Modeling and Preventing Phishing Attacks

Markus Jakobsson*
School of Informatics
Indiana University at Bloomington
Bloomington, IN 47408

Abstract

We introduce tools to model and describe phishing attacks, allowing a visualization and quantification of the threat on a given complex system of web services. We use our new model to describe some new phishing attacks, some of which belong to a new class of abuse introduced herein: the *context aware* phishing attacks. We describe ways of using the model we introduce to quantify the risks of an attack by means of economic analysis, and methods for defending against the attacks described.

Keywords: context aware, identity linking, model, phishing, social engineering, security.

1 Introduction

Traditional security has focused mainly on authentication and encryption, with inroads to topics like privacy, robustness, and security against mobile adversaries. In all such cases, the security modeling has ignored the human factor and the impact on security that such attacks may have when combined with social engineering. The recent tide of so-called *phishing attacks* gives ample evidence that it is necessary to include the human factor in security modeling. These are attacks in which, typically, the victim is deceived to give out secret information such as passwords or other information enabling access to a given resource. Even though most attacks are surprisingly straight-forward – such as point-blank asking a victim for his bank account number and PIN – they are also rather successful. A recent study by Gartner (April, 2004) shows that around 3% of all those surveyed reported giving up financial or personal information in a phishing scam. While it is likely for the very straight-forward attacks to become less successful as public awareness increases, phishing attacks are also likely to become more sophisticated in response. Also, with the development of do-it-yourself kits for phishing [18], most anybody who wants to can become a phisher. Here, we should note that while the term phishing typically is used for automated attacks that are performed en masse, we extend the use of the term to also mean automated attacks that target smaller sets of victims, but which succeeds with much higher probabilities. We believe this is a likely course of events, as it involves taking advantage of partial information of potential victims. Apart from increasing the success ratio, this approach will lower the visibility of the attacks. In particular, by performing targeted attacks, phishers may to a much larger extent avoid phishing honeypots; these are identities and accounts created solely for the purpose of attracting attackers, and are used by service providers to detect from where attacks are performed.

*Part of the work was done while at RSA Laboratories. The author can be reached at markus@indiana.edu, and a copy of this paper can be downloaded from www.markus-jakobsson.com.

Improved public awareness of threats is a necessary component in building a system that is secure against phishing attacks. Not surprisingly, though, public awareness is not sufficient, but must be accompanied by the development and employment of technical security mechanisms. These, in turn, must be based on a solid understanding of the threats – both current ones and potential future ones.

A *first contribution* of this paper is a theoretical yet practically applicable model covering a large set of phishing attacks, aimed towards developing an understanding of threats relating to phishing. We model an attack by a *phishing graph* in which nodes correspond to knowledge or access rights, and (directed) edges correspond to means of obtaining information or access rights from already possessed information or access rights – whether this involves interaction with the victim or not. Edges may also be associated with probabilities, costs, or other measures of the hardness of traversing the graph. This allows us to quantify the effort of traversing a graph from some starting node (corresponding to publicly available information) to a target node that corresponds to access to a resource of the attacker's choice. We discuss how to perform economic analysis on the viability of attacks. A quantification of the economical viability of various attacks allows a pinpointing of weak links for which improved security mechanisms would improve overall system security.

We describe our graph-based model in detail, both in its generic incarnation and using specific examples. Several of these examples correspond to possible phishing attacks against prominent web services and their users. Among these examples, we show how in certain cases, an attacker can mount a man-in-the-middle attack on users who own their own domains, in turn allowing for very effective attacks on most any web service account used by a victim of such an attack. We also discuss how an attacker can obtain access to a newly opened online bank account using a phishing attack.

A *second contribution* of this paper is the description of what we term a *context aware* phishing attack. This is a particularly threatening attack in that it is likely to be successful *not only* against the most gullible computer users (as is supported by experimental results we present.) A context aware attack is mounted using messages that somehow – from their context – are expected (or even welcomed) by the victim. To draw a parallel from the physical world, most current phishing attacks can be described as somebody who knocks on your door and says you have a problem with your phone, and that if you let him in, he will repair it. A context aware phishing attack, on the other hand, can be described by somebody who first cuts your phone lines as they enter your home, waits for you to contact the phone company to ask them to come and fix the problem – and *then* knocks on your door and says he is from the phone company. We can see that observing or manipulating the context allows an attacker to make his victim lower his guards. As a more technical example, we show how to obtain PayPal passwords from eBay users that do not take unusual measures *particularly intended* to avoid this attack. Many of our attacks take advantage of a method we may call *identity linking*. This is a general phishing technique we introduce, by which an attacker determines how identities and email addresses of a victim correspond to each other.

Finally, a *third contribution* is a discussion of how to address the threats we describe – both in their specific and generic shapes.

Outline: We begin by reviewing the related work in section 2. In section 3, we describe our graph-theoretical model of phishing attacks, and describe some novel threats and attacks to illustrate how these are viewed using our model. In section 4, we introduce context aware phishing attacks. We describe these both using our graph-based approach and by the use of examples; these describe attacks on users of eBay/PayPal (subsections 4.1 and 4.2) and online banking (subsection 4.3). In section 5, we address how to measure and mitigate the impact of attacks. We conclude in section 6 by reflecting on what remains to address in order to gain a still better understanding of phishing attacks.

2 Related Work

Phishing can be described as the marriage of technology and social engineering. Whereas phishing attempts can be successful in a situation where one of these components are dominating over the other, it is likely to be the case that the success rate would much increase when the attacker uses both of these components in a strategic manner. This means that in preventing phishing attempts, one should understand both components. We will look at the related work with this in mind.

Many phishing attempts use domain spoofing or homographic attacks [6] as a step towards persuading victims to give out personal information. Preventing such attacks is an important step towards defending against phishing attacks. There are several promising approaches to this problem involving certification, e.g., [12, 14, 16, 20]. A recent and particularly promising solution [9] proposes to combine the technique of standard certificates with a visual indication of correct certification; a site-dependent logo indicating that the certificate was valid would be displayed in a *trusted credentials area* of the browser. Another recent and promising approach [19] detects certain common attack instances, such as attacks in which the images are supplied from one domain while the text resides with another domain, and attacks corresponding to misspellings of URLs of common targets.

At this point, however, none of the above approaches are commonly employed, and all rely on the awareness of the user to some extent. Our approach, in contrast, is to analyze the system security in situations where user awareness may be insufficient, and to make sure that protocols are strengthened in order to avoid system vulnerabilities. The two approaches therefore complement each other well.

A common phishing attack is for an attacker to obtain a victim's authentication information corresponding to one website (that is corrupted by the attacker) and then use this at another site. This is a meaningful attack given that many computer users reuse passwords – whether in verbatim or with only slight modifications. A technique to protect against this type of attack was recently proposed [17], and relies on *local and automated* scrambling of passwords on a site-by-site basis, performed by a plugin.

Large-scale phishing attacks typically rely on spamming of users to reach victims. Therefore, anti-spam methods (see [10] for an overview of proposals) play an important role in defending against phishing attacks. However, when the interaction with the victim is done via a proxy (as the attacks we describe), then standard anti-spam tools do not provide any protection. The attacks we detail herein can employ automated web-based scripts to mount very large-scale attacks.

It should be emphasized that none of the above defense techniques – spoofing detection, password scrambling, or spam filters – will completely make phishing attacks impossible to perpetrate. Instead, they provide valuable but scattered road blocks blocking the way of the attacker.

In terms of the social engineering aspect, it is worth noting that most of today's phishing attempts try to make victims give out personal information by intimidating users and creating fear – a common example is "we need you to confirm your account details or we must shut your account down". Another, but not quite so common, avenue is by baiting the victim with a desired event, e.g., "free premium-upgrade for the first 100 to log in to their account and answer a brief questionnaire." An approach we believe will become more and more common is what we call the context aware attack: this is a more complex approach in that it does not only use threat or enticement, but makes the victim think of the messages as *expected*, and therefore legitimate. We believe that there is a multitude of context aware techniques that could be developed and used by phishers, and that if phishers were to learn from Park Avenue the strategy of convincing consumers, then such attacks would become absolutely vital to defend against. We note that our framework is not specific to any particular delivery method of the attack; in particular, it can be used to describe both HTML emails containing forms, and emails with pointers to spoofed sites – these are the two currently most common delivery methods.

When analysing the viability of an attack, it is meaningful to perform an economic analysis involving the probability of success, the costs of the required actions, and an estimate on the profit that is required to make the attack meaningful. Such analysis [11] has been performed on the viability of spam in situations where proof-of-work [4, 8] is part of the defense mechanisms.

Looking at the big picture, phishing attacks are the result of misalignment of technology to the social environment in which they will be used. This is a topic studied by Anderson [1] – while his work does not address phishing attacks, it is "spiritually connected" to ours in this general sense.

3 Modeling Phishing Attacks

In this section, we introduce a model for capturing the essence of phishing attacks. The main benefit of the model is the possibility of describing a variety of attacks in a uniform and compact manner, along with the overview of potential system vulnerabilities and defense mechanisms that this approach offers us. Apart from describing the model, we illustrate it using two example attacks, which both may be of potential independent value to understand and defend against.

3.1 A Graph Model

Representing access. We will model phishing attacks by graphs, which we will refer to as *phishing graphs*. In such a graph, there are two types of vertices; those corresponding to access to some *information*, and those corresponding to access to some *resource*. For notational simplicity, we will not distinguish between these two types of vertices, but rather let the type of a vertex be implicit from the description associated with it.

Representing single actions and disjunctions. We represent actions by edges in the graph. Two vertices are connected by an edge if there is an action that would allow an adversary with access corresponding to one of the vertices to establish access corresponding to the other of the vertices. In particular, we let there be an edge e_{12} from vertex v_1 to vertex v_2 if there is an action A_{12} that allows a party to obtain access corresponding to v_2 given v_1. Two vertices may be connected by multiple edges, corresponding to different actions allowing the transition between them; this corresponds to a disjunction.

Representing conjunctions of actions. When access corresponding to multiple vertices (corresponding to a set V_1) is needed to perform some action resulting in access to some vertex v_2, then we represent that as follows: We start one directed edge in each one of the vertices of V_1, merging all of these edges into one, ending in v_2. This allows us to represent a conjunction of actions required for a given transition, as is illustrated in figure 1.

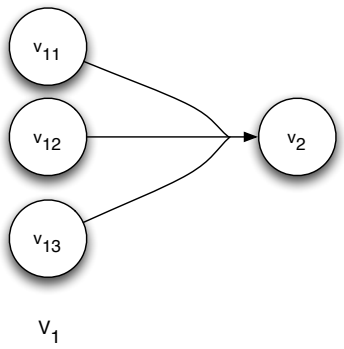

Figure 1: Conjunction: An attacker needs to perform *all* the actions corresponding to the edges leaving $V_1 = \{v_{11}, v_{12}, v_{13}\}$ in order to reach v_2.

Representing a successful attack. We let some set of nodes correspond to possible starting states of attackers, where the state contains all information available to the attacker. (This may simply consist of publicly available information.) We let one node correspond to access to some resource of the attacker's choosing, call this the target node. In order for an attack to be successful, there needs to be a path from a starting state to the target node.

Representing effort, probabilities, and conditions. It is meaningful to label the edges with descriptions of the circumstances under which the action will succeed. One aspect is the *effort* of the action, whether computational; the degree of human involvement necessary; the time the action takes, or similar. Another is the *probability of success* of the action. For example, if the action involves guessing some variable, then we can label the edge with the success probability of correctly guessing it. There may also be *conditions* that are not within the control of the attacker that influence the success of the action; for example, it may be that certain actions only are meaningful for accounts that have been activated but not accessed.

A simplified description. In the next two subsections, we illustrate the model by showing how example attacks can be visualized as traversals of graphs. After having introduced and described context aware attacks in section 4, we turn in section 5 to take a birds-eye view of possible attacks. This more general view of the problem affords us the ability to defend against (known or unknown) attack that falls into the categories captured by the model. For ease of exposition, we present simplified graphs for the attacks we describe, even though this does not quite do justice to our model. We sometimes combine all the nodes and edges of some subgraph into one single node; while this de-emphasizes the manner in which the information of that node is obtained, it also provides us with a more accessible description. We note that a given attack can be described in several ways, where one graph is a simplification of another more detailed one. When the goal is careful analysis of a given attack, the most detailed representation will be used.

3.2 An example: obtaining fraudulent access to a known bank account.

When a person has just opened a bank account, but not yet established access to the on-line bill payment service then the account is vulnerable to attack given the mechanisms used by some banks, such as [2, 5]. In particular, this is the case if the account owner does not establish access for an extended period of time. Namely, many banks allow the account owner to gain initial access to the account by using – instead of a password – the date and amount of the last deposit to the account. The amount could be determined by an attacker under many circumstances:

1. The attacker may know the salary of the victim, and know or guess what percentage of the salary is withheld or contributed to a 401(k) plan. Given the relatively limited number of choices, he has a decent probability of success in guessing the exact amount deposited for each pay period – assuming, of course, that the victim uses direct deposit.
2. At times when all taxpayers are given a refund whose size only depends on the marital status of the taxpayer, then an attacker simply has to know or guess whether the victim is married, and put his hopes in that no other check was deposited at the same time as the refund was; note that this would allow for a simple way of attacking a large number of account holders with a very small effort.
3. The attacker may have performed a payment to the victim, whether by personal check or PayPal; in such situations, he would know the amount and the likely time of deposit.

Similarly, the date can be inferred in many cases, or the attacker could exhaustively try all plausible deposit dates. In all of the instances above, the attacker is assumed to know the account number of the victim, which can be obtained from a check, and which is required for automated payments and credits of many services.

A graphical representation. In figure 2, we show a graphical description of the above attack. Access to the account is represented by vertex v_1. Knowledge of the victim's salary is represented by vertex v_2, and the edge e_{21} corresponds to guessing the level of withholding and percentage of 401(k) contributions. Knowledge of the victim's marital status corresponds to vertex v_3, and the edge e_{31} is labelled with the probability of the tax refund check being deposited alone. (This

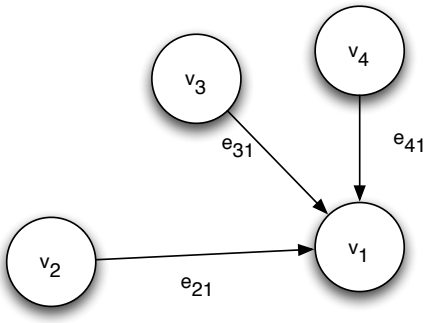

Figure 2: A simplified graphical representation of a phishing attack on a bank account. Nodes v_2, v_3 and v_4 correspond to possible starting states, and v_1 to the target node.

probability is not likely to be known, but could be estimated given access to statistics.) Finally, vertex v_4 corresponds to *access to* performing a payment to the victim (i.e., purchasing something from the appropriate seller, or performing a refund for a purchased item or service.) The edge e_{41} corresponds to the action of performing the payment.

All edges are conditional on the online bill payment service not having been accessed yet, and for all attacks, it is assumed that the attacker knows the account number of the account holder. To model this in more detail, we may represent the knowledge of the account number by a vertex v_5, and change the edges as is shown in figure 3.

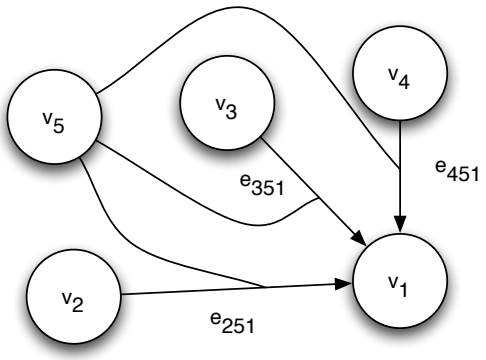

Figure 3: This is the same attack as described in figure 2, but with v_5 representing knowledge of the account number. A more detailed description would also carry information about the cost of performing the actions associated with the different edges.

Remark: We mentioned that v_4 corresponds to the access to performing a payment to the victim, but did not describe how this is achieved. We have also assumed all along that the attacker would know the account number of the victim, and did not describe how this could be obtained. In section 4, we describe attacks that are self-contained in that they are not based on any assumptions on initial knowledge of the attacker.

Connecting personal graphs. In our example above, we have described the graph that corresponds to one particular victim, with vertices all corresponding to access to information or resources associated with one given person: the victim. There are instances, though, where the graphs corresponding to two or more people may be connected. For example, if one of the vertices in a graph corresponds to the knowledge of a person's mother's maiden name, then these vertices of the graphs of two siblings are connected: knowledge of the mother's maiden name of one of them immediately leads to knowledge of the mother's maiden name of the other. There are more – but less obvious – connections of this type.

3.3 An example: Performing a man-in-a-middle attack.

Most domain name registration sites require domain name administrators to authenticate themselves using passwords in order to obtain access to the account information. However, many people forget their passwords. A common approach to deal with this problem is to email the password to the email address associated with the account, but other solutions are also in use. In one case [13], the site instead asks the user to specify what credit card is used to pay the registration fees – since many people use multiple cards, the two last digits are given as a hint. This turns out to open up to an attack. In particular, an attacker can do as follows:

1. Determine the name of the administrative contact associated with a domain name – this is public information, and can be obtained from any domain name registration site.
2. Obtain a list of credit card numbers associated with the administrative contact, and select the one ending in the two digits given as a hint. This can be done in a variety of ways, and may be performed by offering the victim to buy a desirable item at a great price – using a credit card. Some sites request users to give out their credit card numbers under the premise of age verification.
3. Obtain access to the account information, and replace the email forwarding address with an address under the control of the attacker. Now, all emails sent to the victim domain will be forwarded to the attacker.
4. Forward emails (potentially selectively) to the destination they would have been sent to if the rerouting would not have occured, spoofing the sender information to hide the fact that the rerouting took place.

We note that the attacker now can read all the emails sent to users at the victim domain, as well as remove or modify such emails. In effect, he receives all the victim's email, and decides what portion of this that the victim gets to see. That means that the attacker can claim to have forgotten passwords associated with any services to which users of the attacked domain subscribe (as long as they use an email address of the attacked domain to sign up.) Systems that respond to such a request by sending the password to the email address associated with the account will then send this information to the attacker, who will of course not forward this to the real owner, so as to hide that somebody requested the password to be sent out.

A graphical representation. Let us now see how we can represent this attack using a phishing graph. In figure 4, vertex v_1 corresponds to knowledge of the name of the administrative contact of a domain to be attacked. Vertex v_2 corresponds to knowledge of the appropriate credit card number, and vertex v_3 to access to the account. Finally, v_4 corresponds to knowledge of a service for which a user in the attacked domain is registered as the administrative contact, and where passwords are emailed to administrators claiming to have forgotten the passwords, and v_5 to access to the account of such a site. There is an edge e_{12} corresponding to the action of finding out credit card numbers associated with a person with a given name. Edge e_{23} corresponds to the action of using the correct credit card number to authenticate to the site, and edge e_{345} to requesting a forgotten password to be emailed. Note that both v_3 and v_5 may be considered target nodes.

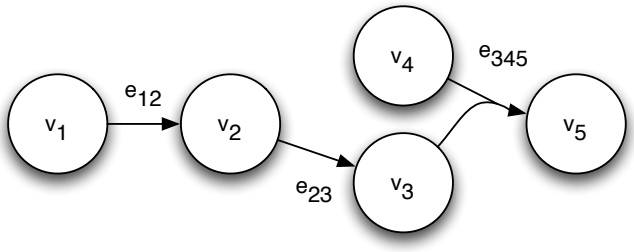

Figure 4: A simplified graphical representation of a man-in-the-middle attack on a domain name server. A detailed representation would also have labels on edges corresponding to effort, probability, and other costs.

Remark: Another common approach to deal with forgotten passwords is to rely on so-called security questions. This is, for example, used at PayPal, where the four possible questions relate to the mother's maiden name; city of birth; last four digits of social security number; and last four digits of drivers license number. The mother's maiden name of a person can be obtained from publicly available documents and services, using a set of clever queries. For example, if a woman has one name when moving to a given address, and another name when moving out, then chances are that the first name is her maiden name, and it would be clear what the mother's maiden name of any of her children is. Consecutive records linking names to addresses or other stable pieces of information can be obtained from records of memberships to organizations, mortgage documents, voters registration records, marriage licences, public litigation files, and other publicly available services, such as [7].. Such records may also be used to determine with a good likelihood the city of birth of a person: by knowing the names of his or her parents, and determining where they lived at the time of the victim's birth. A person's social security number can often be obtained from records of types similar to those from which mothers maiden names can be derived. In addition, if a user enters the answers to any of these questions at a rogue site (for the same purposes: password security questions) then this site has immediate access to the information.

4 Context Aware Attacks

Somewhat vaguely stated, a context aware attack is a phishing attack that is set up in a way that its victims *naturally* will believe in the authenticity of the messages they receive. A little bit more specifically, a context aware attack uses timing and contex to mimic an authentic situation.

In a first phase, the attacker infers or manipulates the context of the victim; in a second phase, he uses this context to make the victim volunteer the target information. The first phase may involve interaction with the victim, but will be of an innocuous nature, and in particular, does not involve the request for any authentication. The messages in the second phase will be indistinguishable by the victim from *expected* messages, i.e., messages that are consistent with the victim's context. We may also describe context aware attacks with a phishing graph where some set of nodes correspond to the first phase, and some other set to the second phase. The edges associated with nodes of the first phase would correspond to actions that are *harmless* in isolation; the edges associated with the second phase would correspond to actions that – by nature of being expected by the victim – do not arouse suspicion.

We will now give a few examples of context aware attacks. We will describe them as if a *person* performed the actions and there is *one* victim of the attack, whereas in reality, the attack may just as well be performed by an agent (or a script) and, therefore, allow a tremendous number of victims to be targeted simultaneously with a negligible effort.

4.1 An example: A context aware attack on an eBay bidder.

By making a bidder in an auction believe he is the winner, the attacker hopes to have the bidder reveal his password by interacting with a website that looks like PayPal. This attack consists of several steps, which we will describe in detail, and can be described by the graph in figure 5.

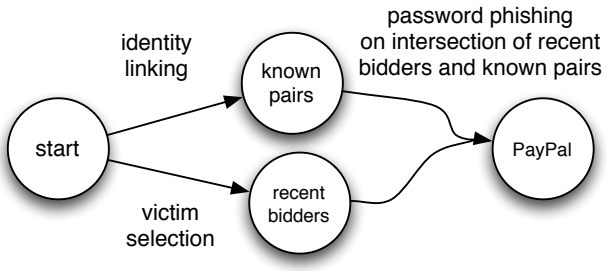

Figure 5: The attacker performs identity linking in order to obtain pairs of email addresses and eBay user identities. He selects victims with high bids close to the end of the auction, then performs password phishing by emailing selected victims a congratulation with included payment instructions. The attack is performed on victims that have been selected both in the recent bidder selection, and for whom identify linking has succeeded.

1. **Identity Linking.** The attacker wishes to learn relationships between the email addresses and eBay user identifiers for a set of potential victims. This set consists of people for whom there is an indication that they may win a given auction which is selected by the attacker;

their likelihood in winning the auction may be established from previous interest in (and bidding on) similar items; current bids for the item of the auction in question; or auxiliary information known by the attacker.

In some cases, it is trivial to establish this link – namely when the victim uses his email address as an eBay user name, or otherwise displays his email address on a page where he is the seller of an item. (Recall that the history of a user is publicly accessible, and contains information about all recent transactions, including information on the identifiers of the seller and winner, as well as a pointer to the item being sold. This allows an attacker to follow such links and obtain user information from the page where the victim is the seller.)

In other cases, this information has to be obtained by means of interacting with the potential victim. We will describe three ways of performing this step; we refer to these as the *inside-out* linking, *outside-in* linking, and *epidemic* linking. For now, and for the simplicity of the disposition, we will assume that the attacker establishes this link for a large set of selected victims of his choosing. We note that once such a link is established, it will, of course be kept. That means that the linking may be performed for another transaction than that for which the second phase of the attack will be performed.

2. **Victim Selection.** Directly after a a targeted auction has ended, the attacker selects a set of victims for whom he has established a link between the user email address and the user eBay identifier. He only selects victims who are *plausible* winners, e.g., who have been the highest bidder at a recent time – this information can be obtained by constant monitoring of the auction page, where it is specified who is the highest bidder at that time. The actual winner is not selected.

3. **Context Aware Password Phishing.** For each selected victim, the attacker sends an email containing a congratulation that the victim won the auction in question. The attacker states a winning bid that is plausible to the victim, e.g., coincides with the latest bid made by the victim. The sender of the email is spoofed[1], and appears to be eBay, and the payment button is associated with a link to a page controlled by the attacker. This page appears just like the PayPal page the user would have seen if he indeed *were* the winner, and he followed the valid link to perform the payment.

In the above description, we have left out how linking is performed. We will now describe this. An inside-out linking attack starts with knowledge of a user identity *inside* a given application (such as an eBay user identity or bank account number) and strives to link this with an *outside* identity (such as an email address, a name or address, social security number, etc.). Conversely, an outside-in attack starts with a publicly known identifier from the outside, and aims to obtain an inside identifier. (Of course, the notion of what is the inside and what is the outside may be subjective, but this only affects the naming of these two attacks, and not their success.) Finally, *epidemic* linking uses the fact that the pairs of eBay identifiers and email addresses are kept in the history of a user's PayPal account. Therefore, each time an attacker successfully compromises the PayPal password of a victim, he also obtains new pairs of identifiers; the more of these he has, the more users he can attack.

[1]Since there are security mechanisms proposed to detect address spoofing in order to defend against phishing attacks, it is worthwhile to point out that the attacks are still likely to succeed even if no spoofing is performed. More precisely, we believe that a large portion of users will attempt to perform a payment at a site with a name entirely unrelated to PayPal, as long as the context makes sense. We have, however, no numbers on the actual percentages.

Inside-out linking. An attacker can obtain the email address of a victim whose eBay user identifier he knows. This can be done in several ways. One is already mentioned: if the victim poses as a seller in some active auction, then the attacker can place a bid for the item, after which he can request the email address of the seller using the available interface. (This can be done in a manner that hides the identity of the attacker, as he plainly can use an eBay account solely created for the purpose of performing this query.) An alternative way is to obtain the history of the victim, then email the victim (using the supplied interface) to ask him a question about a buyer or seller that the history specifies that the victim has done business with. Many people will respond to such a question without using the provided anonymous reply method, thereby immediately providing the attacker with their email address. A victim who has set up the out-of-office automated reply will also *automatically* provide the attacker with the link.

Outside-in linking. An attacker can obtain the eBay identifier of a victim for whom he knows the email address in many ways. An example is as follows: The attacker sends an email to the victim, spoofing the address of the sender to make the email appear to come from eBay. The email plainly informs the user of the importance *never* to enter any authenticating information (such as passwords or mother's maiden name) in an editable field in an email, as these are commonly used by phishers. To acknowledge that the user has read this warning, he is requested to enter his *eBay user identifier* in an editable field (noting that this is *not* a piece of authenticating information.) Alternatively, and perhaps less suspicious, the victim may be asked to go to the verification site pointed to by the email and enter his identifier there. This site will be controlled by the attacker; note that the name of the page pointed to can be unique to a given victim, so the attacker will know what email address the entered information corresponds to even if the *email address* is not entered.

Note that the user will never be asked to enter his password, and will therefore believe that this is an authentic (and reasonable) warning, and that he is acknowledging that he has read the information in a way that is secure. As soon as the attacker obtains the eBay user identifier, he has established the desired link. Note also that this attack works with any type of "inside" information, such as bank account number or full name – anything that the victim will believe is not secret.

4.2 An example: A context aware attack on an eBay seller.

By making a seller believe that he was paid in a manner he does not wish to be paid, the attacker hopes to have the seller reveal his password, again by interacting with a website that appears to be PayPal. This is done as follows:

1. **Victim Selection.** The attacker identifies a seller that accepts PayPal, but does not accept PayPal payments that come from a credit card[2] . (This information is typically available on the page describing the item for sale.) Such a seller will be selected as the victim of the attack.

[2]These are more expensive to the payee, and so, many users do not accept such payments.

2. **Identity Linking.** The attacker obtains the email address of the victim in one of the ways described in the previous subsection, preferably by querying or inside-out linking.

3. **Context Aware Password Phishing.** Right after the end of the auction, the attacker obtains the eBay identifier of the winner of the auction. The attacker then sends an email to the victim, with a spoofed sender appearing to be PayPal. The email states that the winner of the auction has just paid the victim, but that the payment is a credit card backed payment. The victim may either refuse the payment (in which case the buyer has to find an alternative way of paying) or upgrade his PayPal account to accept credit card payments. In either case, of course, the victim has to log in to his account. A link is provided in the email; this leads to a site controlled by the attacker, but appearing just as the PayPal's login page.

Note that it does not matter whether the victim does not attend to the proposed payment refusal or account upgrade before he receives the *real* payment. If this were to occur, then the seller would feel bound to refuse the undesired payment, which completes the attack if performed through the link in the email from the attacker.

4.3 An example: A context aware attack on online banking.

The most common type of phishing attack today does not target eBay/PayPal users, but rather, bank customers who use on-line banking. The typical phishing message appears to come from a bank (that the attacker hopes the victim uses), and requests the user to update his or her information, which requires him or her first to log in. The site to which to log in may, like the email, appear to be legit, but is under the control of the phisher.

In a context aware version of the above attack, the phisher would determine relationships between potential victims, and then send emails to his victims, appearing[3] to originate with friends or colleagues of the victim. The context, therefore, is the knowledge of who knows whom, and as before, identity linking is a crucial part of carrying out the attack. The email that the phisher sends could be along the lines of

> "Hey, I remember that you bank with Citibank. I was down there this morning, and the clerk I spoke to told me to update my account information real quick, because they are updating their security system. They said they would email all account holders later on today, but I wanted to tell you early on, just for your security. I performed my updates, here is the link <obfuscated hyperlink here> in case you don't have it handy. Gotta run, talk to you later!".

In order to be able to mount such an attack, it is sufficient for the phisher to obtain information about who knows whom, at least in terms of their respective email addresses. This can be automatically inferred from public databases, such as Orkut [15]. In order to be able to avoid that the

[3]Spoofing the address to make it appear to come from an acquaintance is, of course, no more difficult than making it appear to come from a bank, and has the additional benefit of being less likely to be automatically labeled as spam by the victim's mail handler.

victim replies to the believed sender of the email (which might make the attack less successful, or at the very least, might make it difficult to also target the latter user), the attacker could specify a reply-to address that makes sure any reply neither is delivered or bounces.

While we have not performed any tests to determine the success rate of an attack of this type, we anticipate that it is going to be substantially higher than the corresponding attack that does not rely on context.

5 Analysis and Defense

5.1 Analysing the General Case

Our special cases only consider particular ways to obtain access to the PayPal account of an eBay bidder or seller. It makes sense to consider these two attacks in conjunction, since many eBay users act both as bidders and sellers over time. Moreover, one should also consider *other* ways for an attacker to obtain access to such a victim's PayPal account, as well as other resources. Let us only consider attacks on PayPal for concreteness and ease of exposition.

Alternative attacks. One alternative way of gaining unauthorized access to a victim's PayPal account was described briefly at the beginning of the paper: to bypass the password by knowing (or successfully guessing) the password security questions. Here, one should recognize that many services may have similar password recovery questions. Thus, even if a user is security conscious and does not reuse *passwords* (or uses a mechanism such as [17]), there is a risk of reuse of *other* authenticating information (such as password security questions) entered into a rogue or corrupted site. This further grows the phishing graph to now incorporate multiple sites and services, and their respective vulnerabilities.

Relations between users. One must also address relations between users. We used the example of mother's maiden names, noting that two siblings would have the same answers to this question, and often also to the question of birth city. This still again grows the phishing graph by adding edges between subgraphs belonging to different users, and we see that one must be concerned not only by leaks of information belonging to a particular user one wants to protect, but of leaks of information belonging to *other* users as well. This is where the graphical representation of the problem is likely to start making a difference in analysing the threat: when the complexity of the threat grows beyond what can be described in a handful of paragraphs.

How can an attacker succeed? For the general case, one should make the best possible attempt to be exhaustive when enumerating the possible attacks. The attacks we have focused on all aim at *obtaining* the password from the victim or *bypassing* the use of passwords (using the password security questions or information such as the last deposits.) One could also consider the possibility of an attacker *setting* the password, as in the example of the domain registration service.

What is the probability of success? When access to a resource depends on some information that can be guessed with a reasonable probability, this can be thought of as a cost of traversal. If the number of possible tries is limited (say, three, after which access is turned off) then the cost is a probability of success; if it is unlimited, then the cost is the computational and communication effort.

In the case of PINs, there is often a limitation on the number of attempts – if three attempts are allowed for a four digit PIN, then this corresponds to a probability of 0.3% of success (in the worst case, when the distribution is uniform.) An attacker can start with a very large pool of potential victims, and then narrow this down based on for which ones he succeeds in guessing the PIN. For these remaining victims, he would then perform the other steps of the attack. Seen this way, the cost can also be seen as the portion of selected victims that remains after a weeding process – the action of guessing their PIN.

Translating limitations into probabilities. In the case of security questions, there is often not a limitation on the number of tries, except that the duration of the attack must be taken into consideration in cases where the entropy of the access information makes exhaustive attacks impractical; this could then be translated into a success probability per time unit of the attack. (We can assume that all attacks are limited in time, given that if they are successful against a large enough number of victims, then defense mechanisms will be put into place.) Similarly, in cases where there are detection mechanisms in place to determine a large number of queries from a particular IP address, one can quantify this in terms of a probability of detection (that may be very hard to estimate) or, again, as a probability of success (before the detection mechanisms are likely to be successful.)

Computing the success probability. We will focus on the costs that can be translated into a probability of success or portion of remaining victims after weeding. We will consider all paths from all starting nodes to the target node, and determine the probability of success associated with each such path. Of course, the probability associated with a conjunction of two or more actions or the sequence of two or more actions is the product of the individual probabilities[4] associated with these actions. Similarly, the probability associated with a disjunction of two actions is the maximum of the two probabilities.

We can determine whether a given attack is feasible by seeing whether the probability of success is sufficiently high that a given minimum number of victims can successfully be found from the population of possible victims.

Economic analysis of threats. When determining whether a given threat must be taken seriously, one should consider the costs of performing the attack and relate these to the likely payoff of performing the attack. It is meaningful to assume that the attacker is rational, and that he will

[4]In cases where the exact probability of some action is not known, one can of course use an estimate on the upper and lower bound of these instead, thereby obtaining an estimated upper and lower bound on the probability of success of the entire attack.

only attempt a given attack if the difference between the expected payoff and the expected cost is greater than zero.

To understand the likelihood of an attack being mounted, one needs to consider the probability of success of the attack or the number of victims for which the attack is expected to succeed. One must also consider the equipment costs related to performing the attack, which are related to the computational costs and communication costs of performing a sufficiently large number of actions for the attack to succeed for a given number of victims. Finally, it is important to take into consideration the costs relating to any potential human involvement (such a performing tasks not suitable for machines, e.g., [3]) and the minimum profit required by the attacker.

Looking back at our graphical representation. A detailed description of attacks on a given resource would have graphical components corresponding to all the possible ways in which the resource could be compromised. The context aware attacks we have described would therefore correspond to subgraphs within this graphical description. There are, of course many other types of attacks. For completeness, the graphical description of the threat against a given user should contain components corresponding to *all* known types of attacks on *each* resource associated with the victim. This is of particular relevance given that he relation between access information used by different services is stronger than what might at first appear – both due to password reuse and commonalities between password security questions. This establishes links between access rights to different services: if an attacker obtains access to a first resource, this often gives him an advantage in obtaining access to a second. Edges corresponding to similarities between passwords and password security questions would be labeled by the probability of successfully guessing one password given knowledge of another. While the exact probabilities may not be known, estimates are meaningful in order to determine whether this introduces a weak link in a given system. To model known types of relations (such as same birth city or mother's maiden name) between users, we simply connect the related subgraphs of such users. All edges will be labeled with their probabilities of success and other costs, such as necessary human involvement. The economic analysis of an attack could then consider relations between services, and between users, and would quantify the cost in terms of the total traversal costs from a starting state of the attacker's choosing to a target node of the attacker's choosing.

5.2 Analysis of One Example Attack

By looking at the graphs corresponding to the attacks we have described, we can analyse various ways of disconnecting the target node from any nodes in the graph that are otherwise reachable by the attacker. We begin by describing experimental results relating to our context aware attack on an eBay bidder (see section 4.1), followed by a discussion on how to defend against this attack.

Experimental results. Using the eBay interface for asking an eBay user a question, we contacted 25 users with the message "Hi! I am thinking of buying an item from XXX, and I saw that you just did this. Did you get your stuff pretty quickly? Would you recommend this seller?". (Here, XXX is the eBay user identifier of a seller from which the recipient had recently bought something.)

We got a reply from 17 of these users, only five of whom used the anonymous reply. Two of the twelve "good" replies were automated out-of-office replies. If this response ratio is characteristic of average users, then 48% of all users would respond to a question of this type in a way that allows for identity linking.

In a survey sent out to colleagues, we got answers to the following two questions from a group of 28 eBay users:

1. Would you be suspicious of domain spoofing when paying following the link to PayPal in the congratulation message for an item you just won?
2. You are bidding on an item, and ten minutes before the end of the auction, If you bid on an item and know you are the highest bidder ten minutes before the end of the auction and if at the end of the auction you get a message from eBay stating that you won, would you be suspicious of its authenticity?

In our small study, only one of the respondents answered yes (and to both questions). If these results also are representative, then more than 96% of all users would be likely to give out their password to a site that looks like PayPal in a context aware attack.

This means that the success ratio of our attack appears to be close to 46%. Given the very low cost of sending email, this makes the attack almost certainly profitable, given that the average amount an attacker can take from a PayPal account is likely to be measured in dollars rather than in fractions of cents. While this result is not statistically significant, and the survey might have resulted in a biased response, it is still an indication of the severity of the attack.

5.3 Defenses Against our Example Attacks

Let us briefly review how one can protect against our example attacks, keeping in mind how this corresponds to a partitioning of the corresponding phishing graph.

- **Protecting the eBay bidder.** Our context aware phishing attack on an eBay bidder relies on being able to determine the email address of a person who is bidding on an item. Currently, the eBay user id of the high bidder is displayed along with the current bid. If only the high bid was displayed, but not the high bidder information, then an attacker would not know what eBay user he needs to obtain the email address for. (This would also protect against DoS attacks on high bidders.)
- **Protecting the eBay seller.** If the attacker is not able to determine the email address of the seller in an auction, then the corresponding context aware phishing attack would fail. Therefore, if all interaction with eBay users would go through eBay, and the associated email addresses would be removed as messages are forwarded, then an attacker would not be able to determine the email address corresponding to an eBay user identity. Of course, if users volunteer their email addresses out of band (such as on the auction site, or in responses) then this line of defense would fail.

- **Protecting the online bank user.** We described how an attacker could use knowledge of personal relations in order to improve the chances of having his victim respond to his request. In order to prevent against this type of attack, one could require that each message be authenticated by the sender, and verified by the receiver. Such an approach requires the existence of an infrastructure for authentication, such as a PKI. Moreover, it requires a higher degree of technical savvy than the average computer user has. A simpler approach would be for the mail programs to verify whether the domain of the originator of the email (as specified by the ISP this party connected to) corresponds to the apparent originator, and alert users of discrepancies. While this allows an attacker in the domain of the claimed sender to cheat, it still limits the threat considerably. If there is a risk that the attacker controls the ISP, the mail program could verify that the entire path corresponds to the *previously observed* entire path for previous emails, and alert users of discrepancies.

6 Future Work

Our understanding of possible threats to a complex system will improve with a more detailed picture of vulnerabilities associated with particular protocols used by service providers. The more we understand such threats, the more edges will we be able to add in an graphical description of access to knowledge and resources, which will build a more complex graph structure with more interdependencies. In turn, as our graphical description of possible phishing threats expand, our ability to analyse threats will decrease. This calls for the use of automated analysis techniques of the types used to analyse individual cryptographic protocols. As our understanding of the costs associated with graph traversal improves, the accuracy of the analysis – whether manual or automated – will improve.

Our work should therefore not be seen as an attempt to fully address the problem of phishing, but rather as a first step in this direction.

Acknowledgements

We want to thank Virgil Griffith, Ari Juels, Filippo Menczer and Susanne Wetzel for helpful discussions and valuable feedback on earlier versions of the paper. Thanks to Virgil and Filippo for suggesting the use of interpersonal relations to improve acceptance of spam.

References

[1] R. Anderson, "Why Cryptosystems Fail," Communications of the ACM, 37(11), pp. 32–40, November 1994.

[2] www.BankOne.com

[3] www.captcha.net

[4] C. Dwork, M. Naor, "Pricing via Processing or Combatting Junk Mail," Proceedings of the 12th Annual International Cryptology Conference on Advances in Cryptology, 1992, pp. 139-147.

[5] www.Fleet.com

[6] E. Gabrilovich and A.Gontmakher, "The Homograph Attack," Communications of the ACM, 45(2):128, February 2002.

[7] www.geneaology.com

[8] www.hashcash.org

[9] A. Herzberg and A. Gbara, "Protecting Naive Web Users," Draft of July 18, 2004.

[10] Internet Mail Consortium, www.imc.org.

[11] B. Laurie and R. Clayton, "'Proof-of-Work' Proves Not to Work", The Third Annual Workshop on Economics and Information Security (WEIS04), 2004.

[12] Microsoft Sender ID Framework, www.microsoft.com/mscorp/twc/privacy/spam_senderid.mspx

[13] www.namesecure.com

[14] S. Olsen, "AOL tests caller ID for e-mail," CNET News.com, January 22, 2004.

[15] www.orkut.com

[16] J. C. Perez, "Yahoo airs antispam initiative," ComputerWeekly.com, December 8, 2003.

[17] B. Ross, D. Boneh, J. C. Mitchell, "A Simple Solution to the Unique Password Problem," crypto.stanford.edu/PwdHash/

[18] www.sophos.com/spaminfo/articles/diyphishing.html

[19] http://crypto.stanford.edu/SpoofGuard/

[20] WholeSecurity Web Caller-ID, www.wholesecurity.com