nature of its conquences for online users and service providers, and its potential for developing protective measures. We argue that the rationales for using deception in phishing studies in general, and for completely waiving consent in naturalistic phishing studies, are entirely consistent with the rationales for such changes in the informed consent process that have been deemed appropriate for social psychological and psychopharmacological research. They are especially well supported by the potential societal benefits of phishing research. However, we question the rationale for the requirement of debriefing subjects in naturalistic phishing studies, and note the unique issues raised in these settings.

When the IRB allows a researcher to use deception in the informed consent process, they typically require that the researcher debrief the subject pursuant to provision 4 of 45 CFR 46 [5], 116(d)(4). The logic for debriefing is twofold. *First*, inherent in the informed consent process is the requirement that the researcher be honest with the subject. Since deception violates this requirement and the underlying ethical principle of respect for persons, the experimenter is required to rectify this violation by explaining that he/she deceived the subject and the rationale for the need to do so. In the debriefing, the experimenter also should address, and be sensitive to, the likelihood that the subject might feel upset about being deceived and assure the subject that no objective damage was done. As noted above, a critical component of this traditional debriefing process is the face-to-face personal contact between researcher and subject. This is where the researcher can fully engage the subject and immediately address his/her unique

concerns to ensure the rectification of the violation in the informed consent process and to ensure that maximal opportunity exists to reduce any discomfort the subject might experience upon finding out about the deception. We argue that online debriefing in a naturalistic phishing study cannot adequately address these goals. A *second* reason for debriefing is that, in some rare cases, the research itself is aimed to study the effects of negative experiences on behavior, and uses false feedback or deceptive means to cause discomfort to the subject, such as rigging a task so the subject fails. In this case, debriefing is required to both alleviate the discomfort and rectify the violation of the informed consent process.

The alteration of informed consent only is allowed under conditions of minimal risk and when the alteration will not adversely affect the rights and welfare of the subjects. Minimal risk means that "*the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life*", as detailed in 45 CFR 46 [5], 102(i). It has been argued that the commonality of certain types of attacks in the wild and experiments with similar apparent functionality (from the point of view of the victim/subject) provides a reason to permit phishing experiments. The phishing attempt used in such a study represents an event that is common to online users, and, in fact, because it is not a genuine attack, does not carry the typical risk inherent in real phishing attacks. This is the case given that the *perceived* risk of the experiments does not substantially differ from the *actual* risks associated with real exploitation. In fact, there is no *actual* risk of exploitation in these experiments. The perceived benefits of a study would have to be quite significant to warrant any actual risk, as opposed to *perceived* risk. The perceived risk is that which subjects *believe* they are exposed to. Namely, it may be that the study is perfectly safe, but it is not possible for subjects to assure themselves that this is so. More in particular, if the subjects would need to place any degree of trust in the procedures of the study, this increases the perceived risk in comparison to a situation in which they can verify by some means that no harm could have been done. In any event, the risks inherent in a phishing study – as long as the researchers can ensure complete security for any information released by the subjects – are lower than those involved in a real phishing attack, to which online users are commonly exposed. In some cases, though, subjects may not have the technical savvy to perform this verification. Still, the perceived risk is lower in cases where an independent party can verify the absence of harm, than in cases where this is not easily done. The ease of verification could in many cases be impacted by the design of the experiment, as will be described in more detail onwards.

This brings us to the question: Does a phishing experiment that deceives a subject and exposes the subject to a fake phishing attack adversely affect the subject's rights or welfare? As noted above, as long as the researcher can ensure the security of any personal information of any information released by the subject (the procedures of which are outlined below), neither a laboratory phishing study nor a naturalistic phishing study should adversely affect the welfare of the subject. However, we question whether the use of debriefing in

naturalistic phishing studies might, in fact, adversely affect the welfare of the subject and propose that this, in part, is justification for not debriefing subjects in these types of phishing studies. In regards to adversely affecting the rights of subjects, the use of deception or waiving consent is not seen as a violation of a personal right, see 45 CFR 46 [5], 116 and [7]. Although laudable, the right to know the truth is not a recognized absolute right. However, the federal regulations and ethicists recognize that it is advisable to address this issue and use debriefing to provide the pertinent information relevant to the truth, when appropriate, see 45 CFR 46 [5], 116(d)4, and [7]. The question we raise is whether using debriefing in a naturalistic phishing study is appropriate.

2.3 Consent and deception in phishing research

Probably the most critical ethical and IRB issue for phishing researchers is that one cannot conduct valid experiments on phishing without either deceiving subjects who have consented to participate in some kind of online activity (where the experimenter will attempt a fake phishing attack), or asking the IRB to waive the consent process entirely. If one was to inform the subject that they may be subject to a fake phishing attack at some time while participating on a study of online activity or while engaging in online activity at a particular site, such as eBay or online banking, most subjects would be wary of this possibility. Thus, they are likely to be watching out for the attack, and will alter aspects of their behavior. This would create an experimental confound and the results of the study would be essentially invalid. Thus, the only way to conduct many phishing experiments is to employ some degree of deception and, in some cases, request a waiver of the informed consent process. In section 3, we outline different types of studies that have been conducted by the second author and his colleagues, with IRB approval from Indiana University.

Two different approaches to altering the consent process have been taken by these studies. The first employs a naturalistic observational design that involves waiving the consent process and allowing the researchers to include deception (a fake phishing attack) to investigate the factors that affect the likelihood of a person falling victim to a real phishing attack. In this approach, subjects are engaging in online economic activity and have no idea that they are subjects in a study. This approach is probably the most valid and ideal manner in which to study phishing, because subjects are behaving in a naturalistic manner. In laboratory studies, subjects may behave differently than they normally would and may alter their behavior simply because they are being observed and evaluated [18, 3], or because of demand characteristics in the experimental setting that lead subjects to alter their behavior [16]. Thus, this represents a reasonable justification that the research could not be conducted without a waiver of consent and the use of deception.

The second approach employs a laboratory experimental design that involves recruiting subjects to participate in a study of online behavior, such as making online purchases, that is conducted in a university setting, and using deception by not fully informing them that some of their online interactions do not origi-

nate from the online retailer. In fact, such interactions may correspond to fake attempts to phish them for personal information, such as usernames and passwords for that site. We call the attempts *fake* here to emphasize that no actual extraction of information and credentials may take place, although the subjects would not know this. This approach, too, appears to meet the requirement that the study could not be done without the use of deception.

2.4 Debriefing - Heals or Hurts?

As noted above, IRBs usually require researchers to debrief subjects who have been deceived, to rectify the requirement that researchers fully inform subjects during the consent process and to address provision 46.116(d)(4) in the federal regulations that pertinent information be provided when appropriate. Similar to typical laboratory social psychological or psychopharmacological research, debriefing in a laboratory phishing study is appropriate, because the face-to-face personal contact between researcher and subject will allow the primary aims of debriefing to be met. However, we question the appropriateness of debriefing in a naturalistic phishing study because it lacks the immediate interpersonal context that is critical to the process. Debriefing is supposed to heal the breach of trust and respect for persons that should have been established by informed consent and to heal any discomfort caused by false negative feedback. This would be, and should be, required whenever subjects consent to participate in a study. However, naturalistic phishing studies that study "real life" online behavior – where the IRB waives informed consent and subjects are not aware that they are being studied – present a somewhat unique case. This is because such studies do not violate the trust or honesty inherent in informed consent, and because debriefing lacks the close interpersonal contact between researcher and subject. Therefore, it may cause more harm than good. Debriefing in these naturalistic studies does not allow the researcher the opportunity to directly interact with the subjects to allay their unique concerns. In fact, our experience is that the only source of risk of harm is a result of debriefing subjects who have been subjected to a fake phishing attack in a naturalistic phishing study. If not debriefed, subjects who are aware of phishing attacks are likely to not be fooled in the study phishing attack, and ignore the attempt as just another of the many attacks they are exposed to on a regular basis. Subjects who are *not* aware of phishing attacks may be fooled, but the information they provide will be discarded by the researchers and no financial or personal harm will result. However, if these latter subjects are debriefed, they may feel upset, anxious, or angry that they were fooled. They may also be upset that they were included in a study without their consent, that they did not have the immediate opportunity to express and discuss their reactions with an authority (i.e., the researcher), and they may incorrectly worry that their personal information has been compromised. Ideally, debriefing could provide a wonderful opportunity to educate users about the dangers and nature of different real phishing attacks. But this is sometimes impossible to guarantee given the lack of control over the debriefing process. For instance, there is no way to ensure that the subject

will read the debriefing message right away; to ensure that the subject understands the debriefing message; or of providing an opportunity of expressing any unique concerns or reactions to the study or the deception with the researcher. Debriefing subjects also will inevitably result in some complaints to the web service provider, which are likely to raise significant public relations concerns for the provider. The dilemma is that phishing research can be of great benefit to both users and providers of online economic services. The fact is that a great majority of online service providers are currently having their services spoofed to dupe customers to reveal personal information. However, the reactions of some users and most providers is to feel threatened by the phishing research. These many issues raise valid concerns about whether debriefing is appropriate in naturalistic phishing studies. Some may conclude that this should then disqualify a naturalistic phishing study from being approved by an IRB and make such a study unethical. We argue that this unique situation should call for debate and further discussion, rather than outright disqualification of these types of studies – given their potential for great benefit and their minimal risk.

2.5 Risk Assessment for Phishing Studies

Risk assessment for any study also includes a risk/benefit analysis to determine if there are any risks and whether those risks are outweighed by the potential benefits of the study. This involves: (i) identifying the potential sources of risk and potential benefits of the study, (ii) determining whether the protocol used by the researchers reduces any potential risks, (iii) determining whether the actual risk in the final protocol is greater than minimal, and if the actual risks are greater than minimal, determining whether the risk management approach taken by the researchers addresses the actual risk, and (iv) assessing whether the potential benefits outweigh any risks. As noted above, there is great potential for benefit in phishing research. Little is known about the true rates of victimization in relation to the type of attach, or the contextual factors (nature of the attack/ clues in emails or fake websites), demographics, and psychological factors that influence whether one falls victim to specific types of attacks. Knowledge about these factors can have a direct and important impact on the financial wellbeing of both internet users and providers.

In naturalistic phishing studies, the first potential risk is that subjects who are fooled by the fake phishing attack will enter their personal information, and that their personal information will get into the wrong hands, making them vulnerable to financial loss and/or identity theft. The second risk is that subjects who know that they are being subject to a phishing attack will be upset due to the perception that someone is trying to take advantage of them. In regards to the first risk, it is relatively easy for researchers to use security procedures to guarantee that any personal information provided by subjects who are fooled by the attack will not be collected or saved by the researchers, or intercepted by anyone else. However, in some cases, as outlined below in section 4, it may be necessary for researchers to request permission from the IRB to have subjects' credentials temporarily available to them. In such cases, researchers would have

to outline to the IRB the duration of availability, the security risks, and how they would protect the confidentiality of that information. Thus, this risk should be easily rendered null. Phishing researchers must pay close attention to this risk and provide details to the IRB as to how they will render this risk null. In regards to the second risk, this risk will remain, but the risk is not greater than minimal because of the high frequency and regularity with which users are subjected to various *real* phishing attacks. In fact, it is likely that many users are so accustomed to be subject to phishing attacks that they are not likely to be upset by the fake attack (not knowing that it, indeed, is fake.) As noted above, the inclusion of a debriefing requirement in a naturalistic phishing study carries a potential for risk that is less easily managed than when using an experimental design, because researchers cannot directly interact with subjects to respond to, and allay, their concerns, and engage the subject in a conversation about phishing and its dangers. In fact, we argue that in a well designed naturalistic phishing study, debriefing is the *only source of risk* that is greater than minimal.

In laboratory studies, the risks are similar to those of a naturalistic study. As noted above, debriefing would be required in such a study because subjects consented to participation expecting to be fully informed. However, debriefing in a laboratory study would carry less overall risk than in a naturalistic study, because researchers have the opportunity to engage subjects in a conversation about their participation, the reasons for using deception, and the dangers of phishing. Debriefing in this context has a better likelihood of allaying concerns and educating the subject.

2.6 Legal Considerations of Phishing Studies

The primary legal issues that phishing research may raise are violations of a provider's *terms of use* for their service and the provider's *intellectual property rights*, such as unauthorized use of trademarks and violation of copyrights. In addition, some state statutes against phishing may include language that might create individual *privacy rights* issues. To the extent that there are state or federal statutes against phishing, or fraudulent use of the internet, that do not have specific language regarding intent of such usage, there is the possibility of lawsuits being filed by individuals or providers alleging a violation of such statutes. Although such statutes may not contain specific language regarding the intent of an alleged phishing attack, fake or actual, it is arguable that the statutes assume that phishing is used with the intent to defraud the public and harm the public by obtaining and illicitly using private information. Clearly, phishing studies have no intent to, and actually do not, defraud or harm the public. Furthermore, critical to such lawsuits is the requirement to demonstrate damages that result from the phishing study, of which there are none, if the security plans used by researchers are thorough and carefully implemented. Specific security procedures are outlined below in section 4, and a more detailed analysis is to be found in [4].

3 Phishing experiments - three case studies

We will describe three phishing experiments, and the process of review associated with these.

3.1 Experiment 1: Social Phishing

The first experiment aimed to understand whether people would be vulnerable to phishing attacks (such as requests to go to certain sites and enter one's password) when the requests appeared to originate with a friend of a subject. The study, which is described in detail in [10] found that this was the case. Indeed, 80% of all subjects went to the webpage indicated in the email they received, in which the sender's address was spoofed to correspond to that of a friend of theirs. This behavior in itself may pose a risk to users, as described in [9, 14]. In addition, 70% of the subjects correctly entered their login credentials at this site, in spite of multiple visual indications that the site was not legitimate. These visual indications were added to add realism to the experiment in that it would mimic a poorly designed phishing site; thus, the real numbers are likely to be higher for a properly designed phishing site.

The first experiment can be broken down in the following manner:

1. Collection of addresses of users who know each other, and selection of subjects.

2. Use of collected addresses to spoof emails to selected subjects, so that the emails appear to come from friends of theirs.

3. Verification of user credentials as the user follows a link in the attack email to take him to a site looking like a proper Indiana University password verification site, but with some visual indications alerting cautious users that the site is not authentic.

4. Debriefing of subjects (both recipients and claimed senders), and discussion of experiment in a blog.

The first step was done in an automated manner using a script to collect information from a popular social networking site. This involved the collection of names, email addresses of users, and of their friends, as publicly indicated on the site. A total of 23000 students at Indiana University at Bloomington had their information harvested in this manner. Out of these, 1731 were selected for the study, all of whom were verified to be at least 18 years old using material from the social networking site. Thus, all the subjects were considered adult, which is an important consideration to the IRB.

The harvesting step required the user agreement of the social network site to be violated; it specifies that this type of data collection may not be performed. However, given the ease for virtually anybody to violate this agreement, the IRB gave permission for this to also be done for the purpose of the study.

The second step involved a form of deception, as subjects were sent spoofed emails. This was approved given the commonality of spoofing, and the negligible actual risks associated with the experiment, as will be described below. While spoofing is not normally done in the wild to make recipients believe they are receiving an email from a friend, but rather from a known institution, there is no technical difference between these two types of spoofing, and both are straightforward to perform.

The third step allowed entry of user names and passwords, and subsequent verification of such pairs. However, the researchers performing the study never had access to these pairs, as a connection was established instead to a university password authentication server, which responded simply with information whether the pair was valid or not. Anybody wishing to verify that the researchers could not have accessed the passwords could have done so by examining the code specifying how the server worked; moreover, it could be verified from logs that there was not switching back and forth between different versions of the code, some of which could have behaved in a malicious manner. None of these measures were needed though, as this part of the study was never scrutinized in the aftermath of its completion.

In the fourth step, the subjects were debriefed and the study explained. Subjects were offered an online discussion forum to vent (which some did) and analyse the importance of the study. No authentication was required to access the blog, mostly due to the fact that this could have caused further anxiety in that subjects might feel the experiment was being repeated. This decision was also made in order to allow the subjects privacy. As a side-effect of this policy, the blog became overloaded with comments from a popular technology site (Slashdot) after a few days, at which time it was deactivated to avoid abuse. Many subjects were angry at the researchers and the fact that the IRB had approved it; however, and tellingly, none admitted having fallen for the attack. Instead, all were angry "on behalf of a friend who was victimized". This gives an interesting insight into the possible stigma associated with being victimized, even in the context of a research study like this. More details surrounding the experimental design, the results, and the user reactions can be found in [10, 8].

3.2 Experiment 2: A Study of eBay Query Features

The second experiment to be described was designed to understand the vulnerability of attacks using HTML markup of links, thereby hiding the content of these. More in detail, the experiment [13] sought to find the success rate of an attack that sends emails to actual eBay users, referring to context [11] of relevance to them and asking them to log in to respond to the query in question. This is a type of attack that was not common in the wild at the time of the experiment design, but which at the time of writing have become one of the more common types of phishing attacks in the context of eBay. The study found, among other things, that this type of attack has a success rate between 7 and 11% (depending on how it was customized, as described in more detail in the full paper [13]). A large portion of the "phishing failures" in the experiment

were due to successful spam filtering, just as in a real attack situation.

The experiment consisted of the following general steps:

1. Collection of eBay user information. This information involves an email address, information about an auction associated with the user, and - where available - the eBay user name. (This is not directly available, but was in many cases possible to extract using automated interaction with the user, as described in more detail in the full paper).

2. Construction and transmission of a spoofed email containing a valid but obfuscated link pointing to an authentic eBay webpage. Users whose email addresses were collected in the previous step were sent such a spoofed email, making it appear that it came from eBay.

3. Verification of credentials in successful instances of the experiment.

As in the first experiment, the first step of the second experiment may have violated the eBay user agreement; permission to do so was given with a similar rationale as in the first experiment. After the onset of the experiment, the researchers realized that they could perform the first step without violating the eBay user terms, simply by using a search engine to get information that is available to anybody (and not only to registered users.) Had the experiment been performed again, this approach might have been favored for collection of all information in the first step.

The second step in the experiment involved spoofing of eBay, and use of eBay logos and trademarks to mimic the general appearance of an actual phishing attack of this sort, which in turn would mimic the actual appearance of real interaction between eBay and one of its users. The use of spoofing was considered acceptable for similar reasons as described for the first experiment above. The use of trademarked logos proved the thorniest issue, in view of potential legal actions by eBay. In the end, it was decided that the large anticipated benefits of the study in comparison with the commonality of unauthorized use of trademarked logos warranted this action, given the lack of risks to the subjects. The construction of the link to the eBay site was done using interaction between registered eBay users (the researchers) and eBay, and may have been in conflict with the user agreement. However, this technique was used to protect users, as it allowed verification of credentials of subjects without allowing the researchers access to the same. For a description of how this was achieved, we again refer the reader to the paper describing the study.

The verification of the credentials of "successfully phished" subjects used eBay authentication servers, but without the knowledge or consent on behalf of eBay. Using a non-invasive technical peculiarity described in [13], the researchers were able to obtain information regarding whether a given login attempt was successful.

The second experiment did not involve any form of debriefing, which was a source of conflict within the IRB, but which finally was approved. The rationale for this decision was twofold: First and foremost, it was determined that the only

real harm that could arise from the study would be associated with debriefing, with no possible way for the researchers to properly explain that no actual harm was done. Secondly, it was argued that debriefing would increase the risk for legal action on behalf of eBay - not because they would become aware of the study - but because its users may feel that eBay is not sufficiently protecting them, thereby potentially causing eBay to lose business. This latter argument, though, can be turned around, instead saying that any security vulnerabilities must be brought to the attention of society.

3.3 Experiment 3: Man-in the-Middle Attacks

The third experiment aimed to study the vulnerability caused by so-called man-in-the-middle attacks. This is a form of attacks in which the phisher poses both as a service provider and as a user, interacting with both victim and service provider in order to supply correct responses to actions, in the manner that the service provider would have reacted. This attack can cause a full emulation of the behavior of the service provider - except of course for the fact that the phisher can obtain access to any transmitted information in the process. This is the case even when end-to-end encryption is performed, given that the victim would encrypt data for the attacker and not the service provider, and the attacker would then encrypt data for the service provider. Similarly, data sent by the service provider would send data encrypted for the attacker, and the attacker would extract the data and send it in an encrypted manner to the victim.

The experiment[1] consists of the following steps:

1. Recruiting of subjects without full disclosure of the goals and means of the study.

2. Interaction with subjects using spoofing, where subjects receive emails appearing to come from eBay, with embedded URL pointers linking to a site acting as a man-in-the-middle attacker.

3. Verification of subject behavior and credentials with special provisions to avoid access of credentials by researchers.

The first step of the experiment therefore involved deception in that subjects were not told what the goals of the study were. This was judged acceptable given the limited risks of the experiments, i.e., given a similar rationale as used for allowing deception on previously described experiments. Similarly, the second step of the experiment involved standard spoofing of emails, which was also approved using this rationale. The verification of credentials was designed to be done by handing off the session to not involve the man-in-the-middle node, simply verifying that the credentials must have been correct by verification of publicly accessible data relating to the auction in question.

[1] For a more detailed description of the setting of the study, we refer to [1, 12].

A thorny issue that the researchers faced was whether to use end-to-end encryption between subjects and researchers (i.e., man-in-the-middle node). Doing so would complicate matters, as the subjects would be given a warning stating that the certificate was not recognized; this in turn would be likely to yield a lower estimate of the real success probabilities. On the other hand, not using SSL at all – which is the likely approach of any attacker – would jeopardize the credentials of our subjects in the face of any eavesdropping on the line. While this is highly unlikely, we did not want to expose our subjects to this risk. It was decided to perform two separate experiments: one in which no SSL encryption was used, but where the web form would not perform any POST of the entered results (and therefore not transmit these to the researchers); and one in which SSL was used, but subjects were coerced to accept the associated certificate during the study enrolment phase. The former version of the experiment does not allow any verification of credentials, but appears to the subject as the likely phishing attack. The second version might have a slightly higher yield than a real attack, due to the use of SSL, but allows the determination of the fraction of correct credentials, which can then be assumed to be the same fraction as that of the first version. In combination, the two versions of the experiment would give a better approximation than either one in isolation. At the same time, harm to subjects would be avoided to the largest extent possible. The researchers were willing to briefly process the subjects' eBay credentials in order to maintain the active man-in-the-middle attack. Thus, in contrast to the first experiment, the credentials would actually be temporarily available on the researchers' machine. However, the researchers were unwilling to obtain even temporary access to the subjects' PayPal passwords, the possible theft of which were a focus of the study. Therefore, the session would be handed over from the man-in-the-middle machine to the real PayPal server by the time the subject was ready to pay. Successful entry of the credential would be determined by verifying with the subject whether the transaction went through.

In spite of favorable views within the IRB, the experiment has - at the time of writing - not been possible to start. The reason is that the FBI demanded that the machine on which the researchers developed and tested the code would be taken off the Internet. This, interestingly enough, was due to the reporting of the man-in-the-middle software running on the researchers machines (but not accessible to others). The software was detected using monitoring software run by eBay and Norton Utilities, causing an automated report of the "offending" machine to eBay, followed by a cease-and-desist order provided by the latter. This peculiar twist of events highlight how researchers and service providers are taking liberties against each other and each other's potential rights, in order to achieve their goals. While neither aims to hurt the other, there is a noticeable difference in goals.

4 Making it Look Like Phishing

In a phishing experiment, it is important to make interaction look like it is phishing, without actually compromising credentials. In the three experiments described above, we have illustrated methods to avoid having to handle credentials, but still being able to verify whether they were correctly entered. In the first experiment, this was achieved by obtaining feedback from a password authentication server that the researchers had access to. In the second experiment, it was done by counting feedback given through the standard eBay interface, to which subjects were routed. In the third experiment, the correctness of PayPal credentials was to be verified by asking subjects whether the transaction went through - this is an example of how one can use a secondary channel to obtain information about the success rate. (While the eBay credentials in the third experiment were not explicitly verified, they still had to be temporarily available to the researchers, marking a departure from the desired approach of not being able to access credentials.)

Technical Pitfalls

The greatest technical difficulty associated with this type of study is that associated with guaranteeing that the measured results are representative of the type of attack under consideration. The difficulty is inherent to this type of research, given that it is not possible to let users know that they are taking part in a study, as this is almost certain to affect the results. This does not only give rise to the ethical dilemma described before, but also poses the researchers with the following technical problem: *How can the study be designed so that all subjects, who would have fallen for the attack, are counted in the study, but only those?* We will draw on the previously described experiments to highlight this issue.

In the first study, the researchers specifically registered a domain that would appear to host the page onto which the subjects were requested to enter their credentials. For reasons of security of the credentials, this site did not collect the credentials, and was in fact hosted on a university server; however, this fact must not be possible for the subject to observe. The reason for this is that subjects were assumed to have a different trust relation with and reaction to a university server and a server at an unknown domain. Whether this is rational or not (i.e., the university server may also be corrupt!) it is imperative to take into consideration when designing the study. Similarly, the information in the login window of the first study was intentionally not looking perfect, but had some clear signs of being associated with a phishing attack. This was done to mimic the level of skill exhibited by an normal phisher, given that we wanted to measure the success rate such a person would have, as opposed to the highest possible success rate (which may require a higher degree of skill and customization, but which is, of course, of independent interest.)

In the second study, the researchers degraded the greeting of the spoofed email to various extents, and even removed the greeting altogether. This was

done to measure the success rates in different potential attack scenarios, corresponding to different degree of knowledge of the victim by the attacker. While degraded content surely did lower the yields, it allowed a reasonable approximation of the corresponding threats to be made. Consistent with this view, no special spamming methods were employed. In particular, the researchers did not use any of the recently observed tricks that a small number of highly skilled spammers use to bypass spam filters. Instead, well known techniques were used. Again, this was done to measure the approximate success rate of a normal instantiation of the attack under consideration, as opposed to the worst-case scenario in which the attacker is highly knowledgeable. Again, this may be an interesting study to perform, but did not correspond to the goals of the researchers, and so, the study design was made accordingly.

In the third study, it was found very difficult to obtain a good estimate on the yield of the attack without exposing subjects to unnecessary risks. Thus, the researchers designed two separate variants of the same experiments, with the goal of allowing a better estimate to be made than if only one of these versions had been used. This points to an interesting type of trade-off between the expected accuracy of the measurements and the potential harm associated with performing the experiment. Such a trade-off was not present in the other two experiments, but is highly likely to occur in other studies onwards. In such cases, it may be of importance to design one experiment to measure a lower bound of the success rates, and one to measure an upper bound of the experiment. It is not trivial to design experiments to make the bounds relatively tight while securing subjects against any unnecessary harm.

5 Subject Reactions

In this section, we will briefly describe subject reactions to phishing experiments. Due to the absence of debriefing in the second study, and of subject interaction in general, we do not know of what the reactions were in that study. We suspect that subjects, who fell for the phishing attempt, have no reaction (as they probably never realized that it was a phishing attack); subjects who did not fall for it, either never saw the email (as it went directly to their spam folder) and therefore had no reaction, or simply classified it as a phishing attack and ignored it. Given that the third study has not been executed, we have no data whatsoever in terms of reactions for that. For the first study (the social networks study), we have ample material, though.

Initially, a large number of subjects in the first study believed that either they or their friends had been affected by malware, causing the offending emails to be sent. Others later believed that the researchers had accessed their machines in order to send the emails in question, often feeling outraged that this had occurred. Thus, whereas many users understand well that it is possible to spoof emails, it was not immediately clear to many that an attacker or researcher can spoof emails from arbitrary senders - including their friends.

Many subjects also felt frustrated that their personal data had been exposed

and used, exhibiting a lack of appreciation for the fact that personal data that is put on publicly accessible forums no longer is private. Furthermore, many subjects did not understand the nuances associated with being able to verify credentials without accessing them, and felt upset that the researchers had collected their passwords. While some subjects saw the educational value of the experience, and appreciated the insights they had gained as a result of being part of the study, there were more users who felt that the study had no value, and felt violated at not having been asked permission before the experiment was performed. (Any explanation that this would invalidate the results of the experiments were seemingly irrelevant in this emotional argument.) Interestingly, none of the subjects admitted to having been fooled by the spoofed email, but all of those who were angry were either angry "on behalf of a friend" who had fallen for it, or upset in rather general terms. This suggests that there is a clear stigma associated with having been victimized (whether any real damage was done or not), which in turn tells us to be suspicious of the results given by surveys of phishing.

Most of the subject reactions were obtained from a blog that was introduced and made available to all 1731 subjects in a debriefing email that was sent to them (whether they were one of the 921 recipients or one of the 810 spoofed senders of an email in the study). A total of 440 posts were collected in three days; this does not take into consideration irrelevant posts (such as advertisements) that were made towards the end of the duration during which posts could be made. In the beginning, all posts appear to have been made by subjects, while after some time, there was a substantial number of posts originating from elsewhere. Not surprisingly, the latter were more supportive of the experiment, and less focused on the perceived damage afflicted by subjects than were the posts made early on. When write access to the blog was cut off, it was due to the overwhelming portion of non-constructive posts. The study involved more than 1700 subjects. Only 30 complaints were filed with the campus support center, and only seven participants demanded that their data be removed from the study (an option everybody was offered in the debriefing statement.) The complete contents of the blog can be accessed at [8].

6 The Issue of Timeliness

A very important issue in this research is the timeliness of the study in regards to the current types of actual phishing attacks, in terms of the IRB review process, and in terms of arriving at an approach that adequately addresses the legal issues. There has been a clear increase in the degree of sophistication in the methods that phishers use to attack consumers. Phishers are continually designing new ways to execute their attacks on users. Phishing research must stay abreast, and ahead, of the scammers in terms of the sophistication and type of phishing strategy, otherwise the research cannot come up with up-to-date ways to defend against these attacks and protect both users and providers. Slow IRB review, or slow legal approval from the researcher's may render the

research obsolete.

Moreover, many studies may depend on some technical aspect of software or hardware involved in the study; if this is updated or replaced, whether fully or in part, then this may render the study meaningless. This is not to say that the studied aspect will be rendered meaningless, as it may only be technical aspects keeping the experiment (and not the attack) from being performed. In the third experiment, as described above, the researchers were faced with frequent changes of the the format of webpages served by eBay, causing a significant increase of the effort to develop and test the software for the study. Given the delay of approval of this study, this effort may again be revisited.

Conclusion

We have outlined ethical and technical intricacies associated with performing research to assess the threats of phishing. Such research, we argue, is important in that it allows the development and testing of hypotheses and countermeasures. The questions it gives rise to, however, are thorny and remain far from addressed by this first effort to understand the complex issues.

References

[1] R. Akavipat and M. Jakobsson, "Understanding man-in-the-middle attacks in electronic commerce," Study 05-10257, Indiana University, Amended September 16, 2005.

[2] The Belmont Report. Office of the Secretary. "Ethical Principles and Guidelines for the Protection of Human Subjects in Research," National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. April, 1979.

[3] L. Berkowitz and E. Donnerstein, "External validity is more than skin deep: Some answers to criticisms of laboratory experiments," American Psychologist, 37, 245-257, 1982.

[4] B. Cate, "Legal Considerations of Phishing Experiments," In Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft, Jakobsson and Myers (Ed.), Wiley, 2006.

[5] Code of Federal Regulations. Title 45: Public Welfare Department of Health and Human Services, Part 46: Protection of Human Subjects. June, 2005.

[6] M. Earleywine and P.R. Finn, "Compensatory response to placebo varies with personality risk for alcoholism and drinking habits," International Journal of Addictions, 29, 1994, pp 583–591.

[7] P.R. Finn, "The Ethics of deception in research, " Contributing author. In R.L. Penslar (Ed.) Research Ethics: Cases & Materials, Indiana University Press, 1995, pp. 87–118.

[8] T. Jagatic, N. Johnson, M. Jakobsson and F. Menczer, www.indiana.edu/ phishing/social-network-experiment/, May 2005.

[9] T. Jagatic, M. Jakobsson and Sid Stamm, www.indiana.edu/ phishing/browser-recon/, July 2005.

[10] T. Jagatic, N. Johnson, M. Jakobsson and F. Menczer, "Social Phishing," To Appear in the Communications of the ACM, 2006.

[11] M. Jakobsson, "Modeling and Preventing Phishing Attacks." Phishing Panel in Financial Cryptography '05. 2005, paper available at www.markus-jakobsson.com

[12] Markus Jakobsson and Steve Myers. "Phishing and Counter-measures: Understanding the Increasing Problem of Electronic Identity Theft." Wiley, 2006.

[13] M. Jakobsson and J. Ratkiewicz, "Designing Ethical Phishing Experiments: A study of (ROT13) rOnl auction query features," WWW '06, 2006.

[14] M. Jakobsson and S. Stamm. "Invasive Browser Sniffing and Countermeasures." WWW '06, 2006.

[15] B. Latane and J.M. Darley, "The Unresponsive Bystander: Why Doesn't He Help?," Prentice-Hall, Englewood Cliffs, 1970.

[16] M.T. Orne, " On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications," American Psychologist, 17, 776-783, 1962.

[17] S. Stewart, P.R. Finn and R.O. Pihl, "A dose-response study of the effects of alcohol on the perception of pain and discomfort due to electric shock in men at high familial-genetic risk for alcoholism," Psychopharmacology 119, 1995, pp.261–267.

[18] E.T. Webb, D.T. Campbell, R.D. Schwartz, L. Sechrest and J.B. Grove, "Nonreactive measures in the social sciences," Boston, MA: Houghton Mifflin, 1983.