Countermeasures: Social Networks

Alla Genkina Jean Camp
Doctoral Candidate Associate Professor
School of Informatics
Indiana University
Bloomington, IN

Phishing is a new category of crime enabled by the lack of verifiable identity information or reliable trust indicators on the web. A phishing attack works when it convinces a person to place trust in a criminally untrustworthy party by masquerading as a trusted party. Better indicators about which parties are trustworthy can enable end users to make informed trust decisions and thus decrease the efficacy of phishing. Physical indicators, as embodied in the marble and brass of a bank, cannot be used on the network. Therefore this chapter describes the theoretical underpinning, design, and human subjects testing of a mechanism to use social networks to create reliable trust indicators.

This chapter will first discuss the role and importance of trust online. Then the chapter will present and evaluate existing technical solutions that attempt to secure trust online in terms of their expectations of user trust behaviors. The chapter concludes with a case study of an application that utilizes social networks as a countermeasure to fraudulent activity online will be introduced.

reference as:"L. Jean Camp & Alla Genkina, "Social Networks", Phishing, Springer-Verlag, eds. M. Jakobsson and S. Myers. (Berlin, DE) 2006."

1 Overview

The core observation in this chapter is that phishing requires an incorrect human trust decision. Understanding phishing requires some understanding of human trust behavior. Therefore this chapter begins with an overview of human trust behavior online. Like trust behaviors offline, people online do not complete a throughly rational calculation of risk before taking action. People apply systematic heuristics. These heuristics (sometimes called rules of thumb) lead to systematic outcomes, both good and bad.

The understanding of the role of human trust behaviors online provided by Section 2 can be used to take a critical look at the currently used mechanism for informing trust behaviors. These technical mechanisms sometimes are at odds with what would be predicted by human trust behaviors. The examination of the dominant methods for trust communication illustrate a need to provide better mechanisms to inform trust behaviors.

One way to provide better trust communication is to leverage social networks. Therefore Section 4 introduces a mechanism for embedding social networks into online trust behaviors. This mechanism is called Net Trust. The goal of the Net Trust system is to inform trust behaviors and undermine the efficacy of all types of fraud. In particular Net Trust uses a reputation system to rate sites. That reputation system will provide positive information only to those sites that have a history of interacting with the social network. Phishing sites are characterized by a lack of history - they appear, attack, and disappear. Net Trust also effectively integrates and communicates third party ratings of websites.

2 The Role of Trust Online

Phishing is hard to block because it preys directly on the absence of reliable identifying information or trust signals online. Absent any information other than an email from a bank, the user must to decide whether to trust a website that looks very nearly identical to the site he or she has used previously. In fact, any online transaction requires a leap of faith on the part of the consumer because the individual must provide the website with personal and financial information in order to complete the transaction. Thus the consumer must completely trust an effectively anonymous party to be both honorable in the intention of protecting information and technically capable of fulfilling that intention. The consumers' trust decision must be made entirely on the basis of web interaction.

In the physical realm, individuals can use visual, geographical, and tactile information to evaluate the authenticity and trustworthiness of a service provider [23]. In the virtual realm, transactions are characterized by spatial, temporal, and social separation. [17, 20]. This separation simplifies masquerade attacks in part by decreasing the cost of constructing a false business facade. While there exists a range of security protocols that are testament to the brilliance of their creators, Internet-based confidence scams continue to increase in profitability and sophistication.

The Federal Trade Commission has reported that in 2004, 53% of all fraud complaints were Internet-related [14] with identity theft at the top of the complaint list with 246,570 complaints, up 15 percent from the previous year [28]. The Pew Internet & American Life Project of the Pew Charitable Trusts (PEW) has reported that 68% of Internet users surveyed were concerned about criminals obtaining their credit card information, while 84% were worried that their personal information would be compromised [19].

Banking institutions, American Federal law, and Basel II [13] capital accords distinguish between these types of fraud risks. There are significant legal distinctions between instantiations of unauthorized use of authenticating information to assert identity in the financial namespace. Yet the risks for the subjects of such information is the same regardless of the mechanisms of disclosure. For example, when Choicepoint exposed information of 145,000 California residents to an identity thief, it was because that thief had created 43 accounts authorized by Choicepoint and purchased the information. The data sharing of Choicepoint was no legal violation, as data brokers have no privacy constraints. The business model, not the security protocols, of Choicepoint is itself the threat to American consumers. The business model of Choicepoint is to obtain all possible information about individuals from purchase records, browsing records, employment records and public records. Choicepoint correlates the information, maintains files on individuals, and resells the repackaged information. In contrast, when Bank of America lost unencrypted back-up tapes containing 1.2M records of personal account information this was a corporate security failure based on a flawed policy. When personnel information was stolen from a laptop at the University of California at Berkeley, it was theft of property and the existence of personal information on the laptop was a violation of university security policy. None of these data exposures were legal violations.

While policymakers and technologists make distinctions between sources of information exposure (e.g., lack of privacy, process failure and security failure) there is no reason for any victim of theft from these incidents to share this perspective. Each information exposure was from different source, but for the subjects of the data the resulting risks were was the same. Similarly, end users have been given a proliferation of tools that identify unacceptable privacy policies (e.g., the Privacy Bird [8]), identify potentially malicious key hierarchies (e.g., PKI [25]), or identify malicious sites of a certain category (e.g., Phishguard [21]). But there has not been a unified system to provide consumers integrated information to evaluate risks in all its forms. And, as just described, users face not only discrete operational risks but also cumulative risks. That is, users do not face the risk of a bad actor or lost privacy from an otherwise good actor or failed technology from an otherwise good actor. Users face a set of risks, and cannot be expected to download a discrete tool for each risk.

Risk information is embedded in physical spaces by the very nature of the physical world. In the physical realm many trust cues are integrated into the single experience of visiting a store: including the location of a business, appearance of employees, demeanor of employees, permanence of physical instantiation, and thus level of investment in business creation. In the virtual realm the consumer is expected to parse particular threats (e.g., secondary use of data, validity of identity assertion) and weigh each threat independently. If a consumer were to exclude the possibility of trusting the vendor without weighing all the possible outcomes of conducting the transaction, the calculation would be so characterized by uncertainty and complexity that the consumer would be unable to act [23]. Trust is inherently simplifying. Therefore, when an individual does make the decision to trust an online entity with their personal information, they inevitably assume an uninformed risk. This assumption of uniformed risk is exacerbated by the fact that there is no mechanism that compiles information across various types of risk. The perceived risk of the transaction can often be mitigated by attacks that leverage out of band sources of trust, such as social networks. In the physical realm, an individual evaluates the perceived risk in a specific context through familiarity of social and physical clues. People infer knowledge about a vendor's values based on their appearance, behavior, or general demeanor, and will proffer trust to those they find significantly similar to themselves [23]. Familiarity "triggers trusting attitudes" because it is perceived as indicating past actions and thus predicting future behaviors [16], which in turn lessens the perceived risk. Brick and mortar businesses must invest in physical infrastructure and trusted physical addresses to obtain high degrees of trust. For example, a business on the fiftieth block of Fifth Avenue (the most expensive real estate in New York and thus America) has invested more in its location than a business in the local mall which has in turn invested more than a roadside stall. The increased investment provides an indicator of past success and potential loss in the case of criminal action. Such investment information is not present on the Internet. For example, "tifany.us" is currently available as a domain name for tens of US dollars. Creating a physical believable version of Tiffany's requires more investment than the online counterpart.

There are four basic difficulties with trust online: cost, identity, signaling and jurisdiction. All these differences between virtual commerce and brick and mortar commerce make trust more difficult.

Cost is the first and foremost difficulty with trust online. The cost of committing fraud is lower online. While fraud attempts have existed since transactions have, information systems make it easier. Con artists from Nigeria have tried to lure victims with advance-fee [1] fraud by phone for decades, but mass email allows them to reach each and every wired citizen on the globe (several times a day in some cases). The diminished cost is more than a function of technological increase in efficiency. Not only is sending an email cheaper than constructing a storefront , but also electronic transactions do not require a social presence. Committing fraud in person demands, at very least, a presentable demeanor and the absence of overt signals of mischief. Committing fraud in person often requires at least temporary residence in a common jurisdiction. Physical fraud requires risk, time, and effort not demanded by online fraud.

Masquerading is easier online because of the lack of reliable identifiers. One reason trust is hard on the internet is that the online world does not support robust, widespread identity mechanisms. Identity mechanisms have proven to be a very complex problem, and are tied to questions of authentication and accountability, which are important properties in any transaction [4]. While offline identity systems are far from perfect, they are still embedded in a social and institutional context. A hand-signed document, for example, can be corroborated against anything from a personal familiarity with the signer to evidence from a phone record that the signer was not in that location at that time. Moreover, if a single signature were to fail, every other document physically signed would no longer be invalid, as is the case with some digital signature schemes. The physical signatures

[1] Advance fee fraud occurs when an individual is offered a great windfall, with examples including participating in money laundering for a percentage of the millions, purchases of crude oil at reduced prices, beneficiary of a will, recipient of an award or paper currency conversion. The recipient/victim is then asked to provide forms, banking information, and of course advance fees for the transaction. If the victim is still in the fraud after the provision of the fees they are encouraged to travel internationally without a visa to collect the funds. After entering a county illegally without a visa, (usually Nigeria) the victims has effectively kidnapped himself and must pay ransom. There is one confirmed death resulting from self-kidnapping in advanced fee fraud in June 1995 and many suspected.

are still embedded in their situational context.

Showing or signaling that a vendor is trustworthy is difficult online. A good party can invest in real estate or fancy physical domains, as described above, in the physical realm. It is very hard for a vendor to prove that it is a reputable vendor with a history and significant capital. Anything an honest vendor can put into a web page, a dishonest vendor can put into a web page. While the internet has been a boon to new forms of expression, the number of actual channels through which one can send information is greatly decreased compared to physical interaction. For example, Tiffany's and a flea market can be compared in many ways: are they equally orderly, are they pleasant, where are they located, how do the places smell, are they crowded, are they loud, etc. Online entities are limited to online media. A new retail bank might use marble and gild to signal trustworthiness. A new online bank cannot use physical mechanisms to signal that they are trustworthy.

Finally, the lack of centralization and diversity in legal jurisdictions mean that fraud in one location often cannot be tied to another party or location, and if connected may not be prosecuted. While some online merchants have tried to leverage cryptography to become what is known as a trusted third party, no single party is trusted by everyone. No single party can inspire trust in part because there is no party that has enforcement mechanisms equal to rule of law across the Internet. Because untrustworthy vendors want to appear trustworthy, they can misrepresent their jurisdictions.

Reliable trust information is difficult to create, and such information is even more difficult in a virtual environment. There are four challenges in securing trust on-line: cost, signaling, identity and jurisdiction. Of course, there is a plethora of mechanisms for securing trust online, despite the difficulties. In the next section the mechanisms for online trust are defined and categorized according to their socio-technical assumptions. Then the efficacy of each type to address these four factors is discussion.

3 Existing Solutions for Securing Trust Online

In the previous section the difficulties of trust in the virtual environment were discussed. The need for trust online was documented. Of course, multiple solutions to the various dimensions of the problem of trust exist. None of these have been completely successful.

In this section we classify the approaches for online trust based on their socio-technical foundations: social and trust networks, third-party certifications, and first-party assertions.

Social and trust networks are system where individuals share information to create a common body of knowledge. Third party certifications include all mechanisms where there is a single party that provides verification to all other individuals. These certification systems may or may not include any mechanism for individual feedback on the quality of the certified party. First party assertions are claims that are made by the party seeking to be trusted. Each of the three are described in more detail below, with examples.

3.1 Reputation Systems and Social Networks

Social networks are a powerful tool that can be used to enrich the online experience. A social network is a map of relationships between individuals, as shown in Figure 1. It is defined by the connections between the individuals, which can range from an informal acquaintance to a close, immediate family member. Individuals may, and often do, have several mutually exclusive social networks; such as a network of co-workers that is not connected to the network of close personal friends or those sharing deep religious convictions. The strength of social networks is a function of the trust between the individuals in the network.

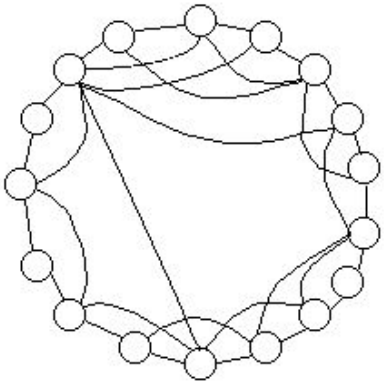

Figure 1: A drawing of a social network, with circles representing people and lines representing connections between people

Figure 1 below shows a social network where each circle or node is a person and each line is a connection. The connection may be strong, such as kinship or friendship, or weak, such as acquaintance. There are commonly attributes of social networks that shape the example below. One attribute is that if two people both know a third person than they are more likely to know each other than if they share only one friendship. This is reflected in the figure below. Another attribute common in social networks illustrated in the figure is that a few people are highly networked (that is, they know many other people) and two highly networked people are very likely to know each other.

On the internet, social networks are implemented through buddy lists, email lists, or even social networking websites. A map of such a social network can be made in the case of email by tracing all email sent from one person to others, and then tracing those as they are forwarded.

Referrals made through an existing social network, such as friends or family, "are the primary means of disseminating market information when the services are particularly complex and difficult to evaluate ... this implies that if one gets positive word-of-mouth referrals on e-commerce from a person with strong personal ties, the consumer may establish higher levels of initial trust in e-commerce' ' (Granovetter as cited in [22]). In 2001, the Pew Internet & American Life Project (PEW) found that 56% of people surveyed said they email a family member or friend for advice about internet vendors before visiting that vendor [26].

Social networks are used in the market today as powerful sources of information for the evaluation of resources. Several commercial websites, such as Overstock.com and Netflix.com, leverage social networks in order to increase customer satisfaction by enabling users to share opinions, merchandise lists, and other information. In fact, Overstock.com attracted more than 25,000 listings in six months after the implementation of its social network functionality. [29].

Mechanisms that leverage social networks include reputation, rating and referral networks. A reputation system measures individuals who interact on the basis of that interaction. A reputation system may be automated or it may depend on feedback from the people in the system. A recommender system has people who rate artifacts. An artifact is a place or thing. An important distinction between rating people and artifacts is that an artifact never rates the person who rated it in return. A recommender systems (often called collaborative filtering systems) takes ratings given by different people, and uses them to make recommendations on materials to those users.

Consider Amazon.com to understand the distinction. Buyers can rate books and offer comments. Buyers can write reviews and rate other peoples' reviews. Buyers develop reputations based on their reviews, as others

rate the reviews offered. If one buyer finds all the reviewers of a second buyer and rates these reviews badly (*not helpful* in the Amazon.com rating scheme), then the second buyer might identify the first and similarly downgrade his reviews. The ratings of comments develop reputations of buyers who review books.

Amazon also tracks every purchase. If these two enemies in the ratings contest have similar buying patterns then Amazon would nonetheless recommend books for one based on the purchases of another. The idea behind the Amazon recommendation system is that people who have been very similar in the past will be very similar in the future.

Books are artifacts that are rated and recommended. Books do not respond when rated. People have reputations based on ratings of reviews. A person who receives a bad rating can retaliate, and can behave in strategic ways to manipulate the reputation. Reputations can be seen as games, where people may or may play fair. Recommendations can be seen as grades or marks.

Recommendation and reputation systems are the mechanisms that are used in social network systems that create ratings. Examples of social network systems include the FaceBook, Friendster, Orkut and LinkedIn. These systems all leverage social networks for rating, with a link in a social network implying validation of the user. LinkedIn is designed for business opportunities including consulting opportunities and employment. Orkut allows people to self-organize in discussion groups, while Friendster connects individuals interested in romantic relationships. The Face Book uses social networks to connect college students for whatever purpose.

Furl and Del.icio.us are recommender systems that use *social browsing*. In Del.icio.us personal browsing information of participants is all sent to Del.icio.us, centralized, and then integrated into a stream of information that is broadcast across all participants. In Furl, individual mirrors of a participants' web browsing is generated on a centralized site. Both systems use their own recommender systems based on individual's browsing to recommend sites. The systems allow users to annotate sites with textual labels. Both systems implement web searching by using the shared participants' browsing history and textual labels. Sites that have been identified as useful by those with similar browsing patterns are implicitly recommended more strongly by being ranked more highly in the results returned to a user from a search.

Besides specifically designed reputation and social network systems, there are other mechanisms that use the underlying dynamics of reputation. Public forums (perhaps on a vendor's site as with Dell Computer) and rating systems provide a natural incentive for vendors to act in an honorable manner or otherwise face economic and social consequences. "Social costs can be imposed on exchange partners that behave in opportunistic ways. For example, a firm that gains a reputation as cheater may bear substantial economic opportunity costs, ..." [2]. Opportunity costs are the literally the costs of lost oportunity. Choosing not to buy a book means that the book cannot be read. Choosing to buy a CD instead of the book implies that the book will not be bought, thus the opportunity cost of the CD is not reading the book. In this case the opportunity costs are the lost sales that the vendor would have fulfilled in a trustworthy manner, but never had a chance to complete because the customer had no trust. The cost of misbehavior can be greater if the sources of the reputation information are trusted, because the value of lost trust is presumably greater.

While "the Internet was originally expected to be a great medium for disintermediation or the elimination of people as intermediaries to sources and services," [24] in fact it altered the nature of intermediation. Rather than relying on centralized dispassionate authorities, social networks are necessary to offer "counsel, guidance to the right sources of information, assessment of quality of the sources, and customized advice" [24].

Reputation systems attempt to enforce cooperative behavior through ratings or credits. The opportunity for retaliation (through ratings or direct punishment) is an important predictor of behavior in repeated trust interactions. [1]

A good reputation (i.e. credit in a reputation system) enables particular actions or provides privileges. The correlation between the ratings and the privileges may be direct or indirect. The reputation system may have

immediate or delayed impact. An example of a reputation system with an immediate impact is the financial consumer credit rating system. A low credit rating score will result in lost opportunities for credit, as well an increased cost to borrow funds. The credit rating system attempts to enforce paying off bills and credit cards in cases where legal enforcement is too costly by assisting lenders in evaluating a credit applicant's risk profiles in a particular transaction. It is not financially worthwhile for an individual lender who is not repaid to retaliate directly. However, the lender can provide information that will reduce the liklihood of any other lender extending credit to the applicant in the future.

The value of ratings can be quite hard to determine. However there is evidence that even when the ratings do not correspond clearly to money, as with the credit rating system, they can still be effective in markets. For example, eBay has a reputation system for vendors and buyers. For vendors a lower reputation corresponds to lower prices and fewer bidders. As eBay accounts are simple to obtain, those who have no history can also receive lower prices and fewer bids for their offerings [27]. Some eBay vendors exclude low-ranked buyers from auctions, but no vendors exclude new buyers. Thus eBay vendor reputations are more useful to vendors than eBay buyer reputations are useful to buyers. As a result, vendors invest more in manipulation of the reputation system.

The design of a reputation system is not trivial. Flawed reputation systems can inadvertently promote opportunistic behavior. Consider the case of cumulative ratings, as on eBay. On eBay, a vendor who has more than twelve transactions but cheats 25% of the time will have a higher rating than a vendor who has been 100% honest but has only ten honest transactions [9]. eBay vendors manipulate the reputation system by refusing to rate customers until the vendors themselves are rated, thus having the implicit threat or retaliation. Note that the vendors know as soon as the buyer's payment has cleared that the buyer has behaved in a trustworthy fashion and could rate accurately at that time. For more than 90% of vendor negative ratings there is a corresponding negative buyer rating, indicating that retaliation is a very real threat. [27] Of course, this manipulation of the reputation system also serves eBay's interest by creating an artificially low rate measure of customer unhappiness, thus encouraging buyers to participate. Essentially the capacity to manipulate the eBay reputation system is a result of eBay's desire to appear to provide a more trustworthy marketplace than it does in fact provide. A reputation system that was more effective for the buyer in indicating reliability of vendors would both create a more trustworthy marketplace and create a higher reported rate on buyer discontent for eBay. Thus a more trustworthy eBay would have a higher apparent fraud rate than the less trustworthy current instantiation.

Reputation systems may be centralized or decentralized. Centralized systems use a single authority to ensure the validity of individual reputations by tracking the reputations and corresponding identities Reputation systems may be distributed, with multiple parties sharing their own reputation information [9][12]. Consider the following example.

While Ebay is centralized, Pretty Good Privacy (PGP) uses a distributed reputation system. The reputation system in PGP is used not to offer claims of honesty in transactions, but rather to confirm the validity of identity claims. PGP allows people to introduce each other and use public key cryptography;s digital certificates to vouch for identity claims. In PGP, a social network instead of centralized authentication is used for verification of claims of identity [15]. PGP has weaknesses just as eBay has flaws. Because there is no gatekeeper there are specious claims of identity; for example, there are multiple fraudulent claimants to the identity Bill Gates. There is also an issue of affiliation in PGP because no one can stop another person from confirming his or her claim of identity. Public keys are public, and anyone can add their verification of the identity claim corresponding to a public key.

The Google search algorithm is based on an implicit reputation system. Google was the first search company to evaluate the relevance of a web page by examining how many other pages link to a page, as well as the text in the linking anchor. Linking to a page is not an explicit vote, and linking was not previously conceived of as a reputation mechanism. Google integrates the information across many pages, and leverages the combined opinions embedded in those pages. The exact Google mechanism for search is a protected corporate secret.

However, there is no question that the implicit reputation mechanism of searching is a critical element of the efficacy of Google search. Google is a centralized evaluation of a distributed reputation. Few individual links have value; however, Google's combination of that reputation information has placed its market value in the billions of American dollars.

Of course, the complexity of the Google algorithm has evolved but the requirements on the user have remained simple. The use of public links to calculate standing make it possible to manipulate Google ratings. In response to the manipulation, the Google algorithm has become more complex. There is a continuous evolution of the Google reputation system as those attempting to manipulate search results create ever more complex attacts. Yet the individual users of Google, both those who contribute links and those who search, interact with an extremely simple interface. It is the complexity of the user interaction not the complexity of the reputation system that drives usability.

Some reputation systems that work in theory do not work in practice. Reputation systems must be simple enough to understand, but not so simple as to be easily subverted. Complexity of the required user interaction can cause a reputation system to fail. Some complex rating systems require that users quantify inherently qualitative information. Multiple designs of otherwise useful reputation and ratings systems require individuals to place numerical ratings of the level of trust of each rater participating in a reputation system. For example, one proposal (named BBK after its inventors) allows individuals to select and rate other individuals who might be providing everything from emails to opinions. BBK functions by constructing a graph of a social network for each user. Each path or link on the social network has a rating between zero and one that indicates if the person is trusted perfectly (one) or not at all(zero). Then recommendations are multiplied by the trust factor before being summed. If a recommendation comes from someone far from the evaluator of trust, all the weighing values in the network are multiplied to obtain the final rating value. The individual extending trust is the one who determines the weights in the graph [3] and thus must be able to meaningfully assign a decimal to each person. The result is a system that is difficult to use particularly for an innumerate computer user.

As has been demonstrated by game theoretic experiments [1], data provided from the Federal Trade Commission (FTC) [14] and PEW [26] social networks and reputation systems encourage but cannot enforce trustworthy behavior. Reputation systems can effectively condense information, share knowledge, and increase accountability. Reputation systems therefore can support individual decision-making about vendors, users, products and websites [18]. However, there is no theoretically ideal rating or reputation system. The design of a reputation system depends upon the abilities and roles of the players, the length of interaction, the potential loss if the system is subverted, and the distribution of that loss. Even the best reputation system design cannot work without careful design of the user interaction.

Recall that the four difficulties of online trust are cost, identity, signaling and jurisdiction. How well do reputation systems and social networks solve these fundamental problems?

The cost of designing and implementing a reputation or ratings system can be quite high. A well designed reputation system can provide effective signaling. Recall that signals are needed to identify trustworthy or untrustworthy vendors and web sites. However, reputation systems do not solve the problem of identity. Reputation systems that are hard to join prevent most from participating. Reputation systems that are easy to join suffer the problem that cheap new identities are not trusted.

Reputation mechanisms can provide censure within their domains, but that is not the equivalent of having jurisdiction, as a bad party can choose to cease participating in a reputation system or create a new identity to lose bad ratings. Recall that there is a problem of preventing bad parties from escaping punishment by obtaining new identities while allowing good parties to join.

3.2 Third Party Certifications

Third party certification systems vary widely in technical validity, from those that are cryptographically secure to those based on easily copied graphics. Third party certification is provided by a range of vendors, from traditional brick and mortar organizations (e.g., Better Business Bureau) to Internet-only organizations (e.g., TRUSTe, Verisign). In any case, the third party organization is paid by the Internet entity that is trying to assure the customer of their inherently trustworthy nature.

The most popular method of securing trust online is through the display of small images called trust seals. Trust seals are meant to facilitate the transfer of trust from an individual to a vendor through a third-party intermediary. Unfortunately, the seals themselves are digital images that can be easily acquired and fraudulently displayed by malicious websites.

Trust seals also often do not inherently expire. Once an online vendor has procured a seal, there cannot be unlimited auditing to ensure that the vendor continues to comply with all the related policies. Compared with ratings or reputations systems, trust seals are generally not continually evaluated and updated. Trust seals are limited in their efficacy because any web server can simply include the graphic and thereby claim the seal. A seal provider in one jurisdiction will have a limited ability to act if a vendor in a different jurisdiction fraudulently displays the provider's seal. A trust seal by a third party will not make a site "automatically trustworthy in the consumer's eyes. The third-party itself must be credible". [2]

Even honestly displayed, the seals can be inherently meaningless. For example, the TRUSTe basic seal confirms only that the vendor complies with the vendor-generated privacy policy. Not only is there no confirmation of the quality of security of the vendors site, but it is also the case that the privacy policy may be exploitive or consist of an assertion of complete rights over customer data.

(a) TRUSTe (b) Child-friendly TRUSTe seal (c) EU TRUSTe seal

Figure 2: Three examples of trust seals that are offered by the dominant seal provider, TRUSTe

For example, the TRUSTe seal (Figure 2, image 2a) indicates only that the site has a privacy policy and follows it. In many cases, this seal corresponds to an exploitive policy declaring users have no privacy; thus indicating that compliance with the policy is a reason to distrust the site.

The TRUSTe Children's On-line Protection Act seal (Figure 2, image 2b) indicates that the site complies with the Children's On-line Protection Act, thus indicating that the site will not knowingly exploit data about minors. Amazon, for example, requires that all book reviewers declare themselves over 13 in order to avoid Children's Online Privacy Act, thereby not reaching the level required of this TRUSTe seal (Figure 2b).

The seal shown in Figure 2, image 2c is the only one of the three presented in Figure 2 with semantic as well as symbolic meaning, as this seal implies that the site complies with the European Privacy Directive. This indicates that the vendor has a reason for compiling each data element, will not resell personally identifiable information data without permission of the subject, and will delete the data when the vendor has no more need for the data.

In effect TRUSTe, the most popular seal provider on the web, is as much a provider of warning signs on the state of privacy practice as a reassurance.

These seals are trivially easy to copy. The seals are graphics with a "click-to-verify" mechanism. The click-to-verify mechanism sends a message to the TRUSTe web site, and the web site pops up a window verifying or denying the legitimacy of a website's claim to be displaying a legitimate seal. The click-to-verify mechanism does not check the name of the referring site, but rather the name of the site that is contained in the query. Therefore, these seals do not provide effective trust communication.

The cryptographically secure widely-used mechanism for third party verification is the Secure Sockets Layer. The Secure Sockets Layer (SSL) can fairly be called the ubiquitous security infrastructure of the Internet [7]. Sites that use SSL have a small lock icon on the lower right hand corner of most browser windows, and are commonly identified as https as opposed to http i n their url.

SSL provides confidentiality while browsing by establishing a cryptographically secure connection between two parties (such as a customer and a vendor) which can then be used to conduct commercial transactions. The SSL connection ensures confidentiality. Confidentiality means that passwords, credit card numbers, and other sensitive personal information cannot be easily compromised by third parties via eavesdropping on the network.

SSL does not shelter the consumer's information from insecure data storage on vendor machines, nor does it prevent the vendor from exploiting the acquired information. SSL cannot indicate the reliability of the merchant in terms of delivering acceptable goods or services. Privacy, data security, and reliance are all elements of trust [4].

While the privacy seals indicate policies, SSL provides confidentiality and indicates claims of identity. SSL confirms vendor claims of identity by providing a certification that asserts that the certificate provider has confirmed the vendor's claim of identity. This claim may only be confirmed by payment mechanisms; for example a credit card used during certificate purchase, so the certification may be cryptographically strong but it is organizationally and socially weak. The link between the key and the corresponding organization is not systemactically verified. The connection is often correct but that does indicate that the connection is strong in organizational terms. The relationship, if any, between the consumer and the vendor is not illustrated by SSL. The name of the field in the certificate to some degree indicates the strength of the claim of identity in an SSL certificate, as it is called "owner's name or alias" in Internet Explorer (IE) documentation.

Once a party has been approved by the user or the browser distributor as a verified issuer of certificates, then SSL certificates issued by that distributor are automatically accepted. Since SSL is the secure mechanisms for identity assertions, one obvious question is the number of entities that are pre-approved by the browser distributor to make those assertions. In January 2004, the number of pr-approved certificate-issuing parties (called trusted roots) included by default in the IE version for Windows XP exceeded one hundred and continues to grow. The listed entities are primarily financial institutions and commercial certificate authorities (CA) but also include a multitude of businesses who have business associations with Microsoft. this illustrates that the list of pre-approved certificate issuers does not correspond to the end users' individual trust decisions off-line.

As with seals, the incentives of the centralized parties who are vendors of the cryptographic identity infrastructure are often not aligned with the incentives of individual users who must evaluate a mass of web resources as good for their own needs or bad.

In the Section 2 the four difficulties of online trust were identified as cost, identity, signaling and jurisdiction. How well do third party assertions solve these fundamental problems?

The cost of third party assertion systems vary widely. Verisign has a cryptographically secure public key infrastructure that is recognized by every browser. Such a mechanism would be very costly to recreate. Protection of the root keys requires significant technical expertise and diligence. In contrast, any individual can offer easy-to-copy seals. An individual or entity might choose to offer seals to verify vendors who meet any standard; for

example, use of union-protected labor or support for particular religious affiliations.

Third party assertions can be partial solutions to the problem of jurisdiction. At the least, third party assertions can provide a centralized locale for complaints and for potential enforcement action. The EU TRUSTe seal indicates an agreement to abide by a set of privacy practices, but does not confirm that the vendor with the seal is in the EU jurisdiction. Yet because the seal is simply an image, any malicious party can copy the seal onto his own site.

Third party systems succeed in verification of identity as long as the context of that identity is well defined. For example, a Better Business Bureau certification that was cryptographically secure could confirm that a business is brick and mortar in a BBB community and in good standing with the local BBB. The currently used, yet easily copied, seal can communicate but not verify that assertion.

Third party assertions are widely used to signal information. Seals can be easily copied they are not optimal. Certificates are secure but they are difficult to understand.

3.3 First Party Assertions

There are a set of trust systems that allow a vendor to assert that his or her behavior is trustworthy, and the consumer of the site is left with the decision to believe those assertions or not. We include privacy policies, and the automated evaluation and negotiation of privacy policies via the Platform for Privacy Preferences (P3P) in this category.

Privacy policies are assertions of trustworthy behavior by vendors. Privacy policies may be difficult to read, and may vary in subtle manners. The Platform for Privacy Preferences (P3P) was developed to enable individuals to more easily evaluate and interpret a website's privacy policy. P3P requires a vendor to generate an XML file that describes the information practices of the website; this file can then be automatically read by the web browser and compared with a user's preset preferences. Microsoft incorporated P3P into Internet Explorer 6.0. Microsoft's implementation is limited, so that P3P primarily functions as a cookie manager. This means that the ability of users to actively negotiate privacy settings for their own data or to automatically identify parties with low-privacy practices is not included in IE 6.0.

In part because browser implementations of P3P failed to take full advantage its power to communicate privacy information, AT&T created a plug-in for Internet Explorer called the "Privacy Bird". The Privacy Bird installation includes a dialogue that converts plain language user policies to XML P3P statements. After that conversion, the Privacy Bird compares the privacy policy of a site with the expressed preferences of the end user. The bird provides simple feedback (e.g., by singing, changing color, issuing cartoon expletives) to end users to enable them to make more informed choices. The Privacy Bird is arguably the most effective user interaction mechanism for evaluating privacy policies to date.

The core problem with P3P is that the protocol relies on the vendor to provide an honest accounting of the information practices on the website, which again forces consumers to place trust in the vendor. The protocol does not have any mechanisms for validation of the vendor's information and so may mislead a consumer to trust an objectionable site.

In the case of IE 6.0, poor implementation of the user interface negated the protocols' attempt to be simple yet informative in both conceptual and usability terms. The IE 6.0's privacy thermostat used a privacy scale from "low" to "high" yet the difference between the settings is not immediately apparent. A user could easily set the privacy rating as "high", but the difference between "high" and "medium" is not easy to determine. Similarly the limit of "low" is not obvious. For example, it is not clear if a low privacy setting allows for the installation of spyware and adware; or if it simply allows websites to place cookies and track browsing. The interface for P3P in IE 6.0 does not include notification of the user if there is difference in a vendor policy and

user preferences. When there are differences between policy and preferences easily understood by the end user, there is no mechanism that translates that difference from XML to natural language. Of course since there is no such information there is no display.

In Section 2 the four difficulties of online trust were defined as cost, identity, signaling and jurisdiction. Do first party assertions address these problems? There is little cost to first party assertions, and no confirmation of identity. By definition first party assertions are decentralized. While first party assertions can provide powerful signals these signals, like third party seals, can easily be falsified. The automated evaluation of privacy polices may be effective in empowering consumers over privacy claims. However, privacy policies may themselves not be trustworthy. A company that offers and follows a strong privacy policy can be more trustworthy. A phishing site with a good privacy policy may simply be more trusted, and therefore more dangerous. Jurisdiction issues are not addressed with first party assertions.

3.4 Existing Solutions for Securing Trust Online

This section has examined the three sources of trust that can be leveraged to enable trust online: social networks, third party assertions, and first party assertions. None alone have proven adequate, but each has the ability to contribute to the fundamental challenges of trust online, first party assertions being the weakest of the three. In the next section, a mechanism that uses both social networks and third party assertions is introduced.

4 Case Study: "Net Trust"

The introduction of this chapter identified phishing as a problem grounded in flawed trust decisions. Section 2 identified the critical problems with trust online as the low cost of masquerade, the near impossibility for a new site to signal itself as trustworthy, the paucity of identification mechanisms and and the lack of centralized authorization or single unifying jurisdiction. These four elements indicate that trust online is both a social and technical problem. Section 3 then classified the existing solutions for securing trust online based on their socio-technical characteristics. Each of three classifications was defined, examples were provided, and the potential for the mechanisms to address the problem of online was discussed. In this section we describe an application for re-enabling trust on the internet. This application is designed to enable the accumulation of signals by trustworthy vendors, increase the difficulty of masquerade attacks, and utilizes social networks to provide contextual information to complement claims of identity.

The range of extant security technologies can solve the problems of impersonation in a technical sense; however, these have failed in the larger commercial and social context. Therefore, the toolbar described here is designed as a socio-technical solution that uses social networks to re-embed information on-line that is implicitly imbued by geography and physical design off-line. The application is called Net Trust [5].

The application is implemented as a toolbar plug-in (see figure 3) for a web browser. Using Net Trust people will be able to get instant feedback from their own social networks while browsing. The application graphically displays the rating of each member of the social network (Figure 4, Image a) as well as aggregate scores (Figure 4, image b). The application automatically signals a site's trustworthiness using the browsing history of other users, as well as allowing users to explicitly rate sites.

Recall that individuals may have multiple social networks including home, family, hobby, political or religious. Regardless of the level of overlap, information created for one social network may be inappropriate for another social network. For example, while I might respect the technical expertise of my colleagues I am less likely to respect their judgments of political sites or parenting advice. In order to support multiple social networks, the application allows a user to have multiple identities (e.g., psuedonyms) coupled with multiple social networks.

Figure 3: The Net Trust Toolbar as it would appear. The far left "alex_work" identifies the users' current pseudonym. The second icon of a hand with a pencil is used to open a window so alex_work can leave comments. The third icon, the pin and note, is used to read other users' comments. The left bar appears red with negative ratings from the social network associated with alex_work for the current site. The right bar is green and similarly indicates any positive reputation rating from the social network associated with alex_work. The third party ratings are provided by broadcasters using icons. The image above shows Google and Paypal providing negative ratings, and the Better Business Bureau providing a positive rating.

The final result is a design that enables a user to make an informed trust decision about a site based on the shared opinions of his or her social network.

There are other toolbars that can bed used to identify phishing sites. (Other chapters in Phishing & Countermeasures provide more details on toolbars as mechanism.) Many of these toolbars use realtime characteristics of the web sites themselves to identify malicious sites; for example, a login box, existence of a SSL certificate, or misdirection in the links. In contrast this toolbar uses features that are not under the control of the malicious agent to prevent phishing: user social network, user history and the history of the user's social network. In addition, the Net Trust toolbar takes advantage of a characteristic of phishing sites to prevent one phishing victim from misdirecting others - the temporal history of phishing sites. Phishing sites go up, are identified, and are taken down. Phishing sites do not stay up over long periods of time. Therefore the impermanence of phishing sites is integrated into the reputation system as described below.

Net Trust is also designed to ensure privacy, in that the users can share selected information; withhold information; and can control with whom they can share information. By leveraging a user's social network as well as user-selected third parties, the application is able to display reputation ratings in a meaningful and useful display. Thus the user will be able to make a contextual decision about a website.

Aggregate scores, which are the average of the scores of all users that have visited a site, are shown by default in the toolbar. This image shows a web site which has both positive and negative ratings from 4 users out of the 14 member social network. This is shown in the context of the toolbar in Figure 3.

This system shares web browsing information in a closed network of peers. In contrast, recall Furl and Del.icio.us. Both are designed to leverage the observation that each user has a unique view of the web informed by their own history and the history of others. In both systems there is significant centralized storage of user browsing and no explicit mechanism for user pseudonyms. Neither of these systems uses the developments in social network systems beyond simple collaborative filtering. Individuals do not select their own peer group. As a result rating information can be inappropriate or even highly polarized; for example, a search for "George W Bush" on Del.icio.us yields images of both a president and of various chimps. Few individuals would be searching for both images.

Del.icio.us and Furl do have a commonality with Net Trust in that they leverage the similarity of browsing patterns. Leveraging a social network requires that the social network provides relevant information. The system will work only if a person has a browsing pattern that is highly correlated to his or her social network. There is little to be learned from a social network that does not visit the same site as the user.

Net Trust enables integration of information from trusted parties. These are not trusted third parties, as is the tradition in security and cryptographic systems. The removal of the word *third* in the traditional trusted third party construct indicates that the individual makes the final trust decision, not the trusted party. There is no root

Figure 4: Users can Provide Explicit Ratings, Implicit Ratings and comments. These can be viewed individually by pushing the pin and note icon shown in Figure 3.

that determines which parties are trusted. In trusted third party systems the browser manufacturer, employer or other third party determines who is a trusted party. In Net Trust certificates are self-signed, and users select trusted information providers. The Net Trust user, not the distributor or developer, makes the final determination of which parties are trusted. Compare this with SSL or Active X, where the user is provided a mechanism to place trust in a third party and after the selection of the third party, the user's decision is informed by the third party, implemented by technical fiat.

4.1 Identity

In this system, identity is not universal. There can be many Bobs so long as there is only one Bob in any particular social network. Effectively, identities are used as handles or buddy names to create a virtual implementation of a pre-existing social network. Identity construction assumes a previous context, so that meaning is derived from the name and context. Essentially each person can construct as many pseudonyms as he or she desires, where each pseudonym corresponds to a distinct user-selected social network.

Pseudonyms engage in disparate social networks. Members of that social network are called "buddies", both to indicate the similarity to other online connections and to indicate that the standard for inclusion may vary. Buddy is also a sufficiently vague term to communicate that there is no standard for strength of the network tie. This design choice reflects the fact that relationships are often contextual. This means not only that people share different information with different people [11], but also that different social networks are associated with different levels of trust. For example, professional colleagues can have much to offer in terms of evaluation of professional websites, but one would not share information with the same level of openness as family. Because of these differences, overlapping contexts can cause a breach in privacy. The use of pseudonyms in this system enables the construction of boundaries between the various roles filled by one person. Figure 5 illustrates the easy method by which users may switch between pseudonyms in Net Trust while web surfing, using the example of work and home pseudonyms.

When a user leaves, departs, or clicks out of a website the url is associated with the pseudonyms visible in the

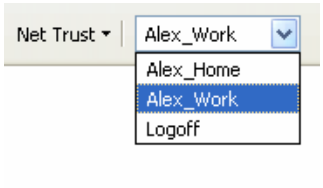

Figure 5: Changing Pseudonyms with Two Choices: Work and Home

toolbar. Choosing to associate a site upon departure instead of arrival allows users to make informed selections of websites. Once a website has been identified as associated with a pseudonym (in the figure shown the pseudonyms is "Alex Work") the user no longer has to select that identity when visiting the associated website. If Alex is in work mode, and then visits a site she has identified as associated with the Alex_home pseudonym Net Trust will change pseudonyms at the site. If Alex wants to share a site across pseudonyms, he has to make a non-zero effort (holding down a control key) to add an additional pseudonyms to a site already associated with one pseudonym. Therefore after a website has been associated with a pseudonym all future visits correspond to that pseudonym, regardless of the website selection at the time of site entry. Thus individuals have to make pseudonym choices only on new websites. Presumably individuals will select a default pseudonym, possibly different pseudonyms for different machines, e.g., a work pseudonym at work or a home pseudonym at home.

4.2 The Buddy List

An essential components of this application is the re-embedding of a user's existing social network into their online browsing experience. In brick and mortar commerce, physical location is inherently associated with social network, as exemplified by the corner store, "regulars" at businesses, and local meeting places. Net Trust uses social networks to capture virtual locality information in a manner analogous to physical information.

Net Trust implements social networks by requiring explicit interaction of the user. The Net Trust user creates a "buddy list" containing the social network associated with a pseudonym. Using the Net Trust application invitation mechanism, a user sends a request to a buddy asking for authorization to add them to their buddy list. Once the buddy approves the request, the user can place the buddy in the social network defined by the appropriate pseudonym. Social networks can be presented for user consideration from importing IM lists, email sent lists, or pre-existing social network tools such as Orkut, Friendster, Face Book, or LinkedIn. This system requires that the individual issuing the invitation to his or her buddy know the email or IM of that buddy. Figure 6 illustrates one option for inviting a person to a social network.

The invitation includes information about the distributed file system, which is described in more detail in the following section. Consider a Net Trust user named Alice who has as an associate a person named Bob.

Before inviting anyone to a network Alice creates a pseudonym. Once the pseudonym is created she creates a set of asymmetric keys, public and private. For simplicity, call the pseudonym Alice@work. The private key allows Alice to confirm that any message from Alice@work came from Alice@work to anyone with the corresponding public key. Alice sends an invitation with a nonce to Bob. The nonce prevents replay attacks and ensures freshness. The public key prevents anyone else from associating themselves with Alice's pseudonyms after the initial introduction.

The example can only continue if Bob agrees to join the system. Because Bob joined the system, Alice will share Alice@work's history with Bob's chosen pseudonym. In Net Trust, the reputation information is con-

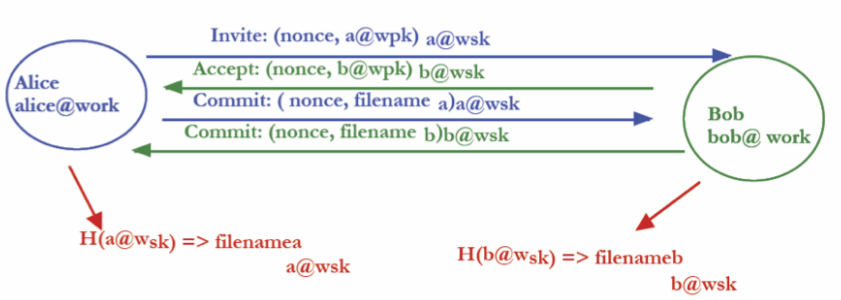

Figure 6: Invitation to Social Network

tained in a file or feed that is identified by a 128 bit random number. The feed or file does not include personally identifiable information.

Since Alice initiated the invitation, she sends Bob her file number and a key used to sign the file. Then Bob has the option to send his file number and a key used to sign his feed. Part of Bob's choice includes filling out information about his affiliation with Alice - her name and his corresponding pseudonym, as well as a review date for her inclusion. Thus interaction is designed to cause joining a stranger's network to cause some cognitive dissonance by demanding unknown information in order to consummate the introduction.

After this introduction ALice and Bob update their own feeds by sending out information. Alice and Bob update their own ratings by periodically downloading each others' published files. The files, designated "filename", include URLs, ratings, and dates. (The reputation system uses the URL with the first level directory, e.g., `www.ljean.com/files` would be distinct from `www.ljean.com/trust`. Bob's ratings are then integrated into Alice's toolbar as Bob's opinions of sites, with Alice's client reading Bob's file and Bob's client reading Alice's file. The data are public and signed, but not linked to any identity excluding via traffic analysis.

In the initial instantiation of Net Trust different individual's opinions are not differently weighed. Implicit weighting is created by segregating individuals into social networks. Recall from section 4 that some systems assume that individuals should be provided with different trust weights because some contribute more than others [10]. This is frequently embedded in reputation systems by creating a a weight of credibility which distinguishes levels of trust within the social network. A common example is that of a grandparent, who might not be as apt at discriminating between legitimate and malicious sites as a computer savvy co-worker. In contrast, our system allows the user to evaluate his or her own context and weigh based on the provision of the information. While the proverbial grandparent may not be computer savvy, this grandparent may have extensive knowledge of hunger-based charities from volunteer work or detailed knowledge of travel locales from personal experience. Therefore, there is no single measure of trust suitable for an individual across all contexts. Figure 7 illustrates the manner in which individual contributions from a social network are easily accessed from the Net Trust toolbar. By simply hitting the icon of "people" as seen up close in Figure 3 , the user will see an enlarged view of their social network and pertinent browsing statistics. User-selected icons are displayed for quicker visual identification and personalization.

Net Trust also allows for the addition of third parties who make assertions about trust. See Figure 8. These ratings are provided by centralized trusted parties. They are called "broadcasters" in this model to emphasize that they distribute but do not collect information. While buddies share information by both sending information and obtaining regular updates, broadcasters only distribute information. Broadcasters use a certificate-based system to distribute their own files, with Boolean ratings. Such lists of "good" and "bad" sites are sometimes called white and black lists or green and red lists. These green lists are stored and searched locally to prevent the need for the Net Trust user to send queries that indicate their browsing habits. Requiring a web query for

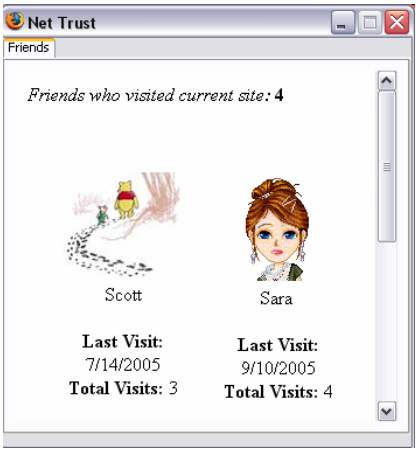

Figure 7: Social Network View

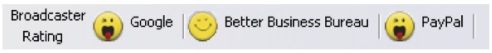

Figure 8: Third Party Trust Assertion with Mr. Yuk, Smiley Face, and Mr. Yuk

searching would create a record of the client's travels across the web, as with Page Rank records on the Google toolbar and Mircosoft anti-phishing toolbar.

Broadcaster's ratings are shown as positive with a happy face, negative as Mr. Yuck, and no opinion as blank. Early user test indicated that a neutral face could be misunderstood by users as a positive or negative assertion. Therefore on a url which is not included in the ratings provided by the broadcaster, the default is to have nothing displayed.

Broadcasters can alter this default by including the appropriate default handling statement in the distributed list. For example, a broadcaster could be the Federal Depository Insurance Corporation for American web browsers. This broadcaster would produce a green list of all valid bank websites. Sites that are not valid banking sites would be indicated with Mr Yuk, as opposed to a simple disappearance of the Smiley Face.

Net Trust users select their own broadcasters. Individuals that can be fooled into downloading false green lists can be undermined by this system. To mitigate the possible harm there is a maximum lifetime for any green list. Broadcasters can be removed, but it is not possible for an attacker to replace one broadcaster with another even if the first one has been removed. (Being able to override that feature requires that an attacker have write permission on a users' drive. At that point, user trust of websites becomes a negligible measure of security.) Thus while the broadcasters provide important information, like any other trust vector subversion of that trust can cause harm. However, since broadcasters only inform trust decisions the harm is limited and if there is bad information the source of the bad information can be detected by the user. Compare this with Active X or the addition of trusted certificate authorities, which alter authorization and thus access on the user's machine. In those cases malicious action from code cannot be determined during regular use by the user. Both of these systems share with Net Trust broadcasters an expiration date.

As explained in Section 4.7 the usability tests indicate that users value peer information more highly than centralized information. This suggests that inclusion of unknown users in the social network of a pseudonym creates a greater risk than that of false broadcasters.

4.3 The Security Policy

The security of this system depends on the ability to identify network participants reliably and prevent leakage of the key used to share history. If histories can be rewritten by attackers the system is a net loss in security terms.

There is no universal identity infrastructure on which this system can depend. Invitations are issued by email and email:identity connection is tenuous in identity terms. Social viruses have long utilized the lack of authentication in email to increase the likelihood of victims taking the necessary action to launch the virus. However, by requiring the ability to respond, this mechanism cannot be subverted by mere deception in the "From" field.

Thus the security policy is one that is based more on economics than on traditional security. The Net Trust system will create value for users and increase the difficulty of masquerade attacks. The next sections describe the rating system, which depends upon the identification of genuine social network members, as opposed to attackers masquerading as members of a social network.

4.4 The Rating System

The application uses two methods to collect ratings: explicit ratings and automatic tracking. A user may explicitly rate a site and add comments to their ratings. There is no specific meaning behind each rating because each user or network may have different requirements of trust, therefore the rating definitions are left up to the user and their network. According to PEW, "26% of adult Internet users in the U.S., more than 33 million people have rated a product, service, or person using an online rating system," [18].

This system also automatically tracks each user's website activity and uses the data as a basis of the rating for a website. If a user repeatedly visits a website, it can be assumed that they find it trustworthy even though they might not have explicitly stated such a fact. According to Dingledine, Freedman, and Molnar, "Simply to observe as much activity as possible and draw conclusions based on this activity is a very effective technique for automated reputation systems" [10].

At each initial visit, an entry is made in the client's local file. Ratings are from 1-4; however, simulations have shown that the constant max rating is itself not important with the following conditions. The opinion range must be large enough to be nontrivially expressive; not so large that a single user can easily dominate the entire networks' opinion; and thus the 1-4 rating is combined with a system requirement that at least five people join a network for it to begin displaying aggregate ratings.

Because of the implicit information on website visitation, the information cannot be centralized and a user must accept an invitation into a social network. Sybil attacks, so that a user adds one other person and then him or her self four times, are possible but unlikely under this system because of the out of band confirmation of user identity. Determining how to prevent Sybil attacks by utilizing additional information, yet not creating a system where this system could itself be used as an identity theft mechanism is a research challenge.

4.5 The Reputation System

User Modeling No user will be correct all the time. Each use will contribute some valuable information and some information with a negative value. Using these observations about value, the reputation design problem can be defined in economic terms. The following paragraphs describe t the reputation model that has been used to evaluate the value of Net Trust.

Without any reputation system, a participants' expected value of using the web is a function of the rewards of visiting a good resource, the harms of visiting a bad resource, and the likelihood of encountering either. The reputation has to provide value for the user to adopt the system, i.e.:

$$E(v_0) = g(1-p) - b(p) \tag{1}$$

where p is the likelihood that any resource is bad, $g$is the utility gained from visiting a good resource. In Equation 1 b is the loss or disutility suffered from visiting a bad resource.

When a Net Trust user obtains reputation information about the quality of a web site he or she adjusts his or her behavior to avoid bad resources and increases use of good resources. In order to make a feasible model, assume that information comes in a single quantity i. Note, however, this information can be both helpful and harmful in an imperfect world. If a piece of information correctly identifies a resource as bad, it is denoted $i+$ because it helps the agent. If a piece of information acts as a *false negative* by conveying information that a resource is bad when it is not, then the agent loses a potential gain. Finally, since the signal can come in many forms, all signals (since the agent is unable to distinguish good from bad) is interpreted with the same decision rule $F(i)$. The expected value of using such a system is:

$$E(v_1) = g(1-p)F(i^+) - b(p)F(i^-) \tag{2}$$

This simply indicates that the improvement of trust decisions by the measure $F(i)$ increases the identification of bad sites by some factor and increases the identification of good sites in a corresponding manner.

To determine whether a given reputation system is beneficial, the decision rule has to produce enough accurate warnings to overcome the harms from the false positives. So the value with Net Trust must be higher than without Net Trust, or

$$E(v_1) - E(V_0) > 0 \tag{3}$$

$$g(1-p)(1-F(i^-)) - b(p)(F(i^+)) > g(1-p) - b(p) \tag{4}$$

$$\frac{g(1-p)}{b(p)} < \frac{1-F(i^+)}{F(i^-)} \tag{5}$$

Notice that Equation 4 separates the reputation and non-reputation parts of the question. It also makes sense - the more frequently a user has to avoid bad sites the less the decision rule has to be correct for the system to be valuable. As there are so many bad sites, avoiding some of them adds value. One way to model the value of the system is to set the initial situation as zero, and then model to see if Net Trust adds positive value or creates a problem with flawed feedback. That can be done by assuming that $E(v0) = 0$. This base case seems to be an appropriate starting point, because it isolates the effective outcome of the reputation system for examination. Since the state of the network isn't known, this also allows modeling the reputation systems with different assumptions about g, p and b.

Assuming that most individuals correctly identify resources half the time and few individuals correctly identify resources almost all the time (95%), then the identification of the sites converge quickly. However, that convergence is greatly accelerated with an initial seeding of information.

This section has described how a model of Net Trust was constructed. To summarize the results, the model defined decision-making with and without the mechanism. Then multiple iterations of the model were run. The results were convergence to correct identification of resources even when the majority of the population has initially no better than average ability to identify resources as good or bad.

The Reputation System The current reputation system has a set of goals. First, the reputation system will combine information from different buddies. Second, if one person in a social network is fooled by a phishing site, that person should not automatically send a false positive signal to the social network. The third goal is in opposition to the second - recent information is likely to be more accurate than historical information and should be weighted accordingly. This section describes the design of the reputation system to address these competing goals. Each particular element is introduced separately but the reputation system depends upon the entire system. This is the reputation system that resulted from repeated modeling of the user as described above.

One way to describe the system is to begin with the installation. There are a series of broadcasters who will provide initial information. Buddies are invited into the system. Then the user visits a site. The rating is affiliated with a pseudonym when the user leaves a site. The initial visit is logged on the basis of domain names. The first visit will create a rating of 1 after a some delay, recall that the delay is to prevent victims of phishing sites from communicating their incorrect positive evaluation to others. The initial delay period is t_0. That is currently set at 120 hours or 5 days.

After the initial delay time (t_0) then the rating is set to 1. That rating will be held constant until time for the rating to decay. The time at which the rating begins to decay is t_d. Currently t_d is set to one month. The rating will decay until it is down to one-half of its value immediately after t_0. For a rating of one, that is obviously 0.5. Each time the website is visited the rating will be an increment of one plus the rating value at t_d. For example, if a site has a rating of one-half, then after a second visit the rating will be 2. Then after t_d the rating will again decay to half its value at t_d, which is in this case 1.

Data compilation from widespread use may result in changes of those variables. The variables and timing were selected based on the user modeling described in 4.5 and from conversations to extract the opinions of experts.

A user may also explicitly rate a site. When a site is explicitly rated, the implicit rating system ceases to function for that site. The explicit score is kept, without decay. A website that receives an explicit negative rating will neither increase nor decrease depending on time and additional visits. Only explicit user action can change a rating based on previous explicit user action.

Therefore for each website that is associated with a pseudonym there is one of the two possible records: either this with explicit rates

1. url, date initail visit, number of visits, last visit¿

or this with explicit rating

2. url, explicit rating.

To restate the reputation in mathematical terms the reputation calculation mechanisms is as follows, with the rating value is R_w.

- For sites with no visits or for a visit less than t_0 previously $R_w = 0$
- For one visit, more than t_0 but less than t_d hours ago $R_w = 1$
- For n visits with a last visit having occurred at $t - t_n > t_d$ $R_w = \min\{4, \max\{n/2, ne^{-c|t-t_n|}\}\}$

Recall from the description above that R_w is the reputation of the website, with a maximum value of 4 in our current model, n is the number of visits, t_n is the date of the most recent visit, t_d is the decay delay parameter, and t is the current date. c is simply a normalizing constant, to prevent too rapid a decay in the reputation after t_d.

Current phishing statistics suggests a value of t_0 of not less than twenty four hours; however, this may change over time. One system design question is if users should be able to easily write t_0 or t_d; i.e., if the system should be designed to allow an easy update if new attack with a greater temporal signature is created. If the value of t_0 it is too low then attack sites could change victims to supporters too quickly. Thus being able to increase the value offers the opportunity for a more secure mechanism. However, the value to alter t_0 can itself become a security risk, as a phisher could convince users to set $t_0 = 0$.

To summarize the reputation system, a single visit yields the initial rating of 1 after some delay. The delay time prevents those who are phished early from becoming agents of infection, and supporting later phishing attacks. Then as the number of visits increases the score itself increases in value. The least value for a visited site that has not been manually rated is zero. The greatest reputationvalue for any site is 4. The least reputation value of any site is -4.

Consider a phishing website. For a phishing website any broadcasters will label the site as bad or neutral. No member of the social network will have ever visited the site. While this may not deter someone from entering a site to shop for something unusual, it is an extremely unlikely outcome for a local bank, PayPal, or Ebay. In order to increase the efficacy of the toolbar against phishing in particular, one element of the project entails boot-strapping all banking sites. Those websites that are operated by FDIC-insured entities are identified by the positive Smiley Face. Those websites that are not FDIC institutions are identified by the Mr Yuk icon. In addition, boot-strapping information can be provided by using bookmarks or a compendium of shared bookmarks as provided by Give-A-Link.

Without the inclusion of the FDIC listing then the Net Trust toolbar has a failure similar to many security mechanisms - that the user is forced to look for what is *not* there. Seals function if they are not only noted as present but also noticed when *missing*. The lock icon on SSL is replaced with red icon, but the user must notice that the *lock is missing*. In email, eBay messages include a header indicating that only messages with the eBay header are to be trusted. Obviously faked emails do not include a flag to indicate that the expected header is missing.

The long term efficacy of the reputation system depends upon how similar social networks are in terms of browsing. Do friends visit the same websites? Do coworkers visit the same website? For example, in the most general reputation system where every user had a given chance of seeing any one page from some distribution, and could correctly judge a bad resource as such with probability p and would mislabel it with the corresponding probability $(1 - p)$ a decision rule could trivially be derived. Every social group will have its own set of shared sites and unique sites and the decision rule could be optimized for that social group. However, that information is not only not available for specific social networks but the data are also not publicly available in general.

Using this reputation system and with the assumption of different degrees of homophily the user modeling as described above indicates that Net Trust would provide a high degree of value in identification of sites. Homophily means that people who are in a social network have browsing habits that are more alike than not. That is, for any site a user visits it is extremely likely that a user in the same social network visits that site. Users are not uniformly distributed across the low-traffic sites on the web. Some sites of are interest only to a small population, such as members of a course at a university, or members of one school or office.

Given that the ideal mechanism cannot be known because social network homophily is not entirely known, the implementation is based on user modeling. Recall that the user modeling indicates that this toolbar will enable a significant increase in the ability of end users to discriminate between resource types. The inclusion of bootstrapping information dramatically increases the ability of end users to discriminate. The modeling of the reputation system indicates that the system as proposed will be valuable in informing trust decisions.

4.6 Privacy Considerations and Anonymity Models

The critical observation of the privacy of Net Trust is that the end user has control over his or her own information. Privacy can be violated when a user makes bad decisions about with whom to share information. However, the system is does not concentrate data nor compell disclosure.

There is a default pseudonym in the system that is shared with no other party. Net Trust comes with three default pseudonyms: user@home, user@work, and private. Websites visited under the private pseudonym are never stored in the Net Trust data structures, and thus never distributed.

Users control the social context under which information are shared. Recall that in all cases, if a user is a logged in under a certain pseudonym, their website activity will only be shared with the social network associated with that pseudonym, not with any other networks that might exist under different pseudonyms. The user may also edit by hand the list of sites that is shared with any particular social network. Subsequently, a user's online activity is only used to inform the buddy view from other user's hand-picked, social network.

Becoming and offering oneself as a broadcaster requires downloading additional software, as well as publishing a public key. The interface for broadcasters in our design accepts only hand-entered URLs and requires notations for each URL entered. Our design is directed at preventing anyone from becoming a broadcaster by accident. There is no implicit rating mechanism for broadcasters.

4.7 Usability Study Results

Net Trust is only useful to the extent that it is usable. Thus Net Trust began with user testing. Twenty-five Indiana University students participated in a usability study of Net Trust. The students were from the School of Informatics both at the graduate and the undergraduate level.

Initially, the participants of the usability study were asked to spend a few minutes investigating three websites. The websites were fabricated especially for the purpose of a usability study, and therefore controlled for content and interface. The participants were asked to indicate if they would trust the sites with their personal and/or credit card information. Next, the toolbar was enabled on in the browser and the participants were instructed to visit each of the three websites again and complete one toolbar task on each site. The tasks included rating a site, adding and removing buddies, as well as switching between buddy and network view. The survey had been previously validated with two tests of undergraduates. For the Net Trust usability test, the toolbars were seeded with reputation information.

Afterward examining the toolbars, the participants were once again prompted to indicate their trust of the three websites taking into account the information provided by the toolbar. For the first two websites, the toolbar showed a large number of "buddies" visiting the site, 6 out of 10 for website 1 and 8 out of 10 for website 2, respectively, as well as positive or neutral ratings for the broadcasters. The last website showed only 2 out of 10 friends visiting the site and negative or neutral rating from the broadcasters. As demonstrated in Table 1, the toolbar significantly increased the propensity to trust a website. The results demonstrate that the toolbar is successful in providing a signal of trust towards a website.

Even when the toolbar showed a significant amount of negative ratings, such as in website 3, the fast that a website had been previously visited by members of a social network increased the propensity to trust. This is supported by the examination of trust mechanisms in Section 3 which argued that social networks are a most powerful mechanism for enabling trust. Of course, designers of malicious code have leveraged the power of social networks since the "I love you" virus leveraged user email address books.

Lastly, the participants of the study did in fact find the user interface to be meaningful, easy to use, enjoyable, and useful. On a scale of one to five, sixty percent of participants supplied a rating of four for "ease of use"

	Without Toolbar	With Toolbar
Website1: Strong Positive	40%	68%
Website2: Strong Positive	48%	76%
Website3: Mixed Messages	20%	24%

Table 1: Percentage of Participants Who Trust Websites

and "usefulness". Moreover, eighty percent indicated that they found the interface meaningful and eighty-four percent said they would enjoy using the system. Several participants saw the value in such a system and asked when it will be available for use.

Thus this chapter has argued that phishing is a problem of uninformed user trust decisions. Then the fundamental challenges for establishing trust online were ennumerated. The chapter described social networks as one of three socio-technical mechanisms for enabling better trust decisions, along with third party certifications and first party assertions. Then Net Trust was introduced as a mechanism that leverages both social networks and third party certification to inform trust decisions. Because of the reliance on social networks, the chapter closes with a brief reminder of the risks of trust-enabling mechanisms.

5 The Risk of Social Networks

Trust-enabling and trust-enhancing mechanisms create risk to the extent that these mechanisms increase trust. Since both the theory and user studies of Net Trust indicate that it can significantly increase extensions of trust by informing trust behaviors, it can also be a source of risk if subverted.

The most significant risk in social networks is the subversion of identifiers. If the identifier of a person in a social network is jeopardized then the pre-existing trust that is associated with that person can be utilized for malicious intents. The "I Love You Virus" was reported to have affected 650 individual sites and over 500,000 individual systems [6]. One reason this virus was so devastating was because as it propagated itself through email it utilized the user's address book. Users' believed that the email was legitimate since they received it from people in their own social networks, and hence opened the malicious files, effectively propagating the virus.

Similarly flooding many mailboxes with a false invitation can open a channel for a phisher to target users after identifying their own web site as trustworthy. By developing an invitation mechanism that requires reciprocation, the system is constructed to create a non-zero threshold to accepting an invitation. However, only user studies will determine if this is adequate. In addition, any social network is limited to fifty participants. Economics would indicate that the scarcity of participant opportunities in a social network increases its value, and thus will itself decrease a willingness to accept invitations thus giving away roles.

Social networks embed trust, and that trust is built on historical interactions. Using that history, leveraging the fact that masquerade sites by definition have no history, and adding third party assertions Net Trust can enable end users to distinguish between true and false claims of identity on the Internet with greater efficacy than with any system that depends only on first party assertion, third party certifications or social networks alone. Yet, as with any technology, the degree to which it is trusted is a measure of the degree of loss should the mechanism be subverted.

6 Conclusion

Phishing is difficult to prevent because it preys directly on the absence of contextual signs in trust decisions online. Absent any information other than an email from a bank, the user must to decide whether to trust a website that looks very nearly identical to the site he or she has used without much consideration. Fraud rates are currently high enough to make consumers more suspicious of e-commerce. Simultaneously, there is very little that an institution can do to show that it is not a masquerade site. If consumers continue to misplace their trust in the information vacuum, losses will accumulate. If they decide to avoid such risks, then the economy loses a valuable commercial channel. Net Trust uses social networks and third party assertions to provide contextual signs for trust decisions online. The context consists of selected third party opinions, the ratings of a particular community, and a specific web site.

In this chapter we presented the idea of leveraging social networks to counter masquerade attacks online. We illustrated the importance of trust online as a mechanism for reducing complexity in uncertain situations. We also analyzed the currently available solutions for securing trust online, which proved to be either technically or socially unsound or inadequate.

Net Trust is an application that combines third party assertions and social network reputation information in a single unified interaction. Net Trust is implemented as a web browser toolbar plug-in. Net Trust was designed first as a user experience, then modeled to test its value for an end user with a set of assumptions. The model illustrated that Net Trust will be valuable under a wide range of assumptions of user competence and network state. The details of the design were described, and the new risks that may be generated by Net Trust enumerated.

The Net Trust toolbar can provide contextual information to inform users in their trust decisions on the Internet. It will not stop all phishing; however, it will provide sufficient primary and secondary value to be enjoyable to use, decrease efficacy of phishing, and increase the information for end user trust decisions.

References

[1] R Axelrod. *The Evolution of Cooperation*. HarperCollins Publishers Inc., New York, NY, 1994.

[2] J. Barney, M. Hansen, and B. Klein. Trustworthiness as a source of competitive advantage. *Strategic Management Journal*, 15:175–190, 1994.

[3] T. Beth, M. Borcherding, and B. Klein. Valuation of trust in open network. In D. Gollman, editor, *Computer Security — ESORICS '94 (Lecture Notes in Computer Science*, NY, NY, 1994. Springer-Verlag.

[4] L Jean Camp. Advances in Information Security. MIT Press, Cambridge, MA, 2000.

[5] L Jean Camp and Allan Friedmand. Peer production of security privacy information. *Telecommunications Policy Research Conference*, 2005.

[6] cert. http://www.cert.org/advisories/ca-2000-04.html.

[7] IBM Corporation. What is the secure sockets layer protocol? 2005.

[8] Lorrie Faith Cranor, Manjula Arjula, and Praveen Guduru. Use of a p3p user agent by early adopters. In *WPES '02: Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society*, pages 1–10, New York, NY, USA, 2002. ACM Press.

[9] R. Dingledine, M. Freedman, and D. Molnar. Peer-to-peer: Harnessing the power of disruptive technologies. In Andy Oram, editor, *Economics of Information Security*, chapter 16. O'Reilly and Associates, Cambridge, MA, 2001.

[10] R Dingledine, M Freedman, and D Molnar. Peer-to-peer: Harnessing the power of disruptive technologies, chapter 16: Accountability. http://freehaven.net/doc/oreilly/accountability-ch16.html, last accessed 10/2004.

[11] J Donath and D. Boyd. Public displays of connection. *BT Technology Journal*, 22(4), October 2004.

[12] M. Feldman, K. Lai, I. Stoica, and J. Chuang. Robust incentive techniques for peer-to-peer networks. *Proc. of EC '04*, May 2004.

[13] Bank for International Settlements. Basel ii: Revised international capital framework. 2005.

[14] FTC. Ftc releases top 10 consumer complain categories for 2004. Technical report, Federal Trade Commission, Washington, DC, February 2005.

[15] Simson Garfinkel, editor. *PGP: Pretty Good Privacy*. O'Reilly Associates, Inc., Sebastopol, CA, 1994.

[16] D. Gefen. E-commerce: the role of familiarity and trust. *The International Journal of Management Science*, 28:725–737, 2000.

[17] S. Grabner-Kraeuter. The role of consumers' trust in online-shopping. *Journal of Business Ethics*, 39, August 2002.

[18] Pew Internet and American Life Project. Trust and privacy online: Why americans want to rewrite the rules. 2004.

[19] Pew Internet and American Life Project. Trust and privacy online: Why americans want to rewrite the rules. *"http://www.pewinternet.org/PPF/r/19/report*$_display.asp(lastaccessed 2/17/2005)$", 2005.

[20] R. Kalakota and A.B. Whinston. *Electronic Commerce*. Addison Wesley,Boston, MA, 251-282, 1997.

[21] Gregg Keize. Do-it-yourself phishing kits lead to more scams. *Information Week*, August 2004.

[22] K. Kim and B. Prabhakar. Initial trust, perceived risk, and the adoption of internet banking. *Proceedings of the Twenty First International Conference on Information Systems*, December 2000.

[23] H. Nissenbaum. Securing trust online: Wisdom or oxymoron. *Boston University Law Review*, 81(3):635–664, June 2001.

[24] J. Olson and G. Olson. i2i trust in e-commerce. *Communications of the ACM*, 43(12):41–44, December 2000.

[25] R. Perlman. An overview of pki trust models. *IEEE Network*, pages 38–43, Nov/Dec 1999.

[26] Pew. What consumers have to say about information privacy. Technical report, Pew Internet & American Life Project, NZY, NY, February 2002.

[27] P Resnick, R Zeckhauser, E Friedman, and K Kuwabara. Reputation systems. *Communications of the ACM*, pages 45–48, December 2000.

[28] Reuters. Identity theft, net scams rose in 04-ftc. 2 2005.

[29] B Tedeschi. Shopping with my friendsters. *The New York Times*, Nov 2004.