2011

A phishing model and its applications to evaluating phishing attacks

Narasimha Shashidhar
Sam Houston State University, Huntsville, TX, USA

Lei Chen
Sam Houston State University, Huntsville, TX, USA

A PHISHING MODEL AND ITS APPLICATIONS TO EVALUATING PHISHING ATTACKS

Narasimha Shashidhar and Lei Chen
Sam Houston State University,
Huntsville, TX, USA

karpoor@shsu.edu, chen@shsu.edu

Abstract

Phishing is a growing threat to Internet users and causes billions of dollars in damage every year. In this paper, we present a theoretical yet practical model to study this threat in a formal manner. While it is folklore knowledge that a successful phishing attack entails creating messages that are indistinguishable from the natural, expected messages by the intended victim, this concept has not been formalized. Our model captures phishing in terms of this indistinguishability between the natural and phishing message distributions. To the best of our knowledge, this is the first study that places phishing on a concrete theoretical framework and offers a new perspective to analyze this threat. We propose metrics to analyze the success probability of a phishing attack taking into account the input used by a phisher and the work involved to create deceptive email messages. Finally, we describe and study a new class of phishing attacks called collaborative spear phishing that may stem from the latest threat posed by the Epsilon email breach in the recent past and point out fundamental flaws in the current email-based marketing business model. In this sense, our study is very timely and presents new and emerging trends in phishing.

Keywords

Phishing, Email fraud, data hiding, identity linking, model, social engineering, security, privacy.

INTRODUCTION

Phishing is an example of a social engineering threat aimed at gleaning sensitive information such as user names, passwords and financial information from unsuspecting victims. Attacks are typically carried out via communication channels such as email or instant messaging by masquerading as legitimate and trustworthy entities. Being a social engineering attack, most studies of this threat have focussed on understanding the techniques used by phishers, devising clever strategies to thwart these attacks and the human factors associated with phishing. In this paper, we deviate from this empirical approach and propose a theoretical yet practical model that captures the dynamics of this threat. A novel feature of our security model is that it captures the inherent human factor and consequently complements the empirical study of phishing.

Our first contribution in this paper is the development of a theoretical framework for phishing. Our model is also very practical and designed to study of a large class of phishing attacks including the non-traditional, but latest threats such as the Android Market fake banking apps (Slashdot, 2010). It is well known that a successful phishing attack entails creating messages that are indistinguishable from the natural, expected messages by the intended victim. Firstly, we formalize this notion in the broadest sense possible to encompass a wide range of attacks. Our model captures the dynamics of phishing in terms of indistinguishability between the natural and phishing message distributions. From the perspective of a phisher, one can view the creation of a phishing message as an attempt to embed a deceptive message within an innocent looking email or instant message. To this end, we treat the problem to be "spiritually" similar to the problem of Steganography. Our motivation stems from the observation that while the goal in Steganography is to create an innocent looking message with a hidden payload without arousing the suspicion of *any* eavesdropper, a phisher tries to create an "innocent" looking message with a hidden (malicious) payload without arousing suspicion even from the recipient. Our work brings out a hidden connection between Steganography and Phishing and we hope that this connection will lead to new perspectives on phishing research.

Secondly, we propose metrics to measure the success probability of a phishing detection algorithm and consequently the success probability of a phishing attempt. We also define the notion of *overhead* as the ratio of the amount of *work* done by a phisher to the *payoff*. This notion of overhead will be useful when we analyse the impact of the Epsilon email breach and the associated payoff for the phishers.

Finally, we describe a new class of phishing attacks, called collaborative spear phishing, an advanced class of spear phishing that may stem from the latest threat posed by the Epsilon email breach in the recent past. A server breach at the Internet marketing company Epsilon, a unit of Alliance Data Systems Corporation, exposed

the names and email addresses of millions of people (News/Technology, 2011) across different organizations. This breach is being described as the worst of its kind by the media (Information Week, 2011), particularly since the breach apparently lasted for months despite warnings of targeted attacks against email service providers. We also point out some of the fundamental flaws in the current email-based marketing business model, which is a by-product of service industrialization. Thus, our study is very timely and presents new and emerging trends in phishing.

PRIOR WORK

Phishing is primarily a social engineering attack and has attracted a lot of research interest in this context. Most studies of phishing have focussed on understanding the techniques used by phishers, devising clever strategies to thwart these attacks and the human factors associated with this threat.

Dhamija et al. (Dhamija et al., 2006) and Downs et al. (Downs et al., 2006) studied the factors affecting the success of different malicious strategies used by phishers in an effort to build systems better capable of thwarting phishing attempts. The impact of social networking websites on phishing was studied by Jagatic et al. (**Error! Reference source not found.**) who found that Internet users may be over four times as likely to become victims if they are solicited by someone appearing to be a known acquaintance. Some of the strategies devised to thwart phishing attacks mentioned in the literature include: Dynamic Security Skins (**Error! Reference source not found.**) that allows a remote web server to prove its identity in a way that is easy for a human user to verify and hard for an attacker to spoof; Detecting phishing emails and websites (Fette et al., 2007) using machine learning techniques; Web Wallet(Wu et al., 2006), a browser sidebar which users can use to submit their sensitive information online; password management and website-login innovations (Yee and Sitaker, 2006) and Cantina, a novel, content-based approach to detecting phishing web sites, based on information retrieval and text mining algorithms (Zhang et al., 2007)]. Another line of research (Wu et al., 2006, Zhang et al., 2007) focuses on the evaluation of anti-phishing tools and their effectiveness.

A graph-theoretic model to analyse the effort expended by a phisher to launch an attack was studied by Jakobsson (Jakobsson, 2005). A phishing attack was modelled using a graph in which nodes correspond to knowledge and edges captured traversal from one node to another. Edges were associated with costs to reflect the effort of the phisher. This paper also defined a new attack approach called the context aware phishing attack using a method called identity linking - determining the correspondence between identities and email addresses of a victim.

Our model is designed to capture the dynamics of every facet of the phishing threat and not isolated to measuring the effort expended by the phisher. Furthermore, we describe attacks such as collaborative spear phishing that are far more complex than the context aware attack and thus subsumes the earlier attack put forth by Jakobsson (Jakobsson, 2005).

Notations and Definitions

For a probability distribution *P* with support *X*, we use the notation *P[x]* to denote the probability that *P* assigns to $x \in X$. A random variable *X* is a function over a sample space Ω, $X : \Omega \to S$, for some set *S* and we say that the random variable *X* takes values in the set *S*. The probability distribution on *S* described by the random variable *X* is denoted by P_X.

Statistical Distance

We use statistical distance as the measure of distance between two random variables and the probability distributions described by these random variables. The statistical distance is the largest possible difference between the probabilities that two probability distributions can assign to the same event. Shoup (Shoup, 2009) presents a detailed treatment of statistical distance and its properties.

Definition

Let *X* and *Y* be random variables which both take values in a finite set *S* with probability distributions P_X and P_Y. The statistical distance between *X* and *Y* is defined as

$$\Delta[X,Y] = \frac{1}{2}\sum_{s\in S}\left|P_X(s) - P_Y(s)\right|.$$

So, two random variables (and the corresponding probability distributions) *X* and *Y* are said to be ϵ-close if $\Delta[X,Y] \leq \epsilon$. This notion of ϵ-closeness will be useful to us when we talk about the two distributions – natural and phishing – being close to each other capturing the notion of indistinguishability.

THE PHISHING MODEL

In this section, we describe our phishing model as depicted by Figure 1 built on the Steganography security model presented by Cachin (Cachin, 1998). We use the notion of a communication channel to capture email, instant and other means of communication. For the purpose of our discussion here, let us use the example of email communication. Let us consider an individual's email inbox. The phishing problem specifies two message distributions corresponding to the two sources of messages that can find their way to that individual's email inbox: The Natural (*N)* and the Phishing (*P)* message distributions. The two source distributions are shown on the left as two black boxes. Typically, we are unaware of the exact probability distributions associated with these input sources and will treat them as such in our description. The individual's inbox normally receives messages from the Natural distribution (switch is set to *0*) corresponding to the phisher being inactive. The natural distribution is meant to capture the distribution of messages that a person *expects* to see. When the phisher is active (switch is set to *1*) s/he receives phishing messages. The algorithm used by the phisher operates on some input stream to create the deceptive messages. The *Distinguisher* algorithm *D* is tasked with being able to distinguish between the messages from these two distributions and essentially protect the user from being phished. Often, the receiver of the email plays the role of the distinguisher *D* although Figure 1 depicts the distinguisher algorithm *D* to be distinct from the receiver. In real life, the receiver along with the software tools, browser toolbar extensions, and spam/phisher filters collectively form the Distinguisher algorithm. The bidirectional arrow between the Distinguisher algorithm D and the receiver is meant to signify this relationship between these entities. The arrow out of the receiver pointing to output/action symbolizes the act of clicking on a link or acting upon the instructions in the received message, whether natural or phishing.

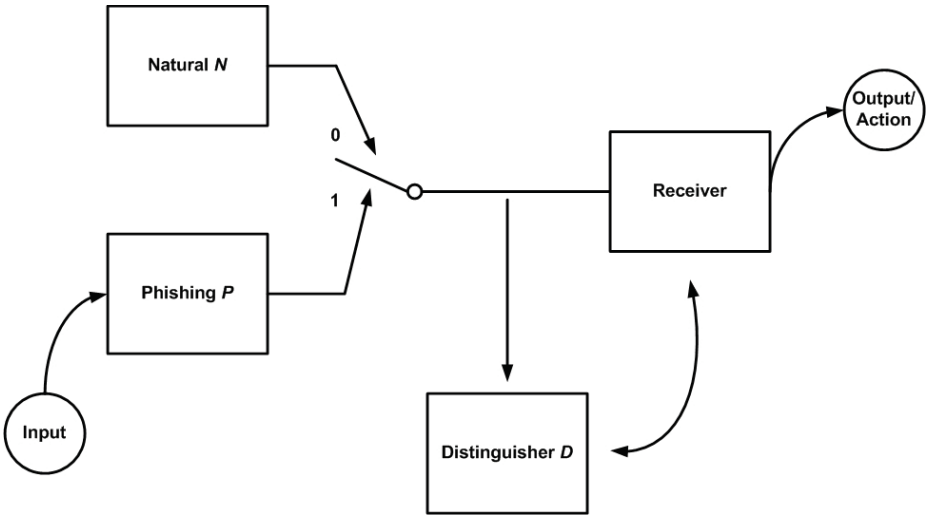

Figure 1: The Phishing Model.

Evaluating the Success of a Phishing Attempt

The success of a phishing attempt is measured by the intended victim's ability to distinguish between "natural" and "phishing" messages over the communication channel. To characterize natural communication we need to define and formalize the communication channel. We follow the standard terminology used in the literature (Hopper et al., 2002, Cachin, 1998) to define communication channels. We let $E = \{e_1, \ldots, e_s\}$ denote an alphabet and treat the *communication channel* as a family of random variables $\mathcal{I} = \{I_h\}_{h \in E^*}$. These channel distributions model a *history-dependent* notion of channel data that captures the notion of real-life communication. As an example, if *E* were to represent the set of the "*email alphabet*" and $h \in E^*$ the history of emails received by a person thus far, then I_h represents the random variable that captures the probability distribution of the person's email inbox at that point in time. In our model, we have captured the history dependence of communication and an individual's expectance to "see" a message in his inbox.

In evaluating the success of a phishing attempt, we need to take into consideration the amount of randomness present in a person's email inbox. We use *min-entropy* as the measure of this randomness. The *min-entropy* of a random variable X, taking values in a set V, is the quantity

$$H_\infty(X) \triangleq \min_{v \in V} (-\log \Pr[X = v]) \ .$$

We say that a communication channel (such as an email inbox) has min-entropy δ if for all $h \in E^*$, $H_\infty(I_h) \geq \delta$, We would like an individual's inbox, for all histories, to have some non-zero randomness, i.e, $\delta > 0$. This randomness parameter is designed to capture the diversity of the messages present in a person's inbox. As an example, if someone were to receive only *one* particular kind of email, then there is no randomness present in this communication scheme. The study of phishing on such a communication channel is not as interesting since the success probability of a phishing attempt in this situation is very small.

Let us now discuss the success probability of the Distinguisher algorithm *D* in being able to detect a phishing message. Let us overload the notation and let *P* denote the phishing algorithm as well as the distribution of the phishing messages produced by it. We now define the *advantage* of the Distinguisher *D* over the phishing algorithm *P* as:

$$\mathbf{Adv}_D^{\mathcal{P}}(m) = \left| \Pr\left[D(m) = \text{success}\right] - \frac{1}{2} \right| \quad , \qquad (1)$$

where *m* is the message to be distinguished and *D(m) = success* is the event that the Distinguisher *D* was successful in identifying a phishing message. Observe that any Distinguisher algorithm has an advantage of ½ in being able to detect a phishing message by merely flipping a fair coin. Hence, we need to look at the absolute value of the difference between the success probability of *D* from ½.

An alternative definition for the *advantage* of the Distinguisher *D* over the phishing algorithm *P* is obtained from the observation that the **total variation distance** between two probability measures *N* and *P* is the largest possible difference between the probabilities that these two probability distributions can assign to the same event, in particular to the event *D(m) = success*.

$$\mathbf{Adv}_D^{\mathcal{P}}(m) = \frac{1}{2} \sum_{m \in M} |N[m] - P[m]| \quad , \qquad (2)$$

where *N* and *P* are the natural and the phishing message distributions respectively and $m \in M$ represents the messages in the message set *M* (the user's inbox). Our model captures phishing in terms of this indistinguishability between the natural and phishing message distributions.

We can now define the *capacity* **C** of an individual to shield him/her from a phishing attack as:

$$\mathbf{C} = \max_{\mathcal{D}} \{\mathbf{Adv}_D^{\mathcal{P}}(m)\}, \qquad (3)$$

this maximum taken over all Distinguisher algorithms *D* available at the individual's disposal. This definition captures the different software tools such as browser toolbars, add-ons and other installed tools that one might use to defend against phishing.

We can now derive the measure for evaluating the success probability $\boldsymbol{S_P}$ of a phishing attempt *P* as:

$$\mathbf{S_P} = \mathbf{1} - \mathbf{C}. \qquad (4)$$

We say that a user is (ϵ, δ)-*secure* from a phishing attack if for all his email-inboxes with min-entropy δ, we have $\mathbf{S_P} \leq \epsilon$.

The *overhead* of a phisher is judged by the relation between the amount of work done by a phisher and the *payoff*. We adopt the ratio *o = w/p* as a measure for overhead. Obviously, if the payoff is high and the work done is low, then the overhead is low. This measure is useful in comparing the *damage* caused by different phishing attacks.

In this section, we discuss the different parameters that contribute towards the *work* done by a phisher. Drake et. al, present an anatomy of a phishing email (Drake et al., 2004) where they enumerate the different tricks used by phishers in an attempt to create deceptive messages that are indistinguishable from the original messages. The most important (and expensive to acquire) of these parameters are the personally identifiable information (PII) such as name, email address, the final four digits of an account number, year of expiration etc. The other costs associated with work are technical in nature, i.e., creating similar sounding domain names such as tax-revenue.com, ebaybuyerprotection.com, creating emails that appear to come from legitimate "From:" email address, designing the structure and content of the email, creating a plausible premise, using Javascript event handlers, redirection, etc. We define work to comprise essentially of two main parts – work done in collecting PII and the technical work, i.e, $w = w_{PII} + w_t$.

COLLABORATIVE SPEAR PHISHING

In this section we introduce a new class of phishing attacks, that we call collaborative spear phishing. We wish to shed light on this new class of phishing attacks that may become popular as a result of the latest server breach at the email marketing giant Epsilon. This attack is an advanced class of spear phishing that a phisher may develop using collaborative filtering techniques described below. In April 2011, a server breach at the Internet marketing company Epsilon, a unit of Alliance Data Systems Corporation, exposed the names and email addresses of millions of people (News/Technology, 2011)]. While a complete list of all the companies affected by the breach is not yet known, roughly 50 companies are said to be on that list, including Best Buy, Citibank, Disney, JPMorgan Chase, The Home Shopping Network, Hilton, Marriott and the College Board. This breach is being described as the worst of its kind by the media (Information Week, 2011).

Collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple data sources. Commonly used to infer purchase statistics by implementing recommendation algorithms for item recommendation by Amazon and other online retailers, this technique can now be used to launch highly advanced phishing attacks. While any breach that leaks personally identifiable information is a blessing to phishers, this particular breach at Epsilon is much more so. In the context of this breach, a phisher might now try to infer potential accounts that an individual *may* have with organizations using information that he already possesses. Furthermore, it gives a plausible premise that a phisher may use to hide his tracks. Observe that the breach at Epsilon leaked much more information than just personally identifiable information – It leaked the relationships that an individual with different organizations. The phisher is able to observe that a particular account is affiliated with a number of organizations and hence is able to filter for more information than s/he could otherwise.

As a quick example, we use a very simple Item-to-Item recommendation algorithm to illustrate this attack. The table below captures Alice, Bob and Emily's relationship with three organizations. A *Yes* in the table below corresponds to the affirmative knowledge that a phisher has obtained (Using the Epsilon database) about an individual's relationship with that organization and *No* (no knowledge) corresponds to the lack of this knowledge.

Table 1: Collaborative Phishing

Name	Best Buy	Citibank	JPMorgan Chase
Alice	Yes	No	Yes
Bob	No	Yes	Yes
Emily	No	Yes	No

The cosine between Best Buy and Citibank is obtained by:

$$\frac{(1,0,0)\cdot(0,1,1)}{||(1,0,0)||||(0,1,1)||} = 0.$$

The cosine between Best Buy and JPMorgan Chase is obtained by:

$$\frac{(1,0,0)\cdot(1,1,0)}{||(1,0,0)||||(1,1,0)||} = \frac{1}{\sqrt{2}}.$$

The cosine between Citibank and JPMorgan Chase is obtained by:

$$\frac{(0,1,1)\cdot(1,1,0)}{||(0,1,1)||||(1,1,0)||} = \frac{1}{2}.$$

Hence, a phisher armed with the knowledge that a particular individual who has an account with Best Buy can make an educated guess that h/she may possibly have an account with JPMorgan Chase as well. This makes good sense because many Best Buy Credit accounts are indeed handled by JPMorgan Chase. While we have used a very elementary algorithm for the sake of exposition, a motivated phisher could use an elaborate collaborative filtering algorithm such as Slope One (Lemire and Maclachlan, 2005) to improve the success of this attack. While the context-aware attack proposed by Jakobsson (Jakobsson, 2005) uses the concept of identity-linking to launch phishing attacks, our proposed attack is not only context-aware but also is capable of extrapolating for information that the phishers don't yet have.

In this paragraph, we point out some of the fundamental flaws in the current email-based marketing business model, which we believe is a by-product of service industrialization - treating services as an industrial process. By placing the personally identifiable information of millions of customers under the control of one organization, such as Epsilon, the *overhead* for the phisher is dramatically reduced – The work is diminished and the payoff is maximised. Furthermore, the phishers can now send targeted emails to their victims thereby making sure that these emails are out of the hands of the phishing research community. They can also ensure guaranteed delivery of their phishing emails by spoofing the correct "From" email addresses that most people have saved in their address books. Gary Warner (CyberCrime and doing Time, 2011) has an elaborate discussion of such targeted phishing attacks.

CONCLUSION

Our primary goal in this paper was to present a treatment of phishing in a formal theoretical framework. Our model captures the dynamics of phishing in terms of indistinguishability between the natural and phishing message distributions. We propose metrics to analyze the success probability of a phishing attack which takes into account the input parameters used by a phisher and the associated work involved to create deceptive email messages. Finally, we present a new class of phishing attacks, called collaborative spear phishing which is an advanced class of spear phishing that may stem from the latest threat posed by the Epsilon email breach in the recent past. We also point out some of the fundamental flaws in the current email-based marketing business model, which is a by-product of service industrialization. In this sense, our study is very timely and presents new and emerging trends in phishing. We hope that our model will help shed some more light on the threats posed by phishing.

REFERENCES

Cachin, C. (1998). An information-theoretic model for steganography. In *Information Hiding*, pages 306–318.

CyberCrime and Doing Time. (2011). The epsilon phishing model. Retrieved 12 Mar, 2011, from http://-garwarner.blogspot.com/2011/04/epsilon-phishing-model.html.

Dhamija, R. and Tygar, J. (2005). The battle against phishing: Dynamic security skins. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 77–88. ACM.

Dhamija, R., Tygar, J., and Hearst, M. (2006). Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM.

Downs, J., Holbrook, M., and Cranor, L. (2006). Decision strategies and susceptibility to phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security*, pages 79–90. ACM.

Drake, C., Oliver, J., and Koontz, E. (2004). Anatomy of a phishing email. In *First Conference on Email and Anti-Spam (CEAS), Mountain View, CA, USA*, pages 2–3. Citeseer.

Fette, I., Sadeh, N., and Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*, pages 649–656. ACM.

Hopper, N., Langford, J., and Von Ahn, L. (2002). Provably secure steganography. *Advances in Cryptology—CRYPTO 2002*, pages 119–123.

InformationWeek (2011). Epsilon fell to spear-phishing attack. Retrieved 15 Mar, 2011, from http://-www.informationweek.com/news/security/attacks/229401372.

Jagatic, T., Johnson, N., Jakobsson, M., and Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10):94–100.

Jakobsson, M. (2005). Modeling and preventing phishing attacks. *LECTURE NOTES IN COMPUTER SCIENCE*, 3570:89.

Lemire, D. and Maclachlan, A. (2005). Slope one predictors for online rating-based collaborative filtering. *Society for Industrial Mathematics*.

News/Technology, ABC News. (2011). Epsilon email breach: What you should know. Retrieved 12 Mar, 2011, from http://abcnews.go.com/Technology/epsilon-email-breach/story?id=13291589.

Shoup, V. (2009). *A computational introduction to number theory and algebra*. Cambridge Univ Pr.

Slashdot (2010). Malicious app in android market. Retrieved 12 Mar, 2011, from http://mobile.slashdot.org/-story/10/01/10/2036222/Malicious-App-In-Android-Market.

Wu, M., Miller, R., and Garfinkel, S. (2006). Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 601–610. ACM.

Wu, M., Miller, R., and Little, G. (2006). Web wallet: preventing phishing attacks by revealing user intentions. In *Proceedings of the second symposium on Usable privacy and security*, pages 102–113. ACM.

Yee, K. and Sitaker, K. (2006). Passpet: convenient password management and phishing protection. In *Proceedings of the second symposium on Usable privacy and security*, pages 32–43. ACM.

Zhang, Y., Egelman, S., Cranor, L., and Hong, J. (2007). Phinding phish: Evaluating anti-phishing tools. In *Proceedings of the 14th annual network and distributed system security symposium (NDSS 2007)*. Citeseer.

Zhang, Y., Hong, J., and Cranor, L. (2007). Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th international conference on World Wide Web*, pages 639–648. ACM.