Warning Time: Optimizing Strategic Signaling for Security Against Boundedly Rational Adversaries

Extended Abstract

Sarah Cooney  
University of Southern California  
Los Angeles, California

Phebe Vayanos  
University of Southern California  
Los Angeles, California

Thanh H. Nguyen  
University of Oregon  
Eugene, Oregon

Cleotilde Gonzalez  
Carnegie Mellon University  
Pittsburgh, Pennsylvania

Christian Lebiere  
Carnegie Mellon University  
Pittsburgh, Pennsylvania

Edward A. Cranford  
Carnegie Mellon University  
Pittsburgh, Pennsylvania

Milind Tambe  
University of Southern California  
Los Angeles, California

ABSTRACT

Defender-attacker Stackelberg security games (SSGs) have been applied for solving many real-world security problems. Recent work in SSGs has incorporated a deceptive signaling scheme into the SSG model, where the defender strategically reveals information about her defensive strategy to the attacker, in order to influence the attacker’s decision making for the defender’s own benefit. In this work, we study the problem of signaling in security games against a boundedly rational attacker.

KEYWORDS

Stackelberg security games; signaling schemes; bounded rationality; behavioral modeling; human subject experiments.

ACM Reference Format:

1 INTRODUCTION

The National Security Strategy released by the White House in 2017 includes defending against cyber attacks as one of its top priorities [10]. A major challenge in cybersecurity is deterring attackers before they can compromise a system. Much of today’s cyber defense is reactive, rather than proactive, and often attacks are not discovered before major damage has been done [11, 21].

Game theory, specifically defender-attacker Stackelberg security games (SSGs), have drawn considerable interest from security agencies for their use in modeling strategic interactions between attackers and defenders, and optimizing defender strategies for real-world applications in physical security domains [19]. Example applications include protecting airport terminals and ports, scheduling air marshals, and planning patrol routes to mitigate poaching activity [1, 6, 12, 17]. Motivated by this success, researchers have taken up the challenge of developing effective game theory-based defense solutions in the cybersecurity domain, including optimally allocating cyber-analyst talent and strategically deploying honeypots on a network [7, 14]. Another important line of game-theoretic research is the strategic exploitation of information by the defender to influence and deceive the adversary. It is formalized in the signaling game model, in which one player acts as a sender and strategically reveals information to another player, known as the receiver [2, 4, 9]. Recent work by Xu et al. incorporates the signaling game model into the SSG model, where the defender strategically reveals information about her defensive strategy to the attacker, in order to influence the attacker’s decision making. They show that using this model improves defender utility against a perfectly rational attacker compared to the traditional SSG model [22]. The work has since been extended to show how to coordinate machine patrollers with signaling capabilities with human patrollers for wildlife protection [23].

We propose that the SSG framework with signaling can be used as a mechanism for proactive defense against cyber attackers. However, in [22] and [23], the benefit to the defender from using this type of signaling scheme relies heavily on the presumption that the attacker will behave according to the assumptions of perfect rationality. Motivated by longstanding research showing that human attackers frequently deviate from the assumptions of perfect rationality [3, 13, 18], we address the use of Xu et al.'s framework in the face of boundedly rational attackers. In the rest of this paper, we briefly introduce a model of two-way deceptive signaling to increase compliance with signals for boundedly rational attackers.

2 COMPUTING A SIGNALING SCHEME

An overview of the classic SSG can be found in [20], and an overview of the framework for a two stage SSG with signaling can be found in [22]. The main difference between two-stage model with signaling and the classic model is that after selecting a target, with some probability the attacker is shown a (possibly deceptive) signal,
stating the target is being protected. (For instance, a sign at the airport indicating extra security checks are occurring.) He then has the choice to continue attacking to or withdraw, which yields a utility of zero to both parties. The goal is to cause the adversary to withdraw his attack upon seeing a signal, even if he knows the target is not always covered when a signal says it is. We will let $T = \{t_1, t_2, \ldots, t_n\}$ be the set of targets the defender is aiming to protect, and denote by $z = (z_t)$ a mixed strategy of the defender in which $z_t \in [0, 1]$ is the defender’s coverage probability at target $t$ [15]. In particular, we have $\sum_t z_t = K$ where $K$ is the number of defender resources, and $K < T$. A Signaling Scheme with respect to $t$ consists of probabilities $(p_t, q_t)$ with $0 \leq p_t \leq z_t$ and $0 \leq q_t \leq 1 - z_t$, such that $p_t$ and $q_t$ are the probabilities of showing a signal given that $t$ is currently covered and uncovered, respectively. Figure 1 illustrates the signaling scheme for a target $t$. A signaling scheme tells the defender how often to warn the attacker, when (1) the warning is true ($p_t$), and (2) it is false ($q_t$). Intuitively, it is the optimal combination of bluffing and truth telling to ensure the attacker always believes the bluff.

Figure 1: The signaling scheme for a target $t$.

It is of note, under the signaling scheme given by [22], hereafter referred to as the pSSE signaling scheme, if overall attacker expected utility ($z_t U^S(t) + (1 - z_t)U^M(t)$) is greater than zero for all $t$, then $p_t = z_t$ [22]. We call this type of model in which $p_t = z_t$ a 1-way deception, because we only deceive the adversary when a warning is shown. When a warning is not shown, it is always true that the target is uncovered. Thus, when no signal is shown the adversary can attack with impunity, resulting in a certain loss for the defender.

Human subject experiments, using an online game based on the scenario of an inside attacker as described in [5], show that boundedly rational attackers frequently ignore signals and continue attacking under the pSSE signaling scheme. Therefore, we introduce a 2-Way Deception signaling scheme that adds uncertainty when no signal is shown and lowers the overall frequency of signaling by proportionally decreasing $p_t$ and $q_t$. Decreasing $p_t$ yields $p_t < z_t$, which adds deception when there is no signal, hence the name 2-way deception. We briefly describe two 2-way signaling scheme algorithms.

**Generalized 2-Way Signaling.** This type of scheme uniformly reduces the frequency of signaling across all targets, and serves as a baseline for 2-way signaling. To compute the signaling scheme, we set $p_t = f z_t$, for some $f \in [0, 1]$ and then solve for $q_t$.

**Behavioral Modeling-Based Signaling** There is a long history of research on improving defender performance against boundedly rational attackers in security games via behavioral models of the adversary to predict the likelihood he will choose each target. Some such models draw on insights from such as quantal response and subjective utility [16, 24], while more recent techniques involve the use of historical data and machine learning techniques [8]. Thus, we also turn to modeling the behavior of the attacker to improve defender performance against boundedly rational adversaries. However, rather than model the choice of target, we model the attacker’s behavior with regard to signaling. For example, using data from the pSSE and a generalized 2-way signaling experiment, we learned a logistic regression model for each of the four sets of targets in our experiment. We used four features of the target—attacker reward ($U^S$), attacker penalty ($U^M$), and coverage probability $z_t$, all held constant, and the frequency of signaling, which is defined as $p_t + q_t$—to predict the probability a subject will attack the given target if shown a signal. We used the iterative method described in [24] to find the signaling scheme that maximizes the defender’s expected utility, which is a non-linear, non-convex expression. Other potential models to explore in future work include support vector machines, classification trees, and neural networks.

**ACKNOWLEDGMENTS**

This research was sponsored by the Army Research Office and accomplished under Grant Number W911NF-17-1-0370.

**REFERENCES**


