Partially Observable Stochastic Games for Cyber Deception against Network Epidemic

Olivier TSEMOGNE[1,2], Yezekael HAYEL[2], Charles KAMHOUA[3], and Gabriel DEUGOUE[1]

[1] University of Dschang, Dschang, Cameroon
[2] CERI/LIA, Avignon Université, France
[3] US Army Research Laboratory, USA

Abstract. A Decentralized Denial of Service is an attack done by an agent capable to control the spread of a malware. This is a combination of epidemiological and conflictual aspects between several decision makers. There exists in the literature papers that study (non oriented) epidemics and papers that study network attacks regardless the epidemiological aspect. We put together the two aspects and provide a new game theoretical model which is part of the family of partially observable stochastic games (POSG) but with particular features. We prove the consistency of heuristic search value iteration (HSVI) based algorithms. Our framework is applied to optimally design a cyber deception technique based on honeypots in order to control an epidemic cyber-attack of a network by a strategic attacker. Some basic simulations are proposed to illustrate the framework described in this work-in-progress paper.

Keywords: Epidemic models, Partially observable stochastic game, Heuristic search value iteration

MSC: 91A80

1 Introduction

Cyber security is becoming an important research area in global security world. First, with the high level of usage of connected devices and equipments, understanding cyber attacks is the most important stage in order to build efficient cyber defense. Second, we are living in a word that is more and more connected, and the effect of networks is no more to prove its impact on our everyday life, particularly in cyber security. For example, despite most internet of things (IoT) providers have improved the security of their devices, the number of IoT devices attacked by distributed denial of service (DDoS) is still increasing [1]. Recently, the combination of tools from game theory (understanding strategic situations with several decision makers), mathematical modelling of infectious disease and network science (understanding interaction structure between decision makers) have demonstrated their strength to build new models that bring interesting cyber defense mechanisms[8] and [9]. An important cyber deception technique in order to mitigate cyber attacks is the use of honeypots. A honeypot is a token that a player can place on an edge of the network to create fake information. Honeypot strategies have been recently theoretically and practically optimized in lateral movement [3] through the heuristic search value iteration (HSVI) [10] discussed in the field of partially observable Markov decision processes (POMDPs). However, the definition of HSVI in game theory applies to the more general concept of one-sided partially observable stochastic games (OS-POSGs). Furthermore, in a context of attack by epidemics, (1) the aim of the network

attacker is to make on every node a transition from his desire, non infected to infected for example, (2) attacker has the true information over the network state. One of such epidemics is Mirai botnet, an epidemic that compromises a maximum number of IoT devices before launching a DDoS using the compromised IoT devices. Consequently, the fight against botnet epidemics can be studied as a zero-sum OS-POSG like in [11]. To the best of our knowledge, there is no work that addresses the stopping of epidemic from the perspective of POSGs. This work-in-progress paper illustrates how to deal with this scientific gap.

We define a game model based on the actions of an attacker trying to compromise and take control of vulnerable nodes in a network, and the actions of a defender trying to mitigate attacker's actions while offering patches to vulnerable nodes. Moreover, each node reacts to both players actions and therefore the global state of the system evolves accordingly. This model associates epidemics and POSG models. The definition of an epidemic model corresponds to the compartmentalisation of the individual states and the possible transitions of an individual from one compartment to another. The Susceptible-Infected-Recovered (SIR) model [2] for instance involves infectious (or infected) nodes (compartment I), who carry the virus, susceptible nodes (compartment S), which are vulnerable but not infected and recovered (or resistant, non vulnerable) nodes (compartment R). Botnet epidemics are SIR epidemics type like in [5], in which, without loss of generality, we consider only $S \to I$ (S to I), $S \to R$ and $I \to S$ transitions. The attacker does not observe defender's actions, which induces a general POSG with partial observations for both sides. Indeed, even-though she knows the game state (perfect information on one side), no player observes the opponent's moves (incomplete information on both sides). Our model considers the vulnerability of nodes and the defender placing honeypots on edges to detect some propagation and cure relevant nodes. The contributions of the paper are threefold:

- a zero-sum stochastic game model is proposed to study optimal deception strategies against virus propagation,
- we study a two-player zero-sum stochastic game in which no player observes the opponent's actions and prove that the heuristic search value iterated can be applied in this model ,
- our model involves two players acting strategically on the system composed of actors acting in a probabilistic way.

The rest of the paper is organized as follows. In the next section, we describe our model and the problem. Then, in section 3, the solution of the problem is given and in section 4, we prove several properties of the dynamic programming operator. After this, we provide some numerical illustrations in section 6 and finally we present a short discussion on further works and a conclusion in section 6. Note that all proofs and a complete related work section are fully described in the long version of this paper [7].

2 Model description

We present the model that describes interactions between the botnet (controlled by an attacker), the devices and the agent (the defender) who intends to prevent the botnet from controlling the network throughout the infected nodes.

2.1 Problem Description

An attacker is trying to take control of a large number of devices of a network and make it a foothold to lunch a fatal attack. This attack may be for example to overload a server with a very

large number of requests. Her strategy consists in silently spreading over the network a worm that ensures her the control of any device. She will propagate the worm until she has taken control of the desired number of devices. Fortunately for attacker, as observed in the Mirai attack [1], many devices do not have customized passwords and are therefore vulnerable, so attacker just has to select these vulnerable devices to spread the worm up to the targeted number. She frequently makes a probe over the network and then knows which nodes are vulnerable, which nodes are infected (and which nodes are resistent). To mitigate this spread, a defender combines two solutions:

1. He offers patches for infected devices and incites them to accept it. He also incites vulnerable, non-infected devices to customize their passwords and therefore become resistant against any attack. However, the result of this incitement is not predictable. Nevertheless, defender knows the decision of any device, i.e. knows if a device has been patched or not.
2. Defender has at his disposal a fixed number of honeypots that he can deploy on edges. The validity of each honeypot is one time-slot. Note the the attacker does not have the honeypots localization knowledge. A honeypot detects any virus propagation that traverses the edge and then the defender strongly incites the device and the newly infected nodes to patch.

This scenario is repeated until attacker has reached the targeted number of infected devices or there is no infected device left in the network (the latter is an absorbing state of the system as the virus has totally disappeared).
Defender's action consists in reallocating a limited number of honeypots (not necessarily all) on edges of the network at any new time-slot of the game. Attacker's action at any time-slot consists in choosing neighbours of infected devices to propagate the worm from, i.e. choosing an edge to propagate the virus from an infected node to a non-infected one. We assume that from each infected device she can chose at most one adjacent node to contaminate. Also, at each time-slot of the game, each node may decide to change his state by applying the patch if the node is infected, or changing his password if the node is susceptible to be infected by the virus. This node decision is made randomly and known by defender. The nodes behavior given by the probability distribution of their choice is known by the attacker and the defender.

2.2 Model

Because of an epidemic spreading over the network and users' actions, there are 3 classes of devices: *infected* devices (I), *susceptible* devices (S), that are vulnerable and non infected, and *resistant* devices (R), that cannot be successfully attacked.[4] For instance, a resistant device has a customized password and we assume that he will be resistant forever. In other words, R is an absorbing state, or there is no transition from R to any other state. For an infected device to become resistant, two transitions are necessary: $I \to S$ then $S \to R$. This is because an infected device must be patched first (transition from state I to state S) to be restored with basic features including default password, and then to change the default password to a customized password (transition from state S to state R). These two transitions cannot be done during a single stage. Indeed, changing the default password is useless while the node is under the control of the botnet. So we consider transition $I \to R$ is not possible at one stage. Consequently, only 3 transitions are possible for any device state dynamic in our framework: $S \to I$, $S \to R$ and $I \to S$ as depicted in figure 1.

[4] Unlike in [6], we consider the use of the world "resistant": (1) instead of "recovered" to keep in thought the non-vulnerability of the device; (2) instead of "removed" to keep in thought that the device is still in the game scenario.

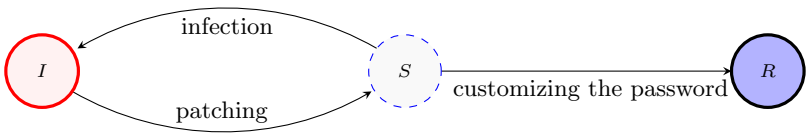


Fig. 1: Possible effective state transitions for each device.

There is a clear conflict between the attacker and the defender, but the information knowledge is not the same for these two decision makers. Attacker knows the global state of network (i.e. the state of each device at any time-slot of the game) but cannot observe defenders' actions while defender, who only knows the decision taken by devices about patching and password changing, has a partial observation of the global state. At the beginning of each time-slot, he only knows nodes who became susceptible or resistant in the past time-slot, but not the ones that are infected.

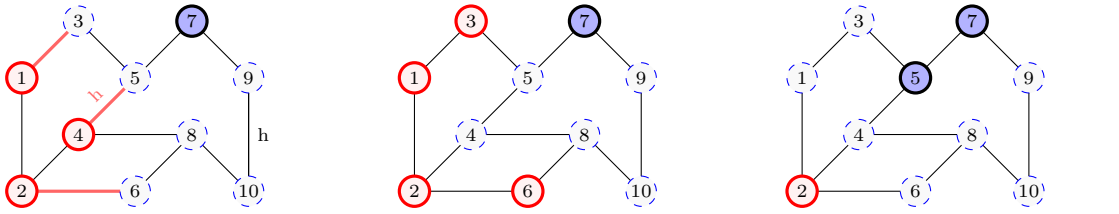
The problem can be modeled with a *two-player zero-sum OS-POSG* concept with private observation in which attacker does not know the actions defender has taken. Such a model is neither a classical POSG ([3]), in which attacker observes defender's actions, nor POSG with private information, in which no player can observe another player's private state [4]. Furthermore, the epidemic aspect brings forward two additional parameters: the endogeneous probabilities of transitions $S \to R$ and $I \to S$. Our model is a stochastic game represented by the tuple:

$$\mathcal{G} = \left(G,\, Z,\, A_1,\, A_2,\, O,\, \varrho,\, \alpha,\, T,\, r,\, b^0\right),$$

where:

- $G = (V,\, E)$ is the network (a finite and non directed graph) with V set of nodes (devices) and E set of edges;
- Z is the set of possible states of the devices, $\Delta(Z)$ is the set of probability distributions over Z. The global state is given by the class (S, I or R) of each node and the total number of states is $3^{|V|}$.
- A_1 and A_2 respectively denote the sets of possible actions for defender and attacker, $A = A_1 \times A_2$ is the set of possible joint actions;
- O is the observation space (of the defender);
- ϱ and α are respectively the probability for a susceptible node to become resistant at the next time step and the probability for an infected node to become susceptible at the next time step (see figure 1);
- $T : Z \times O \times Z \times A \longrightarrow [0,\, 1]$ is an application such that $T(\cdot \,|\, z, a) \in \Delta(Z \times O)$ (i.e. $T(\cdot \,|\, z, a)$ is a probability distribution over $\Delta(Z \times O)$) for any $(z, a) \in Z \times A$. T is called the *transition function*;
- $r : Z \times Z \longrightarrow \mathbb{R}$ is the *reward function* induced from a transition of a node state. $r(z_i, z_i')$ is the reward of defender when state of node i changes from z_i to state z_i';
- $b^0 \in \Delta(Z)$ is the initial belief of defender over the state of the network.

The game is repeated and at each time-slot, each player (attacker and defender) chooses an action as illustrated on figure (2).



Beginning of the time-slot (and of the first stage). Defender places 2 honeypots, on edges $4 \leftrightarrow 5$ and $9 \leftrightarrow 10$; attacker lunches the propagation $1 \to 3$, $2 \to 6$ and $4 \to 5$; the first two ones are undetected but the last one is detected.

End of the first stage (and beginning of the second stage). The propagation on nodes 3 and 6 is not detected and therefore result in two new hosts, whereas defender has intercepted an infection traversing the edge $4 \leftrightarrow 5$. Nodes 4 and 5 therefore accept to patch.

End of the time-slot (and of the second stage). With respect to probabilities α and ϱ, nodes 1, 3 and 6 transition $I \to S$, node 5 transitions $S \to R$ and the other nodes do not effectively change their state.

i = susceptible node; *i* = infected node; *i* = resistant node; ——— = edge; ——— = edge chosen by attacker; h = honeypot edge.

Fig. 2: One time-slot of the game: a possible scenario with 10 nodes

2.3 Model description

For a better understanding of the proposed framework, we bring up the following mathematical notations.

The system To simplify, we say that nodes are indiced by $1, 2, \ldots, i, \ldots, |V|$. An edge is any pair $\{i, j\}$ of connected nodes, i.e. a subset of V of 2 elements. So, $E \subseteq \{e \in 2^V \,|\, |e| = 2\}$. 3 different objects ($S$, I and R) define the possible states of any node at each time-slot of the repeated game. The state of each node defines the global state $z = (z_i)_{i=1}^{|V|}$ of the network with

$$z_i = \begin{cases} S \text{ if node } i \text{ is susceptible} \\ I \text{ if node } i \text{ is infected} \\ R \text{ if node } i \text{ is resistant} \end{cases}.$$

The actions

– Defender's action consists in deploying honeypots on edges. The maximum number h of honeypots is fixed and a honeypot remains on an edge only for one time-slot. Since he knows resistant nodes and also knows that there is no interest for the attacker to attack a resistant node, the defender will place a honeypot only on an edge between non-resistant ends. Hence, a defender's action is any set a_1 of at most h edges that are disjoint from R and has the following properties:

$$\begin{cases} a_1 \subseteq E \\ |a_1| \leq h \\ \forall u \in a_1, \quad u \cap R = \emptyset \end{cases}. \tag{1}$$

– Attacker's action consists in propagating the worm from each infected node through one edge of her choice linking this node to an adjacent, susceptible node if such a node exists. To model this action, we say that any attacker's action is an edge, keeping in mind that : attack is lunched from an infected node; attacker will infect only susceptible nodes; from every infected node, infection will propagate to at most one node. Finally, an attacker's action is a set a_2 of edges such that: each edge contains an infected node and a susceptible node; for all infected node i there exists at most one edge through which an infection is lunched. i.e.:

$$\begin{cases} a_2 \subseteq E \\ \forall u \in a_2, \quad \begin{cases} u \cap I \neq \emptyset \\ u \cap S \neq \emptyset \end{cases} \\ \forall i \in I, \quad |\{u \in a_2 : i \in u\}| \leq 1 \end{cases} . \tag{2}$$

The transition happens in two steps: the joint action $a = (a_1, a_2)$ in the state z makes the network transition to an intermediate state $a(z)$ (Players' action); nodes' probabilistic moves in the state $a(z)$ causes another transition to a state z' for the following time-slot (nodes' actions).

– *Players' action* In case node i is susceptible, his state changes (to infected) if and only if attacker lunches an attack from an infected node to him. In case node i is infected, his state changes (to susceptible) if and only if defender detects an attack lunched from its position through a honeypot to a susceptible node.Remember that resistant nodes remain resistant. We introduce for any collection X of sets, the set $\mathcal{U}(X) = \bigcup_{\omega \in X} \omega$. Note that: a node i is a side of an infection (either the side propagating or receiving) if and only if $i \in \mathcal{U}(a_2)$; node i is a side of an undetected infection if and only if $i \in \mathcal{U}(a_2) \setminus \mathcal{U}(a_1 \cap a_2)$. The transitions due to player's action can be explained as follows: the state of a susceptible node transitions if and only if the node is a side an infection and is not a side of a detected infection; the state of an infected node transitions if and only if the node is a side of a detected infection. We denote by $a(z)_i$ the intermediate state of node i induced by player's actions a when the state of the system is z. Then, for all node i:

$$z_i = S \implies \begin{cases} a(z)_i = I & \iff \quad i \in \mathcal{U}(a_2) \setminus \mathcal{U}(a_1 \cap a_2) \\ a(z)_i = S & \iff \quad i \notin \mathcal{U}(a_2) \setminus \mathcal{U}(a_1 \cap a_2) \end{cases},$$

$$z_i = I \implies \begin{cases} a(z)_i = I & \iff \quad i \notin \mathcal{U}(a_1 \cap a_2) \\ a(z)_i = S & \iff \quad i \in \mathcal{U}(a_1 \cap a_2) \end{cases},$$

$$z_i = R \implies a(z)_i = R.$$

– *nodes' actions* After this first intermediate transition, each node who is still susceptible after player's actions becomes resistant with probability ϱ or remains susceptible; each node who is still infected after player's actions becomes susceptible with probability α or remains infected. The relying probabilities $\mathbb{P}(z_i' \,|\, a(z)_i, z_i)$ are given in the following table:

		z_i'		
		S	I	R
z_i	$a(z)_i$			
S	S	$1-\varrho$	0	ϱ
	I	0	1	0
I	S	1	0	0
	I	α	$1-\alpha$	0
R	R	0	0	1

The observations Defender information concerns the nodes who decide to change their states, possibly under the incitement of defender. Formally, if we consider observation to be the result of such a transition, an observation is a set o such that:

$$\begin{cases} o \subseteq S \cup R \\ i \in o \end{cases} \iff \left(\begin{cases} a(z)_i = S \\ z_i' = R \end{cases} \text{ or } \begin{cases} a(z)_i = I \\ z_i' = S \end{cases} \right). \tag{3}$$

The transition function. The calculation of the probability $T(z', o \mid z, a)$ for a transition from a state z to a state z' is worth done only for the subsequent observation, i.e. for the unique observation $o = o(z, z')$ such that:

$$i \in o(z, z') \iff \left(\begin{cases} z_i = S \\ z_i' = R \end{cases} \text{ or } \begin{cases} z_i = I \\ z_i' = S \end{cases} \right).$$

More explicitly, $T(z', o \mid z, a) = \begin{cases} \mathbb{P}(z_i' \mid a(z)_i, z_i) & \text{if } o = o(z, z') \\ 0 & \text{otherwise} \end{cases}$, where $a(z)_i$ is the intermediate state from the state z to the state $z^{'}$ when the joint action $a = (a_1, a_2)$ is taken.

The rewards The transition of any node's state results in a payoff to defender and exactly the opposite value to attacker. This payoff function of the node states at current and next timeslots and we define it by tree non-negative constants r_1, r_2 and r_3 as shown in figure 3.

		Next state z_i'		
		S	I	R
Current state z_i	S	0	$-r_2$	r_3
	I	r_2	$-r_1$	$--$
	R	$--$	$--$	0

Fig. 3: Defender's payoff for any node i state transition.

Defender's total payoff is defined by:

$$R(z, z') = \sum_{i \in V} r(z_i, z_i'), \tag{4}$$

while his reward is the expected total payoff:

$$\bar{R}(z, a) = \sum_{z' \in Z} \mathbb{P}(z' \mid a(z), z) \times R(z, z') = \sum_{z' \in Z} \sum_{i \in V} \mathbb{P}(z' \mid a(z), z) \times r(z_i, z_i'). \tag{5}$$

3 Solution description

Defender does not know the targeted number of nodes attacker wishes to infect, but attacker's payoff measures how much she is coming close or far to her objective. So we solve the game where defender's objective is to maximize her total (or expected total) reward at infinite horizon. We precise some notion of game theory for a better understanding of the strategy in our particular context.

3.1 Strategies

Attacker may be playing a mixed strategy. Henceforth, any defender's strategy that is optimal in pure strategy is also optimal in mixed strategy. So we are interested only in mixed strategies. At each time-slot of the repeated game, players strategies are called *one-stage strategies*. For defender who does not know the network's state, the strategy π_1 is a probability distribution over the set A_1 of his possible actions. i.e. $\pi_1 \in \Delta(A_1)$. The set of defender one-stage strategies is $\Delta(A_1)$. Attacker's strategy depends on the state z of the network and, for any state z, she plays conditional strategy $\pi_2(\cdot \mid z) \in \Delta(A_2)$. i.e. she plays action a_2 with probability $\pi_2(a_2 \mid z)$. So, attacker's one-stage strategy is a probability vector $\pi_2 : Z \longrightarrow \Delta(A_2)$ that maps a probability distribution $\pi_2(\cdot \mid z)$ to any state z. The set of attacker's one-stage strategies is $\Delta(A_2)^Z$.

Defender updates his belief time-slot after time-slot according to the one-stage strategies. If he has the belief b at current time-slot, plays action a_1, makes observation o while he knows that attacker has played strategy π_2, then he updates his belief to a value $b^{'}$ such that:

$$b_{\pi_2}^{a_1,o}\left(z^{'}\right) = \frac{1}{\mathbb{P}_{b,\pi_2}(o \mid a_1)} \sum_{z \in Z} \sum_{a_2 \in A_2} T\left(z^{'}, o \,\middle|\, z, a_1, a_2\right) b(z)\, \pi_2(a_2 \mid z), \tag{6}$$

where

$$\mathbb{P}_{b,\pi_2}(o \mid a_1) = \sum_{z^{'}, z \in Z} \sum_{a_2 \in A_2} T\left(z^{'}, o \,\middle|\, z, a_1, a_2\right) b(z)\, \pi_2(a_2 \mid z). \tag{7}$$

Yet, each one-stage strategy may follow player's information up to the moment he/she is going to take his/her action. This information is called *history*. The history of defender at time-slot $t \geq 2$ is the sequence $h_1 = \left(a_1^1, o^1, a_1^2, o^2, \ldots, a_1^{t-1}, o^{t-1}\right)$ of observations and defender's actions up to time-slot $t-1$; the history of attacker at time-slot $t \geq 2$ is the sequence $h_2 = \left(z^1, a_2^1, z^2, a_2^2, \ldots, z^{t-1}, a_2^{t-1}, z^t\right)$ of network's states and attacker's actions up to time-slot $t-1$, added to the current state; at time-slot 1 attacker's history is reduced to state of the network and defender has an empty history.

3.2 Utility

Discounting the reward with a factor $\gamma \in [0, 1]$, we consider the total expected reward, denoted utility, at infinite horizon.[5] At any time-slot at which each player i plays strategy π_i in state z, the

[5]Since the number of infected node cannot exceed $|V|$, the probability at each time-slot that all infected nodes become susceptible is greater or equal to $(1-\alpha)^{|V|}$. So, at a certain time-slot, there will be no infected node and later all node will be resistant. There is no payoff from this time-slot and consequently the total expected reward converges even with discount factor $\gamma = 1$.

expected reward (of defender) is given by:

$$\mathbb{E}^{z}_{\pi_1,\pi_2}\left[\bar{R}\right] = \sum_{a_1\in A_1}\sum_{a_2\in A_2}\pi_1\left(a_1\right)\pi_2\left(a_2\,|\,z\right)\bar{R}\left(z,a_1,a_2\right) = \sum_{a\in A}\pi\left(a\,|\,z\right)\bar{R}\left(z,a\right), \tag{8}$$

where $\pi\left(a\,|\,z\right) = \pi_1\left(a_1\right)\pi_2\left(a_2\,|\,z\right)$ is the probability that players play the joint action a. The expected reward (of defender) who plays with belief b over the network state is:

$$\mathbb{E}^{b}_{\pi_1,\pi_2}\left[\bar{R}\right] = \sum_{z\in Z} b\left(z\right)\mathbb{E}^{z}_{\pi_1,\pi_2}\left[\bar{R}\right]. \tag{9}$$

So, if both players are playing a joint strategy $\pi = (\pi_1, \pi_2)$, then, given initial belief b^0, the *utility* (of defender) is given by:

$$\begin{aligned} U_{\pi_1,\pi_2}\left(b^0\right) &= \sum_{z\in Z} b^0\left(z\right)\mathbb{E}^{z}_{\pi^z}\left[\bar{R}\right] + \sum_{t=2}^{\infty}\gamma^{t-1}\times \\ &\quad\times\sum_{z\in Z}\sum_{\substack{z_2,\cdots,z_t\in Z\\ a_1,\cdots,a_{t-1}\in A\\ o_1,\cdots,o_{t-1}\in O}}\left[b^0\left(z\right)\left(\prod_{\tau=2}^{t}\left[T\left(z^\tau,o^{\tau-1}\middle|z^{\tau-1},a^{\tau-1}\right)\pi\left(a^{\tau-1}\middle|z^{\tau-1}\right)\right]\right)\mathbb{E}^{z_t}_{\pi^h}\left[\bar{R}\right]\right] \\ &= \sum_{z\in Z} b^0\left(z\right)\varphi\left(z\right), \end{aligned} \tag{10}$$

where

$$\begin{aligned} \varphi\left(z\right) &= \mathbb{E}^{z}_{\pi^z}\left[\bar{R}\right] + \\ &\quad + \sum_{t=2}^{\infty}\gamma^{t-1}\sum_{\substack{z_2,\cdots,z_t\in Z\\ a_1,\cdots,a_{t-1}\in A\\ o_1,\cdots,o_{t-1}\in O}}\left[\left(\prod_{\tau=2}^{t}\left[T\left(z^\tau,o^{\tau-1}\middle|z^{\tau-1},a^{\tau-1}\right)\pi\left(a^{\tau-1}\middle|z^{\tau-1}\right)\right]\right)\mathbb{E}^{z_t}_{\pi^h}\left[\bar{R}\right]\right]. \end{aligned} \tag{11}$$

3.3 Objectives

Defender's objective is to maximize his utility, which is the opposite for attacker's objective. The solution is then a strategy π_1 which is defender's a best response to some strategy π_2 which is also a best response to π_1. When defender plays a strategy π_1, we should suppose that attacker is best responding to π_1 and consider the utility in this case, referred to as the *value function* v_{π_1} *of strategy* π_1 with initial belief b^0, defined by:

$$\begin{aligned} v_{\pi_1} : \Delta\left(A_1\right) &\longrightarrow \mathbb{R} \\ b^0 &\longmapsto \min_{\pi_2} U_{\pi_1,\pi_2}\left(b^0\right). \end{aligned} \tag{12}$$

Denote

$$U = \sum_{t=0}^{\infty}\gamma^t\left(\min_{z,a_1,a_2}\bar{R}\left(z,a_1,a_2\right)\right) = \frac{\min\limits_{z,a_1,a_2}\bar{R}\left(z,a_1,a_2\right)}{1-\gamma} = \frac{\underline{r}}{1-\gamma} \tag{13}$$

and

$$L = \sum_{t=0}^{\infty} \gamma^t \left(\max_{z,a_1,a_2} \bar{R}(z,a_1,a_2) \right) = \frac{\max\limits_{z,a_1,a_2} \bar{R}(z,a_1,a_2)}{1-\gamma} = \frac{\overline{r}}{1-\gamma}, \tag{14}$$

where $\underline{r} = \min\limits_{z,a_1,a_2} \bar{R}(z,a_1,a_2)$ and $\overline{r} = \max\limits_{z,a_1,a_2} \bar{R}(z,a_1,a_2)$. It is clear each reward is bounded between $\min\limits_{z,a_1,a_2} \bar{R}(z,a_1,a_2)$ and $\max\limits_{z,a_1,a_2} \bar{R}(z,a_1,a_2)$. Consequently, each utility is bounded between L and U and equation (12) is consistent. The goal of defender who has the belief b^0 is to maximize the value of the game. The *optimal value* v^* *of the game* when defender has initial belief b^0 is the application:

$$\begin{aligned} v^* : \Delta(A_1) &\longrightarrow \mathbb{R} \\ b^0 &\longmapsto \max_{\pi_1} v_{\pi_1}\left(b^0\right). \end{aligned} \tag{15}$$

This notation is consistent for the aforementioned reason. The following theorem gives the main result of important properties over the optimal value function v^*.

Theorem 1. *The optimal value function* v^* *of the game is convex and* δ*-Lipschitz continuous.*

Following this theorem, we can prove that the heuristic search with the value iteration (HSVI) procedure holds and can be used to determine solutions of the game.

4 Value Backup Operator

In this section, we prove that the heuristic search with the value iteration (HSVI) procedure holds under our assumptions as well as in the general concept of two-player zero-sum OS-POSG where one player knows everything but not the current action of his opponent. To this end, we review the proof of all relevant properties in the following section. Let us first introduce the value backup operator for stochastic games in which attacker has a complete information and we prove that important results for this operator still hold for our particular stochastic game with partial information. The defender's reward in this time-slot game when strategies π_1 and π_2 are played is

$$U^V_{\pi_1,\pi_2}(b) = \bar{R}^{\text{imm}}_{\pi_1,\pi_2}(b) + \gamma \bar{R}^{b,\text{subs}}_{\pi_1,\pi_2}(V), \tag{16}$$

where

$$\begin{aligned} \bar{R}^{\text{imm}}_{\pi_1,\pi_2}(b) &= \sum_{z\in Z} \sum_{a_1\in A_1} \sum_{a_2\in A_2} b(z)\,\pi_1(a_1)\,\pi_2(a_2 \mid z)\,\bar{R}(z,a_1,a_2) \\ &= \sum_{z\in Z} \sum_{a\in A} b(z)\,\pi(a \mid z)\,\bar{R}(z,a) \end{aligned} \tag{17}$$

is the reward in the one-stage game and

$$\bar{R}^{\text{subs}}_{\pi_1,\pi_2}(b,V) = \sum_{a_1\in A_1} \sum_{o\in O} \pi_1(a_1)\,\mathbb{P}_{b,\pi_2}(o \mid a_1)\,V\left(b^{a_1,o}_{\pi_2}\right) \tag{18}$$

is the reward in the subsequent game. Denote by HV the optimal value function of the stage game, called *value backup* operator, i.e.:

$$[HV](b) = U^V_{\pi_1,\pi_2}(b) = \max_{\pi_1} \min_{\pi_2} \left[\bar{R}^{\text{imm}}_{\pi_1,\pi_2}(b) + \gamma \bar{R}^{\text{subs}}_{\pi_1,\pi_2}(b,V) \right]. \tag{19}$$

After several lemmas and properties described in the full version of the paper in [7], we have the following main theorem which induces that the optimal value v^* of the game is the fix point of the value backup operator H.

Theorem 2. *The operator H is γ-contracting in the space of convex continuous functions $V : \Delta(Z) \longrightarrow \mathbb{R}$ under the max-norm: $\|V\|_\infty = \max_{b \in \Delta(Z)} \|V(b)\|$.*

Henceforth from Banach fix point theorem, operator H admits a fix point V^* to which converges any sequence $(V_n)_{n \in \mathbb{N}^*}$ of convex continuous functions such that $V_{n+1} = HV_n$ for every n. The following theorem states that this fix point is the value v^* of the game, which means that any algorithm that iteratively corrects the value converges to v^*.

Theorem 3. *The value v^* of the game is the fix point of the backup operator H.*

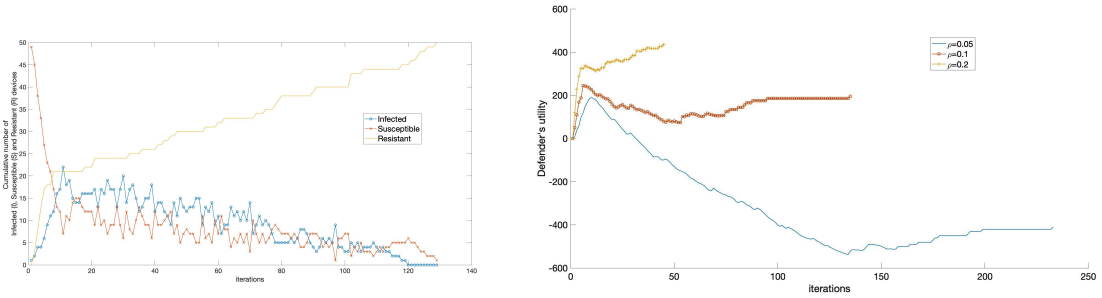
Following this important property and based on dual linear programs, two algorithms are proposed in the longer version of this work-in-progress paper [7] in order to compute the lower bound $V_{\mathrm{UB}}^{\Upsilon}$ and the upper bound V_{LB}^{Γ} of the optimal value v^* of the game. Since $V_{\mathrm{UB}}^{\Upsilon}$ and V_{LB}^{Γ} are iteratively refined upper and lower bounds of the optimal value function at any belief b, it is possible to get a ε-optimal value at any belief b with any precision , i.e: $V_{\mathrm{UB}}^{\Upsilon}(b) - V_{\mathrm{LB}}^{\Gamma}(b) \leq \varepsilon$.

5 Numerical Illustrations

Some simulations of simple strategies for both players are presented in this section. We consider an Erdos-Reyni random graph with 50 nodes and a parameter 0.3 (probability to active each edge). Both players strategy is a fully random strategy without history. Meaning that the attacker chooses randomly a susceptible device from an infected device uniformly, and the defender chooses randomly the edges to allocate honeypots uniformly over the possible edges (edges that connect two not resistant nodes). A single node, chosen randomly, is infected at time-slot 1 and all the other nodes are susceptible. Examples of simulations are depicted on figure (4). Note that on figure (4b) we observe the impact of the probability to change the default password on the defender's utility and also on the extinction time of the virus. The number of honeypots h has also an important impact on these two performance measures. Other simulations run 100 times with $\rho = 0.1$, show that the average utility goes from 83.03 with a 99% confidence interval $[23.38 - 142.68]$ to 385.40 with a 99% confidence interval $[355.23 - 415.58]$, when h goes from 3 to 10.

6 Conclusions and Further Work

This work is related to possible transitions of node states in a network prone to malware attack. Each transition makes attacker loose what defender gains and may be probabilistic or caused by actions of both decision makers. This security problem is part of the large family of cyber security and the solution concept studied here is cyber deception. Particularly, we are interested in honeypots techniques which help to discover infected nodes into a network through observing cyber contamination. Our framework is much more complicated than traditional zero-sum OS-POSG model, because information about player's actions is not fully observable. Even in this complex system, we have been able to prove that the heuristic search value iteration can be applied in order to find lower and upper bound on the optimal value of the game.



(a) Evolution of each state categories when $\rho = 0.1$.

(b) Impact of the reaction of devices to become resistant.

Fig. 4: Output of simulations with $\gamma = 0.99$, $h = 3$, $r_1 = 0.1$, $r_2 = 1$, $r_3 = 10$ and $\alpha = 0.5$.

This work is still in progress and implementation of the algorithms are on going. The epidemiological aspect of this model further makes intricate the non-scalability of algorithms designed for the general model. So we wish to outline efficient algorithm specially designed for this model.

References

1. Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J.A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., Zhou, Y.: Understanding the mirai botnet. In: 26th USENIX Security Symposium (USENIX Security 17), pp. 1093–1110. USENIX Association, Vancouver, BC, Canada (2017). URL https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis
2. Colizza, V., Vespignani, A.: Invasion threshold in heterogeneous meta population networks. Phys Rev Letters **99** (2007)
3. Horák, K., Bošanský, B., Pěchouček, M.: Heuristic search value iteration for one-sided partially observable stochastic games. Proceedings of the 1st International Joint Conference on Artificial Intelligence **31**, 558–564 (2017). DOI 978-1-57735-780-3
4. Kartir, D., Nayyar, A.: Stochastic zero-sum games with asymmetric information (2019)
5. Kim, J., Radhakrishnan, S., Dhall, S.K.: Measurement and analysis of worm propagation on internet network topology (2004)
6. Kiss, I., Miller, J., Simon, P.: Mathematics of Epidemics on Networks, vol. 46 (2017). DOI 10.1007/978-3-319-50806-1
7. O.Tsemogne, Hayel, Y., Kamhoua, C., Degoue, G.: Epidemic model and partially observable stochastic games for cyber deception. draft full version (https://drive.google.com/file/d/1k4Qs0d38cmYXfxV5YE6D1SJqDjqqukE6/view?usp=sharing) (2020)
8. Pawlick, J., Colbert, E., Zhu, Q.: A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. ACM Computing Surveys (2019)
9. Roy, S., Ellis, C., Shiva, S., Dasgupta, D., V. Shandilya, C.W.: A survey of game theory as applied to network security. pp. 1–10 (2010)
10. Smith, T., Simmons, R.: Heuristic search value iteration for pomdps. Proceedings of UAI (2012)
11. Wiggers, A., Oliehoek, F., Roijers, D.: Structure in the value function of zero-sum games of incomplete information (2015)