Multi-Defender Strategic Filtering Against Spear-Phishing Attacks

Aron Laszka
Electrical Engineering and Computer Sciences Dept.
University of California, Berkeley
Berkeley, CA

Jian Lou and **Yevgeniy Vorobeychik**
Institute for Software Integrated Systems
Dept. of Electrical Engineering and Computer Science
Vanderbilt University
Nashville, TN

Abstract

Spear-phishing attacks pose a serious threat to sensitive computer systems, since they sidestep technical security mechanisms by exploiting the carelessness of authorized users. A common way to mitigate such attacks is to use e-mail filters which block e-mails with a maliciousness score above a chosen threshold. Optimal choice of such a threshold involves a tradeoff between the risk from delivered malicious emails and the cost of blocking benign traffic. A further complicating factor is the strategic nature of an attacker, who may selectively target users offering the best value in terms of likelihood of success and resulting access privileges. Previous work on strategic threshold-selection considered a single organization choosing thresholds for all users. In reality, many organizations are potential targets of such attacks, and their incentives need not be well aligned. We therefore consider the problem of strategic threshold-selection by a collection of independent self-interested users. We characterize both Stackelberg multi-defender equilibria, corresponding to short-term strategic dynamics, as well as Nash equilibria of the simultaneous game between all users and the attacker, modeling long-term dynamics, and exhibit a polynomial-time algorithm for computing short-term (Stackelberg) equilibria. We find that while Stackelberg multi-defender equilibrium need not exist, Nash equilibrium always exists, and remarkably, both equilibria are unique and socially optimal.

1 Introduction

A number of high-profile targets have fallen victim to spear-phishing attacks. In 2013, Target, the second largest general merchandise retailer in the US, suffered a massive data breach due to a spear-phishing attack (Smith 2014). As a consequence, Target had to pay Visa issuers $67 million as a reimbursement, and it is reportedly working on a similar deal with MasterCard (Sidel 2015). In 2014, the corporate network of a German steel mill was infiltrated by a spear-phishing attack (Zetter 2015). The attackers manipulated and disrupted control systems, resulting in massive physical damage. Further examples include one of the White House internal networks (McCullagh 2012), computers at the Nuclear Regulatory Commission (Rogers 2014), and Oak Ridge National Laboratory (Zetter 2011).

To mitigate spear-phishing attacks, an organization may set up an e-mail filter, which assigns a maliciousness score to each incoming e-mail and delivers only those that are below a given threshold (Hong 2012). Unfortunately, scoring is inevitably imperfect, and threshold choice must necessarily balance security (risk of delivering malicious e-mails) and usability (blocking of benign traffic).

Unlike non-targeted malicious e-mails, such as spam, spear-phishing e-mails must be customized to their targets, which means that an attacker must spend a substantial amount of effort on each target (Herley and Florêncio 2008). Consequently, attackers can only target a limited number of users in any spear-phishing campaign. This limitation implies that an attacker must select a subset of targets to maximize expected yield from an attack. Moreover, resource limitation on the attacker links the decisions of otherwise independent defenders: filtering decisions by some may result in others being targeted. If a single organization were responsible for setting filtering thresholds for all users, it could optimally account for such interdependencies, as shown in prior work (Laszka, Vorobeychik, and Koutsoukos 2015; Zhao, An, and Kiekintveld 2015). Realistically, however, numerous organizations are typically targeted, and their goals are generally distinct. The externalities that users impose upon one another therefore become strategically significant, and no work to date analyzes the resulting strategic dynamics in the spear-phishing context, even though prior work has considered other, quite different, interdependent security problems (Laszka, Felegyhazi, and Buttyan 2014; Kunreuther and Heal 2003; Chan, Ceyko, and Ortiz 2012; Lou and Vorobeychik 2015).

We address the problem of strategic e-mail threshold selection by a collection of independent users, faced with a threat of both spear-phishing and non-targeted (e.g., spam) malicious e-mail campaigns. We consider short-term strategic dynamics by appealing to a Stackelberg multi-defender equilibrium concept, as well as long-term dynamics using the Nash equilibrium concept. We offer a characterization of both kinds of equilibria, and present a polynomial-time algorithm for computing the Stackelberg multi-defender equilibrium. Remarkably, we demonstrate that while Stackelberg

multi-defender equilibria need not exist, Nash equilibrium always exists. Furthermore, we show that both equilibria are unique, and are socially optimal.

2 Model

Our model is based on the model introduced by Laszka, Vorobeychik, and Koutsoukos (2015), which we now extend for independent and self-interested defenders. For a list of symbols used in our model, see Table 1.

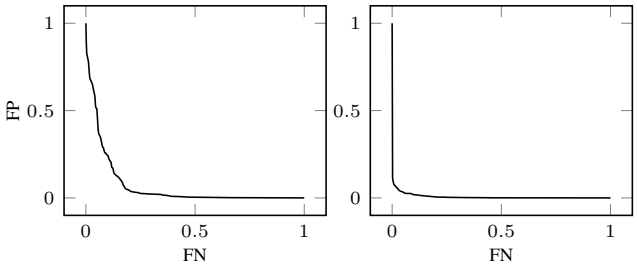
Table 1: List of Symbols

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
$f_u^{a_u}$	optimal false-negative probability of user u given that the user is targeted with probability a_u
$\mathcal{L}_u^{a_u}(f_u)$	expected loss of user u given the user is targeted with probability a_u

We model the strategic interactions of spear-phishing as a game between multiple *users* and a targeting *attacker*. Note that we refer to the defending players as users; however, these players can naturally model groups of users having the same e-mail filtering policy, or even entire organizations.

Users may receive three types of e-mails: *non-malicious*, *malicious non-targeted*, and *malicious targeted*. If a non-malicious e-mail is filtered out, which we call a *false positive* (FP), then the user suffers usability loss. If a malicious e-mail is not filtered out, which we call a *false negative* (FN), then the user might open that e-mail and suffer loss from the attack. We assume that the attainable false-positive and false-negative probability pairs are given by a function $FP : [0, 1] \mapsto [0, 1]$, where $FP(f)$ is the probability of false positives when the the probability of false negatives is f. In any practical e-mail classifier, $FP(f)$ is a non-increasing function of f (see Figure 1 for an illustration). For analytical tractability, we further assume that $FP(f)$ is a continuous, strictly decreasing, and strictly convex function of f. Note that these assumptions hold approximately in practice. For a discussion on how to handle problem instances that violate these assumptions, we refer the reader to the work by Laszka, Vorobeychik, and Koutsoukos (2015).

Malicious e-mails are divided into two categories: targeted and non-targeted. The former includes spear-phishing and whaling e-mails sent by the targeting attacker, while the latter includes spam and non-targeted phishing e-mails. Since the senders of non-targeted e-mails do not choose their targets in a strategic way in practice, we model them as *non-strategic* actors instead of game-theoretic players (see constant N_u below).



(a) UCI dataset (b) Enron dataset

Figure 1: False-negative to false-positive tradeoff curves for the two datasets used by Laszka, Vorobeychik, and Koutsoukos (2015) and Zhao, An, and Kiekintveld (2015).

Strategies A *pure strategy of user* u is a false-negative probability f_u, and we let $\boldsymbol{f}$ denote the strategy profile of the users. Note that we do not have to consider thresholds explicitly in our model, since there is a bijection between false-negative probabilities and thresholds values.

A *pure strategy of the attacker* is a set of users $\mathcal{A}$, who will be attacked. Since targeted e-mails have to be customized, which requires spending a considerable amount of effort on each target, the number of users that can be targeted is limited. Formally, the attacker's strategy is subject to a budget constraint $|\mathcal{A}| \leq A$. For the same reason, we also assume that the attacker is *lazy* in the sense that she does not target a user when she would receive zero payoff for targeting the user.

We will also consider mixed strategies, which are defined naturally: a *mixed strategy of the attacker* is a distribution over subsets of users, while a *mixed strategy of user* u is a distribution over false-negative values from $[0, 1]$.

Payoffs For a given pure-strategy profile $(\boldsymbol{f}, \mathcal{A})$, the attacker's payoff is

$$\mathcal{U} = \sum_{u \in \mathcal{A}} f_u L_u, \tag{1}$$

where $L_u > 0$ is the expected amount of damage when user u falls victim to a targeted attack.

If user u is targeted by the attacker (i.e., if $u \in \mathcal{A}$), then her loss (i.e., inverse payoff) is

$$\mathcal{L}_u^1(f_u) = f_u(L_u + N_u) + FP(f_u)C_u, \tag{2}$$

and if user u is not targeted (i.e., if $u \notin \mathcal{A}$), her loss is

$$\mathcal{L}_u^0(f_u) = f_u N_u + FP(f_u)C_u, \tag{3}$$

where $N_u > 0$ is the loss of user u for delivering non-targeted malicious e-mails, and $C_u > 0$ is the loss for not delivering non-malicious e-mails. Payoffs for mixed-strategies are defined naturally as expectations of these quantities. Note that in contrast with most multi-defender security games, the only externality in our game stems from a strategic attacker.

Solution Concepts In our analysis, we will study both short-term and long-term strategic dynamics of the e-mail filtering problem. As is typical in the literature, we study them using two different solution concepts, *Stackelberg multi-defender equilibrium* and *Nash equilibrium*.

In the short-term model, the game has two stages: in the first stage, the users make their strategic decision simultaneously; while in the second stage, the attacker makes its decision knowing which strategies the users have chosen. We solve this model using the concept of *Stackelberg multi-defender equilibrium* (SMDE), which is defined as follows.

Definition 1 (Stackelberg multi-defender equilibrium)**.** A strategy profile is an SMDE if each user's strategy is a best response to the others, assuming that the attacker will always play a best-response strategy.

In the long-term model, the players make their strategic decisions simultaneously. We solve this model using the concept of *Nash equilibrium* (NE), which is defined as follows.

Definition 2 (Nash equilibrium)**.** A strategy profile $(\boldsymbol{f}, \boldsymbol{a})$ is an NE if every player's strategy is a best response, taking the other players' strategies as given.

3 Analysis

First, in Section 3.1, we provide necessary conditions on the equilibria and introduce additional notation to facilitate our analysis. Then, we study and characterize the Stackelberg multi-defender and Nash equilibria of the game in Sections 3.2 and 3.3, respectively. Finally, we show that these equilibria are socially optimal in Section 3.4.

3.1 Preliminaries

We begin our analysis by providing a necessary condition on the users' mixed-strategy best responses, which applies both to SMDE and NE.

Lemma 1. *The best-response strategy for a user is always a pure strategy.*

As a consequence, for the remainder of this paper, we will consider only pure strategies (i.e., single false-negative values) for the users.

Proof sketch. Suppose that we are given a mixed strategy that is not a pure strategy (i.e., its support consist of more than one false-negative value); then, we show that the expected false-negative value is a better strategy than the distribution. Firstly, it is easy to see that the other players' payoffs (and, hence, their best responses) remain the same if the user changes its strategy from a distribution to the expected value. Secondly, since the function FP is strictly convex, we have from Jensen's inequality that the user's loss is strictly less for the expected value than for the distribution. Therefore, for every mixed strategy that is not a pure strategy, there exists a strictly better pure strategy. □

Next, we introduce a simpler notation for the attacker's mixed strategies. Let a_u be the probability that user u is targeted by the attacker, that is, the probability that u is an element of a subset chosen randomly according to the attacker's mixed strategy. Using this notation, we can express the attacker's expected payoff as

$$\mathcal{U} = \sum_u a_u f_u L_u \tag{4}$$

and user u's expected loss as

$$\mathcal{L}_u^{a_u}(f_u) = f_u(a_u L_u + N_u) + FP(f_u)C_u. \tag{5}$$

For every mixed strategy of the attacker, we can easily compute the corresponding vector of probabilities $\boldsymbol{a}$, which must satisfy $\sum_u a_u \leq A$. Furthermore, it is also easy to see that for every vector of probabilities $\boldsymbol{a}$ satisfying $\sum_u a_u \leq A$, there exists a mixed-strategy whose marginal is $\boldsymbol{a}$. For the remainder of this paper, we will represent the attacker's mixed-strategies as vectors of probabilities.

Now, we introduce some additional notation to facilitate our analysis. Let $f_u^{a_u}$ denote user u's optimal false-negative probability given that the attacker targets it with probability a_u, that is, let $f_u^{a_u}$ be the f_u which minimizes $\mathcal{L}_u^{a_u}(f_u)$. It is easy to see that $f_u^{a_u}$ is well defined for any a_u, and it is a non-increasing and continuous function of a_u.

Finally, consider the value f_u^0, which is the optimal false-negative probability given that the attacker never targets user u (i.e., given $a_u = 0$). If $f_u^0 = 0$ for user u, then it is easy to see that the user will always play the strategy $f_u = 0$, regardless of the other players' strategies or the solution concept used. Furthermore, such users do not affect the other players' strategic choices either, since an attacker will never target user u if $f_u = 0$. Consequently, for the remainder of the paper, we can disregard these users and assume that $f_u^0 > 0$ for every user u.

3.2 Stackelberg Multi-Defender Equilibrium

In this subsection, we characterize the Stackelberg multi-defender equilibrium (SMDE) and design an algorithm to find it. First, we show that in an SMDE, the attacker plays a pure strategy and the users play f_u^1 or f_u^0. Then, we show that the SMDE is unique if it exists, and provide an efficient algorithm for computing it. However, we also find that the SMDE does not necessarily exist, but our algorithm can return "there is no SMDE" if it does not exist.

The following lemma shows that the attacker always plays a pure strategy in an SMDE.

Lemma 2. *A strategy profile is an SMDE* only if *for each user u, either $a_u = 0$ or $a_u = 1$ holds.*

Proof. We prove the claim by contradiction. If $0 < a_u < 1$ for some user u, then her expected loss is

$$\mathcal{L}_u^{a_u}(f_u) = f_u(a_u L_u + N_u) + FP(f_u)C_u. \tag{6}$$

Since the attacker's strategy is a best response, $0 < a_u < 1$ implies that there exists some user $v \neq u$ with $0 < a_v < 1$ and $f_u L_u = f_v L_v$. Hence, if user u changes her strategy to $f_u - \epsilon$ (where ϵ is an arbitrarily small positive number), the attacker will target user v (or some other user) instead of user u. Then, the loss of user u will be

$$\mathcal{L}_u^0(f_u - \epsilon) = (f_u - \epsilon)N_u + FP(f_u - \epsilon)C_u, \tag{7}$$

since she is no longer targeted. We can compute the decrease in her loss due to deviating from her strategy as

$$\begin{aligned}&\mathcal{L}_u^{a_u}(f_u) - \mathcal{L}_u^0(f_u - \epsilon)\\ =&a_u f_u L_u + \epsilon N_u + [FP(f_u) - FP(f_u - \epsilon)]\, C_u.\end{aligned} \tag{8}$$

Clearly, $\mathcal{L}_u^{a_u}(f_u) - \mathcal{L}_u^0(f_u - \epsilon)$ can be greater than 0, as ϵ can be arbitrarily small and $FP(f_u) - FP(f_u - \epsilon)$ can be arbitrarily close to 0. Hence, user u can decrease her loss by deviating from the strategy f_u, which leads to a contradiction with our initial assumption that the strategy profile is an SMDE. Therefore, the claim of the lemma has to hold. □

From Lemmas 1 and 2, we know that in an SMDE, both the users and the attacker play pure strategies. Now, we further constrain the users' equilibrium strategies by showing that user u plays either f_u^1 or f_u^0 in an SMDE.

Lemma 3. *A strategy profile is an SMDE* only if $f_u = f_u^1$ *for every user u who is targeted, and $f_u = f_u^0$ for every user u who is not targeted.*

Proof sketch. We prove the claim by contradiction. Suppose that $\exists u$ such that $f_u \neq f_u^1$ and $f_u \neq f_u^0$. Based on Lemma 2, if the profile is an SMDE, then either $a_u = 0$ or $a_u = 1$, i.e., user u is targeted with probability 1 or 0.

1) First, assume that user u is targeted. Then, we show that strategy f_u^1 is better for user u than strategy f_u. First, if user u is still attacked after she switches to f_u^1, then we have by definition that f_u^1 is better since it minimizes $\mathcal{L}_u^1$. On the other hand, if the attacker no longer targets user u, then we have that the users' loss is even lower: $\mathcal{L}_u^0(f_u^1) \leq \mathcal{L}_u^1(f_u^1) < \mathcal{L}_u^1(f_u)$. Hence, f_u cannot be a best response.
2) If user u is not targeted, there are two cases: $f_u > f_u^0$ or $f_u < f_u^0$. If $f_u > f_u^0$, she can switch to f_u^0 to lower her loss without becoming a target of the attacker. If $f_u < f_u^0$, then we consider another user $v = \operatorname{argmin}_{u\in\mathcal{A}} f_u L_u$, i.e., the user in the targeted set that makes attacker get lower payoff. Using an argument similar to the one used in the proof of Lemma 2, we can show that $f_u L_u < f_v L_v$; otherwise, user v could lower her loss by decreasing f_v with an arbitrarily small value. On one hand, when $f_u^0 L_u < f_v L_v$, user u can switch to f_u^0 to lower her loss without becoming a target of the attacker. On the other hand, when $f_u^0 L_u \geq f_v L_v$, user u can switch to some value f_u' such that $f_u L_u < f_u' L_u < f_v L_v$ and still not be targeted. Then, based on characteristics of $\mathcal{L}_u^0(f_u)$, we have that user u can lower her loss by switching to strategy f_u'.

Consequently, user u has incentives to deviate from her strategy in both cases, which implies that there is no SMDE in which $f_u \neq f_u^1$ and $f_u \neq f_u^0$ for some user u. □

Based on the above results, we first provide a necessary and sufficient condition for a strategy profile being an SMDE, and then present an algorithm to find an SMDE.

Theorem 1. *A strategy profile $(\boldsymbol{f}, \mathcal{A})$ is an SMDE* if and only if

1) $\forall u \in \mathcal{A}$: $f_u = f_u^1$,
2) $\forall u \notin \mathcal{A}$: $f_u = f_u^0$,
3) $F_1 > F_0$,
4) $\forall u \in \mathcal{A}$: $\mathcal{L}_u^1(f_u^1) \leq \mathcal{L}_u^0(\frac{F_0}{L_u})$,

where $F_1 = \min_{u\in\mathcal{A}} f_u^1 L_u$ *and* $F_0 = \max_{u\notin\mathcal{A}} f_u^0 L_u$.

Proof sketch. First, we prove that the conditions of the theorem are necessary. From Lemma 3, we readily have that $f_u = f_u^1$, $\forall u \in \mathcal{A}$, and $f_v = f_v^0$, $\forall v \notin \mathcal{A}$ hold in an SMDE. Next, since the attacker's best response must target the users with the highest $f_u L_u$ values, we also have that $\min_{u\in\mathcal{A}} f_u^1 L_u \geq \max_{u\notin\mathcal{A}} f_u^0 L_u$ has to hold in an SMDE. Furthermore, this inequality has to be strict, otherwise a user in $\mathcal{A}$ could decrease her loss by decreasing her strategy by an arbitrarily small amount. Finally, in an SMDE, users in $\mathcal{A}$ do not have incentive to deviate from their strategy. If some user $u \in \mathcal{A}$ were to deviate, then she would pick a strategy that would divert attacks to another user, that is, she would consider a strategy $f_u \leq \frac{F_0}{L_u}$ (otherwise, following f_u^1 would obviously be better). Since $f_u^0 > f_u^1 > \frac{F_0}{L_u}$, her best choice would have to be $\frac{F_0}{L_u}$. Therefore, if user u has no incentive to deviate, then $\mathcal{L}_u^1(f_u^1) \leq \mathcal{L}_u^0(\frac{F_0}{L_u})$ has to hold.

Second, we prove that the conditions of the theorem are sufficient. For $\forall u \notin \mathcal{A}$, based on characteristics of the functions $\mathcal{L}_u^1(f_u)$ and $\mathcal{L}_u^0(f_u)$, we have that $\mathcal{L}_u^0(f_u^0)$ is the minimal loss user u can ever get, so she has no incentive to deviate. For $\forall u \in \mathcal{A}$, $\mathcal{L}_u^1(f_u^1)$ is the minimal loss of user u given that she is targeted. Hence, the only way that she could decrease her loss is to avoid being targeted by the attacker. In order to avoid being targeted, she has to pick a strategy $f_u \leq F_0/L_u$. From the convexity of $\mathcal{L}_u^0$ and $f_u^0 > f_u^1 > F_0/L_u$, we have that $\mathcal{L}_u^0(f_u)$ is a decreasing function when $f_u < F_0/L_u$. Hence, her best strategy that avoids being targeted is F_0/L_u; however, it follows from $\forall u \in \mathcal{A}$: $\mathcal{L}_u^1(f_u^1) \leq \mathcal{L}_u^0(F_0/L_u)$ that this is inferior to f_u^1. Therefore, the users' strategies are best responses under the conditions of the theorem. Finally, it follows readily from $F_1 > F_0$ that the attacker's strategy is also best response. □

In Theorem 1, we provided conditions for determining whether targeting a given set of users is an SMDE. In order to find an equilibrium, we could enumerate every subset $\mathcal{A}$ of users subject to $|\mathcal{A}| = A$, and check whether targeting $\mathcal{A}$ is an SMDE using Theorem 1. However, the running time of this approach grows exponentially as a function of A, and quickly becomes prohibitively large. We now provide a rather strong and surprising result which states that in an SMDE, the attacker will always target the set of A users with highest value of $f_u^1 L_u$.

Lemma 4. *Let $\mathcal{A}$ be a subset of users such that $|\mathcal{A}| = A$ and $\min_{u\in\mathcal{A}} f_u^1 L_u > \max_{u\notin\mathcal{A}} f_u^1 L_u$. In an SMDE, all of the users in $\mathcal{A}$ will be targeted.*

Proof. We prove the claim by contradiction. Suppose that there is an SMDE such that some user $v \notin \mathcal{A}$ is targeted. Then, user v plays f_v^1, and there exists some user $w \in \mathcal{A}$

who is not targeted and plays f_w^0. Since the attacker's strategy is a best response, we have that $f_v^1 L_v \geq f_w^0 L_w$. From $\min_{u \in \mathcal{A}} f_u^1 L_u > \max_{u \notin \mathcal{A}} f_u^1 L_u$, we obtain $f_v^1 L_v < f_w^1 L_w$. However, since $\forall u : f_u^1 \leq f_u^0$, we also have $f_v^1 L_v < f_w^1 L_w \leq f_w^0 L_w$, which contradicts $f_v^1 L_v \geq f_w^0 L_w$. Hence, the original claim must hold. □

Algorithm 1 Find a Stackelberg Multi-Defender Equilibrium (SMDE)

input: a set of users $\mathbf{U}$, L_u, $\mathcal{L}_u^1(f_u)$ and $\mathcal{L}_u^0(f_u)$ for every user u, and A for attacker
return: a SMDE or "there is no SMDE"

for each user u **do**
  compute f_u^1 and f_u^0 based on $\mathcal{L}_u^1(f_u)$ and $\mathcal{L}_u^0(f_u)$
end for
if $|\mathbf{U}| \leq A$ **then**
  $\mathcal{A} \leftarrow \mathbf{U}$
  $F_0 \leftarrow 0$
else
  $\mathcal{A} \leftarrow$ the set of A users with highest $f_u^1 L_u$ value
  $F_0 \leftarrow \max_{u \notin \mathcal{A}} f_u^0 L_u$
end if
$F_1 \leftarrow \min_{u \in \mathcal{A}} f_u^1 L_u$
if $F_1 > F_0$ and $\forall u \in \mathcal{A}$, $\mathcal{L}_u^1(f_u^1) \leq \mathcal{L}_u^0(\frac{F_0}{L_u})$ **then**
  return profile $(\boldsymbol{f}, \mathcal{A})$ in which $\forall u \in \mathcal{A}$: $f_u = f_u^1$, o.w. $f_u = f_u^0$
else
  return "there is no SMDE"
end if

Then, based on Theorem 1 and Lemma 4, we propose Algorithm 1 for finding an SMDE. We also find that an SMDE may not necessarily exist, and we provide an example for this case below. Furthermore, we find that the SMDE is unique if it exists. To see this, recall that in an SMDE, the attacker always plays a pure strategy targeting the set of users with highest values of $f_u^1 L_u$, and this set is obviously unique. Algorithm 1 always finds the unique SMDE if it exists, and returns "no SMDE" if there is no SMDE. Finally, it is also easy to see that the running time of the algorithm is polynomial in the number of users.

Numerical Example Consider a game consisting of two users (user 1 and user 2) and an attacker, who can target only a single user (i.e., $A = 1$). Let $L_1 = L_2 = 1$, $N_1 = N_2 = \frac{1}{2}$, $C_1 = 1$, and $C_2 = 2$. Finally, let $FP(f) = (1-f)^2$, which obviously satisfies our assumptions about FP. Then, $f_1^1 = \frac{1}{4}$, $f_1^0 = \frac{3}{4}$, $f_2^1 = \frac{5}{8}$, and $f_2^0 = \frac{7}{8}$.

Now, we show that the game does not have an SMDE. From Lemma 2, we know that the attacker will target either user 1 or user 2. First, suppose that the attacker targets user 1 (i.e., $\mathcal{A} = \{1\}$). Then, from Theorem 1, we have that the users' strategies must be $f_1 = f_1^1 = \frac{1}{4}$ and $f_2 = f_2^0 = \frac{7}{8}$. However, this contradicts that $\mathcal{A}$ is a best response, since $f_1 L_1 = \frac{1}{4} < \frac{7}{8} = f_2 L_2$. Second, suppose that the attacker targets user 2. Then, the user's strategies must be $f_1 = f_1^0 = \frac{3}{4}$ and $f_2 = f_2^1 = \frac{5}{8}$, which contradicts that $\mathcal{A}$ is a best response, since $f_1 L_1 = \frac{3}{4} > \frac{5}{8} = f_2 L_2$.

3.3 Nash Equilibrium

We begin our analysis of long-term dynamics by providing a necessary and sufficient conditions on the existence of a pure-strategy Nash equilibrium.

Lemma 5. *The game has a pure-strategy NE if and only if there exists a set of A users $\mathcal{A}$ such that $\min_{u \in \mathcal{A}} f_u^1 L_u \geq \max_{u \notin \mathcal{A}} f_u^0 L_u$. If a pure-strategy NE exists, then it is unique, and the attacker's equilibrium strategy is $\mathcal{A}$.*

Proof sketch. First, it easy to see that there can exist at most one set $\mathcal{A}$ satisfying the condition in the lemma, since $\min_{u \in \mathcal{A}} f_u^1 L_u \geq \max_{u \notin \mathcal{A}} f_u^0 L_u$ implies $\min_{u \in \mathcal{A}} f_u^1 L_u > \max_{u \notin \mathcal{A}} f_u^1 L_u$. Second, if such a set $\mathcal{A}$ exist, then the attacker targeting $\mathcal{A}$ and the users playing their best responses f_u^1 or f_u^0 is obviously an equilibrium. Finally, in any pure strategy equilibrium, the set of users targeted by the attacker must satisfy the condition in the lemma; otherwise, the attacker's strategy would not be a best response. □

Next, we extend our analysis by considering both pure and mixed strategies, and we show that a unique Nash equilibrium always exists in this case.

Theorem 2. *There always exists a unique Nash equilibrium.*

Proof. Before we begin, observe that a strategy profile $(\boldsymbol{f}, \boldsymbol{a})$ forms a Nash equilibrium if and only if there exists a value Λ such that for every user u,

- $a_u = 0 \Longrightarrow f_u = f_u^0$ and $f_u^0 L_u \leq \Lambda$;
- $0 < a_u < 1 \Longrightarrow f_u = f_u^{a_u}$ and $f_u L_u = \Lambda$;
- $a_u = 1 \Longrightarrow f_u = f_u^1$ and $f_u^1 L_u \geq \Lambda$.

We provide a constructive proof, that is, we show how to find a mixed-strategy Nash equilibrium. First, we define

$$f_u(\Lambda) = \begin{cases} f_u^0 & \text{if } f_u^0 L_u \leq \Lambda, \\ f_u^1 & \text{if } f_u^1 L_u \geq \Lambda, \\ \frac{\Lambda}{L_u} & \text{otherwise.} \end{cases} \tag{9}$$

It is easy to see that these functions are well-defined. Furthermore, they are continuous and non-decreasing in Λ. Also notice that when the value of function $f_u(\Lambda)$ is strictly greater than f_u^1 and strictly less than f_u^0, the function is strictly increasing.

Next, we define

$$a_u(\Lambda) = \begin{cases} 0 & \text{if } f_u(\Lambda) = f_u^0, \\ 1 & \text{if } f_u(\Lambda) = f_u^1, \\ a^* \text{ such that } f_u(\Lambda) = f_u^{a^*} & \text{otherwise.} \end{cases} \tag{10}$$

Notice that these functions are also well-defined, since $f_u^1 \leq f_u(\Lambda) \leq f_u^0$ always holds and $f_u(\Lambda)$ is continuous. From the characteristics of $f_u(\Lambda)$, we also have that each $a_u(\Lambda)$ is a continuous and non-increasing function of Λ. Furthermore, when the value of a function is strictly greater than 0 and strictly less than 1, the function is strictly decreasing.

Now, we let

$$Err(\Lambda) = A - \sum_u a_u(\Lambda)\ . \tag{11}$$

It is easy to see that $Err(\Lambda)$ is a non-decreasing and continuous function of Λ. Furthermore, for $\Lambda = 0$, we have $Err(\Lambda) = A -$ number of users < 0; and when Λ is high enough, we have $Err(\Lambda) = A - 0 > 0$. Consequently, we can always find a value Λ^* such that $Err(\Lambda^*) = 0$. On the one hand, if the game has a pure strategy equilibrium, then there is a gap between the Ath highest $f_u^1 L_u$ value and the $(A+1)$th highest $f_u^0 L_u$ value. Since the value of $Err(\Lambda)$ is 0 whenever Λ is in this gap, Λ^* is not unique; however, the strategies given by $f_u(\Lambda^*)$ and $a_u(\Lambda^*)$ are unique. On the other hand, if the game does not have a pure strategy equilibrium, then Λ^* is unique.

Finally, it is easy to see that the strategy profile given by $f_u(\Lambda^*)$ and $a_u(\Lambda^*)$ forms a Nash equilibrium, since they satisfy the conditions established at the beginning of the proof. Furthermore, for any Λ^* value, the only strategy profile that satisfies the conditions is the one given by $f_u(\Lambda^*)$ and $a_u(\Lambda^*)$. Since there is a unique Λ^* value (or in the case of pure-strategy equilibrium, a range of Λ^* values) that results in a feasible attacker strategy, the equilibrium strategy is unique. □

Numerical Example In the previous subsection, we introduced a numerical example, for which no SMDE exists. Using the constructive proof of Theorem 2, we can find the unique mixed-strategy NE for this example: $\left(f_1 = \frac{2}{3}, f_2 = \frac{2}{3}, a_1 = \frac{1}{6}, a_2 = \frac{5}{6}\right)$.

It is easy to verify that these strategies form an NE. First, since $f_1 L_1 = \frac{2}{3} = f_2 L_2$, the attacker's strategy is obviously a best response. Second, user 1's best-response strategy minimizes the loss $\mathcal{L}_1^{a_1}(f_1) = f_1(a_1 L_1 + N_1) + FP(f_1)C_1 = f_1 \frac{4}{6} + (1 - f_1)^2 = (f_1)^2 - \frac{4}{3} f_1 + 1 = \left(f_1 - \frac{2}{3}\right)^2 + \frac{5}{9}$, whose minimum is clearly attained at $f_1 = \frac{2}{3}$. Finally, we can verify that user 2's strategy is a best response in the same way.

3.4 Social Optimum

In the preceding subsections, we have characterized SMDE and NE, which are formed by the users' selfish decisions. A crucial question regarding these equilibria is how close they are to the social optimum, i.e., to the strategies chosen by a social planner who is interested in minimizing the players losses. We define the social optimum formally as follows.

Definition 3 (Social Optimum)**.** The users' strategies constitute a social optimum if they minimize the sum of the users' losses, given that the attacker always plays a best response.

The following theorem establishes a very surprising result: even though users act in their self-interest, the Nash equilibrium that they reach is a social optimum.

Theorem 3. *The Nash equilibrium is a social optimum.*

Before proving the theorem, we first have to establish the following lemma.

Lemma 6. *If $\boldsymbol{f}$ is a social optimum, then $f_u^1 \le f_u \le f_u^0$ for every u.*

Proof sketch. For the sake of contradiction, suppose that the claim does not hold, that is, suppose that there exists a social optimum strategy $\boldsymbol{f}$ such that $f_u^1 > f_u$ or $f_u > f_u^0$ for some user u.

First, suppose that $f_u^1 > f_u$. If there are less than A users v who have greater $f_v L_v$ values than u does, then the attacker's best response may include u. If there are multiple users with the same value $f_u L_u$, then the attacker can choose any of them, without changing the social cost. Hence, we can assume that user u is targeted by the attacker. In this case, increasing f_u to f_u^1 will obviously be a socially better strategy, since it does not change the attacker's best response and we have by definition that f_u^1 minimizes the loss for user u.

If there are at least A users v who have greater $f_v L_v$ values than u, then the attacker's best response cannot include u. In this case, increasing f_u by a small amount such that $f_u L_u$ does not exceed the Ath greatest value will be a socially better strategy, since it does not change the attacker's best response, but decreases the loss for user u (by increasing f_u towards f_u^0). Hence, $f_u^1 > f_u$ must lead to a contradiction.

Second, suppose that $f_u L_u > f_u^0 L_u$. Then, we can use a similar argument as above to show that there exists a socially better strategy, regardless of whether user u is targeted or not. Therefore, the claim of the lemma has to hold. □

Now, we can prove Theorem 3.

Proof sketch of Theorem 3. We prove the claim of the theorem by showing that for any social optimum strategy $\boldsymbol{f}$, there exists an attacker strategy $\boldsymbol{a}$ such that the $(\boldsymbol{f}, \boldsymbol{a})$ is a Nash equilibrium. Since the equilibrium is unique, this will prove that the social optimum is also unique and it is in fact equivalent to the equilibrium.

First, suppose that the A highest $f_u L_u$ values are strictly greater than the remaining $f_u L_u$ values. Then, it is easy to see that for the A users with the highest values, $f_u = f_u^1$ must hold; otherwise, $\boldsymbol{f}$ cannot be a social optimum. Next, it is also clear that for the remaining users, $f_u = f_u^0$ must hold; otherwise, $\boldsymbol{f}$ cannot be a social optimum. Finally, it is easy to verify that the attacker strategy which targets the first group of users forms an equilibrium with $\boldsymbol{f}$.

Now, suppose that the Ath highest $f_u L_u$ value is equal to the (number of users $- A$)th lowest $f_u L_u$ value, and let this value be denoted by Λ. Then, for users u such that $f_u L_u < \Lambda$, $f_u = f_u^0$ must hold; otherwise, $\boldsymbol{f}$ cannot be a social optimum since changing f_u towards f_u^0 would decrease social cost. For these users, let $a_u = 0$, which ensures that f_u is a best response. Next, for users u such that $f_u L_u > \Lambda$, $f_u = f_u^1$ must hold; otherwise, $\boldsymbol{f}$ cannot be a social optimum since changing f_u towards f_u^1 would decrease social cost. For these users, let $a_u = 1$, which ensures that f_u is a best response. Finally, for users u such that $f_u L_u = \Lambda$, let a_u such that $f_u = f_u^{a_u}$, which again ensures that f_u is a

best response. Note that in order for these a_u values to exist, f_u must be between f_u^1 and f_u^0, which we have shown in Lemma 6.

From the construction of $\boldsymbol{a}$, we readily have that the users' strategies are best responses. Since the attacker is indifferent between users with $f_u L_u = \Lambda$, we also have that $\boldsymbol{a}$ is a best response given that it is a feasible strategy. Hence, it remains to show that $\boldsymbol{a}$ is feasible, i.e., the sum of the probabilities is A. Suppose for the sake of contradiction that this is not true, i.e., the sum of the probabilities assigned to users with $f_u L_u = \Lambda$ is not equal to A minus the number of users with $f_u = f_u^1$. This means that the users f_u values are either higher or lower than what would be the best response to the sum of the actual probabilities. Hence, by either decreasing or increasing Λ, the sum of the users losses could be decreased to the optimal value for the sum of the actual probabilities. Since this would contradict the assumption that $\boldsymbol{f}$ is socially optimal, $\boldsymbol{a}$ must be a feasible strategy, which concludes our proof. □

Finally, we show that if an SMDE exists, then it is also an NE, which proves that the SMDE is also socially optimal.

Theorem 4. *The SMDE is a social optimum.*

Proof. We prove that an SMDE is an NE by showing that if a strategy profile satisfies Theorem 1, then the players' strategies are best responses. First, it follows from the first condition that users in $\mathcal{A}$ play their best responses. Second, it follows from the second condition that remaining users play their best responses. Finally, it follows from the third condition that the attacker's strategy is a best response. Consequently, an SMDE is also an NE, which proves that it is a social optimum. □

4 Related Work

Strategic selection of thresholds for filtering spear-phishing e-mails has been considered in the research literature only very recently. Laszka, Vorobeychik, and Koutsoukos (2015) model the decision problem faced by a single defender who has to protect multiple users against targeted and non-targeted malicious e-mail. Their work focuses on characterizing and computing optimal defense strategies, and they use numerical results to demonstrate that strategic threshold selection can substantially decrease losses compared to naïve thresholds. Zhao, An, and Kiekintveld (2015) study a variant of the previous model: they assume that the targeting attacker can launch an unlimited number of costly spear-phishing attacks in order to learn a secret, which only a subset of the users know. Their work also focuses on the computational aspects of finding an optimal defense strategy; however, their variant of the model does not consider non-targeted malicious e-mails.

Similar problems in adversarial classification have been studied in the research literature earlier. For example, Dritsoula, Loiseau, and Musacchio (2012) consider the problem of choosing a threshold for classifying an attacker into two categories, spammer and spy, based on its intrusion attempts. More recently, Lisỳ, Kessl, and Pevnỳ (2014) study adversarial classification in a general model, which can be applied to e-mail filtering, intrusion detection, steganalysis, etc., and analyze Nash and Stackelberg equilibria based on the ROC curve of the classifier. However, both of these research efforts consider only a single defender.

Strategic interactions between multiple, interdependent defenders have been extensively studied (Laszka, Felegyhazi, and Buttyan 2014). The two models that are most similar to ours are by Bachrach, Draief, and Goyal (2013) and Lou and Vorobeychik (2015). In the former, the attacker may target only one defender, and the NE may be arbitrarily worse than the social optimum. In the latter, the defenders' strategies are discrete and the utility function is linear, whereas our model involves non-linear utilities. Our model has similarities to congestion games as well, which also consider interdependent players. Briefly, the central difference from these games is the presence of a strategic attacker, which significantly changes strategic dynamics.

Finally, the problem of designing e-mails classifiers for detecting spam and phishing has also been extensively studied (Blanzieri and Bryl 2008). Note that these results are complementary to the strategic threshold-selection problem, since the latter builds on an exogenously given classifier. Potentially malicious e-mails can be classified based on many attributes. For example, Fette, Sadeh, and Tomasic (2007) build a classifier for detecting phishing e-mails using a variety of features, such as the number of links in the e-mail and the age of the linked-to domain names. When evaluated on a real-world dataset, the false negative rate of the classifier was less than 4%, while its false positive rate was around 0.1%. As another example, Bergholz et al. (2010) design an e-mail classifier for detecting spam and phishing e-mails, and they describe a number of novel features, such as design elements of known brands and intentional distortion of content not perceivable by the reader.

5 Conclusion

In order to mitigate the serious threat posed by spear-phishing attacks, defenders can deploy e-mail filters. However, the strategic nature of these attacks and the independent configuration of the e-mail filters may lead to a coordination problem. In this paper, we studied this coordination problem by extending previous work on strategic threshold-selection. We considered both short-term and long-term dynamics, and found that the defenders' selfish choices need not to lead to an equilibrium on the short term, but they definitely lead to one on the long-term. Finally, we found that – quite remarkably – these equilibria are socially optimal.

These results are in stark contrast with the majority of multi-defender and interdependent security games. Usually, these games exhibit socially suboptimal equilibria, in which defenders "free-ride" on others, whereas equilibria are socially optimal in our model. The explanation for these differences lies in the problem-specific assumptions on which the exact formulations of the players' utility functions are based. These differences are very important from a practical perspective (e.g., devising regulations), and provide a motivation for our alternative model.

There are multiple natural future research directions. In this paper, we have shown that the defenders may reach an

NE on the long term; however, it remains an open question how quickly they converge to the equilibrium. This questions is especially interesting in a setting where the parameters of the game and, hence, the NE evolve over time. As another direction, one could consider alternative core assumptions, such as incomplete or imperfect information, which might lead to qualitatively different results.

Acknowledgment

This work was supported in part by FORCES (Foundations Of Resilient CybEr-Physical Systems), which receives support from the National Science Foundation (NSF award numbers CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166), and by National Science Foundation under Award IIS-1526860, the Air Force Research Laboratory under Award FA8750-14-2-0180, the Office of Naval Research under Award N00014-15-1-2621, and Sandia National Laboratories.

References

Bachrach, Y.; Draief, M.; and Goyal, S. 2013. Contagion and observability in security domains. In *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1364–1371. IEEE.

Bergholz, A.; De Beer, J.; Glahn, S.; Moens, M.-F.; Paaß, G.; and Strobel, S. 2010. New filtering approaches for phishing email. *Journal of Computer Security* 18(1):7–35.

Blanzieri, E., and Bryl, A. 2008. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review* 29(1):63–92.

Chan, H.; Ceyko, M.; and Ortiz, L. E. 2012. Interdependent defense games: Modeling interdependent security under deliberate attacks. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 152–162.

Dritsoula, L.; Loiseau, P.; and Musacchio, J. 2012. Computing the Nash equilibria of intruder classification games. In *Proceedings of the 3rd International Conference on Decision and Game Theory for Security (GameSec)*. Springer. 78–97.

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)*, 649–656. ACM.

Herley, C., and Florêncio, D. 2008. A profitless endeavor: Phishing as tragedy of the commons. In *Proceedings of the 2008 New Security Paradigms Workshop (NSPW)*, 59–70. ACM.

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

Kunreuther, H., and Heal, G. 2003. Interdependent security. *Journal of Risk and Uncertainty* 26(2-3):231–249.

Laszka, A.; Felegyhazi, M.; and Buttyan, L. 2014. A survey of interdependent information security games. *ACM Computing Surveys* 47(2):23:1–23:38.

Laszka, A.; Vorobeychik, Y.; and Koutsoukos, X. 2015. Optimal personalized filtering against spear-phishing attacks. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, 958–964.

Lisỳ, V.; Kessl, R.; and Pevnỳ, T. 2014. Randomized operating point selection in adversarial classification. In *Proceedings of the 2014 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Part II*. Springer. 240–255.

Lou, J., and Vorobeychik, Y. 2015. Equilibrium analysis of multi-defender security games. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 596–602.

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/`.

Sidel, R. 2015. Target to settle claims over data breach. The Wall Street Journal, `http://www.wsj.com/articles/target-reaches-settlement-with-visa-over-2013-data-breach-1439912013`.

Smith, G. 2014. Massive Target hack traced back to phishing email. Huffpost Business, `http://www.huffingtonpost.com/2014/02/12/target-hack_n_4775640.html`.

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

Zetter, K. 2015. A cyberattack has caused confirmed physical damage for the second time ever. Wired, `http://www.wired.com/2015/01/german-steel-mill-hack-destruction/`.

Zhao, M.; An, B.; and Kiekintveld, C. 2015. An initial study on personalized filtering thresholds in defending sequential spear phishing attacks. In *Proceedings of the 2015 IJCAI Workshop on Behavioral, Economic and Computational Intelligence for Security*.