Model-based Response Planning Strategies for Autonomic Intrusion Protection

STEFANO IANNUCCI, Mississippi State University, USA
SHERIF ABDELWAHED, Virginia Commonwealth University, USA

The continuous increase in the quantity and sophistication of cyber attacks is making it more difficult and error-prone for the system administrators to handle the alerts generated by Intrusion Detection Systems (IDSs). To deal with this problem, several Intrusion Response Systems (IRSs) have been proposed lately. IRSs extend the IDSs by providing an automatic response to the detected attack. Such a response is usually selected either with a static attack-response mapping or by quantitatively evaluating all the available responses, given a set of pre-defined criteria. In this paper, we introduce a probabilistic model-based IRS built on the Markov Decision Process (MDP) framework. In contrast with most existing approaches to intrusion response, the proposed IRS effectively captures the dynamics of both the defended system and the attacker and is able to compose atomic response actions to plan optimal multi-objective long-term response policies to protect the system. We evaluate the effectiveness of the proposed IRS by showing that long-term response planning always outperforms short-term planning and we conduct a thorough performance assessment to show that the proposed IRS can be adopted to protect large distributed systems at run-time.

CCS Concepts: • **Security and privacy → Artificial immune systems**;

Additional Key Words and Phrases: Intrusion Response System, Autonomic Intrusion Protection

ACM Reference Format:
Stefano Iannucci and Sherif Abdelwahed. 2018. Model-based Response Planning Strategies for Autonomic Intrusion Protection. *ACM Trans. Autonom. Adapt. Syst.* 1, 1 (April 2018), 23 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

According to the Akamai's state of the Internet 2015 Q3 Report [2], there has been a 179.66% increase in total DDoS attacks with respect to the same period of 2014. Security mechanisms, such as firewalls, encryption and properly configured access control policies have quickly shifted from being *the* defense mechanisms to being just *the first line* of defense [24]. The second line of defense is usually represented by signature-based or anomaly-based Network Intrusion Detection Systems (IDSs). The former are able to scan the content of the network packets looking for signatures of known attacks, but are unable to identify unknown (0-days) attacks. To this end, anomaly-based IDSs [5] have recently gained interest. They use machine learning or deep learning techniques [4] in order to find whether the protected system does not comply with the expected behavior and/or whether there is some anomaly in the network traffic flow.

Authors' addresses: Stefano Iannucci, Mississippi State University, 665 George Perry Street, Mississippi State, MS, 39762, USA, stefano@dasi.msstate.edu; Sherif Abdelwahed, Virginia Commonwealth University, Richmond, VA, USA, sabdelwahed@vcu.edu.

1556-4665/2018/4-ART $15.00
https://doi.org/0000001.0000001

The increasing number of cyber-attacks makes it difficult and error-prone for the system administrators to manually handle all the alerts generated by the IDSs. Intrusion Response Systems (IRSs) try to address this problem by automatically selecting the responses to the attacks detected by the IDSs [45]. Two main types of IRSs have been proposed so far [36]: (i) static mapping and (ii) dynamic evaluation (e.g., [37, 41, 44]). With the static mapping type the system administrators are expected to manually associate each category of detectable attacks with a prospective response action. However, periodically upgrading the response mapping can be overwhelming, given the massive amount of day by day newly discovered attacks and the ability of the attackers to bypass known protection mechanisms. The dynamic evaluation approach tries to overcome this limitation by letting the system administrator associate each single category of attack to a set of response actions. The response action is then chosen among the others according to an underlying system model and some evaluation criteria (such as the resolution time, cost and impact) by solving a multi-objective optimization problem (e.g., [16]) or by ranking the alternatives (e.g., [9, 11, 16, 17, 35, 37, 41]).

Most of the works proposed so far either try to model the behavior of the attacker using attack graphs (e.g., [17]) or to model the dependencies between the system components (e.g., [44]), but only a few of them introduce a comprehensive model able to describe the attacker behavior, the defender (IRS) behavior and the actual system dynamics (e.g., [47]). Having a full model of the system associated with a control framework has several advantages, among them the possibility to simulate the behavior of the controlled system and to estimate its evolution over time [1].

In this paper we use the Markov Decision Process (MDP) framework to model a system controlled by an IRS. Unlike other approaches, we do not select a single short-term optimal response action, rather we produce an optimal long-term policy, that is, an optimal sequence of response actions able to drive the system from its initial (under attack) state to a set of target (desired) states. We use the model to simulate the behavior of the system and we show that long-term policies always outperform short-term policies by estimating average values and confidence intervals of the attack resolution time, cost and impact for a system subject to real-world attacks. In addition, as an extension to our previous works [21, 22] we observe that being able to proactively react to a potential threat before it occurs is often better than waiting for it to reveal. To this end, we extend the previous Single-Agent MDP formulation, which only describes the system and the IRS behavior, by adding the attacker behavior to the model, using a competitive Multiple-Agent MDP implemented as a stochastic game [7]. We show that, when this model is adopted, the IRS is either able to fully prevent an attack or at least to anticipate some defense actions before the attack actually occurs.

Since the MDP state space grows exponentially with the number of features used to describe the system and the attacker states, we propose an algorithm able to instantiate the minimal MDP, that is, the MDP characterized by the minimum number of attributes and actions needed to drive the system to a secure state. Furthermore, we compare both the performance and the effectiveness of the Single-Agent formulation using state of the art optimal and a sub-optimal MDP planners implemented in the BURLAP library[1]. In addition, we extend the BURLAP library with a parallel Java implementation of the Value Iteration algorithm (VI) [3], which scales linearly with respect to the number of worker threads. The latter complements our previous work [23] on the performance evaluation of VI on Intel MIC architecture [13]. Finally, we evaluate the performance of an optimal planner in the Multiple-Agent case.

The remainder of the paper is organized as follows: in Section 5 we present a discussion of related works. The meta-modeling framework and the system models are presented in Section 2. In Section 3 we compare the performance and the effectiveness of optimal and sub-optimal MDP

[1]Brown-UMBC Reinforcement Learning and Planning (BURLAP), http://burlap.cs.brown.edu/

solvers and we propose an algorithm for reducing the MDP state space; detailed response policies evaluations are presented in Section 4. Section 6 concludes the work.

2 SYSTEM MODEL

The proposed system model derives from the MDP framework. In the following, we provide an overview of the theoretical foundations of the Single-Agent and Multiple-Agent MDPs and we explain how we use this framework to build a system model representing the behavior of an IRS when trying to protect a system from cyber attacks.

2.1 The MDP Modeling Framework

A Single-Agent Discrete-Time MDP is a stateful probabilistic approach to model the behavior and the run-time dynamics of a system. An MDP [3] is a tuple $\langle S, A, P, R, \gamma \rangle$, where S represents the state space that the agent can navigate and $s_k \in S$ represents the agent state at discrete time k. Even if not explicitly considered in the MDP framework, a common practice is to characterize each state with a number of attributes. A is the finite set of actions available to the agent to navigate the state space. Specifically, by executing at time k an action $a_k \in A$ in the current state $s_k \in S$, the agent moves to a successor state $s_{k+1} \in S$. The transition dynamics from the current to the next state are given by the transition probability function P. This function specifies, for each source state $s_k \in S$, for each destination state $s_{k+1} \in S$, and for each action $a_k \in A$, the value $P(s_k, a_k, s_{k+1})$, that is, the probability value that by executing the action a in state s at time k, the resulting state will be s_{k+1}. When the transition probability function is time independent (*stationary*) we have $\forall k.P(s_k, a_k, s_{k+1}) = P(s_k, a, s_{k+1})$. Every time an action is executed, the MDP agent is rewarded with a bonus (or penalized with a cost), according to the reward function R. That is, $R_k = R(s_k, a_k, s_{k+1})$ represents the reward that the agent will earn (or the cost the agent will pay) for executing at time k the action a in state s_k and being taken to some state s_{k+1}. Some MDP models use a different reward function, based only on s_k and a_k and not considering s_{k+1}. When the reward function is time independent (*stationary*) we have $\forall k.R(s_k, a_k, s_{k+1}) = R(s_k, a, s_{k+1})$. γ is the discount factor, usually defined in the interval $[0, 1]$, which specifies how much short-term rewards are preferred over long-term rewards.

The overall behavior of the agent is described by a deterministic or stochastic policy π. When π is deterministic it specifies, for each s_k, the action a_k that the agent must execute. When π is probabilistic it specifies a probability distribution such that $\pi : S \times A \rightarrow [0, 1]$. The objective of the agent is to find a policy π^* such that the discounted reward $R_k = \sum_{j=0}^{\infty} \gamma^j R_{k+j+1}$ is maximized.

Several optimal and sub-optimal algorithms for solving MDPs have been proposed (e.g., [3, 27, 29, 31, 38]), but one of the most commonly used remains the VI algorithm [3] because of its simplicity. It is based on the concept of *state-value* function $V_\pi(s_k) = \mathbb{E}_\pi[R_k | s_k]$, that is, the expected reward achievable by the agent starting from state s_k and then following policy π. The base step of the algorithm is to assign an initial random state-value V^0 to all the states and then to execute the iterative refinement process described in [6]:

$$V^{i+1}(s_k) = \max_{a_k \in A} R(s_k) + \gamma \sum_{s_{k+1} \in S} P(s_k, a_k, s_{k+1}) V^i(s_{k+1}) \tag{1}$$

The sequence of functions V^i converges linearly to the optimal value V^* in the limit and provides thus the expected maximum reward obtainable by following the optimal policy π^* from state s_k.

A Multi-Agent Discrete-Time Markov Decision Process (MA-MDP), also known as *Stochastic Game*, is an extension of the Single-Agent MDP. A MA-MDP with n agents is defined by the tuple $\langle S, A_1, ..., A_n, P, R_1, ..., R_n, \gamma \rangle$. With this formulation, different agents have possibly different sets of

available actions and can have potentially different reward function. The transition from s_k to s_{k+1} is therefore given by the joint action of all the agents: $a_k = [a_{k,1}, ..., a_{k,n}]$. According to the reward functions definitions, the agents could cooperate to reach a common objective or could behave selfishly, i.e., trying to achieve personal goals at the expense of the other agents. The former case is described in detail in [6], while the competitive case is described in [7].

2.2 System and IRS Modeling with MDP

In this work we use the Single Agent and stationary MDP modeling framework to model the behavior of a system responding to an attack. The agent represents the IRS, whose objective is to find an optimal policy to drive the system from a dangerous (under attack) state to a final (desired) state. We introduce two extensions to the general MDP framework: a *termination function* T and the *pre-condition function* PC. The termination function $T : S \rightarrow \{true, false\}$ is used to define the subset $S_{tgt} = \{s \in S | T(s) = true\}$ of the target states, which represents the set of the states where the agent stops its execution; the pre-condition function $PC : S \times A \rightarrow \{true, false\}$ is instead used to define whether an action $a \in A$ is executable in state $s \in S$. T and PC are introduced to simplify notations. They can be expressed directly in the base model with a target state $s_k \in S_{tgt}$ formulated as a state where $\forall a.P(s_k, a, s_{k+1}) = 0$, while a non-executable action a has the characteristic $\forall s_{k+1}.P(s_k, a, s_{k+1}) = 0$. The objective of the MDP agent is to drive the system from a starting state s to a target state $s' \in S_{tgt}$ such that the path between s and s' maximizes its reward.

In the following we describe how we modeled a simple system, characterized by 14 system attributes and that could be subject to 7 different attacks.

2.2.1 States Characterization. We use the Object-Oriented MDP representation introduced in [12] in which each state is characterized by a number attributes. Specifically, we compose the states by joining 2 macro-attributes: (i) the attack vector $\mathbf{p}$ and (ii) the system variables $\mathbf{v}$. The former contains as many variables as the number of attacks detectable by the IDSs and each variable $p_i \in \mathbf{p}$ represents the probability value that the system is currently under attack i. The latter represent the current system status.

We consider 7 different attacks and 14 system attributes. The attacks are modeled by the attributes $p_{scan}, p_{vsftpd}, p_{smbd}, p_{phpcgi}, p_{ircd}, p_{distccd}, p_{rmi}$, which represent the probability that the controlled system is being attacked, respectively by: a portscan, an exploit on the `vsftpd` daemon (OSVBD-73753), an exploit on the `smbd` daemon (CVE-2007-2447), an exploit on the execution of PHP as a CGI application (CVE-2012-1823), an exploit on the `ircd` daemon (CVE-2010-2075), an exploit on the `distccd` daemon (CVE-2004-2687) and finally and exploit on the `rmi` Java daemon (CVE-2011-3556). We specifically chose these attacks because their respective vulnerabilities are exposed by `metasploitable`[2], an intentionally vulnerable Linux Virtual Machine that can be used to conduct security training, test security tools and practice common penetration testing techniques. We consider the following system attributes:

- $firewall \in \{true, false\}$ represents whether the system firewall is active.
- $\{blocked\_ips\}$ represents the set of currently blocked source IP addresses from the firewall of the considered system.
- $\{flowlimit\_ips\}$ represents the set of currently throughput-limited source IP addresses.
- $alert \in \{true, false\}$ represents whether the system administrator has been alerted about the ongoing attack.

[2]Virtual Machine to test Metasploit https://information.rapid7.com/metasploitable-download.html

- $\{honeypot\_ips\}$ represents the set of IP addresses whose traffic is currently being redirected to an honeypot.
- $logVerb \in \{0, 1, 2, 3, 4, 5\}$ represents the currently configured logging verbosity of the applications installed on the considered system.
- $active \in \{true, false\}$ represents whether the considered system is currently active and serving requests or if it has been shut down.
- $quarantined \in \{true, false\}$ represents whether the considered system is currently active and serving requests or if it has been isolated from the network.
- $rebooted \in \{true, false\}$ represents whether the considered system has been rebooted during the execution of the current policy.
- $backup \in \{true, false\}$ represents whether the considered system has ever been backed up during the execution of the current policy.
- $updated \in \{true, false\}$ represents whether the software installed on the controlled system is updated.
- $manuallySolved \in \{true, false\}$ represents whether there has been a manual intervention during the execution of the current policy.
- $everQuarantined \in \{true, false\}$ represents whether the system has been quarantined during the execution of the current policy.
- $everShutDown \in \{true, false\}$ represents whether the system has been shut down during the execution of the current policy.

2.2.2 Reward Function. Although several works aim at addressing the problem of evaluating the response cost to counter or mitigate an intrusion, a standard methodology has not been decided yet [43]. A common approach for cost evaluation is to take into consideration the effectiveness of the response action as in [37] or to deal with the negative impact that it can have on the system [15, 18]. A third parameter usually contemplated for this evaluation is the operational cost that must be sustained to pay for hardware, software and human resources needed to counter the attack [30, 43]. However, there are also other factors that should be considered when planning a defense policy, among the others: the Confidentiality, Integrity, Availability (CIA) triad [10, 40] and the Service Level Agreement (SLA) [34] that the system is supposed to meet. The former is related to the data, which could be released, modified or made inaccessible without authorization, causing respectively a confidentiality, integrity and availability issue. The latter is instead related to the Quality of Service (QoS) that must be provided to the end-users in terms of non-functional requirements such as applications response time and system reliability [8].

In this work, we consider all the aforementioned attributes with the exception of the CIA triad for the defender's reward function. The latter is instead considered for the attacker's reward function and evaluated as discussed in Section 2.3.3. Furthermore, we also consider the response time, that is, the expected time needed to execute the defense policy on the target system. Specifically, we characterize the MDP reward function as a penalty score on the actions considered for inclusion in the defense policy. The reward function evaluates the response actions according to the following criteria:

- **Response Time** $T(x) \in \mathbb{R}$**,** represents the time needed to apply the response action x.
- **Cost** $C(x) \in \mathbb{R}$**,** represents the operational cost of applying the response action x.
- **Impact index** $I(x) \in [0, 1]$**,** represents the impact index of the response action x on the system and it is computed as follows:

$$\begin{cases} I(x) = w_R \frac{R(x)-R_0}{R_{max}} + w_D \frac{D_0-D(x)}{D_{min}}, & R(x) \leq R_{max}, D(x) \geq D_{min} \\ I(x) = 1, & otherwise \end{cases}$$

where $R(x)$ and $D(x)$ are the actual applications' execution time and reliability after the execution of the action x, R_0 and D_0 are respectively the applications' execution time and reliability during normal operations and $\forall x.R(x) \geq R_0, D(x) \leq D_0$; R_{max} and D_{min} are respectively the SLA upper limit for the response time and the lower limit for the reliability, w_R and w_D are custom weights with $w_R + w_D = 1$.

The reward function is then defined as follows.

$$R_{irs} = -w_t \frac{T(x)}{T_{max}} - w_c \frac{C(x)}{C_{max}} - w_i I(x) \tag{2}$$

where $w_t, w_c, w_i \in [0, 1]$ are custom weights used to balance the importance of the criteria in the multi-criteria optimization problem. T_{max} and C_{max} represent respectively the maximum response time and the maximum cost over all the considered response actions and are used to normalize their values. It is worth noting that in the reward function we did not consider the effectiveness of the response action, because it is tightly integrated with the model of the system in the form of actions' post-conditions, as described in Section 2.2.3.

2.2.3 Response Actions. In order to avoid activating potentially disruptive response actions when the system is not under severe attack and to better deal with the stochastic nature of the IDS inputs, we introduce two thresholds on the attack probability attributes, namely T_1 and T_2, $T_1 < T_2$. Thus, given an attack probability p, it can belong to one of the following 4 stages: ($p < T_1$) the IDSs have detected an insignificant anomaly that should be considered as noise and no response actions should be triggered. With ($T_1 \leq p < T_2$) the IDSs have detected a significant anomaly, which cannot be classified as an attack. However, the system can start planning some response action in order to prevent possible attacks. With ($T_2 \leq p < 1$) the anomaly detected by the IDSs is considered to be an unidentified attack, therefore the response plan generated by the IRS can only contain generic responses. When ($p = 1$) the attack has been identified and a specific response plan can be computed.

In the following we describe some of the response actions that our IRS prototype is able to apply on the controlled system. For each of them, we provide a description of its behavior and the response time R, cost C and impact I attributes, needed to compute the expected reward when planning the optimal policy. Each response is characterized by pre-conditions and post-conditions. The former identify a subset of the states in which the actions can be executed; the latter are used instead to compute the state in which the system will be after the execution of the considered action. It is worth noting that, although in this work we only consider statically defined transition probabilities as post-conditions, it is possible to establish a feedback loop between the system and the IRS so that they could be updated at run-time to better mimic the actual system behavior. Eventual dependencies between response actions are not directly modeled: indeed, using pre-conditions, we are able to model the eventual dependency of a response action on a given subset of states, which in turn could imply that some dependent actions have been executed prior to the execution of the current action. Table 1 summarizes response time, cost and impact attributes for the considered actions.

Firewall Activation. Starts the system's firewall in case it was not started previously. Its characteristics are:

- **Reward Attributes:** $T = 2, C = 1, I = 0$
- **Pre-Conditions:** $(p_{scan} \geq T_1 \vee p_{vsftpd} \geq T_1 \vee p_{smbd} \geq T_1 \vee p_{phpcgid} \geq T_1 \vee p_{distccd} \geq T_1 \vee p_{rmi} \geq T_1 \vee p_{ircd} \geq T_1) \wedge \neg firewall \wedge \neg quarantined \wedge active \wedge logVerb > 0$
- **Post-Conditions:** $Prob = 1, firewall = 1$

This action can be executed when at least one entry of the attack probability vector **p** is greater than or equal to T_1, the firewall itself has not been activated yet, the system is active and it has not been quarantined and the log verbosity is at least equal to 1. The resulting state after the execution of the action will be reached with probability 1 and is identical to the current state, but with the $firewall$ attribute set to $true$.

Block Source IP `badIP`. Configures the system's firewall in order to drop IP packets originated by the IP `badIP`. Its characteristics are:

- **Reward Attributes:** $T = 1, C = 3, I = 0.3$
- **Pre-Conditions:** $p_{scan} \geq T_2 \wedge firewall \wedge \neg quarantined \wedge active \wedge badIP \notin blocked\_ips \wedge alert \wedge logVerb > 1$
- **Post-Conditions:** $Prob = 1, blocked\_ips = blocked\_ips \cup \{badIP\}, p_{scan} = 0$

This action can be executed when the port-scan attack probability is greater than or equal to T_2 and the firewall has been previously activated. Furthermore, it is required that the system is active and that it has not been quarantined and that its log verbosity is at least equal to 2. Finally, the system administrator must have been previously alerted and the IP address of the attacker must not yet belong to the set of the blocked IPs. The resulting state after the execution of the action is identical to the current state, but with the `badIP` included into the set of the blocked IPs and with p_{scan} attribute set to 0. Setting the probability of an attack to zero for the next state means that the expected result in executing the given action is to certainly stop the attack.

Flow Rate Limit `badIP`. Configures the system's firewall to limit the traffic rate of IP packets originated by the IP `badIP`. Its characteristics are:

- **Reward Attributes:** $T = 3, C = 1, I = 0.2$
- **Pre-Conditions:** $p_{scan} \geq T_1 \wedge firewall \wedge badIp \notin flowlimit\_ips \wedge \neg quarantined \wedge active \wedge logVerb > 0$
- **Post-Conditions:**

$$\begin{cases} Prob = 0.5, & limited\_ips = limited\_ips \cup \{badIP\}, \\ & p_{scan} = 0 \\ Prob = 0.5, & limited\_ips = limited\_ips \cup \{badIP\} \end{cases}$$

This action can be executed when a port-scan attack probability is greater than or equal to T_1 and the firewall has been previously activated. Furthermore, it is required that the system is active and that it has not been quarantined and that its log verbosity is at least equal to 1. Finally, the IP address of the attacker must not belong to the set of the flow rate limited IPs. This action can drive the system to two different resulting states, with probability 0.5 each. In one case the action is able to stop the attacker and therefore we have $p_{scan} = 0$ together with the attacker IP address included in the set of flow rate limited IPs. In the other case the action is unable to stop the attacker and therefore we only obtain to limit the flow rate of the attacker's IP by adding it to the set of the flow rate limited IPs.

2.2.4 Termination Function. The policy planning terminates when the system reaches a target state $S_{tgt} = S_a \cup S_c$, where S_a is the subset of states in which the anomaly is harmless, while S_c is the subset of states representing a fully clean system. Both the subsets are identified with a Boolean expression on the state attributes, but for space reason we only report the Boolean condition representing the fully clean system state:

Action Name	Resp. Time	Cost	Impact
Generate Alert	1	1	0
Firewall Activation	2	1	0
Block Source IP	1	3	0.3
Unblock Source IP	1	3	0
Flow Rate Limit	3	1	0.2
Unlimit Flow Rate	3	1	0
Redirect to Honeypot	3	3	0.1
Un-honeypot	3	3	0
Increase Log Verbosity	2	1	0.05
Decrease Log Verbosity	1	1	0
Quarantine Host	5	5	1
Unquarantine Host	5	5	0
Manual Resolution	3600	200	0
System Reboot	60	6	0.7
System Shutdown	30	6	1
System Start	30	6	0
Backup Host	3600	10	0.1
Software Update	600	300	0.1

Table 1. Response Actions Parameter Summary

$S_c = \{s \in S | p_{scan} < T_1 \wedge p_{vsftpd} < T_1 \wedge p_{smbd} < T_1 \wedge p_{phpcgi} < T_1 \wedge p_{irc} < T_1 \wedge p_{distcc} < T_1 \wedge p_{rmi} < T_1 \wedge blocked\_ips = \emptyset \wedge flowlimited\_ips = \emptyset \wedge honeypot\_ips = \emptyset \wedge logVerb = 0 \wedge active \wedge \neg quarantined\}$.

A clean system state is represented by an attack probability vector whose values are all under the T_1 threshold and there are no firewall limitation configured.

2.3 Attacker Modeling with MA-MDP

The model described so far is able to capture the dynamics of the underlying system and can be used to plan optimal long-term policies to defend the system against an attack. However, even if the long-term policies always outperform short-term policies (more details in Section 4), an IRS built on such a model is not able to anticipate, and thus to prevent, a possible multi-step attack because the model does not describe the attacker behavior.

The competitive multi-agent extension of the model aims at introducing a proactive defense mechanism by describing the system and its dynamics when subject to control actions executed by both the IRS and the attacker. Knowing what actions are available to the attacker and their interdependencies allows for the planning of proactive long-term response policies, able to block ongoing attacks and to prevent an attack escalation.

In this extended model, each attack is characterized by three factors: (i) an *attack belief*, representing the probability that the attacker will launch a specific attack in the future; (ii) an attack action, based on pre-conditions that can specify dependencies on other attacks or on targets on the system; (iii) the effects that the attack has on the system attributes.

The multi-agent extension of the model, based on a two-agents stochastic game, inherits all the characteristics of the single-agent model and extends (i) the set of attributes, (ii) the set of the available actions and (iii) the reward function, which is now based on a joint action model.

2.3.1 Extended Attributes. Among the new attributes, the most important are the *timer* and the *nextAttackThreshold* attributes. These are used to take into account the execution time of both attacks and responses in the system, in order to simulate the coordinated attack-response behavior. Specifically, the *timer* attribute has been added to the IRS model, while *nextAttackThreshold* has been added to the attacker model. The main system timer is kept by *timer*. Its value is incremented each time a response action is executed IRS-side. From the attacker side, a newly launched attack increases *nextAttackThreshold* with the expected time needed to complete the attack. The attacker will not be able to launch new attacks until its threshold is greater than the system timer.

Attack beliefs are characterized by probability values. We add to the attacker agent model as many attributes as the number of executable attacks.

2.3.2 Extended Actions. Unlike the single-agent model, the multi-agent one is based on the concept of *joint action*. $(x, y)_k$ is a joint action for a two-agent stochastic game, where $x \in A_{irs}$ and $y \in A_{attacker}$ represent the actions chosen at time k respectively by the IRS and by the attacker. The new set of actions $A_{attacker}$, available only to the attacker agent, contains: *attackVsftpd, attackSmbd, attackPhpcgi, attackIrcd, attackDistccd, attackRmi, noOp*. The first 6 actions model actual attacks towards the system, while the last action represents a *void* attack, used to describe an attacker waiting for a running attack to complete or for the pre-conditions of some attack to become true.

Due to space limitation, we describe only the *attackVsftpd* and the *noOp* actions, being the others similar to the *attackVsftpd* action. Similarly to the response actions, the attack actions are characterized by pre-conditions and by post-conditions. Specifically, Boolean pre-conditions can be used to model multi-stage attacks, where the next stage can be subject to the achievement of some previous step. Each attack action is characterized by a response time, i.e., the time needed for the attack to complete.

attackVsftpd. Exploits the vulnerability OSVBD-73753 to attack the *vsftp* daemon.

- **Pre-Conditions:** $p_{vsftpd} < T_1 \wedge$
  $p_{scan} \geq T2 \wedge \neg softwareUpToDate \wedge irsTimer \geq nextActionTimer$
- **Post-Conditions:** $Prob = 1, p_{vsftpd} = 1$

The pre-conditions illustrate that the attack is executable by the attacker if it is not currently being executed ($p_{vsftpd} < T_1$) and when the portscan has been completed successfully (having the attribute $p_{scan} \geq T2$ at the end of the port-scan means that the IRS was unable to run any action to counter the portscan). Furthermore, in order to successfully exploit the vulnerability, the software must not be updated ($\neg softwareUpToDate$) and finally the attack can be executed only when any eventual previous attack has been completed ($irsTimer \geq nextActionTimer$). When all the pre-conditions are verified and the attack is launched, the system gets compromised with probability 1 and therefore its p_{vsftpd} attribute is set to 1.

noOp. Models an attacker currently unable to run an attack because either (i) no pre-conditions are currently verified for any of the attack actions or (ii) an attack is already running.

- **Pre-Conditions:** $true$
- **Post-Conditions:** $\emptyset$

Being just a *void* action, *noOp* can always be executed by the attacker and it does not have any post-condition because it does not have any effect on the system.

The extended model also requires some modification to all the IRS response actions, in order to make them able (i) to manage the time concept adding to the *timer* attribute the response time needed for their execution and (ii) to be executed even when an attack has not been detected yet.

To this end, among the others, we change the pre-conditions of *backup* and *softwareUpdate*. We only present here the former, but the same considerations apply to the latter.

backup. The purpose of this action is to model a system executing a backup, needed as a pre-condition for executing the *softwareUpdate* action.

- **Pre-Conditions:** $((p_{scan} \geq T1 \vee p_{vsftpd} \geq T1 \vee p_{smbd} \geq T1 \vee p_{phpcgid} \geq T1 \vee p_{ircd} \geq T1 \vee p_{distccd} \geq T1 \vee p_{rmi} \geq T1)$ $\boldsymbol{\vee attP_{vsftpd} \geq T2 \vee attP_{smbd} \geq T2 \vee attP_{phpcgid} \geq T2 \vee attP_{ircd} \geq T2}$ $\boldsymbol{\vee attP_{distccd} \geq T2 \vee attP_{rmi} \geq T2)} \wedge \neg quarantined \wedge active \wedge alerted \wedge logVerb > 1 \wedge backup \wedge \neg softwareUpToDate$
- **Post-Conditions:** $Prob = 1, timer+ = responseTime(backup), backup = true$

The bold symbols represent the newly added preconditions. Thus, in the extended model the action is executable either when any of the currently detected attack attributes are at least equal to $T1$ or when there is any attack belief greater than or equal to $T2$. The post-condition increments the timer with the time needed to perform the backup, as defined in Table 1, and sets the *backup* attribute to true.

2.3.3 Joint Reward Function. A *joint reward function* $R_k = (R_{k,irs}, R_{k,attacker})$ is used to model the reward of the agents in the stochastic game, where $R_{k,irs}$ represents the reward achieved by the IRS and $R_{k,attacker}$ represents the reward achieved by the attacker, both at discrete time step k. The IRS reward is computed with the same reward function described in Section 2.2.2, while the reward of the attacker is evaluated according to the Common Vulnerability Scoring System (CVSS) [33] using the CIA triad as follows:

$$R_{k,attacker} = w_{sc} Score_C + w_{si} Score_I + w_{sa} Score_A \tag{3}$$

where $w_{sc}, w_{si}, w_{sa} \in [0, 1]$, $w_{sc} + w_{si} + w_{sa} = 1$ are custom weights and $Score_C, Score_I, Score_A \in \{0, 0.5, 1\}$ are respectively the Confidentiality, Integrity and Availability scores related to the attack action. A score equal to 0 means that the attack does not have any impact on the system; 0.5 means that the attack action has a partial impact on the system (e.g., considerable informational disclosure, modification of some system files and reduced information availability); 1 means that the attack can completely compromise the target system, by achieving either a total information disclosure or the ability to modify any file or the total unavailability of the system. A complete discussion on the evaluation of the CIA triad is reported in [33].

For the purpose of this paper, we attribute the rewards 0, 0.5, 1 respectively to the *noOp, portScanAttack, attackVsftpd* actions. The stochastic game solver, based on a multi-agent version of the VI algorithm, is set to maximize the disjunct reward. As a result, we simulate two selfish systems where each one does not know the internals of the other. An alternative would be to simulate two selfish systems that exactly know the counterpart using a zero-sum stochastic game, where $R_{k,irs} = -R_{k,attacker}$. In the latter case, maximizing a reward of a system means minimizing the reward of the other.

3 PERFORMANCE EVALUATION

VI is one of the mostly used algorithm to plan an optimal policy for Single-Agent and Multi-Agent MDPs. It produces successive approximations of the optimal value function until the expected objective value is stable for all the MDP states. Unfortunately, even if each iteration can be performed in $O(|A||S|^2)$ steps [26], the number of states composing the MDP grows exponentially with the number of the defined attributes. The BURLAP library provides an implementation for both Single-Agent and Multi-Agent VI, as well as an implementation of a sub-optimal rollout-based Monte-Carlo

planning algorithm named UCT [29]. However, since all the provided implementations are single-threaded, we extended the library by adding a multi-threaded implementation of the Single-Agent VI algorithm.

In this section we compare the performance, intended as planning time and reward gap in comparison to the optimal case, of the following algorithms: (i) single-threaded, Single-Agent VI; (ii) multi-threaded, Single-Agent VI (in the following, Parallel-VI); (iii) single-threaded UCT; (iv) single-threaded Multi-Agent VI. We show that, while obviously suffering the exponential state growth like the single-threaded implementation, Parallel-VI is able to scale almost linearly with the number of available cores. For systems where a small reward loss is acceptable, instead, the UCT algorithm provides the best performances, improving the planning time by more than 3 orders of magnitude. Finally, the single-threaded Multi-Agent VI algorithm is the one requiring the longest planning time.

The policy planners have been applied on a system characterized by up to 1000 Boolean state attributes and up to 1000 response actions. Each action is bound to one attribute and it changes its Boolean value when executed, in order to generate the full state space. The termination condition is based on an additional *termination* attribute that can be set to *true* by any action with probability 1/10. The reward function assigns the reward -1 to the actions with an even index and -2 to the actions with an odd index. All the tests have been executed on a single compute node of the Shadow supercomputer at Mississippi State University, characterized by 20 cores and 512 GB of RAM. Only a single core has been used for the single-threaded VI and for UCT, while up to 10 threads have been run for Parallel-VI.

The Multi-Agent game inherits the same structure of the Single-Agent case, but with the following changes: (i) the *termination* attribute is replaced by an integer attack counter and all the response actions are capable of decreasing the counter of a single unit with probability 0.1; (ii) a second agent modeling the attacker has been added to the system. This is capable of executing two actions: *noOp* and *attack*. The former does not provide the agent with any reward, while the latter provides the maximum reward. When the *attack* action is successful (with probability 0.05) it increases the attack counter by a single unit. The game ends when the IRS agent succeeds in zeroing the attack counter.

Figure 1 compares the planning time of all the aforementioned planning algorithms. Specifically, all the VI-based algorithms have been configured with $\gamma = 0.9$, while the UCT algorithm has been configured to perform 10, 20 or 30 roll-outs and with a look-ahead of 10 steps. Results highlight that UCT is able to scale linearly with the number of states, while the planning time of both VI and Parallel-VI grows exponentially, as well as the Multi-Agent VI. Figure 2 shows that the speedup obtained by Parallel-VI is almost linear according to the number of threads. We used only half of the cores provided by the compute node, in order to focus on the algorithm speedup avoiding architectural bottlenecks.

Figure 3 compares the VI and UCT rewards in the Single-Agent case. The rewards provided by the optimal planner VI are used as a baseline to compare the rewards provided by UCT. As expected, the average reward obtained by VI is close to -10, specifically -10.07 because it always chooses the response actions characterized with the highest reward. By contrast, the UCT algorithm with 30 roll-outs produces an average reward of -10.86.

The memory usage is exponential in the number of attributes in the case of VI, reaching a peak of 71GB with 50 attributes and 50 actions. The trend is instead linear for all the UCT configurations, resulting in a maximum usage of 5GB for UCT-30 with 1000 attributes and 1000 actions.

We observe that the planning time of all the planners strictly depends on the cardinality of the state space, which in turn depends on the number of the attributes. Therefore, regardless of the chosen planner, it is important to limit their number as much as possible in order to reduce the

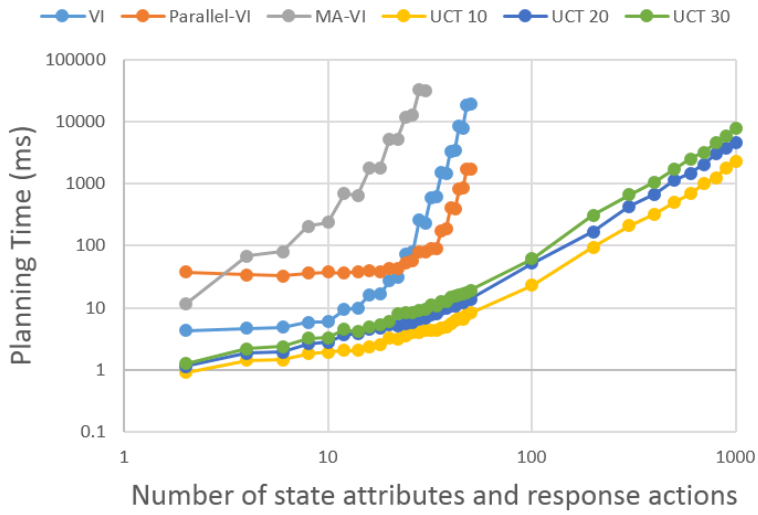

Fig. 1. Planning Time Comparison

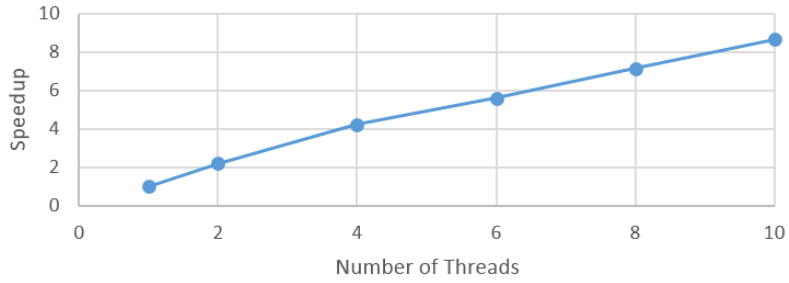

Fig. 2. Parallel-VI speedup

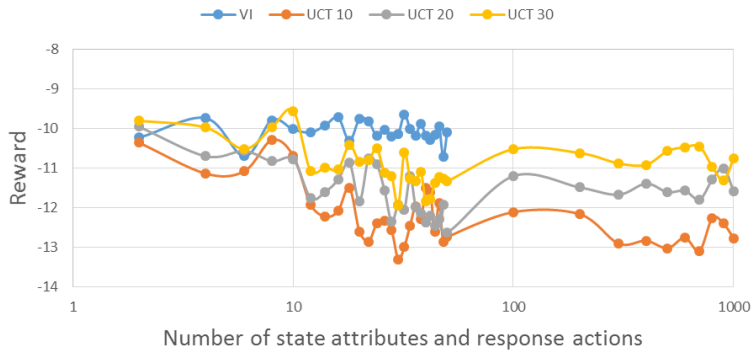

Fig. 3. Rewards comparison

Input: abnormalAttributes
Output: attributes, actions
1: $Set\ attributes = \emptyset, Set\ actions = \emptyset$
2: $attributes \leftarrow abnormalAttributes$
3: **for** $Attribute\ att \in attributes$ **do**
4: $Set\ newActions = actLookup(att)$
5: $actions \leftarrow actions \cup newActions$
6: **for** $Action\ act \in newActions$ **do**
7: $Set\ newAttributes \leftarrow attLookup(act)$
8: $attributes \leftarrow attributes \cup newAttributes$
9: **end for**
10: **end for**
11: **return** $attributes, actions$

Fig. 4. Attributes and Actions Dynamic Selection Algorithm

planning time. We also observe that not necessarily the entire set of attributes and the entire set of actions must be included in the MDP problem: while countering any threat, we only need to consider the attributes and the actions that, directly or indirectly, help in facing the threat. The rationale is that specific threats are supposed to impact only specific system attributes and not all of them.

To this end, we designed and implemented in the proposed IRS a dynamic attributes and actions selection engine, which is in charge of instantiating the MDP problem with the minimum number of attributes and actions.

Figure 4 describes the algorithm used to generate the minimum set of attributes and actions. It takes in input the set of abnormal attributes *abnormalAttributes*, that is, the set of attributes whose values differ from the values of the attributes belonging to the final states and returns the minimal sets of attributes and actions. For each abnormal attribute, the algorithm retrieves the set of actions that refer to it in its pre- and post-conditions (line 4) and adds it to the output set of actions. For each newly discovered action, it then retrieves the list of attributes used as pre-conditions or post-conditions for the considered action (line 7) and adds them to the output set of attributes. The proposed algorithm can be executed in $O(|Att| \times |A|)$, being $|Att|$ the number of defined attributes and it introduces a negligible overhead in the overall planning time.

The proposed attributes and actions selection engine breaks the connection between the number of attributes required to describe the system and the planning time, which is now only dependent on the maximum cardinality of attributes impacted by a threat. As a consequence, the proposed IRS is able to compute optimal response policies in less than 2 seconds using the Parallel-VI algorithm for threats impacting up to 50 system attributes and that require up to 50 different response actions to be countered, thus we believe it possible to use it at run-time to protect large systems. Whether the attack should impact more attributes, the UCT algorithm provides a planning time which makes it feasible to run it at run-time, with an eventual reward degradation. In the Multi-Agent case, the proposed IRS is able to deal at run-time with threats impacting up to 20 attributes.

4 EXPERIMENTAL RESULTS

In this section we describe the experiments we carried out to validate the proposed approach and to demonstrate that long-term planning outperforms short-term planning.

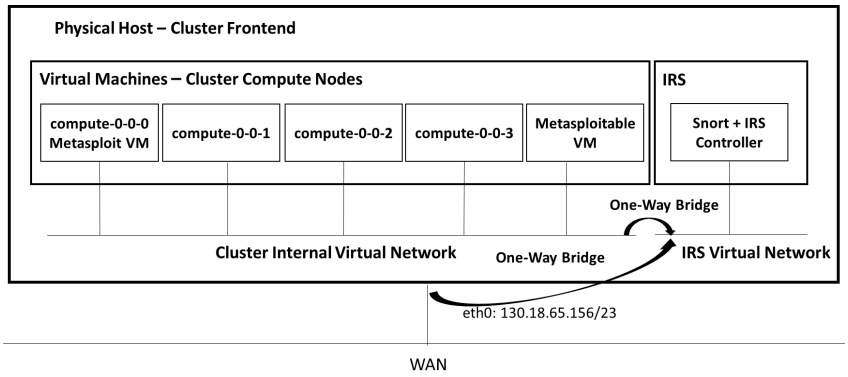

Fig. 5. Testbed Architecture

We set up a system composed by: an HPC cluster based on Rocks[3], a Snort IDS [39] and the IRS controller described in Section 2, as shown in Figure 5. The testbed is composed of a single physical machine which hosts two separate virtual networks, namely: (i) Cluster Internal Virtual Network and (ii) IRS Virtual Network. The former is attached to all the compute nodes of the Rocks cluster, while the latter is attached to Snort and to the IRS. The two networks are constituted by two different layer-2 segments and, while the first is also bridged to physical WAN interface `eth0`, the IRS Virtual Network is instead isolated from external traffic. We run on the physical host two instances of the tool `daemonlogger`, which is used to mirror the traffic from the Cluster Internal Virtual Network and from `eth0` to the IRS Virtual Network. Traffic mirroring is accomplished at layer 2 and it is one-way, that is, frames captured on `eth0` or on the Cluster Internal Virtual Network are forwarded to the IRS Virtual Network but not vice-versa.

We simulate a scenario in which an attacker already compromised one compute node in the cluster and is trying to exploit OSVDB[4] and CVE[5] vulnerabilities exposed by another compute node, namely: OSVBD-73753, CVE-2007-2447, CVE-2012-1823, CVE-2010-2075, CVE-2004-2687, CVE-2011-3556. To this end, we set up 5 compute nodes: `compute-0-0-1` to `compute-0-0-3` are healthy VMs; `compute-0-0-0` is the VM compromised by the attacker and finally `metasploitable` is a vulnerable, but not yet compromised compute node, target of the attacks. The compromised compute node is a VM in which we installed the Metasploit software [32]. We use this VM to scan the internal network and to launch attacks towards the vulnerable VM `metasploitable`.

4.1 Single-Agent Policies Evaluation

In the following we compare the policies generated by the Single-Agent model-based IRS using both the VI and the UCT algorithms. Specifically, we configure the VI algorithm to run with two different settings: $\gamma = 0.9$ and $\gamma = 0$ (in the following, respectively VI-0.9 and VI-0); the UCT algorithm is instead configured with a lookahead of 30 steps (in the following, UCT-30). VI-0.9 fully exploits the MDP features, by planning response policies considering both immediate and future rewards. VI-0, by contrast, is only able to select the best short-term action as in [9] and [37]. Therefore, we compare it to VI-0.9 to show the performance improvement achievable with long-term planning against short-term planning. Finally, UCT-30 is a sub-optimal planner configured to consider long-term rewards.

[3] http://www.rocksclusters.org
[4] https://blog.osvdb.org/
[5] https://cve.mitre.org/

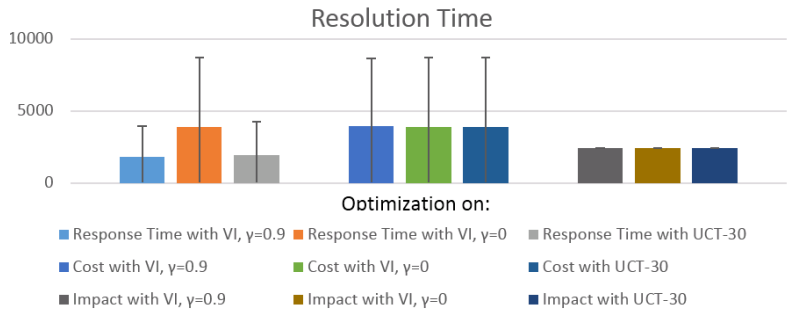

Fig. 6. Exploit Resolution Time Comparison

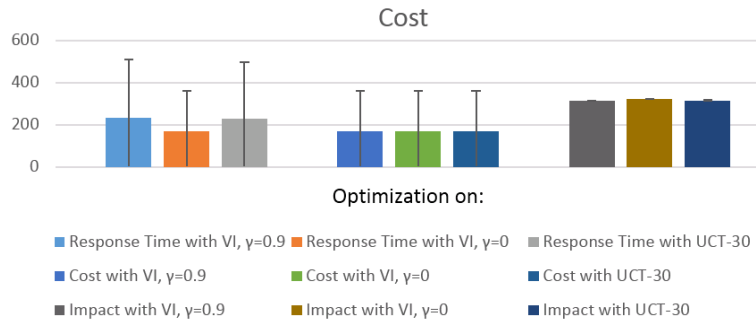

Fig. 7. Exploit Cost Comparison

We ran two different sets of experiments: (i) a vulnerability exploit and (ii) a combination of portscan attack and vulnerability exploit. All the experiments have been repeated 10000 times and the output of each single experiment is a response policy applicable on a real system, characterized by its own resolution time, cost and impact, given by the sum of the respective attributes of the component response actions. For each metric, we compare the average values and the 95% confidence intervals. The reward function has been configured to optimize the response policy exclusively either on response time ($w_r = 1, w_c = 0, w_i = 0$), or cost ($w_r = 0, w_c = 1, w_i = 0$), or impact ($w_r = 0, w_c = 0, w_i = 1$).

4.1.1 Vulnerability Attack. Figure 6 compares the average resolution times and confidence intervals of the planned response policies. The first set of columns compares the resolution times obtained by VI-0.9, VI-0 and UCT-30 while optimizing on response time; the second set of columns compares the resolution times while optimizing on cost; finally, the third set of columns compares the resolution time while optimizing on impact. The lowest resolution time has been obtained by VI-0.9 with optimization on response time, while the worst result has been obtained with VI-0, with a 115% overhead; the UCT-30 overhead is instead 6%. The following is the most frequently planned response policy with VI-0.9 and optimization on response time: *generateAlert, increaseLogVerb, activateFirewall, increaseLogVerb, increaseLogVerb, increaseLogVerb, increaseLogVerb, systemReboot, backup, softwareUpdate, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb*. It is interesting to note how the planner tries to minimize the average resolution time in the planned

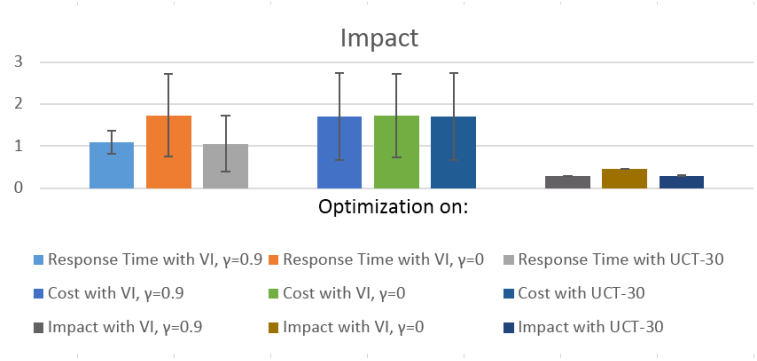

Fig. 8. Exploit Impact Comparison

the policy: the latter can be indeed easily split in four phases: (i) preparation, (ii) first defense attempt, (iii) second defense attempt, (iv) conclusion. The first response attempt (*systemReboot*) has a response time equal to 60 seconds and it is able to face the attack with probability 0.3. Therefore, even if most of times rebooting the machine would not be a successful resolution of the attack, it offers a very good alternative that can be used to lower the average resolution time which, in case of a backup and software update would always be equal to 4200 sec.

Figure 7 compares the costs incurred by the IRS to face the vulnerability exploit. In this case the performance obtained by the three planning algorithms with optimization on cost are perfectly comparable. This happened because most of time the locally optimal policy was also the global optimal policy. The following is the most frequently planned response policy with VI-0.9 and optimization on cost: *generateAlert, increaseLogVerb, increaseLogVerb, increaseLogVerb, systemReboot, activateFirewall, increaseLogVerb, increaseLogVerb, quarantineSystem, backup, manualResolution, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb.* Like in the optimization on response time case, here the IRS tries to lower the average cost by preferring the *systemReboot* action over the *quarantineSystem, backup, manualResolution.*

In Figure 8 a comparison of the impacts produced by the computed policy is shown. The lowest impact on the real system has been obtained with VI-0.9 and optimization on impact; VI-0 introduced an impact overhead of 50%, while UCT-30 did not introduce any impact overhead. The following is the only planned response policy with VI-0.9 and optimization on impact: *generateAlert, increaseLogVerb, activateFirewall, increaseLogVerb, backup, softwareUpdate, decreaseLogVerb, decreaseLogVerb.* Here the *systemReboot* action is not taken into consideration because it has a high impact on the system. Instead, *backup* and *softwareUpdate* are always choosen.

In conclusion, VI-0.9 always outperformed both VI-0 and UCT-30 and the latter outperformed VI-0 in all the experiments.

4.1.2 Simultaneous Portscan And Vulnerability Attack. Figure 9 compares the resolution times obtained with the different planning algorithms and different optimization strategies. The lowest resolution time has been obtained with VI-0.9 configured to optimize on response time. UCT-30 introduced a 6% overhead, outperforming VI-0 which introduced a 116% overhead. The following is the most frequently planned response policy with VI-0.9 and optimization on response time: *generateAlert, increaseLogVerb, activateFirewall, increaseLogVerb, blockSrcIP, increaseLogVerb, increaseLogVerb, increaseLogVerb, systemReboot, backup, softwareUpdate, unblockSrcip, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb.* To counter the portscan, the *blockSrcIP* main defense

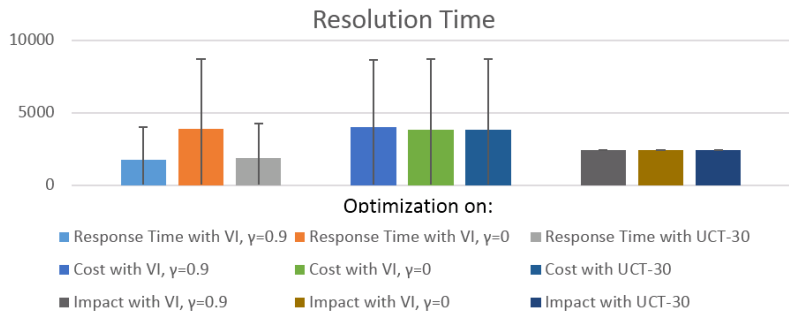

Fig. 9. Combined Attack Resolution Time Comparison

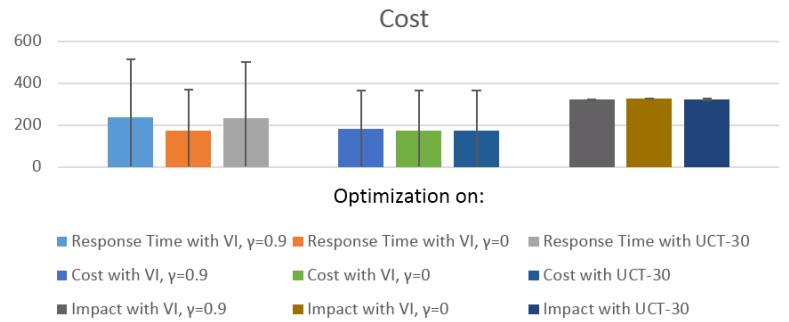

Fig. 10. Combined Attack Cost Comparison

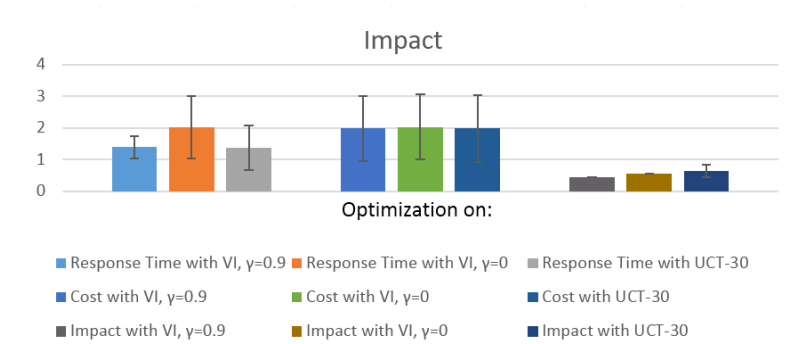

Fig. 11. Combined Attack Impact Comparison

action has been taken because it is the one with the lowest response time; to counter the vulnerability exploit, instead, a *systemReboot* followed by *backup* and *softwareUpdate* have been planned, where the *systemReboot* has been used to lower the average resolution time.

The execution costs obtained by the three planning algorithms with optimization on cost and shown in Figure 10 are perfectly comparable, therefore evidencing that most of times the locally optimal policy is also the best globally optimal one. The following is the most frequently planned

response policy with VI-0.9 and optimization on cost: *generateAlert, increaseLogVerb, increaseLogVerb, increaseLogVerb, increaseLogVerb, systemReboot, increaseLogVerb, quarantineSystem, backup, manualResolution, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb.* Unlike the previous case, here `blockSrcIP` has not been added to the policy because the chosen *manualResolution* response action, which is the one with lowest cost to face the vulnerability exploit, is also able to deal with the portscan attack.

Finally, Figure 11 compares the impacts produced by the computed policy. The lowest impact has been obtained by VI-0.9 with optimization on impact, while VI-0 introduced an impact overhead of 22% and UCT-30 an impact overhead of 39%. The following is the most frequently planned response policy with VI-0.9 and optimization on impact: *generateAlert, increaseLogVerb, activateFirewall, increaseLogVerb, increaseLogVerb, redirectToHoneypot, backup, softwareUpdate, disableHoneypot, decreaseLogVerb, decreaseLogVerb, decreaseLogVerb.* In conclusion, VI-0.9 always outperformed both VI-0 and UCT-30.

4.2 Multi-Agent Policies Evaluation

In this section we comment the defense policies generated executing the IRS configured with the multi-agent model. Specifically, we model an attack consisting of a portscan followed by an exploit of the vulnerability OSVBD-73753 of the *vsftp* daemon. We compare the evolution of the multi-agent system with the evolution of the single-agent one. In the single-agent case, the behavior of the attacker is not modeled. Therefore, when the attacker executes the portscan, the IRS reacts by planning a policy similar to *increaseLogVerb, generateAlert, activateFirewall, increaseLogVerb, blockSrcIP, unblockSrcip, decreaseLogVerb, decreaseLogVerb.* Depending on the time needed by the attacker to complete the portscan and on the time needed by the IRS to counter the attack, it could happen that the attacker manages to complete the portscan before the IRS could complete the deployment of the response actions. Being able to complete the portscan allows the attacker to discover the vulnerability and, as a consequence, it immediately launches the exploit on the *vsftp* daemon. At this point the IRS reacts again by executing one of the policies described in Section 4.1.1. We observe that: (i) the disjoint execution of the policies brings an overhead in terms of repeated actions and that (ii) the execution of the second policy happens when the system has already been compromised. Since in the multiple-agent case the behavior of the attacker is instead modeled, when the attacker launches the portscan, the IRS can use the information of the attacker's attack belief to guess what is the next attack and to proactively deploy a response policy. In the following we show two games between the attacker and the IRS, represented as a list of stages of the form *(attackerAction / irsAction).* In the first game, the portscan attack is assumed to require 60 seconds, enough for the IRS to deploy all the needed countermeasures. In the second game, instead, the portscan attack is assumed to require only 5 seconds. Due to limited space, we removed all the stages of the game regarding exclusively log verbosity increase or decrease.

4.2.1 Game 1: Full Prevention.

(1) portScanAttack / generateAlert
(2) noOp / activateFirewall
(3) noOp / blockSrcIP
(4) noOp / unblockSrcIP
(5) noOp / backup
(6) noOp / softwareUpdate

We observe that the attacker, after having launched a 60 seconds portscan attack, waits for it to finish with a series of noOp. During the waiting, the IRS is able not only to generate a policy to protect against the portscan (with the *blockSrcIp*), but also to proactively address the prospective vulnerability exploit. This prevents the attacker from being able to launch further attacks.

4.2.2 *Game 2: Reaction and Prevention.*

(1) portScanAttack / generateAlert
(2) noOp / activateFirewall
(3) attackVsftpd / increaseLogVerbosity
(4) noOp / blockSrcIP
(5) noOp / quarantineSystem
(6) noOp / backup
(7) noOp / manualResolution
(8) noOp / unblockSrcIP
(9) noOp / softwareUpdate

In this case the attacker manages to complete the portscan before its IP gets blocked by the firewall. The *blockSrcIP* does not have effect on the attack because it is based on a reverse shell and therefore the firewall rule just added does not work. Afterwards, the IRS planned to counter the attack with a manual resolution on a quarantined system, after having executed a backup. This step solves the ongoing attack, but leaves the system vulnerable to other threats because the software has not been updated yet. Therefore, as a last step, the IRS plans to update the software.

5 RELATED WORKS

Autonomic systems and, specifically, self-protecting systems, is an already established research field [19, 28]. The research produced so far on self-protection is mostly focused on intrusion detection rather than protection [45]. However, the increasing amount cyber attacks [2] makes it infeasible to manually handle all the generated alerts. IRSs try to address this problem by selecting the appropriate responses to the detected attacks.

Existing works on dynamic IRS can be classified according to the following dimensions: (i) multi-objective planning, that is, the ability of an IRS to select the optimal response action (or response plan) according to a custom set of weighted criteria; (ii) IDS uncertainty, that is, the ability of the IRS to deal with stochastic IDS alerts; (iii) long-term policies, that is, the ability of the IRS of planning policies that span over a long period of time rather than just selecting the immediate optimal response; (iv) system model, that is, whether the IRS is based on a model of the system, which can be used to describe the system dynamics due to an ongoing attack or due to the application of a response action (plan); (v) attacker model, that is, whether the IRS is based on a model describing the attacker behaviour and the potential graph of achievable targets.

In [44] the authors introduce a network model, consisting of resources, system users, network topology and firewall rules. The model is then used to specify direct and indirect dependencies among the resources and between the users and the resources, in order to be able to predict the impact of a service unavailability on dependent services and on dependent users. Such a model is used by the IRS to choose the response action able to avert a certain threat in order to minimize the overall impact on the system and, ultimately, on the users.

ADEPTS [17] focuses on attack containment, that is, on restricting the effect of the intrusion to a subset of the services. The presented approach maximizes the availability of the overall system at the expenses of the features compromised by the attack, which are isolated from the rest of the system. Unlike [44], in which the proposed model is system-centric, in this work the authors propose an attack-centric model, based on intrusion graphs. The latter support OR, AND and QUORUM nodes and can be used to easily represent attack propagation and escalation. The Compromised Confidence Index (CCI) is used to compute the confidence of the detected alert and, by traversing the graph, the confidence of a particular system breach. The response action is then selected from a response repository by evaluating the effectiveness and the potential disruptiveness of all the available responses.

In [42] the authors propose an IRS that takes into consideration the stochastic nature of the detections made by the IDS and the response action is only triggered if the confidence level of the detected attack is greater than a specified threshold. In [16] an optimal response selection is proposed based on financial cost, reputation loss and processing resource. A modified version of the classical genetic algorithm is used to represent the association between each response action with the system resources affected by the execution of the action.

A long-term response planning is presented in [36]. The work uses a Hierarchical Task Network (HTN) to model the IRS goal, the high-level response actions, and the mechanisms to enable response actions. This work uses a fixed set of goals and statically maps each goal to a sequence of high-level response actions.

A Partially Observable MDP (POMDP) with a single-objective reward function is used in [46] to model a IRS able to plan optimal response policies. Since the POMDP is subject to an exponential growth of the states according to the number of the considered attributes, the authors propose a hierarchical decomposition to reduce the computational complexity.

The authors of [35] use a Bayesian Direct Acyclic Graph (DAG) [25] to model attacker behavior. The DAG nodes describe system assets and their dependencies, while edges represent possible exploitation paths. Responses are evaluated according to the Confidentiality Integrity Availability (CIA) triad preferring confidentiality and integrity over availability.

In [47] the authors propose a game-theoretic model named RRE, based on a non-zero sum stochastic game. The core of the work is represented by Attack Response Trees (ART). A leaf node of an ART represents a binary system attribute, which is set to 1 if an IDS alert that include such an attribute is triggered or to 0 otherwise. Binary attributes are then combined using `AND` and `OR` logical ports, in order to define a path towards the impairment of the system's functionalities and, ultimately, towards the global system when the impairment reaches the root node. Each node of the tree, with the exception of the leaves, can be labelled with a *response tag*. The latter represents a response action that, when executed, is able to set to 0 all the attributes specified by the leaves in the corresponding subtree. In the same way as our approach, the ART model is not built on the basis of the attack itself, but on the consequences that the attack has on the system. RRE includes a component which is in charge of computing the security level of the system based on a set of `if-then` rules manually defined by the system administrator, according to his personal system knowledge. The computed security level is represented as a string such as `low, medium, high`.

All the reviewed works, with the exception of [16] and [36] make use of either a system or an attacker model to compute the optimal response action. However [16] and [47] are the only works considering a multi-objective optimization, which is fundamental for a fine-tuning of the produced response plans. [35, 36, 47] are the only works considering a long-term planning. However, the first only introduces static long-term plan templates, while the second only produces the immediate optimal response action evaluated with infinite look-ahead. [47], instead, proposes a long-term response plan based on the evaluation of a stochastic game between the attacker and the IRS. The authors deal with the exponential growth of the state space using approximation techniques, but they not provide hints about the gap that the approximated policies have with respect to the optimal policies.

This work aims at providing the entire set of features. Specifically, we introduce a reward function based on the Simple Additive Weighting (SAW) technique [20] to support multi-objective planning; we handle IDS uncertainty by modeling attack probabilities as state attributes; we use the MDP framework to produce optimal stochastic long-term policies; we provide both the system and the attacker model. The former is statically described with state attributes and dynamically described with the state transitions; the latter is described as a multi-agent stochastic game.

6 CONCLUSIONS AND FUTURE WORKS

In this work we presented an IRS, based on the MDP framework, that supports multi-objective long-term planning. The proposed approach models both the attacker and the defended system behavior and takes into accounts the uncertainty of IDS detections. We presented the MDP framework as a model for building reactive and proactive IRSs. We used a Single-Agent MDP to model the behavior of a reactive IRS applied to a system subject to several attacks and we used a competitive Multiple-Agent MDP to model a game between a proactive IRS and an attacker. Since the state space of the models grows exponentially according to the number of the attributes used to describe the protected system, we introduced a dynamic attributes and actions selection algorithm, which is able to instantiate the minimal MDP problem given the currently ongoing threat. The performance assessment showed that the proposed IRS is able to optimally plan response policies at run-time with threats affecting up to 50 system attributes and requiring up to 50 different response actions to be countered with the proposed parallel version of the VI algorithm. Should the threat involve more attributes or actions, the IRS is anyway able to solve the MDP and to drive the system towards a protected state in a sub-optimal way. Finally, a thorough effectiveness validation showed that long-term policies always outperform short-term ones and that stochastic games can be effectively used to proactively protect a system.

As a future work we plan to establish a feedback loop between the controller and the managed system, in order to let the actionâĂŹs post-conditions probabilities evolve according to the real system evolution. Furthermore, we plan to consider non-deterministic MDPs [14] in order to produce a set of near-optimal decision policies from which the system administrators could pick the best one according to his/her personal knowledge. Such a semi-automatic behavior could be particularly useful in industrial control systems (SCADA), which are used extensively in critical infrastructures.

ACKNOWLEDGMENTS

This work is partially supported by the Pacific Northwest National Laboratory, under U.S. Department of Energy Contract DE-AC05-76RL01830.

The authors would like to thank Dr. Qian Chen of Savannah State University for her comments that improved the overall quality of the manuscript.

REFERENCES

[1] Sherif Abdelwahed, Jia Bai, Rong Su, and Nagarajan Kandasamy. 2009. On the application of predictive control techniques for adaptive performance management of computing systems. *Network and Service Management, IEEE Transactions on* 6, 4 (2009), 212–225.

[2] Akamai. 2015. Akamai's State of the Internet: Q3 2015 Report. https://www.stateoftheinternet.com/resources-cloud-security-2015-q3-web-security-report.html. (2015).

[3] RE Bellman. 1957. Dynamic Programming. Princeton, NJ: Princeton UniversityPress. *BellmanDynamic Programming1957* (1957).

[4] Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning* 2, 1 (2009), 1–127.

[5] Monowar H Bhuyan, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. 2014. Network anomaly detection: methods, systems and tools. *Communications Surveys & Tutorials, IEEE* 16, 1 (2014), 303–336.

[6] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*. Morgan Kaufmann Publishers Inc., 195–210.

[7] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 38, 2 (2008), 156–172.

[8] Valeria Cardellini, Emiliano Casalicchio, Vincenzo Grassi, Stefano Iannucci, Francesco Lo Presti, and Raffaela Mirandola. 2012. Moses: A framework for qos driven runtime adaptation of service-oriented systems. *IEEE Transactions on Software Engineering* 38, 5 (2012), 1138–1159.

[9] Qian Chen, Sherif Abdelwahed, and Abdelkarim Erradi. 2014. A Model-Based Validated Autonomic Approach to Self-Protect Computing Systems. *Internet of Things Journal, IEEE* 1, 5 (2014), 446–460.
[10] Yulia Cherdantseva and Jeremy Hilton. 2013. A reference model of information assurance & security. In *Availability, reliability and security (ares), 2013 eighth international conference on*. IEEE, 546–555.
[11] Chun-Jen Chung, Pankaj Khatkar, Tianyi Xing, Jeongkeun Lee, and Dijiang Huang. 2013. NICE: Network intrusion detection and countermeasure selection in virtual network systems. *Dependable and Secure Computing, IEEE Transactions on* 10, 4 (2013), 198–211.
[12] Carlos Diuk, Andre Cohen, and Michael L Littman. 2008. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 240–247.
[13] Jianbin Fang, Henk Sips, Lilun Zhang, Chuanfu Xu, Yonggang Che, and Ana Lucia Varbanescu. 2014. Test-driving intel xeon phi. In *Proceedings of the 5th ACM/SPEC international conference on Performance engineering*. ACM, 137–148.
[14] Mahdi Milani Fard and Joelle Pineau. 2011. Non-Deterministic Policies in Markovian Decision Processes. *J. Artif. Intell. Res.(JAIR)* 40 (2011), 1–24.
[15] Ahmed Fawaz, Robin Berthier, and William H Sanders. 2016. A Response Cost Model for Advanced Metering Infrastructures. *IEEE Transactions on Smart Grid* 7, 2 (2016), 543–553.
[16] Boutheina A Fessi, Salah Benabdallah, Noureddine Boudriga, and M Hamdi. 2014. A multi-attribute decision model for intrusion response system. *Information Sciences* 270 (2014), 237–254.
[17] Bingrui Foo, Yu-Sung Wu, Yu-Chun Mao, Saurabh Bagchi, and Eugene Spafford. 2005. ADEPTS: adaptive intrusion response using attack graphs in an e-commerce environment. In *Dependable Systems and Networks, 2005. DSN 2005. Proceedings. International Conference on*. IEEE, 508–517.
[18] Mansoureh Ghasemi, Hassan Asgharian, and Ahmad Akbari. 2016. A cost-sensitive automated response system for SIP-based applications. In *Electrical Engineering (ICEE), 2016 24th Iranian Conference on*. IEEE, 1142–1147.
[19] Salim Hariri, Bithika Khargharia, Houping Chen, Jingmei Yang, Yeliang Zhang, Manish Parashar, and Hua Liu. 2006. The autonomic computing paradigm. *Cluster Computing* 9, 1 (2006), 5–17.
[20] C.L. Hwang and K. Yoon. 1981. *Multiple Criteria Decision Making, Lecture Notes in Economics and Mathematical Systems*. Springer.
[21] Stefano Iannucci and Sherif Abdelwahed. 2016. A Probabilistic Approach to Autonomic Security Management. In *Proceedings of the 13th IEEE International Conference on Autonomic Computing (ICAC)*.
[22] Stefano Iannucci and Sherif Abdelwahed. 2016. Towards Autonomic Intrusion Response Systems. (2016).
[23] Stefano Iannucci, Qian Chen, and Sherif Abdelwahed. 2016. High-Performance Intrusion Response Planning on Many-Core Architectures. (2016).
[24] Zakira Inayat, Abdullah Gani, Nor Badrul Anuar, Muhammad Khuram Khan, and Shahid Anwar. 2016. Intrusion response systems: Foundations, design, and challenges. *Journal of Network and Computer Applications* 62 (2016), 53–74.
[25] Finn V Jensen. 1996. *An introduction to Bayesian networks*. Vol. 210. UCL press London.
[26] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* (1996), 237–285.
[27] Michael Kearns, Yishay Mansour, and Andrew Y Ng. 2002. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning* 49, 2-3 (2002), 193–208.
[28] J. O. Kephart and D. M. Chess. 2003. The Vision of Autonomic Computing. *IEEE Computer* 36, 1 (2003), 41–50.
[29] Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*. Springer, 282–293.
[30] Wenke Lee, Wei Fan, Matthew Miller, Salvatore J Stolfo, and Erez Zadok. 2002. Toward cost-sensitive modeling for intrusion detection and response. *Journal of computer security* 10, 1-2 (2002), 5–22.
[31] Lihong Li, Michael L Littman, and L Littman. 2008. Prioritized sweeping converges to the optimal value function. (2008).
[32] Carlos Joshua Marquez. 2010. An Analysis of the IDS Penetration Tool: Metasploit. *The InfoSec Writers Text Library, Dec* 9 (2010).
[33] Peter Mell, Karen Scarfone, and Sasha Romanosky. 2007. A complete guide to the common vulnerability scoring system version 2.0. In *Published by FIRST-Forum of Incident Response and Security Teams*, Vol. 1. 23.
[34] Daniel A Menascé. 2002. QoS issues in web services. *IEEE internet computing* 6, 6 (2002), 72–75.
[35] Erik Miehling, Mohammad Rasouli, and Demosthenis Teneketzis. 2015. Optimal Defense Policies for Partially Observable Spreading Processes on Bayesian Attack Graphs. In *Proceedings of the Second ACM Workshop on Moving Target Defense*. ACM, 67–76.
[36] Chengpo Mu and Yingjiu Li. 2010. An intrusion response decision-making model based on hierarchical task network planning. *Expert systems with applications* 37, 3 (2010), 2465–2472.
[37] Sven Ossenbuhl, Jessica Steinberger, and Harald Baier. 2015. Towards automated incident handling: How to select an appropriate response against a network-based attack?. In *IT Security Incident Management & IT Forensics (IMF), 2015*

Ninth International Conference on. IEEE, 51–67.
[38] Martin L Puterman and Moon Chirl Shin. 1978. Modified policy iteration algorithms for discounted Markov decision problems. *Management Science* 24, 11 (1978), 1127–1137.
[39] Martin Roesch et al. 1999. Snort: Lightweight Intrusion Detection for Networks.. In *LISA*, Vol. 99. 229–238.
[40] Jerome H Saltzer and Michael D Schroeder. 1975. The protection of information in computer systems. *Proc. IEEE* 63, 9 (1975), 1278–1308.
[41] Alireza Shameli-Sendi and Michel Dagenais. 2015. ORCEF: Online response cost evaluation framework for intrusion response system. *Journal of Network and Computer Applications* (2015).
[42] Natalia Stakhanova, Samik Basu, and Johnny Wong. 2007. A Cost-Sensitive Model for Preemptive Intrusion Response Systems.. In *AINA*, Vol. 7. 428–435.
[43] Christopher Roy Strasburg, Natalia Stakhanova, Samik Basu, and Johnny S Wong. 2008. The methodology for evaluating response cost for intrusion response systems. (2008).
[44] Thomas Toth and Christopher Kruegel. 2002. Evaluating the impact of automated intrusion response mechanisms. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*. IEEE, 301–310.
[45] Eric Yuan, Naeem Esfahani, and Sam Malek. 2014. A systematic survey of self-protecting software systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 8, 4 (2014), 17.
[46] Xin Zan, Feng Gao, Jiuqiang Han, Xiaoyong Liu, and Jiaping Zhou. 2010. A hierarchical and factored POMDP based automated intrusion response framework. In *Software Technology and Engineering (ICSTE), 2010 2nd International Conference on*, Vol. 2. IEEE, V2–410.
[47] Saman A Zonouz, Himanshu Khurana, William H Sanders, and Timothy M Yardley. 2014. RRE: A game-theoretic intrusion response and recovery engine. *Parallel and Distributed Systems, IEEE Transactions on* 25, 2 (2014), 395–406.

Received September 2016; revised April 2017; accepted November 2017