Optimal Personalized Filtering Against Spear-Phishing Attacks

Aron Laszka and **Yevgeniy Vorobeychik** and **Xenofon Koutsoukos**
Institute for Software Integrated Systems
Department of Electrical Engineering and Computer Science
Vanderbilt University
Nashville, TN

Abstract

To penetrate sensitive computer networks, attackers can use spear phishing to sidestep technical security mechanisms by exploiting the privileges of careless users. In order to maximize their success probability, attackers have to target the users that constitute the weakest links of the system. The optimal selection of these target users takes into account both the damage that can be caused by a user and the probability of a malicious e-mail being delivered to and opened by a user. Since attackers select their targets in a strategic way, the optimal mitigation of these attacks requires the defender to also personalize the e-mail filters by taking into account the users' properties.

In this paper, we assume that a learned classifier is given and propose strategic per-user filtering thresholds for mitigating spear-phishing attacks. We formulate the problem of filtering targeted and non-targeted malicious e-mails as a Stackelberg security game. We characterize the optimal filtering strategies and show how to compute them in practice. Finally, we evaluate our results using two real-world datasets and demonstrate that the proposed thresholds lead to lower losses than non-strategic thresholds.

1 Introduction

To successfully breach highly secure systems, attackers have to focus on the weakest link in the chain of security, which is often the users (Sasse, Brostoff, and Weirich 2001). One particularly pernicious form of attack on users is spear phishing, that is, targeting specific users (or classes of users) through malicious e-mail, making use of their individual characteristics, such as who their bosses or friends are, to build trust (Hong 2012). In recent years, we have seen several spear-phishing attacks that successfully breached highly secure organizations. For example, in 2011, the Oak Ridge National Laboratory, which conducts classified and unclassified energy and national security work, was breached by a spear-phishing attack (Zetter 2011). In this incident, the attackers sent an e-mail, which claimed to be from human resources, to the lab employees. This e-mail contained a link to a malicious website, which infected the employees' computers with a malware that subsequently stole sensitive data and sent it to an unknown destination. As another example, in 2012, one of the White House internal networks was breached using spear phishing (McCullagh 2012). The attackers, who are believed to have used servers in China, were allegedly able to access the network of the president's military office, which is in charge of, for example, strategic nuclear commands. Finally, computers at the Nuclear Regulatory Commission (NRC) of the U.S., which contain sensitive information that could be used for surveillance or sabotage, were breached three times in the past three years (Rogers 2014). In the most recent incident, the attackers first compromised an NRC employee's personal e-mail account, which they then used to send e-mails to 16 other employees. The e-mail contained a malicious PDF attachment, which infected the computer of an employee who opened the attachment (Rosenblatt 2014).

The defining characteristic of spear-phishing attacks which differentiates them from regular phishing or spam is that they are targeted at specific, carefully chosen individuals or groups. Since sending a large number of similar e-mails (e.g., with the same malicious attachment) would almost certainly raise an alarm, the attackers focus on a subset of the users who constitute the weakest links of the system. Moreover, the emergence of digital and social media has made it easier for attackers to know much about their prospective targets, such as where they work, what they are interested in, and who their friends are (McAfee Labs 2014; Jagatic et al. 2007).

Typical mitigation for phishing attacks is the same as for spam: there is an e-mail filtering system, often based in part on machine learning, which computes a risk score for each e-mail and filters those for which the risk score exceeds some pre-specified threshold. The value of this filtering threshold has to be carefully chosen, since overzealous filtering may also remove many non-malicious e-mails. Hence, defenders have to find the right balance between security and usability (Sheng et al. 2009). Furthermore, these thresholds can be personalized, as different users have different levels of carefullness and different potential to cause damage. For example, a recent report found – based on a large-scale experiment – that the departments which hold the most sensitive data in a business, such as HR, accounting, and finance, are the worst at detecting fraud (McAfee Labs 2014).

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

However, the targeted nature of spear phishing makes the problem qualitatively different: since the attacker selects the target users by taking into account both their individual properties and their filtering thresholds, the defender has to set the thresholds in a strategic way. In this paper, we investigate the problem of optimally setting personalized filtering thresholds against spear-phishing attacks, given an e-mail classifier with its associated false-negative / false-positive probability tradeoff. Specifically, we model this problem as a Stackelberg game, characterize the optimal filtering strategies, and show how these filtering strategies can be computed in practice at scale. We also evaluate the proposed filtering strategies using real e-mail data, demonstrating that our approach leads to better outcomes for the defender.

The remainder of this paper is organized as follows. In Section 2, we discuss related work on filtering malicious e-mails. In Section 3, we introduce our game-theoretic model. In Section 4, we present analytical results on our model. In Section 5, we present numerical results. Finally, in Section 6, we give our concluding remarks.

2 Related Work

There are many research results on measuring users' susceptibility to phishing attacks and the detection and classification of potentially malicious e-mails. These results are complementary to ours, since we assume that the users' susceptibility has been measured and a classifier has been trained, and we build our model on these assumptions.

Several experiments have been conducted to measure individuals' susceptibility to phishing attacks. For example, the authors of (Jagatic et al. 2007) performed an experimental study at Indiana University to measure individuals' probabilities of falling victim to phishing. To measure these probabilities, the authors launched an actual (but harmless) phishing attack targeting students and utilizing publicly available acquaintance data mined from social-network websites. The results show that certain characteristics of the targeted students, such as gender, academic major, and grade, have a significant effect on the probabilities. As another example, the authors of (Sheng et al. 2010) performed an online survey to study the relationship between demographic and phishing susceptibility. The study, which was based on an online roleplaying task, found that certain factors, such as gender and age, have a significant effect.

The problem of detecting malicious e-mails has also been extensively studied. For example, the authors of (Fette, Sadeh, and Tomasic 2007) apply machine learning to a feature set designed to highlight user-targeted deception. When evaluated on a real-world dataset, their method correctly identified over 96% of the phishing emails while misclassifying only approximately 0.1% of the non-malicious e-mails. More recently, the problem of classifying malicious e-mails has also been studied as an adversarial data-mining problem. In adversarial data mining (or adversarial machine learning), the classification problem is viewed as a game between the classifier and an adversary, who manipulates the instances to be classified in order to increase the number of false negatives (Dalvi et al. 2004). For example, the authors of (L'Huillier, Weber, and Figueroa 2009) build an adversary-aware classifier for detecting phishing e-mails using an online version of Weighted Margin Support Vector Machines, and they present experimental results showing that it is highly competitive compared to previous online classification algorithms.

Besides their textual content, phishing e-mails can also often be identified by detecting links to malicious websites, which can initiate a drive-by download or install. The authors of (Ma et al. 2009) study the problem of detecting malicious websites and propose a website classifier, which uses statistical methods, lexical features of the URL, and host-based features, such as WHOIS and geographic properties. As another example, the authors of (Choi, Zhu, and Lee 2011) propose a method using machine learning to detect malicious URLs and to identify the nature of the attack. The proposed method uses a variety of features, including lexical features of the URL, link popularity of the website, content features of the webpage, and DNS features.

3 Model

Now, we introduce our game-theoretic model of filtering targeted and non-targeted malicious e-mails. For a list of symbols used in this paper, see Table 1.

Table 1: List of Symbols

Symbol	Description
$FP(f)$	false-positive probability given that the false-negative probability is f
A	number of users targeted by the attacker
L_u	expected damage for delivering targeted malicious e-mails to user u
N_u	expected damage for delivering non-targeted malicious e-mails to user u
C_u	expected loss from filtering out non-malicious e-mails to user u
f_u^T	optimal false-negative probability of user u given that the user is targeted
f_u^N	optimal false-negative probability of user u given that the user is not targeted

We assume that the e-mail classifier of the organization outputs a maliciousness score for each received e-mail, and an e-mail is delivered to the recipient if and only if the score is below a given threshold. We call misclassified malicious e-mails *false negatives* (i.e., when a malicious e-mail is below the threshold) and misclassified non-malicious e-mails *false positives* (i.e., when a non-malicious e-mail is above the threshold). By adjusting the filtering threshold, the organization can increase the probability of false positives and decrease the probability of false negatives, or vice versa.

We represent the attainable false-positive and false-negative probability pairs using a function $FP : [0, 1] \mapsto [0, 1]$, where $FP(FN)$ is the probability of false positives when the the probability of false negatives is FN. In any practical classifier, FP is a non-increasing function of FN. For analytical tractability, we further assume that FP is a

continuous, strictly decreasing, and strictly convex function of FN. Note that, in Section 5, we show that our results can be applied successfully to FP functions that do not satisfy these additional assumptions.

We let L_u denote the expected amount of damage (i.e., loss) that the organization sustains for delivering malicious targeted e-mails to user u. This amount L_u can be computed as

$$\begin{aligned} L_u = &\,\mathrm{E}[\text{damage to organization} \mid \text{user } u \text{ falls victim}] \\ &\times \Pr[\text{user } u \text{ falls victim} \mid \text{e-mail is delivered}] \\ &\times \text{rate of targeted attacks.} \end{aligned} \tag{1}$$

In practice, any organization that aims to be prepared against cyber-attacks needs to have some estimate of its cyber-assets' value and the expected frequency of attack attempts; hence, it should be able to estimate the first and the third factors. Moreover, the second factor (i.e., the probability of falling victim) can be measured by sending probe e-mails to the users.

Besides spear phishing, the organization also receives non-targeted malicious e-mails. We let N_u denote the loss that the organization sustains for delivering malicious non-targeted e-mails to user u. Finally, an organization also has to take into account the production and usability loss sustained when a non-malicious e-mail is filtered out. We let C_u denote the amount of loss sustained for not delivering non-malicious e-mails addressed to user u.

Attacker-Defender Game

We model the conflict between the targeting attacker and the organization as a Stackelberg security game, where the defender's role is played by the organization.

The attacker's strategic choice is to select a subset of users $\mathcal{A}$ to whom she sends malicious e-mails. Since a large number of e-mails containing the same malware or linking to websites distributing the same malware could easily be detected, the attacker tries to stay covert by sending only a limited number of e-mails. Formally, we model this limitation by assuming that the attacker's strategy has to satisfy $|\mathcal{A}| \leq A$, where A is a constant.

The defender's strategic choice is to select the false-negative probability f_u for each user u. Recall that the resulting false-positive probability for user u is $FP(f_u)$.

For a given strategy profile $(\boldsymbol{f}, \mathcal{A})$, the players' payoffs are defined as follows. The attacker's payoff is

$$\mathcal{U}_{\text{attacker}} = \sum_{u \in \mathcal{A}} f_u L_u \ , \tag{2}$$

and the defender's loss (i.e., inverse payoff) is

$$\mathcal{L}_{\text{defender}} = \mathcal{U}_{\text{attacker}} + \sum_{u} f_u N_u + FP(f_u) C_u \tag{3}$$

$$= \sum_{u \in \mathcal{A}} f_u L_u + \sum_{u} f_u N_u + FP(f_u) C_u \ . \tag{4}$$

In the analysis, our goal will be to find the attacker's best response and the defender's optimal strategies, which are defined as follows.

Definition 1. An attacker strategy is a *best response* if it maximizes the attacker's payoff, taking the defense strategy as given.

As is typical in the security literature, we consider subgame perfect Nash equilibria as our solution concept (Korzhyk et al. 2011). We will refer to the defender's equilibrium strategies as optimal strategies for the remainder of the paper. Note that, as we will discuss at the beginning of Section 4, our model allows the attacker to break ties between multiple best-response strategies in an arbitrary manner.

Definition 2. We call a defense strategy *optimal* if it maximizes the defender's payoff given that the attacker will always play a best-response strategy.

4 Analysis

We begin our analysis with characterizing the attacker's best-response strategies and then study the problem of finding an optimal defense strategy.

From Equation (2), it follows immediately that, against a given defense strategy $\boldsymbol{f}$, the targeting attacker's best-response strategy is to choose the set of A users with the highest $f_u L_u$ values. Furthermore, if there are multiple best-response strategies (i.e., multiple sets of users attaining the same sum), then these strategies all yield the same payoff to the defender as well, since the defender's payoff depends on the attacker's strategy only through the attacker's payoff (see first term in Equation (4)). In other words, the attacker can break ties between best responses in an arbitrary way.

To facilitate our analysis, we now introduce some additional notation. Let f_u^T denote the optimal value of f_u given that $u \in \mathcal{A}$, and let f_u^N denote the optimal value of f_u given that $u \notin \mathcal{A}$. Formally, for each user u, f_u^T and f_u^N are the values at which the minima of

$$f_u(L_u + N_u) + FP(f_u) C_u \tag{5}$$

and

$$f_u N_u + FP(f_u) C_u \tag{6}$$

are attained, respectively. Note that it is fairly easy to show that these values are well-defined and unique for each user.

Optimal Defense Subproblem

First, we study an important subproblem of finding an optimal defense strategy. Suppose that a set of users $\mathcal{A}$ is given, and we are looking for the optimal defense strategy against which $\mathcal{A}$ is a best-response strategy for the attacker. In other words, we restrict our search space to defense strategies in which the users of $\mathcal{A}$ have the highest $f_u L_u$ values.

We begin with a special case, in which the parameter values of the users in $\mathcal{A}$ differ substantially from those of the remaining users.

Proposition 1. *Suppose that a set of users $\mathcal{A}$ is given, and the defender's choice is restricted to strategies against which $\mathcal{A}$ is a best response. If $\min_{u \in \mathcal{A}} f_u^T L_u \geq \max_{u \notin \mathcal{A}} f_u^N L_u$, then choosing f_u^T for every $u \in \mathcal{A}$ and choosing f_u^N for every $u \notin \mathcal{A}$ is the optimal defense strategy.*

Proof. (Sketch.) Firstly, $\mathcal{A}$ is a best response for the attacker, since the users in $\mathcal{A}$ have the highest $f_u L_u$ values. Secondly, for each $u \in \mathcal{A}$, $f_u = f_u^T$ is optimal by definition, and for each $u \notin \mathcal{A}$, $f_u = f_u^N$ is also optimal by definition. Then, as the defender's loss is the sum of the losses for the individual users, the strategy $\boldsymbol{f}$ must also be optimal for the given $\mathcal{A}$. □

It is noteworthy that this strategy profile (i.e., the defender's strategy $\boldsymbol{f}$ given by Proposition 1 and the attacker's strategy $\mathcal{A}$) would actually be a unique Nash equilibrium in a simultaneous version of the game, as both players' strategies are best responses.[1] However, this Nash equilibrium is not necessarily a subgame perfect Nash equilibrium in our Stackelberg game.

The above proposition provides a complete characterization of the optimal defense strategy for a special case. Next, we consider the general case, where the condition of Proposition 1 might not hold, and provide necessary conditions on the optimal defense strategy.

Theorem 1. *Suppose that a set of users $\mathcal{A}$ is given, and the defender's choice is restricted to strategies against which $\mathcal{A}$ is a best response. Then, in an optimal defense strategy, there exists a value Λ such that*

- *for every $u \in \mathcal{A}$, $f_u L_u = \Lambda$ if $f_u^T L_u < \Lambda$, and $f_u = f_u^T$ otherwise,*
- *for every $u \notin \mathcal{A}$, $f_u L_u = \Lambda$ if $f_u^N L_u > \Lambda$, and $f_u = f_u^N$ otherwise.*

Intuitively, the above theorem states that, in an optimal defense, users u in $\mathcal{A}$ with a sufficiently high $f_u^T L_u$ will have $f_u L_u = f_u^T L_u$, users u not in $\mathcal{A}$ with a sufficiently low $f_u^N L_u$ will have $f_u L_u = f_u^N L_u$, and all other users u will have a uniform $f_u L_u$ value, which we let be denoted by Λ. See Figure 1 for an illustration.

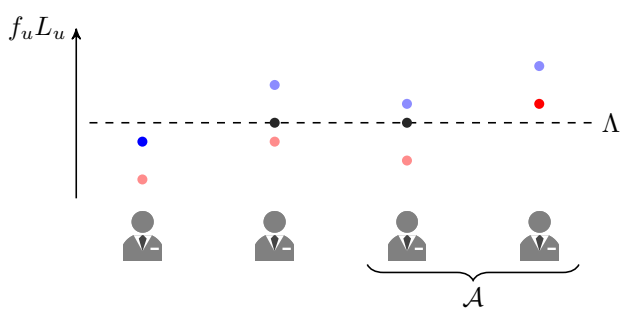


Figure 1: Illustration for Theorem 1 with four users and $A = 2$. Blue dots represent $f_u^N L_u$ values, and red dots represent $f_u^T L_u$ values.

[1]The uniqueness of the equilibrium follows from the observation that, if a set $\mathcal{A}$ satisfies the condition of the above lemma, then no other set can satisfy it. Furthermore, it can easily be shown that the game has a Nash equilibrium only if there exists a set $\mathcal{A}$ satisfying the condition of the lemma.

Proof. (Sketch.) It is obvious that $\min_{u \in \mathcal{A}} f_u L_u \geq \max_{u \notin \mathcal{A}} f_u L_u$ is a necessary and sufficient condition for $\mathcal{A}$ to be a best response. Then, given an optimal defense strategy $\boldsymbol{f}$, let Λ be $\max_{u \notin \mathcal{A}} f_u L_u$. We have to show that each $f_u L_u$ takes the value $f_u^T L_u$, $f_u^N L_u$, or Λ given by the lemma.

First, if $u \in \mathcal{A}$, then the optimal value for $f_u L_u$ would be $f_u^T L_u$; however, the actual value cannot be lower than Λ, since $\mathcal{A}$ would not be a best response otherwise. Using the convexity of $FP(f_u)$, it can then be shown that $f_u L_u = \Lambda$ is an optimal choice whenever $f_u^T L_u < \Lambda$.

Second, if $u \notin \mathcal{A}$, then the optimal value for $f_u L_u$ would be $f_u^N L_u$; however, the actual value cannot be higher than Λ by definition (recall that we let $\Lambda = \max_{u \notin \mathcal{A}} f_u L_u$ for the proof). Again, using the convexity of $FP(f_u)$, it can be shown that $f_u L_u = \Lambda$ is an optimal choice whenever $f_u^N L_u > \Lambda$. □

Note that, based on the above theorem, we can easily find the optimal Λ value for any given set $\mathcal{A}$ using, for example, a binary search. Consequently, we can find an optimal defense strategy by iterating over all A-sized subsets of the users and solving each defense subproblem.

Generally, this approach is not feasible in practice, as the number of possible subsets increases exponentially with A. However, if the number of users that can be targeted by the attacker is very limited, we can find the attacker's best response using an exhaustive search. In the case $A = 1$, this simply means iterating over the set of users. For the general case, we provide an efficient approach in the following subsection.

Optimal Defense

The previous theorem establishes that, in an optimal defense strategy, the users' f_u values are either f_u^T, f_u^N, or some $\frac{\Lambda}{L_u}$. Now, we discuss how this observation can be used to find an optimal defense strategy. The following theorem shows how to find an optimal strategy for a given Λ value. Note that this differs from the assignments in Theorem 1, where the set $\mathcal{A}$ was given.

Theorem 2. *Suppose that we are given a constant Λ, and the defender's choice is restricted to strategies where $\max_{u \notin \mathcal{A}} f_u L_u \leq \Lambda$ and $\min_{u \in \mathcal{A}} f_u L_u \geq \Lambda$ for a best response $\mathcal{A}$[2]. Then, the output of the following algorithm is an optimal defense strategy:*

1. *For each user u, compute the loss of user u when it is not targeted as follows: if $f_u^N L_u < \Lambda$, then the loss is $f_u^N N_u + FP(f_u^N) C_u$; otherwise, the loss is $\frac{\Lambda}{L_u} N_u + FP(\frac{\Lambda}{L_u}) C_u$.*
2. *For each user u, compute the loss of user u when it is targeted as follows: if $f_u^T L_u > \Lambda$, then the loss is $f_u^T (L_u + N_u) + FP(f_u^T) C_u$; otherwise, the loss is $\frac{\Lambda}{L_u}(L_u + N_u) + FP(\frac{\Lambda}{L_u}) C_u$.*

[2]Recall that the attacker always targets the A users with the highest $f_u L_u$ values; hence, both $\max_{u \notin \mathcal{A}} f_u L_u$ and $\min_{u \in \mathcal{A}} f_u L_u$ are uniform over the best responses.

3. *For each user u, let the cost of user u being targeted be the difference between the above computed loss values.*
4. *Select a set $\mathcal{A}$ of A users with the lowest costs of being targeted.*
5. *For every $u \in \mathcal{A}$, let $f_u = f_u^T$ if $f_u^T L_u > \Lambda$, and let $f_u = \frac{\Lambda}{L_u}$ otherwise.*
6. *For every $u \notin \mathcal{A}$, let $f_u = f_u^N$ if $f_u^N L_u < \Lambda$, and let $f_u = \frac{\Lambda}{L_u}$ otherwise.*
7. *Output the strategy $\boldsymbol{f}$.*

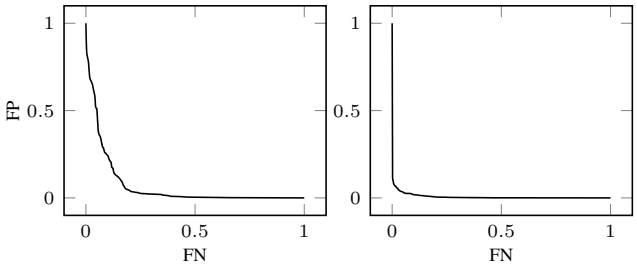
Proof. (Sketch.) First, suppose that – besides Λ – a best response $\mathcal{A}$ is also given. In other words, the defender's choice is restricted to strategies against which $\mathcal{A}$ is a best response, $\max_{u \notin \mathcal{A}} f_u L_u \leq \Lambda$, and $\min_{u \in \mathcal{A}} f_u L_u \geq \Lambda$. Then, we can show that Steps 5 and 6 of the above algorithm are optimal using an argument similar to the one in the proof of Theorem 1.

Second, we show that Steps 1 to 4 yield an optimal set $\mathcal{A}$. For the sake of contradiction, suppose that for some instance of the game, there exists a set $\mathcal{A}^*$ that leads to lower expected loss for the defender. Note that, since we already have that Steps 5 and 6 give an optimal assignment for any set, we can assume that the defense strategies corresponding to the sets $\mathcal{A}$ and $\mathcal{A}^*$ are given by Steps 5 and 6. Now, let u^+ be a user that is in $\mathcal{A}^*$ but not in $\mathcal{A}$, and let u^- be a user that is in $\mathcal{A}$ but not in $\mathcal{A}^*$. By removing u^+ and adding u^- to $\mathcal{A}^*$, the defender's expected loss is decreased by the difference between the costs of u^+ and u^- being targeted. Since $\mathcal{A}$ consists of the A users with the lowest costs of being targeted (see Step 4), this difference has to be non-negative; hence, the expected loss is not increased by such changes to $\mathcal{A}^*$. Then, using at most A such changes, we can transform $\mathcal{A}^*$ into $\mathcal{A}$, without increasing the expected loss. However, this contradicts the assumption that $\mathcal{A}^*$ leads to lower expected loss than $\mathcal{A}$; therefore, the original claim must hold. □

Efficient Search Let $\mathcal{L}_{\text{defender}}(\Lambda)$ denote the minimum loss that the defender can achieve for a given Λ value (i.e., the defender's loss for the defense strategy output by the algorithm of Theorem 2 and the attacker's best response against it). Then, finding an optimal defense strategy is equivalent to finding $\operatorname{argmin}_\Lambda \mathcal{L}_{\text{defender}}(\Lambda)$ (see Figures 3(a) and 3(b) for an illustration). Hence, we reduced the problem of finding an optimal defense strategy to the problem of optimizing a single scalar value.

5 Experiments

In this section, we evaluate our model using real-world datasets and compare our optimal strategic thresholds to uniform thresholds. Please note that the goal of these experiments is not to find a classifier that performs better than other classifiers in the literature, since our model assumes that a classifier and the resulting false positive / false negative curves are given. The goal of these experiments is to demonstrate the practical feasibility of our approach for setting the classification thresholds and to show that it outperforms non-strategic solutions.



(a) UCI dataset (b) Enron dataset

Figure 2: False-positive probability as a function of false-negative probability.

Datasets

We used two publicly available datasets for our numerical examples. For both datasets, we trained a naïve Bayes classifier.

UCI The first dataset is from the UCI Machine Learning Repository (Bache and Lichman 2013), which is a labeled collection of 4601 e-mail messages. Each e-mail has 57 features, most of which indicate frequencies of particular words or characters. We used 80% of the collection for training our classifier, and the remaining 20% for testing it, that is, for obtaining the false negative / false positive trade-off curve.

Enron The second dataset is the Enron e-mail dataset[3] (Klimt and Yang 2004). For each message, we computed 500 features based on the content of the message. We used 12 thousand e-mails from the dataset for training our classifier and 1500 e-mails for testing it.

Figures 2(a) and 2(b) show the false-positive probability (FP) as a function of the false-negative probability (FN) for the UCI and Enron datasets, respectively.

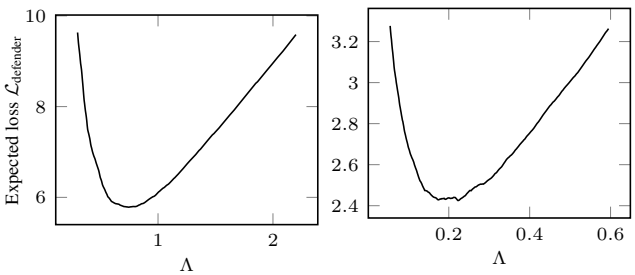
Finding Optimal Strategies

Recall that, in Section 3, we assumed the function $FP(FN)$ to be strictly convex. However, the actual curves shown by Figures 2(a) and 2(b) are only approximately convex, since they have a number of smaller irregularities. We now discuss how the overcome the challenges posed by these irregularities to the application of our theoretical results.

First, the values of f_u^N and f_u^T might be non-unique if the function $FP(FN)$ is not strictly convex. However, in practice, the probability of multiple global minima is negligible.[4] Nevertheless, if there were multiple values minimizing the defender's loss for user u when $u \in \mathcal{A}$, we could simply define f_u^T to be the maximal value. It is easy to see that this is the best choice, since it will allow us to use the optimal value $f_u^T L_u$ instead of Λ as long as possible. Similarly, we can define f_u^N to be the minimal value that minimizes the defender's loss for user u when $u \notin \mathcal{A}$.

[3] `http://www.cs.cmu.edu/~./enron/`

[4] In any case, we are limited by the size of the classifier's testing set and the actual precision of floating point numbers, so the existence of multiple global minima is mostly a peculiarity of the implementation.



(a) UCI dataset (b) Enron dataset

Figure 3: Expected loss as a function of Λ for $A = 3$.

Second, finding the values of f_u^N and f_u^T could be challenging, since the defender's loss can be a non-convex function of f_u. However, in practice, the function $FN(FP)$ is actually given by a set of datapoints, whose cardinality is necessarily upper bounded by the cardinality of the testing set of the classifier. Consequently, even a simple exhaustive search is feasible, since its running time will be linear in the size of the input.

Finally, finding the optimal value of Λ could also be challenging, since the objective function (i.e., the defender's expected loss) can be a non-convex function of Λ. Figures 3(a) and 3(b) show the defender's expected loss as functions of Λ for strategies computed using the algorithm of Theorem 2 for the UCI and Enron datasets, respectively. However, we can see that the objective function is relatively smooth in practice, it has only a few local minima, all of which are in the vicinity of the global minimum. Furthermore, we can even use an exhaustive search, since the function $\mathcal{L}_{\text{defender}}(\Lambda)$ is again given by a set of data points, whose cardinality is upper bounded by the number of users $\times$ cardinality of the testing set of the classifier. Hence, the running time of an exhaustive search will be quadratic in the size of the input.

Comparison with Non-Strategic Thresholds

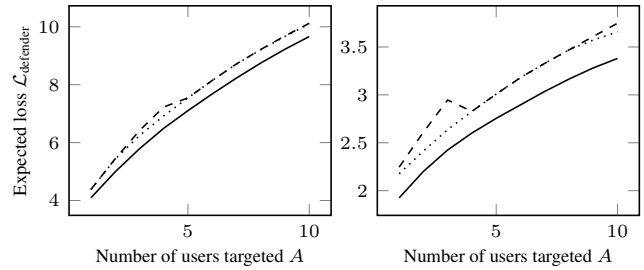
Now, we study the main question regarding our results: can the strategic setting of thresholds decrease the expected amount of losses? To answer this question, we compare our strategic thresholds with two non-strategic, uniform thresholds. These uniform thresholds do not expect the attacker to select the targets in a strategic manner, but they are otherwise optimal (i.e., minimize the expected losses).

Uniform Threshold #1 The first baseline assumes that the attacker targets the users uniformly at random; hence, the uniform false-negative probability f is computed as

$$\operatorname*{argmin}_f f\left(\sum_u N_u + \frac{A}{\sum_u 1}\sum_u L_u\right) + FP(f)\sum_u C_u\ .$$

Uniform Threshold #2 The second baseline assumes that the attacker targets those users who have the most potential to cause losses (i.e., have the highest L_u values); hence, the uniform false-negative probability f is computed as

$$\operatorname*{argmin}_f f\left(\sum_u N_u + \max_{\mathcal{A}:\ |\mathcal{A}|=A}\sum_{u\in\mathcal{A}} L_u\right) + FP(f)\sum_u C_u\ .$$



(a) UCI dataset (b) Enron dataset

Figure 4: Expected loss as a function of A for the optimal strategy (solid line) and uniform thresholds (dashed and dotted lines).

For the numerical examples, we generated a set of 31 users as follows:

- For every user, potential losses due to undelivered non-malicious and delivered targeted malicious e-mails are approximately ten times higher than losses due to delivered non-targeted e-mails. Formally, for each user u, $L_u, C_u \approx 10 \times N_u$. The motivation behind this choice is the standard assumption that undelivered non-malicious e-mails are much worse than delivered non-targeted malicious e-mails, such as spam. Furthermore, based on examples of spear-phishing attacks, it is reasonable to assume that targeted malicious e-mails are also much worse.
- The potential damage values L_u, C_u, and N_u follow a power law distribution. Formally, the number of users with damage values between some l and $l+1$ is approximately twice as much as the number of users with values between $l+1$ and $l+2$. Finally, the value of L_u ranges from 0.5 to 5.5. The motivation behind modeling the potential damage values with a power law distribution is the typical hierarchical structure of organizations, where the number of employees at a higher level is typically smaller.

Figures 4(a) and 4(b) compare our strategic solution to uniform thresholds at various attack sizes for the UCI and Enron datasets, respectively. The solid line (—) shows the defender's expected loss for our optimal strategy, the dashed line (- -) shows the loss for uniform threshold #1, and the dotted line (····) shows the loss for uniform threshold #2. Note that, for every threshold, we computed the defender's loss based on the attacker's best response in our model, as the goal is to compare how different thresholds perform against a targeting attacker.

We can see that the proposed strategic solution is clearly superior in every case. Furthermore, the improvement over the non-strategic thresholds is quite stable with respect to A, that is, the improvement does not diminish as the attacker targets more and more users. Finally, by comparing the results for the two datasets, we can see that the relative

improvement is higher for the more detailed dataset (i.e., Enron), which suggests that it is possible that our solution could lead to even higher improvements for more detailed datasets.

6 Conclusion

Since the weakest link in the chain of security is often human behavior, thwarting spear-phishing attacks is a crucial problem for any organization that aims to attain a high level of security. Besides user education, the most typical defense against phishing attacks is the filtering of malicious e-mails. In this paper, we focused on the problem of finding optimal filtering thresholds against targeted and non-targeted malicious e-mails.

The targeted, strategic nature of spear-phishing attacks presents an interesting problem, which we modeled as a Stackelberg security game. While characterizing the attacker's best response is trivial, characterizing and finding the defender's optimal strategy is much more challenging. However, using Theorem 2, we can reduce this problem to a much simpler scalar optimization, which – as we discussed in Section 5 – can be efficiently solved in practice, even for large datasets.

Finally, we evaluated our theoretical results using two real-world datasets, which result in typical false-negative / false-positive curves. We compared our strategic thresholds to two non-strategic thresholds, and found that our strategic thresholds are clearly superior. Furthermore, we also found that the improvement over the non-strategic thresholds is higher for the more detailed dataset and it does not diminish as the number targeted users increases. This shows that our method scales well not only computationally, but also performance-wise.

Acknowledgment

This work was supported in part by the National Science Foundation under Award CNS-1238959, by the Air Force Research Laboratory under Award FA8750-14-2-0180, and by Sandia National Laboratories.

References

Bache, K., and Lichman, M. 2013. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`.

Choi, H.; Zhu, B. B.; and Lee, H. 2011. Detecting malicious web links and identifying their attack types. In *Proceedings of the 2nd USENIX Conference on Web Application Development*, 1–12. USENIX Association.

Dalvi, N.; Domingos, P.; Sanghai, S.; Verma, D.; et al. 2004. Adversarial classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 99–108.

Fette, I.; Sadeh, N.; and Tomasic, A. 2007. Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, 649–656. New York, NY, USA: ACM.

Hong, J. 2012. The state of phishing attacks. *Communications of the ACM* 55(1):74–81.

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

Klimt, B., and Yang, Y. 2004. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning (ECML)*, 217–226. Springer.

Korzhyk, D.; Yin, Z.; Kiekintveld, C.; Conitzer, V.; and Tambe, M. 2011. Stackelberg vs. Nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research* 41(2):297–327.

L'Huillier, G.; Weber, R.; and Figueroa, N. 2009. Online phishing classification using adversarial data mining and signaling games. In *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics*, 33–42. ACM.

Ma, J.; Saul, L. K.; Savage, S.; and Voelker, G. M. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1254. ACM.

McAfee Labs. 2014. Threats report - August 2014. `http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q2-2014.pdf`.

McCullagh, D. 2012. White House confirms 'spearphishing' intrusion. CNET, `http://www.cnet.com/news/white-house-confirms-spearphishing-intrusion/`.

Rogers, J. 2014. Hackers attack Nuclear Regulatory Commission 3 times in 3 years. FOX News, `http://www.foxnews.com/tech/2014/08/20/hackers-attack-nuclear-regulatory-commission/`.

Rosenblatt, S. 2014. Nuclear regulator hacked 3 times in 3 years. CNET, `http://www.cnet.com/news/nuclear-commission-hacked-3-times-in-3-years/`.

Sasse, M. A.; Brostoff, S.; and Weirich, D. 2001. Transforming the 'weakest link' – A human/computer interaction approach to usable and effective security. *BT Technology Journal* 19(3):122–131.

Sheng, S.; Kumaraguru, P.; Acquisti, A.; Cranor, L.; and Hong, J. 2009. Improving phishing countermeasures: An analysis of expert interviews. *Proceedings of the 4th APWG eCrime Researchers Summit*.

Sheng, S.; Holbrook, M.; Kumaraguru, P.; Cranor, L. F.; and Downs, J. 2010. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 373–382. ACM.

Zetter, K. 2011. Top federal lab hacked in spear-phishing attack. WIRED, `http://www.wired.com/2011/04/oak-ridge-lab-hack/`.