Robust Bayesian Methods for Stackelberg Security Games (Extended Abstract)

Christopher Kiekintveld and Milind Tambe
University of Southern California
{kiekintv, tambe}@usc.edu

Janusz Marecki
IBM T.J. Watson Research
janusz.marecki@gmail.com

ABSTRACT

Recent work has applied game-theoretic models to real-world security problems at the Los Angeles International Airport (LAX) and Federal Air Marshals Service (FAMS). The analysis of these domains is based on input from domain experts intended to capture the best available intelligence information about potential terrorist activities and possible security countermeasures. Nevertheless, these models are subject to significant uncertainty—especially in security domains where intelligence about adversary capabilities and preferences is very difficult to gather. This uncertainty presents significant challenges for applying game-theoretic analysis in these domains. Our experimental results show that standard solution methods based on perfect information assumptions are very sensitive to payoff uncertainty, resulting in low payoffs for the defender. We describe a model of Bayesian Stackelberg games that allows for general distributional uncertainty over the attacker's payoffs. We conduct an experimental analysis of two algorithms for approximating equilibria of these games, and show that the resulting solutions give much better results than the standard approach when there is payoff uncertainty.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed AI—*Intelligent agents, Multiagent systems*

General Terms

Algorithms, Experimentation

Keywords

Game theory, security, robustness, Bayesian, Stackelberg, optimization, replicator dynamics

1. INTRODUCTION

Game-theoretic modeling is an increasingly important tool in real-world security applications. Two deployed software systems for the Los Angeles International Airport (LAX) since August 2007 [3] and the Federal Air Marshals Service (FAMS) [5] apply game theory to generate optimal randomized schedules for security resources. A critical element of these applications is the game model, which specifies the possible actions for the security forces and attackers, as well as the payoffs for each possible outcome. These models are typically based on the analysis of domain experts, which it turn is based on available intelligence information. Unfortunately, in most cases this information is highly uncertain, particularly with regards to the preferences and capabilities of potential adversaries. In this work we investigate security games where the defender has substantial uncertainty about the attacker's payoffs.

Previous work on Bayesian Stackelberg games has considered uncertainty over a finite number of distinct attacker types. Unfortunately, existing algorithms (such as DOBSS [2]) do not scale well as the number of attacker types increases, so this approach is generally limited to a small number of attacker types. We introduce a more general model with a continuous space of attacker types. Conceptually, this model replaces the point estimates of an attacker's payoffs with a continuous distribution of possible payoffs, such as a Gaussian or Uniform distribution. This provides a natural way for domain experts to express their uncertainty about attacker payoffs, which are key parameters of the game model. Finding Bayesian Stackelberg equilibria of these games is challenging, so we introduce two approximation techniques. We compare these methods empirically with a perfect-information approach that uses point estimates of the payoffs, and show that the approximation algorithms dramatically outperform the baseline method.

Cite as: Robust Bayesian Methods for Stackelberg Security Games (Extended Abstract), Christopher Kiekintveld, Janusz Marecki and Milind Tambe, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1467-1468

2. BAYESIAN SECURITY GAMES

We define a new class of Bayesian security games, extending the security game model defined by Kiekintveld et. al. [1] to include continuous uncertainty about the attacker's payoffs. A security game has two players, a *defender*, Θ, and an *attacker*, Ψ, a set of *targets* $T = \{t_1, \ldots, t_n\}$ that the defender wants to protect and a set of identical *resources* $R = \{r_1, \ldots, r_m\}$ (e.g., police officers) that the defender may deploy to protect the targets. The defender's pure strategy, denoted by σ_Θ, is a subset of targets from T with size less than or equal to m. The attacker's pure strategy, denoted by σ_Ψ, is exactly one target from T.

We consider a Bayesian Stackelberg game where the defender has uncertainty about the attacker's payoffs. The defender has two possible payoffs if target t is attacked. If t is covered by a resource, the payoff is $U^c_\Theta(t)$, and if it is not covered the payoff is $U^u_\Theta(t)$. An attacker's payoffs are dependent on the attacker's type $\omega \in \Omega$. For any target t each attacker type ω receives a payoff of $U^u_\Psi(t, \omega)$ if t is uncovered, and $U^c_\Psi(t, \omega)$ if t is covered. The attacker's type (and therefore payoffs) is determined by nature at the start of the game. The defender then commits to a mixed strategy for covering the targets, and the attacker observes both this strategy and it's type before selecting a strategy. We apply the standard solution concept of Bayesian Stackelberg equilibrium, in which both players play a best-response. Every attacker type optimizes against the

(known) defender strategy, and the defender chooses an optimal strategy given the distribution of attacker types.

3. SOLUTION METHODS

We briefly describe two approaches for computing approximate solutions to continuous Bayesian Stackelberg security games. Both methods apply Monte-Carlo sampling from the space of attacker types to estimate the probability that a target will be attacked for a given defender strategy.

Sampled Bayesian ERASER: The idea of the first method is to generate a finite Bayesian Stackelberg game to approximate the infinite game with continuous attacker payoff distributions. Each sample attacker type has two payoffs for each target, depending on whether the target is covered or not. We construct a finite Bayesian Stackelberg game using these sample types, each occurring with equal probability. The resulting game can be solved using DOBSS [2], an optimal mixed-integer program for finite Bayesian Stackelberg games. We improve the speed of this method by incorporating insights from ERASER [1] for the case of multiple defender resources. We call this method *Sampled Bayesian ERASER* (SBE), where SBE-x denotes the number of sample types x.

Sampled Replicator Dynamics: The SBE method computes an exact optimal solution for the defender based on a sampling of the possible attacker types. Here we introduce a method that approximates the defender strategy, in addition to sampling the attacker types. This approach is based on replicator dynamics [4]. The algorithm begins by drawing a set of sample attacker types. For any fixed defender strategy, we can compute the best-response for each sample type. The number of types that choose to attack each target gives an approximation of the probability that each target will be attacked, which in turn allows the expected payoff for the defender to be computed for this coverage strategy. Replicator dynamics is used to search over the space of possible defender coverage strategies. We call this method *Sampled Bayesian ERASER* (SBE) and use SBE-x to denote this methods with x sample attacker types.

4. EVALUATION

We omit the majority of our evaluation due to space constraints, but present one result demonstrating the importance of modeling uncertainty rather than using a perfect-information approximation. We generate 500 random game instances with 5 targets and 1 defender resource. The defender's payoffs for a covered target are drawn from $U[0, 100]$, and the uncovered payoffs from $U[-100, 0]$. The attacker's payoffs are represented by Gaussian distributions, with mean values drawn from $U[-100, 0]$ for covered targets and $U[0, 100]$ for uncovered targets; we vary the standard deviation. A sample attacker type is defined by drawing one value from each of these distributions (two values for each target).

The baseline algorithm uses a single point to estimate each payoff, rather than a distribution. This is motivated by the standard methodology for eliciting game models from domain experts, where no information about the uncertainty of the parameters is included in the model. We model this with a perfect-information model where the attacker has only one type, corresponding to the mean value for each payoff distribution. This can be solved exactly using the SBE algorithm with a single attacker type, which we refer to as "SBE-Mean."

Figure 1 presents results for the solution quality for SBE-Mean, SBE, and SRD. We vary payoff uncertainty along the x-axis, measured by the standard deviation of the Gaussian distributions for the attacker payoffs (in the same units as the payoffs). We run each algorithm to generate a coverage strategy for the defender, and evaluate this coverage strategy against the true distribution of attacker types. Since we do not have a closed-from solution to compute this exactly, we rely on a very close approximation generated by sampling 10000 attacker types to evaluate the payoffs for each algorithm. The expected payoffs are shown on the vertical axis. We run SBE with up to 7 sample types and SRD with up to 1000 due to large differences in the computational scalability of the algorithms. With only 7 types, SBE takes roughly 2 seconds to run, while SRD runs in less than half a second with 1000 types and 5000 search iterations.

In Figure 1 we see that the solution quality for both SBE and SRD is dramatically higher than the SBE-Mean baseline when there is payoff uncertainty, even if the uncertainty is relatively small. SBE and SRD show improvements over the baseline even with very small numbers of sample attacker types, with diminishing returns as the number of types increases. This is a strong indication that the perfect-information approach is not a good approximation for security games with uncertainty about the attacker's payoffs. SBE and SRD represent the first steps towards more robust methods that give high-quality solutions even when there is payoff uncertainty.

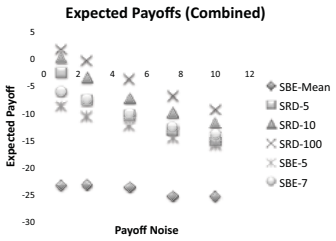

Figure 1: Expected payoffs for SBE-Mean, SBE, and SRD with varying numbers of sample attacker types.

Acknowledgement

This research was supported by the United States Department of Homeland Security through the Center for Risk and Economic Analysis of Terrorism Events (CREATE) under grant number 2007- ST-061-000001. However, any opinions, conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the Department of Homeland Security.

5. REFERENCES

[1] C. Kiekintveld, M. Jain, J. Tsai, J. Pita, F. Ordóñez, and M. Tambe. Computing optimal randomized resource allocations for massive security games. In *AAMAS*, 2009.

[2] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, and S. Kraus. Playing games with security: An efficient exact algorithm for Bayesian Stackelberg games. In *AAMAS*, pages 895–902, 2008.

[3] J. Pita, M. Jain, C. Western, C. Portway, M. Tambe, F. Ordonez, S. Kraus, and P. Paruchuri. Depoloyed ARMOR protection: The application of a game-theoretic model for security at the Los Angeles International Airport. In *AAMAS (Industry Track)*, 2008.

[4] P. Taylor and L. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 16:76–83, 1978.

[5] J. Tsai, S. Rathi, C. Kiekintveld, F. Ordóñez, and M. Tambe. IRIS - A tools for strategic security allocation in transportation networks. In *AAMAS (Industry Track)*, 2009.