On the Tradeoff between Privacy and Utility in Collaborative Intrusion Detection Systems-A Game Theoretical Approach

Richeng Jin
Department of ECE
North Carolina State University
rjin2@ncsu.edu

Xiaofan He
Department of EE
Lamar University
xhe1@lamar.edu

Huaiyu Dai
Department of ECE
North Carolina State University
hdai@ncsu.edu

ABSTRACT

Intrusion Detection Systems (IDSs) are crucial security mechanisms widely deployed for critical network protection. However, conventional IDSs become incompetent due to the rapid growth in network size and the sophistication of large scale attacks. To mitigate this problem, Collaborative IDSs (CIDSs) have been proposed in literature. In CIDSs, a number of IDSs exchange their intrusion alerts and other relevant data so as to achieve better intrusion detection performance. Nevertheless, the required information exchange may result in privacy leakage, especially when these IDSs belong to different self-interested organizations. In order to obtain a quantitative understanding of the fundamental tradeoff between the intrusion detection accuracy and the organizations' privacy, a repeated two-layer single-leader multi-follower game is proposed in this work. Based on our game-theoretic analysis, we are able to derive the expected behaviors of both the attacker and the IDSs and obtain the utility-privacy tradeoff curve. In addition, the existence of Nash equilibrium (NE) is proved and an asynchronous dynamic update algorithm is proposed to compute the optimal collaboration strategies of IDSs. Finally, simulation results are shown to validate the analysis.

KEYWORDS

Collaborative Intrusion Detection Systems; Utility-privacy tradeoff; Game theory

1 INTRODUCTION

Considering that complete prevention of cyber-attacks is extremely difficult, if not impossible, Intrusion Detection Systems (IDSs) have been introduced as an effective second line of defense to minimize the damage caused by these attacks. However, conventional IDSs are not scalable to large networks due to the huge amount of traffic activities. In the meantime, the development of sophisticated large-scale attacks renders the performance of an individual IDS rarely satisfactory. To mitigate this problem, Collaborative IDSs (CIDSs) have been proposed in literature (see, e.g., [7, 9] and the references therein).

HotSoS'17, Hanover, MD, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

A CIDS consists of a group of IDSs that monitor different (and possibly partially overlapped) sub-networks and jointly detect potential attacks. In such collaborative environments, IDSs are expected to exchange their intrusion alerts and other relevant data. Considering that some confidential information may be leaked in such information sharing procedure, some techniques have been proposed to protect the privacy in CIDSs [2, 4, 5, 10–12], at the cost of utility loss (i.e., a detection performance degradation). However, there are two major limitations in these pioneering works. Firstly, it is often difficult to quantify the amount of preserved privacy and utility loss in the existing methods. Secondly, the existing methods do not have the flexibility of properly adjusting the collaboration strategies in response to a given privacy requirement.

In this work, a new privacy-preserving collaboration scheme is proposed for CIDS, which is amenable to the quantitative utility-privacy tradeoff analysis and flexible in meeting the pre-specified privacy requirement. Considering the self-interestedness of the organizations and the intelligence of the attacker (a super attacker which combines the joint efforts of multiple distributed attackers is assumed), a game-theoretic approach is taken in this work. More specifically, the interaction among the attacker and the group of collaborative IDSs is modeled as a two-layer game. The first-layer focuses on the interaction between the attacker and each individual IDS. Particularly, the influence of the privacy requirement on the IDSs' responding strategies and the overall detection performance is explored, and based on which, the corresponding utility-privacy tradeoff curve is obtained. The second-layer focuses on the interaction among IDSs themselves and based on which, the optimal collaboration strategies of the IDSs in different scenarios are derived.

The remainder of this paper is organized as follows. Section 2 formulates the utility-privacy tradeoff problem. The proposed two-layer game model is presented in Section 3. The proposed game is solved in Section 4 and the theoretical analysis is validated through simulations in Section 5. The limitations are discussed in Section 6. Conclusions and future works are presented in Section 7.

2 PROBLEM FORMULATION

In this work, a network that consists of N different self-interested organizations (each having an IDS) is considered, denoted by $\mathcal{N} = \{1, 2, ..., N\}$.

2.1 Attacker Model

A smart attacker that can infer the possible responding strategies of IDSs and choose its optimal attacking strategy accordingly is considered. It is assumed that the attacker can launch attacks on different organizations independently and the objective of the

attacker is to attack the target organizations in the network without being detected.

Furthermore, for the ease of presentation, it is assumed that the attacker will only launch one type of attack (e.g., DDoS) on each organization.[1] If an IDS responds to the attacks (e.g., identify the attacker) successfully, the attacker would not attack the corresponding organization again (in the time frame of interest) using the same type of attack. This assumption makes sense because an attacker usually launches an attack by exploiting the vulnerabilities of the system and once the attack is detected and identified, the vulnerabilities will be fixed and relevant signatures be recorded, which makes the same attack ineffective. If the attacker switches to a new type of attack, it is equivalent to start a new game in our model, which is hence not considered here for simplicity.

The action space of the attacker against each organization is $\mathcal{U}^{\mathcal{A}} = \{u_1^A, u_2^A\}$, where u_1^A corresponds to "attack" and u_2^A corresponds to "no attack". The mixed strategy chosen by the attacker against organization i at time t is denoted by $\boldsymbol{p}_{i,t}^{\boldsymbol{A}} = [p_{i,t}^A(u_1^A), p_{i,t}^A(u_2^A)]$, in which $p_{i,t}^A(u_1^A)$ and $p_{i,t}^A(u_2^A)$ are the probabilities that the attacker takes action u_1^A and u_2^A against organization i at time t, respectively.

2.2 Defender Model

For each IDS i, the objective is to respond to the attacks properly when the corresponding organization is under attack. The action space of IDS i is $\mathcal{U}^I = \{u_1^I, u_2^I\}$, where u_1^I corresponds to "respond" and u_2^I corresponds to "do nothing". The mixed strategy chosen by IDS i is denoted by $\boldsymbol{p}_{i,t}^I = [p_{i,t}^I(u_1^I), p_{i,t}^I(u_2^I)]$, in which $p_{i,t}^I(u_1^I)$ and $p_{i,t}^I(u_2^I)$ are the probabilities that IDS i takes action u_1^I and u_2^I at time t, respectively.

In addition, it is assumed that when the attacker launches an attack on an organization, the IDSs at different organizations will have correlated observations. This is a valid assumption for many realistic scenarios (e.g., the organizations and their IDSs are within the same network). In this paper, it is assumed that when the attacker launches an attack on organization i, each IDS j will observe abnormal traffics with probability q_{ji}, and when the attacker does not launch any attack on organization i, each IDS j will observe normal traffics with probability $q'_{ji} = 1$ for $i = 1, 2, ..., N$. For example, when the attacker launches DDoS attack or spreads certain forms of malware, the unusual traffic flows generated by the attack may be observed by the IDSs in the network with different probabilities, depending on their locations. Without loss of generality, it is assumed that $q_{ji} > 0.5, \forall j, i \in \mathcal{N}$. After observing the traffics, each IDS will independently run the intrusion detection algorithms and set an alert if intrusions are detected. Considering that an intrusion alert does not necessarily indicate intrusions due to the possible false alarm of the IDS, it needs to further decide whether to respond or not based on its own detection results as well as the detection results shared by others.

This work considers the scenario in which the IDSs in the network can collaborate for better detection and response performance against the attacker. The incentive of collaboration lies in the fact that an organization will suffer a potential loss if other organizations were taken down. For example, the malware injected in one organization may spread to other organizations due to the shared network environment. However, sharing the detection results may lead to potential privacy leakage for each IDS. For instance, if IDS i successfully detects an attack on IDS j, it will realize that the attacker may launch the same attack on itself and therefore be better prepared for this type of attack. By knowing the detection results, the attacker can infer the security state (e.g., whether the IDS knows the existence of the attack) of the corresponding IDS and therefore choose a better attacking strategy. Moreover, an intrusion alert usually contains some private information (e.g., IP address, processing time), which may raise big privacy concerns for the IDSs. As a result, the IDSs should also balance their utilities and privacy concerns so as to choose proper collaboration strategies.

Table 1: Payoff matrix of the game for organization i

	Respond	Do nothing
Attack	$(1-2b_i)W_i - C_{a,i}W_i$, $-(1-2b_i)W_i - C_{r,i}W_i$	$W_i - C_{a,i}W_i$, $-W_i$
No attack	0, $-C_{r,i}W_i$	0, 0

2.3 General Settings

It is assumed that each organization i processes W_i security asset, representing the loss of security when IDS i fails to successfully respond to the attacks [1]. In practice, the security assets of the organizations depend on their roles in the network and the data or information they hold. If IDS i fails to respond to the attacker successfully, the attacker gets a payoff W_i and IDS i gets a payoff $-W_i$. Otherwise, the payoffs for the attacker and IDS i are $-W_i$ and W_i, respectively.

Table 1 illustrates the payoff matrix of the attacker/IDS interaction on organization i, in which the first entry and second entry in each cell denote the payoffs of the attacker and the IDS, respectively. In the matrix, $b_i \in [0, 1]$ denotes the possibility of successful response for the IDS i; similar to [1], the cost of attacking and responding are assumed to be proportional to the security asset of organization i, denoted by $C_{a,i}W_i$ and $C_{r,i}W_i$, respectively, in which $C_{a,i}$ and $C_{r,i}$ denote the corresponding cost coefficients. When the attacker chooses to attack and IDS i chooses to respond at the same time, the probability of successful response for IDS i is b_i, which means IDS i will get payoff $W_i - C_{r,i}W_i$ and $-W_i - C_{r,i}W_i$ with probability b_i and $1 - b_i$, respectively. Therefore, the payoff of IDS i in expectation is $-(1-2b_i)W_i - C_{r,i}W_i$. Similarly, the payoff of the attacker is $(1-2b_i)W_i - C_{a,i}W_i$. The payoffs of both the attacker and IDS i in other cases are defined similarly. Note that when the IDS chooses "do nothing", the payoff of the attacker choosing "attack" should be higher than that of choosing "no attack" (otherwise, the attacker has no incentive to attack), which indicates $C_{a,i} < 1$. Similarly, $C_{r,i} < 1$.

[1] When multiple types of attacks are available to the attacker, multiple independent games can be formed, each corresponding to a different type of attack.

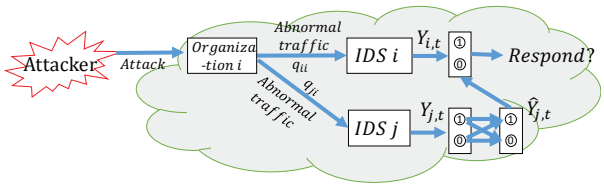

Figure 1: Diagram of the game model.

3 COLLABORATIVE INTRUSION DETECTION GAME MODEL

In this section, the problem is modeled as a repeated two-layer single-leader multi-follower game, in which the attacker acts as the leader and the IDSs act as followers. The first-layer game models the interaction between the attacker and each of the IDSs, respectively, while the second-layer game models the collaborative information sharing among the IDSs themselves. Figure 1 depicts a special case of the game model in which there are only two IDSs. More specifically, the problem is solved in two steps: first of all, the first-layer game between the attacker and each IDS is solved, which determines the optimal payoffs of both the attacker and IDSs as functions of the collaboration strategies of the IDSs. Then, based on the payoff functions from the first-layer game, the IDSs further determine their optimal collaboration strategies given their privacy requirements in the second-layer game.

3.1 The First-layer Leader-follower Game

In the first layer game, it is assumed that the follower plays a myopic best-response strategy to the leader's strategy at each time t [6]. Note that since it is not possible for the IDSs to know the future strategies of the attacker and the future detection results of themselves, the myopic strategy is actually the best strategy that an IDS can take.

3.1.1 The Followers' Problem. Given the attacker's strategy $\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}$ and its own detection result $Y_{i,t}$ at time t (with $Y_{i,t}=1$ and $Y_{i,t}=0$ denoting alert and no alert, respectively), each IDS i first estimates the probability that the attacker actually launches an attack, which is given by

$$Q^i(u_1^A|Y_{i,t}) = \begin{cases} \dfrac{p^A_{i,t}(u_1^A)p(Y_{i,t}=1|u_1^A)}{p(Y_{i,t}=1)}, & \text{for } Y_{i,t}=1, \\ \dfrac{p^A_{i,t}(u_1^A)p(Y_{i,t}=0|u_1^A)}{p(Y_{i,t}=0)}, & \text{otherwise,} \end{cases} \tag{1}$$

where $p(Y_{i,t}=1|u_1^A)$ is the probability that the detection result of IDS i at time t is $Y_{i,t}=1$ given that the attacker launches an attack; $p(Y_{i,t}=0|u_1^A)$ is the probability that the detection result of IDS i at time t is $Y_{i,t}=0$ given that the attacker launches an attack; $p(Y_{i,t}=1)$ and $p(Y_{i,t}=0)$ are the probabilities that the detection results of IDS i at time t are $Y_{i,t}=1$ and $Y_{i,t}=0$, respectively. They are given by

$$p(Y_{i,t}=1|u_1^A) = q_{ii}p_i^d + (1-q_{ii})p_i^{fp}, \tag{2}$$

$$p(Y_{i,t}=0|u_1^A) = 1 - p(Y_{i,t}=1|u_1^A), \tag{3}$$

$$\begin{aligned} p(Y_{i,t}=1) = {} & p^A_{i,t}(u_1^A)[q_{ii}p_i^d + (1-q_{ii})p_i^{fp}] \\ & + p^A_{i,t}(u_2^A)[q'_{ii}p_i^{fp} + (1-q'_{ii})p_i^d], \end{aligned} \tag{4}$$

$$p(Y_{i,t}=0) = 1 - p(Y_{i,t}=1), \tag{5}$$

in which q_{ii} is the probability that IDS i observes an abnormal traffic pattern when the attacker launches attacks on organization i; q'_{ii} is the probability that IDS i observes a normal traffic pattern when the attacker does not launch any attack on organization i; and p_i^d, p_i^{fp} denote the detection rate and the false positive rate of IDS i.

Then, each IDS i finds its optimal strategy by solving the following optimization problem:

$$\boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, Y_{i,t}) = \underset{\boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}}{\operatorname{argmax}} \quad \boldsymbol{U}^{\boldsymbol{I}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}, \boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, Y_{i,t}). \tag{6}$$

The payoff function at IDS i $\boldsymbol{U}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}, \boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, Y_{i,t})$ is given by

$$\begin{aligned} \boldsymbol{U}^{\boldsymbol{I}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}, \boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, Y_{i,t}) = {} & Q^i(u_1^A|Y_{i,t})p^I_{i,t}(u_1^I)[-(1-2b_i)W_i - C_{r,i}W_i] \\ & - Q^i(u_1^A|Y_{i,t})p^I_{i,t}(u_2^I)W_i - [1-Q^i(u_1^A|Y_{i,t})]p^I_{i,t}(u_1^I)C_{r,i}W_i, \end{aligned} \tag{7}$$

where $Q^i(u_1^A|Y_{i,t})p^I_{i,t}(u_1^I)$ is the probability of the case that the attacker launches an attack and IDS i chooses to respond given the detection result $Y_{i,t}$, and $-(1-2b_i)W_i - C_{r,i}W_i$ is the payoff of IDS i in this case; $Q^i(u_1^A|Y_{i,t})p^I_{i,t}(u_2^I)$ is the probability of the case that the attacker launches an attack and IDS i chooses to do nothing given the detection result $Y_{i,t}$, and $-W_i$ is the payoff of IDS i in this case; $[1-Q^i(u_1^A|Y_{i,t})]p^I_{i,t}(u_1^I)$ is the probability of the case that the attacker does not launch an attack and IDS i chooses to respond given the detection result $Y_{i,t}$, and $-C_{r,i}W_i$ is the payoff of IDS i in this case.

3.1.2 The Leader's Problem. As the attacker knows that the followers will choose their strategies to maximize their corresponding payoffs, it will choose the strategy that maximizes its own payoff. However, since the attacker does not know the detection results of IDS i, it has to maximize the expected payoff corresponding to the distribution $p(Y_{i,t})$ which can be obtained by (4) and (5) given its chosen strategy. As a result, the attacker finds its optimal strategy against IDS i by solving the following optimization problem:

$$\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}) = \underset{\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}}{\operatorname{argmax}} \sum_{t=1}^{T_e^i} \boldsymbol{U}^{\boldsymbol{A}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, \boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}})), \tag{8}$$

where T_e^i is the time when IDS i successfully responds to the attacker and $\boldsymbol{U}^{\boldsymbol{A}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, \boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}))$ is given by

$$\begin{aligned} & \boldsymbol{U}^{\boldsymbol{A}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}}, \boldsymbol{p}^{\boldsymbol{I}}_{\boldsymbol{i,t}}(\boldsymbol{p}^{\boldsymbol{A}}_{\boldsymbol{i,t}})) = \\ & \sum_{j\in\{0,1\}} \big[p(Y_{i,t}=j)p^A_{i,t}(u_1^A)p^I_{i,t}(u_2^I|Y_{i,t}=j)(W_i - C_{a,i}W_i) \\ & + p(Y_{i,t}=j)p^A_{i,t}(u_1^A)p^I_{i,t}(u_1^I|Y_{i,t}=j)[(1-2b_i-C_{a,i})W_i]\big], \end{aligned} \tag{9}$$

where $p^A_{i,t}(u_1^A)p^I_{i,t}(u_2^I|Y_{i,t}=j)$ is the probability of the case that the attacker launches an attack and IDS i chooses to do nothing given the detection result $Y_{i,t}=j$ and $W_i - C_{a,i}W_i$ is the payoff of the attacker in this case; $p^A_{i,t}(u_1^A)p^I_{i,t}(u_1^I|Y_{i,t}=j)$ is the probability

of the case that the attacker launches an attack and IDS i chooses to respond given the detection result $Y_{i,t} = j$ and $(1 - 2b_i - C_{a,i})W_i$ is the payoff of the attacker in this case.

3.1.3 Collaborative IDS Case. In the previous subsections, it is assumed that each IDS works independently. In practice, however, the IDSs can share their detection results $Y_{i,t}$'s with others so as to help improve the performance of other IDSs, which will in return enhance the security of the whole network. However, sharing these detection results will lead to the risk of private information leakage (e.g., security state). As a result, each IDS i is assumed to share an obfuscated version of $Y_{i,t}$ with others, denoted by $\hat{Y}_{i,t}$. In this work, it is assumed that each IDS i will misreport its true detection result to other IDSs with probability $p^c_{i,t}$ and the preserved privacy is measured by the entropy introduced by $p^c_{i,t}$ [3], given as follows:

$$H(p^c_{i,t}) = -p^c_{i,t}\log_2(p^c_{i,t}) - (1-p^c_{i,t})\log_2(1-p^c_{i,t}). \tag{10}$$

In this case, each IDS i finds its optimal strategy by solving the following modified optimization problem:

$$\boldsymbol{p}^I_{i,t}(\boldsymbol{p}^A_{i,t}, Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t}) = \underset{\boldsymbol{p}^I_{i,t}}{\text{argmax}} \quad \boldsymbol{U}^I_{i,t}(\boldsymbol{p}^I_{i,t}, \boldsymbol{p}^A_{i,t}, Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t}), \tag{11}$$

where $\hat{\boldsymbol{Y}}_{-i,t}$ denotes the obfuscated detection results shared by other IDSs at time t, and $\boldsymbol{U}^I_{i,t}(\boldsymbol{p}^I_{i,t}, \boldsymbol{p}^A_{i,t}, Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t})$ is given by

$$\begin{aligned}&\boldsymbol{U}^I_{i,t}(\boldsymbol{p}^I_{i,t}, \boldsymbol{p}^A_{i,t}, Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t}) = \\ &Q^i(u^A_1|Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t})[p^I_{i,t}(u^I_1)(2b_i - 1 - C_{r,i})W_i - p^I_{i,t}(u^I_2)W_i] \\ &- [1 - Q^i(u^A_1|Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t})]p^I_{i,t}(u^I_1)C_{r,i}W_i,\end{aligned} \tag{12}$$

where $Q^i(u^A_1|Y_{i,t}, \hat{\boldsymbol{Y}}_{-i,t})$ could be obtained similarly as in the non-collaboration case.

3.2 The Second-layer Game

The second layer game models the interaction among the IDSs themselves. In this game, an action of each IDS i is a probability $p^c_{i,t} \in [c_i, 0.5]$[2] with which the IDS i would send out wrong detection results in order to protect its own privacy, and c_i depends on the privacy policy of each organization. The utility function of each IDS i is given as follows:

$$\begin{aligned}&\boldsymbol{U}^{I,2}_{i,t}(\boldsymbol{p}^c_t) = \sum_{j\neq i}\beta_{i,j}[R^{est}_{i,t}(\boldsymbol{p}^c_t) - R^{est}_{i,t}(\boldsymbol{p}^c_{-j,t}, p^c_{j,t} = 0.5)]\times \\ &[R^{est}_{j,t}(\boldsymbol{p}^c_t) - R^{est}_{j,t}(\boldsymbol{p}^c_{-i,t}, p^c_{i,t} = 0.5)] - \lambda_i P_L(p^c_{i,t}),\end{aligned} \tag{13}$$

where $\boldsymbol{p}^c_t = (p^c_{1,t}, p^c_{2,t}, \cdots, p^c_{N,t})$ is a vector which denotes the misreport probabilities of all the IDSs; $\boldsymbol{p}^c_{-i,t}$ denotes the misreport probabilities of all the IDSs other than IDS i; $R^{est}_{i,t}(\boldsymbol{p}^c_t)$ denotes the estimated payoff of IDS i given $\boldsymbol{p}^c_t$, which will be discussed in Section 4; $R^{est}_{i,t}(\boldsymbol{p}^c_{-j,t}, p^c_{j,t} = 0.5)$ denotes the estimated reward of IDS i when IDS j randomly reports its detection result (i.e., $p^c_{j,t} = 0.5$), and therefore $R^{est}_{i,t}(\boldsymbol{p}^c_t) - R^{est}_{i,t}(\boldsymbol{p}^c_{-j,t}, p^c_{j,t} = 0.5)$ measures IDS i's estimated payoff improvement due to the shared detection result from IDS j; $\beta_{i,j}$ and λ_i are constants that measure the importance of payoff improvement and privacy loss, respectively. The privacy loss $P_L(p^c_{i,t})$ is given by

$$P_L(p^c_{i,t}) = 1 - H(p^c_{i,t}), \tag{14}$$

where $H(p^c_i)$ denotes the entropy introduce by $p^c_{i,t}$. As a result, each IDS i has to solve the following optimization problem:

$$\begin{aligned}\max_{p^c_{i,t}} \quad & \boldsymbol{U}^{I,2}_{i,t}(\boldsymbol{p}^c_t) \\ \text{s.t.} \quad & c_i \leqslant p^c_{i,t} \leqslant 0.5.\end{aligned} \tag{15}$$

[2]In this work, it is assumed that the misreporting probabilities are common knowledge for all the IDSs. Therefore, it is equivalent for an IDS to misreport with probability $p^c_{i,t}$ or $1 - p^c_{i,t}$.

4 SOLVING THE GAME

Note that the optimal strategies of both the attacker and the IDSs have the same expressions at different time slots. Therefore, the subscript t will be omitted in this section for the ease of presentation. In this work, we focus on the scenario where $p^d_i > 0.5$ and $p^{fp}_i < 0.5$ for all i without loss of generality.

4.1 The First-layer Leader-follower Game

The leader-follower game is often solved by backward induction. First, solve the follower's problem for every possible strategy taken by the leader. The solution consists of the best response strategy of the follower as a function of the leader's strategy. Then, the leader decides its optimal strategy according to the followers' best responses. The obtained solution is often referred to as a Stackelberg-Nash equilibrium (SNE) [8].

4.1.1 Non-collaborative IDS Case. In this case, by performing backward induction, the best response of IDS i can be solved as

$$p^I_i(u^I_1) = \begin{cases}1 & \text{if } Q^i(u^A_1|Y_i) > \frac{C_{r,i}}{2b_i}, \\ \in [0,1] & \text{if } Q^i(u^A_1|Y_i) = \frac{C_{r,i}}{2b_i}, \\ 0 & \text{if } Q^i(u^A_1|Y_i) < \frac{C_{r,i}}{2b_i}.\end{cases}$$

Combing the payoff function of the attacker, the SNE of the attacker and IDS i can be obtained as follows:

$$\begin{cases}p^A_{i,*}(u^A_1) = \frac{C_{r,i}p(Y_i=1|u^A_2)}{(2b_i - C_{r,i})p(Y_i=1|u^A_1) + C_{r,i}p(Y_i=1|u^A_2)}, \\ p^I_{i,*}(u^I_1) = 0.\end{cases}$$

Remark 1. *The SNE obtained above is a weak equilibrium since when* $p^A_i(u^A_1) = \frac{C_{r,i}p(Y_i=1|u^A_2)}{(2b_i - C_{r,i})p(Y_i=1|u^A_1) + C_{r,i}p(Y_i=1|u^A_2)}$, *for any* $p^I_i(u^I_1) \in [0,1]$, *IDS i will receive the same payoff. To push IDS i to choose its desired strategy (i.e.,* $p^I_{i,*}(u^I_1) = 0$*), the attacker will set*

$$p^A_i(u^A_1) = \frac{C_{r,i}p(Y_i = 1|u^A_2)}{(2b_i - C_{r,i})p(Y_i = 1|u^A_1) + C_{r,i}p(Y_i = 1|u^A_2)} - \epsilon,$$

where ϵ *is a small positive number. In this case, the corresponding payoff is only slightly less than the desired SNE obtained above when* ϵ *is sufficiently small, which is acceptable for the attacker. For the ease of discussion,* ϵ *is set to be 0 in the following analysis, but the results obtained still hold when* $\epsilon > 0$*, as long as it is sufficiently small.*

Remark 2. *At the SNE obtained above, the optimal strategy of IDS i is to respond with probability* $p^I_{i,*}(u^I_1) = 0$. *This is because the attacker is modeled as the leader in the game and thus can take the advantage*

and choose a strategy to force the IDS not to respond. Nonetheless, since both $p(Y_i = 1|u_2^A)$ and $p(Y_i = 1|u_1^A)$ are functions of p_i^d and p_i^{fp} which measure the detecting capability of IDS i, the existence of IDS renders the attacker to choose a low attacking probability.

The corresponding payoffs of the attacker and IDS i at the above SNE are given as follows:

$$\begin{cases} \boldsymbol{U_{i,*}^A} = \frac{C_{r,i}p(Y_i=1|u_2^A)}{(2b_i-C_{r,i})p(Y_i=1|u_1^A)+C_{r,i}p(Y_i=1|u_2^A)}(1-C_{a,i})W_i, \\ \boldsymbol{U_{i,*}^I} = -\frac{C_{r,i}p(Y_i=1|u_2^A)}{(2b_i-C_{r,i})p(Y_i=1|u_1^A)+C_{r,i}p(Y_i=1|u_2^A)}W_i. \end{cases}$$

4.1.2 Collaborative IDS Case. Again, by performing backward induction, the best response of IDS i can be solved as

$$p_i^I(u_1^I) = \begin{cases} 1 & \text{if } Q^i(u_1^A|Y_i, \hat{\boldsymbol{Y}}_{-\boldsymbol{i}}) > \frac{C_{r,i}}{2b_i}, \\ \in [0,1] & \text{if } Q^i(u_1^A|Y_i, \hat{\boldsymbol{Y}}_{-\boldsymbol{i}}) = \frac{C_{r,i}}{2b_i}, \\ 0 & \text{if } Q^i(u_1^A|Y_i, \hat{\boldsymbol{Y}}_{-\boldsymbol{i}}) < \frac{C_{r,i}}{2b_i}. \end{cases}$$

Combing the payoff function of the attacker, the SNE of the attacker and IDS i can be obtained as follows:

$$\begin{cases} p_{i,*}^{A,c}(u_1^A) = \frac{C_{r,i}p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_2^A)}{(2b_i-C_{r,i})p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_1^A)+C_{r,i}p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_2^A)}. \\ p_{i,*}^{I,c}(u_1^I) = 0. \end{cases}$$

The corresponding payoffs of the attacker and IDS i at the above SNE are given as follows:

$$\begin{cases} \boldsymbol{U_{i,*}^{A,c}} = \frac{C_{r,i}p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_2^A)(1-C_{a,i})W_i}{(2b_i-C_{r,i})p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_1^A)+C_{r,i}p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_2^A)}. \\ \boldsymbol{U_{i,*}^{I,c}} = -\frac{C_{r,i}p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_2^A)W_i}{(2b_i-C_{r,i})p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_1^A)+C_{r,i}p(Y_i=1,\hat{\boldsymbol{Y}}_{-\boldsymbol{i}}=\boldsymbol{1}|u_2^A)}. \end{cases}$$

Note that in this case, $p(Y_i = 1, \hat{\boldsymbol{Y}}_{-\boldsymbol{i}} = \boldsymbol{1}|u_1^A)$ and $p(Y_i = 1, \hat{\boldsymbol{Y}}_{-\boldsymbol{i}} = \boldsymbol{1}|u_2^A)$ are functions of misreporting probabilities $p_j^c, \forall j \in \{1, 2, \cdots, N\} \cap \{j \neq i\}$, and when $p_j^c = 0.5, \forall j$, the optimal strategies of both the attacker and IDS i agree with those of the non-collaborative IDS case, respectively.

Proposition 1. *The collaboration scheme (i.e., IDS j shares $\hat{Y}_j$ with IDS i) will always give a better payoff for IDS i, for $i = 1, 2, \cdots, N$.*

4.2 The Second-layer Game

Given the payoff functions of the IDSs in both non-collaborative and collaborative cases for all possible collaboration strategies, the payoff at SNE is used as the estimate, and hence $R_i^{est}(\boldsymbol{p^c}) = \boldsymbol{U_{i,*}^{I,c}}(\boldsymbol{p^c})$, and then the utility function of IDS i is given by

$$\begin{aligned} \boldsymbol{U_i^{I,2}}(\boldsymbol{p^c}) = \sum_{j\neq i} \beta_{i,j}[\boldsymbol{U_{i,*}^{I,c}}(\boldsymbol{p^c}) - \boldsymbol{U_{i,*}^{I,c}}(\boldsymbol{p_{-j}^c}, p_j^c = 0.5)]\times \\ [\boldsymbol{U_{j,*}^{I,c}}(\boldsymbol{p^c}) - \boldsymbol{U_{j,*}^{I}}(\boldsymbol{p_{-i}^c}, p_i^c = 0.5)] - \lambda_i P_L(p_i^c). \end{aligned} \tag{16}$$

In addition, the action set of IDS i is given by $A_i = \{p_i^c | c_i \leqslant p_i^c \leqslant 0.5\}$. Given the utility function and the action set of all the IDSs, we can prove that the second-layer game admits a pure strategy Nash equilibrium (NE) in certain conditions.

Proposition 2. *The second layer game admits a Nash equilibrium in pure strategy when the following condition holds:*[3]

$$A(i) < B(i,j), \forall i, j \in \{1, 2, \cdots, N\} \cap \{j \neq i\}, \tag{17}$$

where

$$A(i) = \frac{p(Y_i = 0|u_2^A) - p(Y_i = 1|u_2^A)}{p(Y_i = 1|u_1^A) - p(Y_i = 0|u_1^A)}, \tag{18}$$

$$B(i,j) = \frac{(2b_j - C_{r,j})p(Y_j = 1|u_1^A)}{C_{r,j}p(Y_j = 1|u_2^A)} \prod_{k\neq i,j} \frac{p(\hat{Y}_k = 1|u_1^A)}{p(\hat{Y}_k = 1|u_2^A)}. \tag{19}$$

Note that the concavity of the utility function makes problem (15) a convex optimization problem, which is easy to solve numerically. Suppose that all the IDSs solve the corresponding convex optimization problems asynchronously and broadcast their collaboration strategies using their own timescale. Let T_u^i denote the set of times that IDS i update its misreport probability, and assume that these sets are infinite for all the IDSs (i.e., all the IDSs will update infinitely often), an asynchronous dynamic update algorithm is proposed to compute the NE of the second layer game as in Algorithm 1.

Algorithm 1 Asynchronous Dynamic Update Algorithm

```
Initialization: set t = 0, p_i^c = 0 for i = 1, 2, ··· , N
repeat
  for all t = 0, 1, ..., N do
    if t ∈ T_u^i then
      IDS i solves the convex optimization problem and updates
      p_i^c(t).
    else
      p_i^c(t) = p_i^c(t − 1)
    end if
  end for
  t=t+1
until converged
```

5 NUMERICAL STUDY

In this section, numerical study is performed to validate the analytical results.

5.1 Utility-privacy Tradeoff

In this subsection, we consider a network consisting of N target systems protected by N corresponding IDSs, and it is assumed that all of them have high security requirements. In such a scenario, the IDSs would have more powerful response capability and the cost of response is considered to be small (i.e., b_i is large and $C_{r,i}$ is small, $\forall i \in \{1, 2, \cdots, N\}$). Considering these, we set $C_{a,i} = C_{r,i} = 0.1, W_i = 1000, b_i = 0.9, p_i^d = 0.7, p_i^{fp} = 0.3$ for $i = 1, 2, \cdots, N$. In addition, it is assumed that when the attacker launches an attack on organization i, IDS i will observe abnormal traffic with probability $q_{ii} = 1$, while the other IDS j will observe abnormal traffic with probability $q_{ji} = 0.8, \forall j \neq i$.

Figure 2 shows the tradeoff between the average payoff improvement (i.e., the difference of the utility of the collaborative scheme and that of the non-collaborative one in of the first-layer game) and the preserved privacy (i.e., $1 - P_L(p_i^c)$) of all the IDSs. It can be seen that in all the examined scenarios, the collaborative scheme

[3]Note that this condition always hold when the network is large enough, i.e., $N \to \infty$.

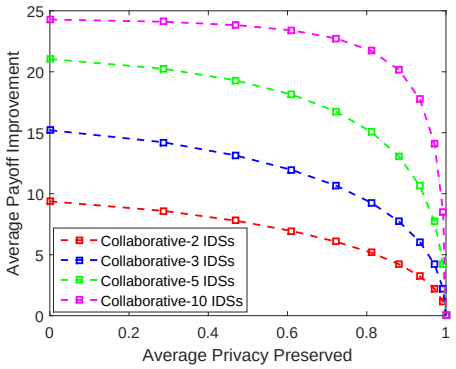

Figure 2: Utility-Privacy tradeoff curve.

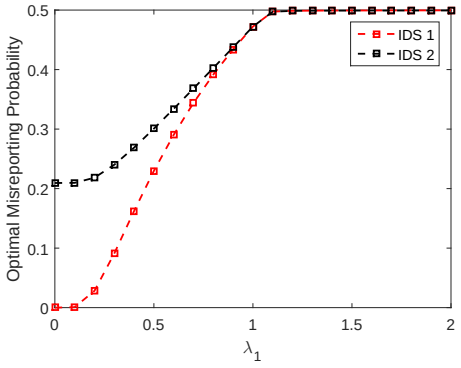

Figure 3: Misreporting probability against λ_1.

always enhances the performance, which justifies Proposition 1. In addition, the payoff improvement achieves its highest value when the preserved privacy is 0 (i.e., all the IDSs share their detecting results with others honestly) and the payoff improvement vanishes to 0 when the preserved privacy attains 1 (i.e., all the IDSs randomly send out their detection results with probability 0.5). Furthermore, when the number of collaborative IDSs increases, the IDSs can reserve more privacy while achieving the same the payoff improvement. Intuitively, when there are more collaborative IDSs, an IDS can gather more information about whether the attacker has launched an attack or not, given all the shared detection results. As a result, once the attacker launches an attack, the probability of being detected and triggering the IDSs to respond is higher, which in turn decreases the attacker's attacking probability.

5.2 Collaboration Strategies

In this subsection, the optimal collaboration strategies of IDSs in the two collaborative IDSs case are examined (similar results are observed for the cases of $N > 2$). The parameter $\beta_{i,j}$ is chosen to be $[\boldsymbol{U}_{i*}^{\boldsymbol{I},\boldsymbol{c}}(0) - \boldsymbol{U}_{i*}^{\boldsymbol{I}}][\boldsymbol{U}_{j*}^{\boldsymbol{I},\boldsymbol{c}}(0) - \boldsymbol{U}_{j*}^{\boldsymbol{I}}]$ for normalization.

Assuming $\lambda_2 = 1$, Figure 3 shows how λ_1 will influence the collaboration strategies of both IDSs. In our model, λ_1 determines how important privacy is for IDS 1, and is thus closely related to its privacy requirement. It can be seen that with different λ_1, not only the misreport probability of IDS 1 changes but also that of IDS 2. More specifically, when λ_1 becomes larger, both IDSs would collaborate with higher misreport probability. This may be explained as follows: a larger λ_1 implies that IDS 1 emphasizes more on privacy, and hence it would prefer to increase its misreport probability. In the meantime, a higher misreport probability of IDS 1 also decreases IDS 2's willingness to collaborate. As a result, IDS 2 will increase its misreport probability in response.

In addition, it is worth mentioning that for different privacy requirements (i.e., different λ_1 and λ_2), our model is able to guide the IDSs in finding optimal collaboration strategies that can achieve a suitable balance between utility and privacy.

6 LIMITATIONS

In this work, it is assumed that the collaborative IDSs are all trustworthy, that is, they will report their misreport probabilities honestly. In real scenarios, however, some selfish IDSs may break the rule by sending out wrong misreport probabilities in order to better protect their own privacy. Even worse, when there are some compromised IDSs in the network, they may broadcast wrong detection results to mislead others, and therefore threaten the effective collaboration. In these cases, the performance of the proposed approach may degrade.

7 CONCLUSIONS AND FUTURE WORKS

In this work, the utility-privacy tradeoff problem in CIDSs is formulated as a repeated two-layer single-leader multi-follower game which ends once the IDSs respond to the attacker successfully. By solving the first layer leader-follower game, the utility-privacy tradeoff curve for given collaboration strategies depending on the privacy policies of different organizations is obtained. By solving the second layer game, the collaborative strategies for the IDSs at NE can be computed. In addition, the existence of NE of the second-layer game is proved and an asynchronous dynamic update algorithm is developed to compute the NE. Further extending this work to dynamic settings or multiple possible attacks settings constitute interesting future directions.

REFERENCES

[1] L. Chen and J. Leneutre. 2009. A Game Theoretical Framework on Intrusion Detection in Heterogeneous Networks. *IEEE Transactions on Information Forensics and Security* 4, 2 (June 2009), 165–178. DOI: http://dx.doi.org/10.1109/TIFS.2009.2019154

[2] H. G. Do and W. K. Ng. 2015. Privacy-preserving approach for sharing and processing intrusion alert data. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*. 1–6. DOI: http://dx.doi.org/10.1109/ISSNIP.2015.7106911

[3] E. T. Jaynes. 1957. *Information theory and statistical mechanics.* Physical review 106.4: 620.

[4] P. Lincoln, P. A. Porras, and V. Shmatikov. 2004. Privacy-preserving sharing and correlation of security alerts. In *USENIX Security Symposium*. 239–254.

[5] M. E. Locasto, J. J. Parekh, A. D. Keromytis, and S. J. Stolfo. 2005. Towards collaborative security and P2P intrusion detection. In *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop*. 333–339. DOI: http://dx.doi.org/10.1109/IAW.2005.1495971

[6] J. Marecki, G. Tesauro, and R. Segal. 2012. Playing Repeated Stackelberg Games with Unknown Opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2 (AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 821–828. http://dl.acm.org/citation.cfm?id=2343776.2343814

[7] G. Meng, Y. Liu, J. Zhang, A. Pokluda, and R. Boutaba. 2015. Collaborative Security: A Survey and Taxonomy. *ACM Comput. Surv.* 48, 1 (Jul. 2015), 1:1–1:42.

[8] M. J. Osborne and A. Rubinstein. 1994. *A Course in Game Theory.* Cambridge, MA: MIT Press.

[9] E. Vasilomanolakis, S. Karuppayah, M. Mühlhäuser, and M. Fischer. 2015. Taxonomy and Survey of Collaborative Intrusion Detection. *ACM Comput. Surv.* 47, 4 (May 2015), 55:1–55:33.

[10] E. Vasilomanolakis, M. Krügl, C. G. Cordero, M. Mühlhäuser, and M. Fischer. 2015. SkipMon: A locality-aware Collaborative Intrusion Detection System. In *2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC)*. 1–8. DOI: http://dx.doi.org/10.1109/PCCC.2015.7410282

[11] D. Xu and P. Ning. 2005. Privacy-preserving alert correlation: a concept hierarchy based approach. In *21st Annual Computer Security Applications Conference (ACSAC'05)*. DOI: http://dx.doi.org/10.1109/CSAC.2005.45

[12] D. Xu and P. Ning. 2006. A Flexible Approach to Intrusion Alert Anonymization and Correlation. In *Securecomm and Workshops, 2006*. 1–10. DOI: http://dx.doi.org/10.1109/SECCOMW.2006.359544