Insider Threat Modeling: an Adversarial Risk Analysis Approach

Chaitanya Joshi, David Rios Insua, and Jesus Rios,

***Abstract*—Insider threats entail major security issues in geopolitics, cyber security and business organizations. Most earlier work in the field has focused on standard game theoretic approaches. We provide here two alternative, more realistic models based on adversarial risk analysis (ARA). ARA does not assume common knowledge and solves the problem from the point of view of just one of the players, the defender (typically), taking into account their knowledge and uncertainties regarding the choices available to them, to their adversaries, the possible outcomes, their payoffs/utilities and their opponents payoffs/utilities. The first model depicts the problem as a standard Defend-Attack-Defend model. The second model segments the set of involved agents in three classes of users and considers both sequential as well as simultaneous actions. A data security example illustrates the discussion.**

***Index Terms*—Insider threat, Game theory, Adversarial Risk Analysis.**

I. Introduction

INSIDER threats are encountered in international security, geo-politics, business, cyber security and so on. They are not only widely perceived to be significant ([1], [2]), but also often considered to be more damaging and more likely than outsider attacks ([1], [3]). Moreover, it is feared that the impact of the insider threat problem actually known is only the tip of an iceberg as many organizations are choosing not to report such incidents unless required to do so by law ([4]). As described in [5], it is a field in which little data is available, specially in the cyber security domain. Protection from insider threats is challenging as the perpetrators might have access to sensitive resources and privileged system accounts. Solutions to insider threat problems are considered to be complex ([6]). Technical solutions do not suffice since insider threats are fundamentally a people issue, as thoroughly discussed in [7] and [8].

In its simplest form, it is natural to view the insider threat problem as a two player game. We may call the first player *the organization* (which could refer to a single business or military unit or a similar entity, but also to a whole country or a coalition of entities or countries) and the second one, *the employee* (which could refer to one or more employees, contractors, or persons who have significant access to the organization and have been trusted with such access). A typical scenario would be as follows: since insider threats are a well-known phenomena, it will frequently be the case that several measures would have already been implemented by the organization (at least, in case of sufficiently mature organizations) to prevent or deter an insider attack. As an example, [9] provide a catalog of best practices against insider threats in cybersecurity. The employee will typically be aware of the measures in place and plans an attack accordingly. Once the attack has been carried out and detected, the organization will undertake actions to end the attack and mitigate any damage caused, possibly based on the resources deployed at the first stage. This type of interactions have been named sequential Defend-Attack-Defend games, see e.g. [10].

It is therefore natural that game-theoretic models of the insider threats phenomenon have been explored. For example, [11] model the problem as a two-player, zero-sum dynamic game. At each discrete time point, both players make decisions resulting in a change of state and opposite (given the zero-sum property) rewards to them. The authors then look for *Nash equilibria* (NE). This model is oversimplified in several respects. For example, there could be multiple attackers, the attacker pay-offs might not be immediate to obtain and the game might not be zero-sum. Also, in most cases, the defender would have already employed measures to prevent an insider attack and, therefore, the problem should be modeled as a sequential Defend-Attack-Defend game instead and not as a simultaneous one.

A more realistic approach is described in [12] who consider an insider threat problem in cybersecurity, trying to model the continuous interactions between an intruder and an intrusion detection system (IDS). They assume bounded rationality on them, use quantal response equilibria instead of the standard NE and assign pay-offs through utilities to assess the outcomes. However, their model focuses on a particular application and is not immediately generalizable. Moreover, the game does not consider multiple players and carries on even after detecting an attack as the detection causes the attack to be stopped, but does not eliminate the attacker from the game. [13] also model insider threats to IT systems considering bounded rationality and combine game theory with an information fusion algorithm to improve upon traditional IDS based methods by being able to consider various types of information. [14] and [15] propose three player games to model the use of Advanced Persistent Threats (APT) by a malicious insider. They employ a two layer game and show the existence of NE.

While game theory has been the typical choice to model interactions between two or more strategic adversaries, limitations of such theory, e.g. [16], [17], or [18], have long been pointed out, focusing on common knowledge assumption

C. Joshi is with the Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand e-mail: (cjoshi@waikato.ac.nz).

D. Rios Insua is with ICMAT, Madrid, Spain and J. Rios is with IBM Research, NY, USA.

Manuscript submitted February 14, 2019.

and the conservative nature of its solutions. Limitations of conventional risk analysis in security have been pointed out as well; [19] and [20] warn that it is inappropriate to model, say, terrorist actions in the same way as hurricanes. Therefore, in this paper, we shall propose adversarial risk analysis (ARA), [21], approaches to insider threats. ARA does not assume common knowledge and solves the problem from the point of view of just one of the players, typically, the defender, taking into account their knowledge and uncertainties regarding the choices available to them, to their adversaries, the possible outcomes, their payoffs/utilities and their opponents payoffs/utilities. Since its introduction, it has been used to model a variety of problems such as network routing for insurgency ([22]), international piracy ([23]), counter-terrorism ([24]), autonomous social agents ([25], and urban security resource allocation ([26]). ARA takes into account the expected utilities for the defender as well as the random expected utilities for the opponents, incorporating uncertainty regarding the strategic reasoning of the opponents. However, an ARA solution to insider threats has not yet been developed.

The structure of the paper is as follows. We first deal with the problem through an ARA Defend-Attack-Defend model between the organization and the employee. We then segment the employees in three classes (good, inadvertent and malicious insiders) considering more sophisticated ARA models. Finally, we illustrate the concepts with a numerical example and end up with some discussion and ideas for further work.

II. A Defend-Attack-Defend model for the insider threat problem

We start with a Defend-Attack-Defend model to deal with the insider threat problem, which considers a defender D (the organization, she) and an agent A (the employee, he). Our model is based upon the graphical framework described in [27]. Figure 1 presents the problem using a bi-agent influence diagram (BAID) where decisions are represented by square nodes, uncertainties using circular nodes and utilities with hexagonal nodes. Nodes corresponding to D are not shaded; those corresponding to A are diagonally shaded; and, finally, the shared chance node S is shaded using horizontal dashed lines. Dashed arrows indicate that the involved decisions are made with the corresponding agent knowing the values of the preceding nodes, whereas solid arrows indicate probabilistic or value dependence of the corresponding node with respect to the predecessors.

The action and outcome sets are as follows. Initially, the organization must choose one of the available preventive measures d_1 in the set $\mathcal{D}_1$. Having observed the preventive measure taken, the employee will adopt one of the actions a in $\mathcal{A}$; this set could consist of either 'no attack' or 'attack' or different types/intensities of attacks or other attack options. The set $\mathcal{S}$ consists of the possible outcomes s that can occur as a result of the preventive measure d_1 and the attack a adopted. Once the attack has been detected, the organization will choose to carry out one of the possible actions d_2 in the set $\mathcal{D}_2$ to end the attack, limit any damage and possibly pre-empt

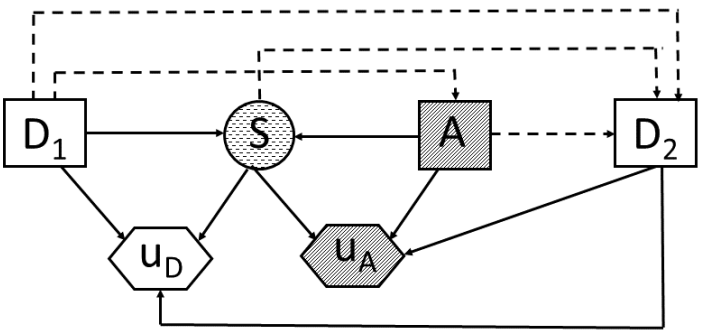

Fig. 1. BAID for the Defend-Attack-Defend insider threat game

future attacks leading to the final outcomes of both agents, respectively, evaluated through their utility functions u_D and u_A. Note that all three sets $\mathcal{D}_1$, $\mathcal{A}$ and $\mathcal{D}_2$ could contain a *do nothing* action.

For its solution, the defender must first quantify the following:

1) The distribution $p_D(a|d_1)$ modeling her beliefs about the attack a chosen at node A by the employee given the chosen defense d_1.
2) The distribution $p_D(s|d_1, a)$ modeling her beliefs about the outcome s of the attack, given a and d_1.
3) Her utility function $u_D(d_1, s, d_2)$ which evaluates the consequences associated with their first (d_1) and second (d_2) defensive actions as well as the outcome s of the attack.

Given these assessments, the defender first seeks to find the action $d_2^*(d_1, s)$ maximizing her utility

$$d_2^*(d_1, s) = \arg \max_{d_2 \in \mathcal{D}_2} u_D(d_1, s, d_2), \tag{1}$$

leading to the best second defense when the first one was d_1 and the outcome was s. Then, they seek to compute the expected utility $\psi_D(d_1, a)$ for each $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$ as

$$\psi_D(d_1, a) = \int u_D(d_1, s, d_2^*(d_1, s)) p_D(s|d_1, a)\, ds. \tag{2}$$

Moving backwards, she computes her expected utility for each $d_1 \in \mathcal{D}_1$ using the predictive distribution $p_D(a|d_1)$ through

$$\psi_D(d_1) = \int \psi_D(d_1, a) p_D(a|d_1)\, da. \tag{3}$$

Finally, the defender has to find her maximum expected utility decision $d_1^* = \arg\max_{d_1 \in D_1} \psi_D(d_1)$. This backward induction shows that the defender's optimal strategy is to first choose d_1^* and, then, after having observed s, choose $d_2^*(d_1^*, s)$.

The above analysis requires the defender to elicit $p_D(a|d_1)$. This can either be done using risk analysis based approaches such as [28] or using the ordinal judgment procedure by [29] or by modeling the strategic analysis process of the attacker. The defender could model the attacker's strategic analysis by assuming that the attacker will perform an analysis similar to hers to find their optimal attack a^*. To do so, the defender should assess the attacker's utility function $u_A(a, s, d_2)$ and probability distributions $p_A(s|a, d_1)$ and $p_A(d_2|d_1, a, s)$.

However, since the corresponding judgments will not be available to the defender, we could model her uncertainty about them through a random utility function $U_A(a, s, d_2)$ and random probability distributions $P_A(s|a, d_1)$ and $P_A(d_2|d_1, a, s)$. Once these random quantities are elicited, the defender solves the attacker's decision problem using backward induction. This is done by following a process similar to how they solved their own decision problem but taking into account the randomness in judgments. First, the defender finds the random expected utility for each $d_2 \in \mathcal{D}_2$

$$\mathbf{\Psi_A}(d_1, a, s) = \int U_A(a, s, d_2) P_A(d_2|d_1, a, s)\, dd_2. \quad (4)$$

Then, they find the random expected utility for each pair $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$

$$\mathbf{\Psi_A}(d_1, a) = \int \mathbf{\Psi_A}(d_1, a, s) P_A(s|d_1, a)\, ds, \quad (5)$$

and compute the random optimal attack $A^*(d_1)$ given the defense d_1

$$A^*(d_1) = \arg\max_{a \in \mathcal{A}} \mathbf{\Psi_A}(d_1, a). \quad (6)$$

Finally, once the defender assesses $A^*(d_1)$, she is able to solve her decision problem. The desired predictive distribution by the defender about the attack chosen a given the initial defense d_1 is

$$p_D(a|d_1) = p_D(A^* = a|d_1) \text{ and}$$
$$p_D[A^* \leq a|d_1] = \int_0^a P_D(A^* = x|d_1)\, dx. \quad (7)$$

Note that, in the above analysis, we have assumed that all the involved quantities are continuous. Should some of the quantities be discrete, the corresponding integrals would be replaced by sums. Further, in Section IV, we illustrate how $P_D(a|d_1)$ can be approximated using Monte-Carlo methods.

III. An ARA model for the insider threat problem with segmented employees

The sequence of interactions between an organization and an employee could be more complex for various reasons. Firstly, it has been described ([30], [31], [32]) that the measures in D_1 can have unintended negative consequences. If the employee feels that the measures introduced by the organization to mitigate insider threats are intrusive or micro-managing or even aggressive, that could lead him to react in unintended ways. This could include not reporting suspicious activities or misusing the reporting processes either accidentally or intentionally. At worst, it could even motivate an employee to go rogue. Secondly, although we have treated the group employee as a single entity, in reality, this group could typically include a large number of people and therefore, the organization may be faced with multiple actors taking multiple actions. Note that, usually, a majority of employees will not take any action that would harm the organization. In fact, some of them would actively help prevent an insider attack. For example, one of the possible insider actions in A could be to correctly follow the processes or measures set out by the organization possibly resulting in the successful prevention of the imminent attack altogether. Finally, the actions by employees could be dependent (sequential) or independent (simultaneous).

We shall focus on considering the issue of modeling different types of employees. [31] provide a segmentation with inadvertent and malicious insiders. We shall classify the employees as A_1 (*the good*), A_2 (*the bad*) and A_3 (*the ugly*), with S_1, S_2 and S_3 being the corresponding outcome sets. Each group of employees generates a relevant game as shown in Figure 2. Specifically, we consider that:

- A_1 are the employees who correctly and promptly perform their duties including following any procedures to prevent insider attacks. They have a positive impact on the productivity and work culture of the organization and will correctly report any suspicious activity, thus helping the organization to protect itself. Therefore, their actions will be positive to it.
- A_2 are the employees who, while not intentionally working to harm the organization, will help to create an environment which could increase the chances of an insider attack through their accidental or deliberate actions. For example, they could misuse the defensive procedures, creating a culture of mis-trust and loss in productivity. This, in turn, could lead to employees not feeling safe to report suspicious activities and even potentially motivate others to go rogue and plan an insider attack. Therefore, their actions will be negative to the organization.
- A_3 are the employees who will actively aim at harming the organization. They are the ones who intend to launch an insider attack. Their actions will therefore be very negative to the organization. Actions by A_1 may reduce the chance of insider attacks as well as the chance of one of them succeeding. Similarly, actions by A_2 may increase the chance of an insider attack as well as the chance of one of them succeeding.

First, we solve this game assuming that employees act in a sequential manner: at any given time, only one type of employees take an action. We then solve this game for a situation in which two or all three types of employees could act simultaneously.

A. Sequential action

The game in Figure 2[a] refers to the role played by the 'good' employee. Their action set $\mathcal{A}_1$ includes correctly implementing defensive procedures and whistle blowing suspicious activities through appropriate channels. The outcome set $\mathcal{S}_1$ consists of 'attack prevented' or 'attack not prevented'. In the first case, we assume that no further action is required from any of the players and, hence, the game ends. However, if the attack was not prevented, the attacker, the 'ugly' employee, will proceed with their chosen action $\mathcal{A}_3$, resulting in the outcome set $\mathcal{S}_3$ consisting of damage at various levels. Upon detection, the organization will take whatever actions $\mathcal{D}_2$ necessary to end the attack and contain any damage.

The game represented in Figure 2[b] considers the role played by the 'bad' employee. Their action set $\mathcal{A}_2$ includes intentional or unintentional misuse of defensive procedures, possibly leading that suspicious activities are either not reported

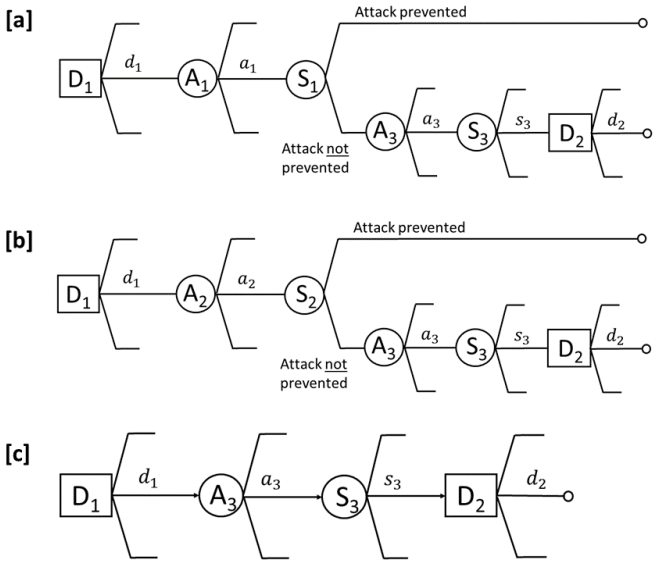

Fig. 2. Decision trees for the three games in the insider threat problem with segmented employees.

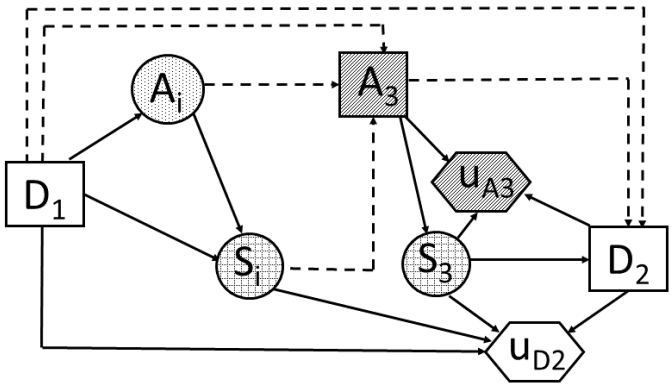

Fig. 3. MAID for decision trees [a] and [b] in the segmented employees insider threat game

or reported through external/unauthorized channels which, in turn, could cause significant harm to the organization. At worst, such a culture could even motivate an employee to launch an insider attack. The outcome set $\mathcal{S}_2$ consists of the same events as in $\mathcal{S}_1$. In case that the attack was prevented, we assume that no further action was required from any of the players and hence the game ends. However, in the event that the attack was not prevented, the 'ugly' employee will proceed with their chosen action $\mathcal{A}_3$ which could consist of an attack of a certain level resulting in the outcome set $\mathcal{S}_3$. Upon detection, the organization will take whatever actions $\mathcal{D}_2$ necessary to end the attack and contain any damage.

It may be possible that the 'ugly' employee is able to carry out their operation without being affected by the actions of the other groups of employees. This scenario is represented by the ID in Figure 2[c]. This game is identical to the model considered in Section II.

For the first two games (Figs. 2[a] and [b]), the ARA will consist of identical sets of steps. Henceforth, we use A_i, $i = 1, 2$ and S_i, $i = 1, 2$. The MAID for the segmented employee game for both cases is depicted in Figure 3. Note that we differentiate between node A_i, which is uncertain, and node A_3, which is a decision node but belonging to a different decision maker, as this last one is strategic. The defender must first quantify the following.

1) Her predictive distribution $p_D(a_i|d_1)$ about the action that will be chosen at node A_i given the defense d_1.
2) Her predictive distribution $p_D(s_i|d_1, a_i)$ about the outcome of such action, given a_i and d_1.
3) Her predictive distribution $p_D(a_3|d_1, a_i, s_i)$ about the attack that will be chosen at note A_3 given the outcome s_i and actions a_i and d_1.
4) Her predictive distribution $p_D(s_3|d_1, a_i, s_i, a_3)$ about the outcome of the attack, given outcome s_i, actions a_3, a_i and d_1.
5) The utility function $u_D(d_1, a_i, s_i, a_3, s_3, d_2)$ given their first and second defensive actions, the outcomes of the attack s_3 and s_i and the actions a_3 and a_i.

Given these, the defender works backwards along the decision trees in Figure 2 [a] or [b]. First, they seek to find the action $d_2^*(d_1, a_i, s_i, a_3, s_3)$ maximizing their utility

$$d_2^*(d_1, a_i, s_i, a_3, s_3) = \arg\max_{d_2 \in \mathcal{D}_2} u_D(d_1, a_i, s_i, a_3, s_3, d_2). \tag{8}$$

Then, for each $(d_1, a_i, s_i, a_3) \in \mathcal{D}_1 \times \mathcal{A}_i \times \mathcal{S}_i \times \mathcal{A}_3$, they seek to compute the expected utility $\psi_D(d_1, a_i, s_i, a_3)$ through

$$\psi_D(d_1, a_i, s_i, a_3) = \int u_D(d_1, a_i, s_i, a_3, s_3, d_2^*(d_1, a_i, s_i, a_3, s_3)) \, p_D(s_3|d_1, a_i, s_i, a_3)\, ds_3. \tag{9}$$

Next, they compute the expected utility $\psi_D(d_1, a_i, s_i)$ for each (d_1, a_i, s_i) through

$$\psi_D(d_1, a_i, s_i) = \int \psi_D(d_1, a_i, s_i, a_3) p_D(a_3|d_1, a_i, s_i)\, da_3. \tag{10}$$

They then find the expected utility $\psi_D(d_1, a_i)$ for each (d_1, a_i), as

$$\psi_D(d_1, a_i) = \int \psi_D(d_1, a_i, s_i) p_D(s_i|d_1, a_i)\, ds_i, \tag{11}$$

and their expected utility for each $d_1 \in \mathcal{D}_1$ using their predictive distribution $p_D(a|d_1)$

$$\psi_D(d_1) = \int \psi_D(d_1, a_i) p_D(a_i|d_1)\, da_i. \tag{12}$$

Finally, the defender finds their maximum utility decision as $d_1^* = \arg\max_{d_1 \in \mathcal{D}_1} \psi_D(d_1)$. This backward induction shows that the defender's optimal strategy is to first choose d_1^* and then, after having observed a_i, s_i, a_3 and s_3, choose action $d_2^*(d_1^*, a_i, s_i, a_3, s_3)$.

The above analysis requires the defender to elicit $p_D(a_3|d_1, a_i, s_i)$ and $p_D(a_i|d_1)$. Of these, $p_D(a_i|d_1)$ refers to the actions by the *good* or *bad* employees, neither of whom intend to strategically harm the organization *per se*. Therefore, action A_i can be considered to be non-strategic. For this reason, A_i is represented as a random node in the MAID in Figure 3. Further, A_i being non-strategic, $p_D(a_i|d_1)$

can be elicited using historical data/research on employee behavior, where available. Eliciting $p_D(a_3|d_1, a_i, s_i)$ is, however, less straightforward. The defender could model the attacker's strategic analysis process by assuming that the attacker will perform an analysis similar to the defender to find their optimal action a_3^*. The attack A_3 will only go ahead if the outcome s_i has not resulted in it being prevented. Provided that attack A_3 can take place, while A_i and S_i may have an effect on the probability of an attack, we assume that the choice of an attack depends only on the defender action d_1, that is, $p_D(a_3|d_1, a_i, s_i) = p_D(a_3|d_1)$. To elicit it, the defender must assess $U_A(a_3, s_3, d_2)$, $P_A(s_3|a_3, d_1)$ and $P_A(d_2|d_1, a_3, s_3)$. These random utilities and distributions could be elicited in several ways, outlined in [33]. Once elicited, the defender solves the attacker's decision problem using backward induction - similar to how they solved their own decision problem. First, the defender finds the random expected utilities for each action $d_2 \in \mathcal{D}_2$

$$\mathbf{\Psi_A}(d_1, a_3, s_3) = \int U_A(a_3, s_3, d_2) P_A(d_2|d_1, a_3, s_3)\, dd_2. \tag{13}$$

Then, they find the random expected utilities integrating out $s_3 \in \mathcal{S}_3$

$$\mathbf{\Psi_A}(d_1, a_3) = \int \mathbf{\Psi_A}(d_1, a_3, s_3) P_A(s_3|d_1, a_3)\, ds_3, \tag{14}$$

and, finally, compute the random optimal attack

$$A_3^*(d_1) = \arg\max_{a_3 \in \mathcal{A}_3} \mathbf{\Psi_A}(d_1, a_3). \tag{15}$$

The desired predictive distribution by the defender about the attack chosen a_3 given the initial defense d_1 is

$$p_D(a_3|d_1) = p_D(A_3^* = a_3|d_1) \text{ and}$$
$$p_D[A_3^* \leq a_3|d_1] = \int_0^{a_3} P_D(A_3 = a|d_1)\, da. \tag{16}$$

Note that, in the above analysis, we have assumed that all the involved quantities are continuous. Should some of the quantities be discrete, the corresponding integrals would be replaced by sums. Further, in Section IV, we illustrate how $P_D(a|d_1)$ can be approximated using Monte-Carlo methods.

B. Simultaneous actions

We now consider the following scenarios where one or more types of employees act simultaneously. The MAIDs for these games are shown in Figures 4 and 5.

Figure 4 is the MAID for a game in which both the *good* and the *bad* employees act simultaneously and after having observed D_1. Similar to Section III-A, these actions are considered as non-strategic and hence are represented as a joint random node A_1A_2 in the MAID. These actions result in the random outcome S_{12}. The *ugly* employee observes these actions and the outcome before launching their attack. For this game, the ARA solution proceeds in an identical manner to the solution described in Section III-A. The decision maker first seeks to find action $d_2^*(d_1, a_1, a_2, s_{12}, a_3, s_3)$ which will maximize their utility

$$d_2^*(d_1, a_1, a_2, s_{12}, a_3, s_3) = \arg\max_{d_2 \in \mathcal{D}_2} u_D(\cdot, d_2), \tag{17}$$

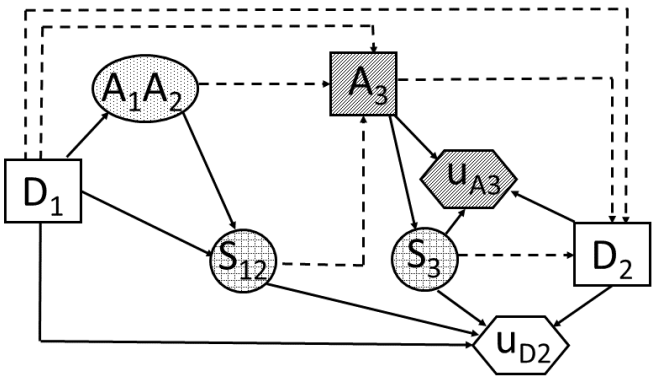

Fig. 4. MAID for the game where A_1 and A_2 act simultaneously followed by A_3.

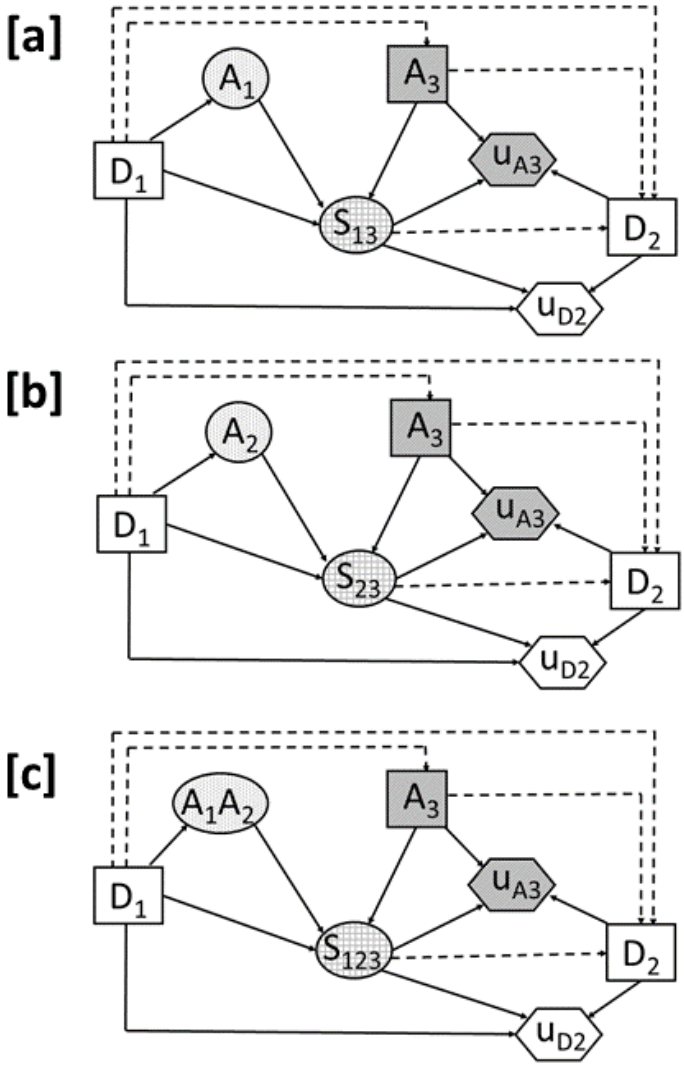

Fig. 5. MAIDs for games where A_3 acts simultaneously with either A_1 or A_2 or both act simultaneously followed by A_3.

where $u_D(\cdot, d_2) = u_D(d_1, a_1, a_2, s_{12}, a_3, s_3, d_2)$. Note that this utility function is dependent on both a_1 and a_2 and on the random outcome s_{12}. Then, they seek to compute the expected

utility $\psi_D(d_1, a_1, a_2, s_{12}, a_3)$ through

$$\psi_D(d_1, a_1, a_2, s_{12}, a_3) = \int u_D(d_1, a_1, a_2, s_{12}, a_3, s_3, d_2^*) \, p_D(s_3|s_{12}, a_3)\, ds_3. \quad (18)$$

Next, they compute the expected utility $\psi_D(d_1, a_1, a_2, s_{12})$ for each (d_1, a_1, a_2, s_{12}) through

$$\psi_D(d_1, a_1, a_2, s_{12}) = \int \psi_D(d_1, a_1, a_2, s_{12}, a_3) \, p_D(a_3|a_1, a_2, s_{12})\, da_3. \quad (19)$$

They then compute the expected utility $\psi_D(d_1, a_1, a_2)$ for each (d_1, a_1, a_2), as

$$\psi_D(d_1, a_1, a_2) = \int \psi_D(d_1, a_1, a_2, s_{12}) p_D(s_{12}|d_1, a_1, a_2)\, ds_{12}, \quad (20)$$

and their expected utility for each $d_1 \in \mathcal{D}_1$ using their predictive joint distribution $p_D(a_1, a_2|d_1)$ about what the *good* and the *bad* employees may do

$$\psi_D(d_1) = \int \psi_D(d_1, a_1, a_2) p_D(a_1, a_2|d_1)\, da_1 da_2. \quad (21)$$

Finally, the defender finds their maximum utility decision as $d_1^* = \arg\max_{d_1 \in \mathcal{D}_1} \psi_D(d_1)$.

This backward induction shows that the defender's optimal strategy is to first choose d_1^* and then, after having observed a_1, a_2, s_{12}, a_3 and s_3, choose action $d_2^*(d_1^*, a_1, a_2, s_{12}, a_3, s_3)$. The above analysis requires the defender to elicit $p_D(a_3|a_1, a_2, s_{12})$ and $p_D(a_1, a_2|d_1)$. Of these, $p_D(a_1, a_2|d_1)$ refers to the actions by the *good* or *bad* employees, neither of whom intend to strategically harm the organization *per se*. Further, it is reasonable to believe that their respective actions are independent of the actions by the other, and therefore, $p_D(a_1, a_2|d_1) = p_D(a_1|d_1) \times p_D(a_2|d_1)$. Action A_1 and A_2 can be considered to be non-strategic and both $p_D(a_1|d_1)$ as well as $p_D(a_2|d_1)$ can be elicited using historical data/research on employee behavior, where available. $p_D(a_3|d_1, a_1, a_2, s_{12})$ can be elicited by following Eqs. 13 to 16 after replacing a_i with (a_1, a_2) and s_i with s_{12}.

The MAIDs in Figure 5 correspond to extensions of the Defend-Attack-Defend game described in Section II. In these games either two (a, b) or all three (c) types of employees act simultaneously after observing D_1 and the ARA solution proceeds in a similar manner to the solution described in Section II. The main difference is that now the Defender must quantify the joint probabilities for the actions of two or all three types of employees concerned. For example, for the MAID in Figure 5 [a], the defender must quantify the distributions $p_D(a_1, a_3|d_1)$ and the distribution $p_D(s_{13}|d_1, a_1, a_3)$. If the actions of the *good* and *ugly* employees can be considered to be independent given d_1 then, $p_D(a_1, a_3|d_1) = p_D(a_1|d_1) \times p_D(a_3|d_1)$, where, $p_D(a_1|d_1)$ can be elicited using historical data and research on employee behavior (again, actions by the *good* and *bad* employees are considered as non-strategic and therefore represented by a chance node in the MAIDs) and $p_D(a_3|d_1)$ can be elicited by modeling the strategic analysis process of the *ugly* employee as detailed in Section II.

C. Model uncertainty

We have seen how one can find an optimal action, that is, the ARA solution for the organization given a specific game/model. The expected utility $\psi_D(d_1)$ that we find in each of the models is, in fact, $\psi_D(d_1|M)$, where M refers to the game under consideration. In reality though, the exact scenario will be unknown. It will not be known if the *ugly* employee is able to act without being affected by the actions of the *good* and/or the *bad* employees and if so, whether such interaction is sequential or simultaneous. A Bayesian approach allows the organization to incorporate model uncertainty into the analysis and identify the expected utility taking model uncertainty into account [34].

The organization starts by listing the set $\mathcal{M}$ of possible models, which will contain a subset of or all of the models considered above. Then, they must elicit a prior distribution $p_D(M), \forall M \in \mathcal{M}$. The defender then performs the ARA analysis on each of those models to obtain their expected utilities $\psi_D(d_1|M), \forall M \in \mathcal{M}$. Their expected utility taking into account the model uncertainty is then given by

$$\psi_D(d_1) = \sum_{M \in \mathcal{M}} p_D(M) \psi_D(d_1|M). \quad (22)$$

Their their maximum utility decision then is $d_1^* = \arg\max_{d_1 \in \mathcal{D}_1} \psi_D(d_1)$.

IV. Example

We consider an insider threat scenario motivated by [32] in which the malicious insider attempts to harm the incumbent organization without getting caught. The organization focuses on information/data collection and needs to protect itself against both insider and outsider attacks. It already has its sites and IT systems protected so that only authorized personnel are able to access them. However, anticipating attacks, the organization is considering implementing an additional security layer to defend itself. The defensive actions (D_1) under consideration are

1) anomaly detection/data provenance tools;
2) information security measures and employee training; and
3) carrying out random audits.

The malicious insider's aim could be financial fraud, data theft, espionage or whistle blowing. Regardless of the exact nature of the attack, we assume that the attacker's options (A) refer to its scale, say *small*, *medium* or *large*. For simplicity, we assume that the attack will either fully succeed (*S*) or fail (*F*). Once the attack has been carried out, irrespective of whether it is successful or not, we assume that the attack will be detected at some point, either through their own inspections or outside sources. In the wake of the detection, the organization can choose to carry out one of the following defensive actions (D_2):

1) major upgrade of defenses;
2) minor upgrade of defenses; or
3) no upgrade.

A. Using the defend-attack-defend model

We first analyze the problem using the model in Section II. We start by assessing the defender's utility function $u_D(d_1, a, s, d_2)$. We assume here that the defender's utilities depend not only on the outcome s (and d_1 and d_2) but also on the attack a. Indeed, we assume that u_D aggregates the monetary costs $c(d_1)$ and $c(d_2)$ associated with actions d_1 and d_2 respectively and the monetized perceived utilities associated with every (a, s) combination through

$$u_D(d_1, a, s, d_2) = c(d_1) + c(d_2) + u(a, s). \quad (23)$$

The costs and perceived utilities are listed in Tables I and II. They are scaled from -100 to 100. The utility u_D can then be computed for each combination; for example, $u_D(d_1 = \text{random audit}, a = \text{medium}, s = \text{not successful}, d_2 = \text{no upgrade}) = -50 + 60 + 0 = 10$.

We next elicit the probabilities $p_D(s|d_1, a)$. Suppose that they are as listed in Table III, with probabilities of failed attacks obtained through $p_D(\text{not successful}\,|d_1, a) = 1 - p_D(\text{successful}\,|d_1, a)$. In order to implement the ARA solution, the defender must first identify the action $d_2^*(d_1, s)$ maximizing their utility. In this case, d_2^* turns out to be 'no upgrade', being the cheapest option and will therefore maximize u_D. Then, they must compute the expected utility $\psi_D(d_1, a)$ using Eq. (2). The expected utility $\psi_D(d_1, a)$ can now be computed, for example, $\psi_D(\text{random audits, medium}) = -100 \times 0.4 + 10 \times 0.6 = -34$. The $\psi_D(d_1, a)$ values are given in Table IV. Then, we need to compute the expected utility for each $d_1 \in D_1$ using Eq. (3) and the predictive distribution $p_D(a|d_1)$ about what the malicious insider may do. Assume first that the defender has elicited $p_D(a|d_1)$ using her own beliefs as in Table V. The defender's expected utility $\psi_D(d_1)$ for each action d_1 is computed using Eq. (3). For example, $\psi_D(\text{random audits}) = -47.5 \times 0.5 - 34 \times 0.4 - 24 \times 0.1 = -39.75$. Similarly, the expected utility for *anomaly detection and data provenance* is -69.005, whereas for *information security and training* it is -29. This implies that the optimal option for the organization is to invest in information security and staff training and, if the attack was to happen, then, irrespective of whether it was successful or not, the optimal follow-up action would be not to upgrade their defenses. Recall that utilities defined in Eq. (23) only consider the monetary costs of defensive actions and not their potential benefits. This has been done for simplicity here, but in real life, the utilities defined should take into account both the costs as well as the potential benefits.

We now illustrate how $p_D(a|d_1)$ could be elicited by modeling the attacker's strategic analysis process using Eqs. (4) to (7). The defender could model the attacker's strategic analysis by assuming that the attacker will perform an analysis similar to the defender to find their optimal attack a^*. To do this, the defender must elicit the attacker's random utilities and probabilities $U_A(a, s, d_2)$, $P_A(s|a, d_1)$ and $P_A(d_2|d_1, a, s)$, using any information the defender might have, as well as considering the possible motivations for the attackers and their skill level.

Table VI lists the distributions elicited for $U_A(a, s, d_2)$ by the defender, with utilities between -100 and $+100$. We assume the defender thinks that the attacker believes that the defender has a short-sighted view and their utilities are a direct function of the costs involved in establishing the upgrades D_2. Whereas they will find it less valuable to upgrade their defensive mechanisms if the attack had, in fact, failed, no upgrade will on average be the least attractive option given that an attack was detected (whether successful or not). Table VII lists the distributions elicited for $P_A(d_2|d_1, a, s)$ by the defender, consistent with the utilities $U_A(a, s, d_2)$. For example, since an upgrade is considered to be less valuable in the event of a failed attack, the defender is unlikely to upgrade in the wake of a failed attack, reflected in the $Dir(1, 9, 90)$ distribution elicited for it. On the other hand, the $Dir(5, 4.9, 0.1)$ indicates that an upgrade is considered almost certainly likely in the wake of a successful large attack (irrespective of D_1). In Table VII, for each action A, the first row corresponds to, D_1 = anomaly detection & data provenance; the second to, D_1 = information security and training; and third to, D_1 = random audits. For each Dirichlet distribution, $Dir(\alpha_1, \alpha_2, \alpha_3)$, α_1 relates with probability of major upgrade, α_2 to minor and α_3 to no upgrade. Finally, Table VIII lists the distributions elicited for $P_A(\text{successful}\,|d_1, a)$ by the defender. For example, she believes that the attacker thinks that an attack is much more likely to succeed if D_1 is *information security and training* compared to the other options. Also, they believe that the attacker thinks that a small attack is much more likely to succeed than a medium or large attack. For each combination of d_1, a and s, we simulate $N = 1000$ samples from $U_A(a, s, d_2)$ and $P_A(d_2|d_1, a, s)$ to obtain samples from the attacker's expected utility $\mathbf{\Psi_A}(d_1, a, s)$ using Eq. (4) and, then, a sample of the attacker's expected utility $\mathbf{\Psi_A}(d_1, a)$ using Eq. (5). Then, for each of the simulations, we find the optimal defense d_1 maximizing $\mathbf{\Psi_A}(d_1, a)$ and, finally, estimates $p_D(a|d_1)$ by counting how many times (out of N) would the attacker choose a particular attack given d_1. These are presented in Table IX. We can now use these estimates of $p_D(a|d_1)$ to compute the defender's expected utility $\psi_D(d_1)$ using Eq. (3). The expected utility for the *anomaly detection and data provenance* comes out to be -36.115, for *information security and training* be -39.051, and, finally, for the *random audits* be -34.566. This implies that in this case, the optimal option for the organization is to invest in conducting random audits and, if the attack was to happen irrespective of whether it was successful or not, the optimal follow-up action would be not to upgrade their existing defenses.

Observe, therefore, that the $p_D(a|d_1)$ elicited by modeling the attacker's strategic analysis (Table IX) turns out to be quite different from that elicited using their own belief and knowledge (Table V), leading to different optimal decisions.

B. Using the segmented employees model

We now analyze this problem using the model discussed in Section III-A by assuming three types of employees. Again, we start by assessing the defender's utility function

$u_D(d_1, a_i, s_i, a_3, s_3, d_2)$. Just like we did with the model in Section IV-A, we assume that u_D adopts the form

$$u_D(d_1, a_i, s_i, a_3, s_3, d_2) = c(d_1)+u(a_i, s_i)+u(a_3, s_3)+c(d_2), \quad (24)$$

where $c(d_1)$ and $c(d_2)$ are as defined in Table I and $u(a_3, s_3)$ is the same as $u(a, s)$ defined in Table II. To define $u(a_i, s_i)$, we first define the values that A_i and S_i can take for $i = 1, 2$. As described in Section III, the outcome sets are $\mathcal{S}_i = \{$attack prevented, not prevented$\}$. In reality, the set $\mathcal{A}_1$ could consist of various actions that a good employee can take, for example, $\mathcal{A}_1 = \{$diligently perform all tasks, follow appropriate processes, be vigilant, etc.$\}$. Similarly, $\mathcal{A}_2 = \{$ misuse of policies, incorrectly following processes, actions affecting culture of organization, actions affecting productivity, etc.$\}$. The exact actions undertaken will affect the likelihood of an attack being prevented or not. Also, the utility $u(a_i, s_i)$ could depend on every combination of the actions a_i and outcomes s_i. However, for the sake of simplicity, we assume that individual actions do not affect the outcome or the utilities, but only the nature of the actions (desirable or not) does. Therefore, we do not distinguish between different desirable actions and consider them to be represented by a single a_1 and similarly, represent all non desirable actions using a single a_2. The $u(a_i, s_i)$ values are thus elicited as in Table X, which indicates a preference to desirable actions irrespective of the outcome. We are now able to calculate $u_D(d_1, a_i, s_i, a_3, s_3, d_2)$ using Tables I, II and X. In order to implement the ARA solution to this problem, the defender must use backward induction and first identify action $d_2^*(d_1, a_i, s_i, a_3, s_3)$ which will maximize their utility. In this case, again, d_2^* turns out to be 'No upgrade', as it is the cheapest option and will therefore maximize u_D. Next, we need to elicit the probabilities $p_D(s_3|d_1, a_i, s_i, a_3)$. Note that these probabilities are only defined when $s_i =$ *not prevented*. We further assume that if the attack could not be prevented; then, the probabilities of its success are irrespective of the actions a_i encountered. Under that assumption, $p_D(s_3|d_1, a_i, s_i = \text{not prevented}, a_3) = p_D(s_3|d_1, \text{not prevented}, a_3) = p_D(s_3|d_1, a_3)$. Therefore, these are considered to coincide with $p_D(s|d_1, a)$ in Table III. We are now able to compute $\psi_D(d_1, a_i, \text{not prevented}, a_3)$ using (9). These are listed in Table XI. We next seek to compute the expected utility $\psi_D(d_1, a_i, \text{not prevented})$, which requires us to elicit $p_D(a_3|d_1, a_i, \text{not prevented})$. This can be elicited either using the defender's knowledge, experience or guess or by modeling the malicious insider's strategic analysis process using Eqs. (13) to (16). Consider the first case; assume that the nature of the attack is independent of the type of employee A_1 or A_2 encountered earlier. Under this assumption, $p_D(a_3|d_1, a_i, \text{not prevented})$ is considered to coincide with $p_D(a|d_1)$ in Table V. $\psi_D(d_1, a_i, \text{not prevented})$, thus calculated, is listed in Table XII. The defender now seeks to compute the expected utility $\psi_D(d_1, a_i)$ by integrating out $p_D(s_i|d_1, a_i)$. Note that $\psi_D(d_1, a_i, \text{prevented}) = u_D(d_1, a_i, \text{prevented})$, since the game does not proceed any further if the attack was indeed prevented. $u_D(d_1, a_i, \text{prevented})$ are also listed in Table XII. Suppose that the defender considers that the probability of the attack being prevented only depends on the type of employee encountered and is independent of d_1. Suppose that the chances of preventing the attack was considered to be 50% if the attacker encountered the *good* employees and just 10% is the attacker encountered by the *bad* employees, that is, $p_D(S_1 = \text{prevented}|d_1, a_1) = 0.5$ and $p_D(S_2 = \text{prevented}|d_1, a_2) = 0.1$. $\psi_D(d_1, a_i)$ thus computed, is listed in Table XIII. Finally, the defender needs to integrate out $p_D(a_i|d_1)$ to compute the expected utility $\psi_D(d_1)$ of his defensive actions D_1 so as to identify the optimal action d_1^* that will maximize this expected utility. Suppose the defender believes that the *good* and the *bad* employees are randomly and evenly spread throughout their entire workforce and, therefore, $p_D(a_i|d_1)$ is independent of d_1. Suppose the defender guesses that 80% of the employees are good ones and the rest are bad. Then, $\psi_D(\text{anomaly detection \& data provenance}) = -71.2229$, $\psi_D(\text{information security \& training}) = -31.22$ and $\psi_D(\text{random audits}) = -33.255$. Thus, based on the elicited utilities and probabilities, the optimal defensive action is to *invest in Information security* and *training of the staff*.

We now consider the case where $p_D(a_3|d_1, a_i, \text{not prevented})$ is elicited by modeling the attacker's strategic thinking process. As discussed earlier, given that the attack has not been prevented, the choice of the attack will be independent of the type of employee (A_1 or A_2) encountered. We assume that the choice of an attack depends only on the defender actions d_1 and d_2. That is, $p_D(a_3|d_1, a_i, s_i) = p_D(a_3|d_1)$; to elicit it, the defender must assess $U_A(a_3, s_3, d_2)$, $P_A(s_3|a_3, d_1)$ and $P_A(d_2|d_1, a_3, s_3)$. It is also reasonable to assume that attacker's preferences and uncertainties are also independent of the type of employee encountered. Therefore, we consider $U_A(a_3, s_3, d_2)$ to coincide with $U_A(a, s, d_2)$ in Table VI, $P_A(s_3|a_3, d_1)$ to be exactly same as $P_A(s|a, d_1)$ elicited in Table VIII and, finally, $P_A(d_2|d_1, a_3, s_3)$ with $P_A(d_2|d_1, a, s)$ elicited in Table VII. We follow Eqs. (13) to (16) to compute $P_D(A_3 = a_3|d_1, a_i, \text{not prevented})$, which, as expected turns out to be $p_D(A = a|d_1)$ elicited in Table IX. We now compute the random expected utility $\mathbf{\Psi_D}(d_1, a_i, s_i)$, using Eq. (10) and proceed to compute the random expected utility $\mathbf{\Psi_D}(d_1)$ using Eqs. (11) and (12). In this case, we have $\psi_D(\text{anomaly detection \& data provenance}) = -52.147$, $\psi_D(\text{information security \& training}) = -37.049$ and $\psi_D(\text{random audits}) = -30.248$. Therefore, based on the elicited utilities and probabilities, the optimal defensive action is to invest in performing random audits.

Thus, similar to the earlier case, eliciting $p_D(a_3|d_1)$ by modeling the attacker's strategic thinking process yields a different optimal decision for the defender.

C. Model uncertainty

Suppose now that the defender is not certain if the malicious insider will be able to act on their own or whether his actions will be affected by other employees. He decides

to consider two models: M_1, the model in Section IV-A and M_2, that in Section IV-B. He elicits a prior probability $p_D(M_1) = 0.3$, which implies that $p_D(M_2) = 0.7$. He would then perform the ARA analysis and arrive at his expected utilities $\psi_D(d_1|M_1) = (-36.115, -39.051, -34.556)$, for the three options when $p_D(a|d_1)$ is elicited by modeling the attacker's strategic thinking, as illustrated in Section IV-A. Similarly, he arrives at his expected utilities $\psi_D(d_1|M_2) = (-52.147, -37.049, -30.248)$, as illustrated in Section IV-B. Then using (22), the expected utilities $\psi_D(d_1)$ taking into account the the model uncertainty are $\psi_D(d_1 = (-47.337, -37.65, -31.541)$. Thus, investing in random audits is the optimal strategy for the defender taking into account their model uncertainty.

V. Discussion and Further Work

Insider threats constitute a major security problem worldwide. We have develped ARA based models to determine optimum strategies to counter insider threats and illustrated using a data security application.

In general, as in with almost any security application, interactions between the defender and the attacker will expand over several time periods and they will respectively evolve their defenses and attacks so as to effectively counter their adversarial actions. This can be modeled using a Markov decision process (MDP). However, a general ARA solution to MDPs has not been developed yet, thus being a promising area for further research. We could then provide a specific MDP solution to the insider threat problem. This approach could also provide an ARA solution to support the advanced persistent threat (APT) problem, being a persistent and long term threat.

Insider threats come in many different forms. The problem considered here is probably the most obvious, two player version where the malicious insider seeks to harm the organization directly. However, there are more complicated three player versions. The three player versions could consist of two attackers and one defender or the other way around or even an attacker, a defender and a victim (the victim being a third party). For example, a three player case consisting of a malicious insider, the APT and the organization consists of two attackers and a defender. But the malicious insider could be also be someone who uses their privileges to exploit, abuse or harm a third party. This third party could be clients, customers, students, patients, etc. A recent well known example of this type is that of the USA gymnastic team osteopathic physician Dr. Larry Nassar who was convicted for sexual abuse of young athletes under the pretext of treating them for their injuries. Therefore, an important extension would be to develop ARA solutions to such complex three player games. This could provide a much more realistic alternative to the game theoretic models proposed in [14] and [15].

Players are not always entirely rational and hence incorporating bounded rationality may make the model more realistic. ARA is naturally equipped to incorporate attackers with different reasonings, such as non-strategic thinking, Nash equilibrium, level-k thinking and the mirror equilibrium ([27]). However, a general ARA solution using the bounded rationality has not yet been developed. Developing such a solution will enable a bounded rationality ARA solution to the insider threat problem.

ARA relies on the elicitation of the adversarys utilities and probabilities. Robustness analysis of ARA to these elicitations is necessary, but has yet to be developed. [35] highlight the need and illustrate how a robustness analysis can be performed in principle for ARA. It is important to be able to investigate the sensitivity of the ARA outcome - the optimal strategy - to any errors or mis-specifications in the utilities and the probabilities elicited for the analysis.

Appendix A
Tables for Examples in Section IV

TABLE I
Costs associated with defensive actions d_1 and d_2.

d_1	$c(d_1)$	d_2	$c(d_2)$
Anom. det. & Data prov.	-100	Major upgrade	-100
Info. Sec.& train.	-60	Minor upgrade	-25
Random audits	-50	No upgrade	0

TABLE II
Monetized Perceived Utility for Every Combination (a, s).

a	s	$u(a, s)$
Small	Success	-25
Small	Fail	30
Medium	Success	-50
Medium	Fail	60
Large	Success	-100
Large	Fail	80

Acknowledgment

The work of CJ was supported by the Strategic Investment funding provided by the University of Waikato. The work of DRI is supported by the AXA-ICMAT Chair on Adversarial Risk Analysis, the Spanish Ministry of Economy and Innovation program MTM2017-86875-C3-1-R and project MTM2015-72907-EXP. Work supported by the EU's Horizon 2020 project 740920 CYBECO (Supporting Cyberinsurance from a Behavioural Choice Perspective).

References

[1] H. Schulze, *Insider Threat, 2018 report*. ca Technologies, 2018. [Online]. Available: https://www.ca.com/content/dam/ca/us/files/ebook/insider-threat-report.pdf

[2] B. Ware, *Insider Attacks, 2017 insider threat study*. Haystax, 2017. [Online]. Available: https://haystax.com/blog/whitepapers/insider-attacks-industry-survey/

[3] CERT, *2012 Cyber Security Watch Survey. How Bad is the Insider Threat?* Software Engineering Institute, Carnegie Mellon.

TABLE III
PROBABILITIES $p_D(S = \text{SUCCESSFUL} \mid d_1, a)$ ELICITED FOR EVERY (d_1, a).

d_1	$a =$ small	$a =$ med.	$a =$ large
Anom. det. & Data prov.	0.1	0.07	0.05
Info. sec. & train.	0.3	0.25	0.2
Random audits	0.5	0.4	0.3

TABLE IV
$\psi_D(d_1, a)$ FOR EVERY COMBINATION OF a AND s.

D_1	$A =$ small	$A =$ med.	$A =$ large
Anom. det. & Data prov.	-75.5	-47.7	-29
Info. sec. & train.	-46.5	-27.5	-16
Random audits	-47.5	-34	-24

TABLE V
$p_D(a|d_1)$ ELICITED BY DEFENDER USING THEIR BELIEFS

D_1	$A =$ small	$A =$ med.	$A =$ large
Anom. det. & Data prov.	0.8	0.15	0.05
Info. sec. & train.	0.2	0.6	0.2
Random audits	0.5	0.4	0.1

TABLE VI
DISTRIBUTIONS $U_A(a, s, d_2)$ ELICITED BY DEFENDER.

	$A =$ small	
D_2	Succ.	Fail
Maj.upgr.	$N(-80, 5)$	$N(-90, 2)$
Min.upgr.	$N(-50, 10)$	$N(-60, 5)$
No upgr.	$100 - Exp(5)$	$100 - Exp(5)$
	$A =$ medium	
D_2	Succ.	Fail
Maj.upgr.	$N(-80, 5)$	$N(-90, 2)$
Min.upgr.	$N(-40, 10)$	$N(-60, 5)$
No upgr.	$100 - Exp(3)$	$100 - Exp(3)$
	$A =$ large	
D_2	Succ.	Fail
Maj.upgr.	$N(-80, 5)$	$N(-90, 2)$
Min.upgr.	$N(-30, 10)$	$N(-60, 5)$
No upgr.	$100 - Exp(1)$	$100 - Exp(1)$

TABLE VII
DISTRIBUTIONS $P_A(d_2|d_1, a, s)$ ELICITED BY DEFENDER.

$A =$ small	
Success	Fail
$Dir(1, 3, 6)$	$Dir(1, 9, 90)$
$Dir(1, 5, 4)$	$Dir(1, 9, 90)$
$Dir(1, 4, 5)$	$Dir(1, 9, 90)$
$A =$ medium	
Success	Fail
$Dir(2.5, 7, 0.5)$	$Dir(1, 9, 90)$
$Dir(1.5, 8, 0.5)$	$Dir(1, 9, 90)$
$Dir(2, 6, 2)$	$Dir(1, 9, 90)$
$A =$ large	
Success	Fail
$Dir(5, 4.9, 0.1)$	$Dir(1, 9, 90)$
$Dir(5, 4.9, 0.1)$	$Dir(1, 9, 90)$
$Dir(5, 4.9, 0.1)$	$Dir(1, 9, 90)$

TABLE VIII
$P_A(S = \text{SUCCESSFUL} \mid d_1, a)$ ELICITED BY DEFENDER FOR EVERY COMBINATION (a, s).

D_1	$A =$ small	$A =$ medium	$A =$ large
Anom. det. & Data prov.	$Beta(4, 6)$	$Beta(2, 8)$	$Beta(0.5, 9.5)$
Inf.sec. & train.	$Beta(9, 1)$	$Beta(8, 2)$	$Beta(7, 3)$
Random audits	$Beta(7, 3)$	$Beta(6, 4)$	$Beta(3, 7)$

TABLE IX
$P_D(a|d_1)$ ELICITED BY DEFENDER MODELING THE STRATEGIC ANALYSIS OF THE ATTACKER

D_1	$A =$ small	$A =$ med.	$A =$ large
Anom.det. & Data prov.	0.112	0.102	0.786
Info.sec. & train.	0.706	0.132	0.162
Random audits	0.399	0.119	0.482

TABLE X
$u(a_i, s_i)$ FOR EVERY COMBINATION OF a_i AND s_i.

A_i	$S_i =$ prev.	$S_i =$ not prev.
a_1	50	-10
a_2	10	-30

TABLE XI
$\psi_D(d_1, a_i, \text{NOT PREVENTED}, a_3)$ VALUES FOR A_1 AND A_2.

	A_1			A_2		
D_1	$A_3 =$ Small	$A_3 =$ Med.	$A_3 =$ Large	$A_3 =$ Small	$A_3 =$ Med.	$A_3 =$ Large
Anom.det. & prov.	-85.5	-57.7	-39	-105.5	-77.7	-59
Info.sec. & train.	-56.5	-37.5	-26	-76.5	-57.5	-46
Random audits	-57.5	-44	-34	-77.5	-64	-54

[4] P. Wood, B. Nahorney, K. Chandrasekar, S. Wallace, and K. Haley, *Internet Security Threat Report*. Symantec, 2016, vol. 21. [Online]. Available: https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf

[5] J. Hunker and C. Probst, "Insiders and insider threats: An overview of definitions and mitigation techniques," *Journal Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 2, no. 1, pp. 4–27, 2009.

[6] W. Lee and B. Rotoloni, *Emerging Cyber Threat Report 2015*. Georgia Tech Information Security Centre and Georgia Tech Research Institute, 2015.

[7] K. Sarkar, "Assessing insider threats to information security using technical, behavioural and organisational measures," *Info. Sec. Tech. Rep.*, vol. 15, pp. 112–133, 2010.

[8] F. Greitzer, A. Dalton, L. Kangas, C. Noonan, and R. Hohimer, "Identifying at-risk employees: Modeling psychosocial precursors of potential insider threats," *Proc. 25th HICSS*, 2012.

[9] G. Silowash, D. Cappelli, A. Moore, R. Trzeciak, T. Shimeall, and L. Flynn, "Common sense guide to mitigating insider threats," *Def. Tech. Inf. Center Tech. Report*, 2012.

[10] G. Brown, M. Carlyle, J. Salmeron, and R. Wood, "Defending critical infrastructure," *Interfaces*, vol. 36, pp. 530–544, 2006.

[11] D. Liu, X. Wang, and J. Camp, "Game-theoretic modeling and analysis of insider threats," *International Journal of Critical Infrastructure Protection*, vol. I, pp. 75 – 80, 2008.

[12] I. Kantzavelou and S. Katsikas, "A game-based intrusion detection mechanism to confront internal attackers," *Computers & Security*, vol. 29, no. 8, pp. 859 – 874, 2010.

[13] K. Tang, M. Zhao, and M. Zhou, "Cyber insider threats situation awareness using game theory and information fusion-based user behavior

TABLE XII
$\psi_D(d_1, a_i, \text{NOT PREVENTED})$ ELICITED BY DEFENDER

	Not prevented		Prevented	
D_1	A_1	A_2	A_1	A_2
Anom.det. & provenance.	-79.005	-99.005	-50	-90
Info.sec. & train.	-39	-59	-10	-50
Random audits	-49.75	-69.75	0	-40

TABLE XIII
$\psi_D(d_1, a_i)$ COMPUTED BY THE DEFENDER.

D_1	A_1	A_2
Anom.det. & Data prov.	-64.5025	-98.1045
Info.sec. & train.	-24.5	-58.1
Random audits	-24.875	-66.775

predicting algorithm," *Journal of Information & Computational Science*, vol. 8, no. 3, pp. 529 – 545, 2011.
[14] X. Feng, Z. Zheng, P. Hu, D. Cansever, and P. Mohapatra, "Stealthy attacks meets insider threats: A three-player game model," in *MILCOM 2015 - 2015 IEEE Military Communications Conference*, Oct 2015, pp. 25–30.
[15] P. Hu, H. Li, H. Fu, D. Cansever, and P. Mohapatra, "Dynamic defense strategy against advanced persistent threat with insiders," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 747–755.
[16] H. Gintis, *The Bounds of Reason: Game Theory and the Unification of Behavioural Sciences.* Princeton University Press, 2009.
[17] C. Camerer, *Behavioural Game Theory*. Princeton University Press, 2003.
[18] H. Raiffa, J. Richardson, and D. Metcalfe, *Negotiation Analysis*. Havard University Press, 2002.
[19] L. A. Cox, Jr., "Game theory and risk analysis," *Risk Analysis*, vol. 29, no. 8, pp. 1062–1068, 2009.
[20] G. G. Brown and L. A. Cox, Jr., "Making terrorism risk analysis less harmful and more useful: Another try," *Risk Analysis*, vol. 31, no. 2, pp. 193–195, 2011.
[21] I. R. Insua, J. Rios, and D. Banks, "Adversarial risk analysis," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 841–854, 2009.
[22] S. Wang and D. Banks, "Network routing for insurgency: An adversarial risk analysis framework," *Naval Research Logistics (NRL)*, vol. 58, no. 6, pp. 595–607, 2011.
[23] J. C. Sevillano, D. R. Insua, and J. Rios, "Adversarial Risk Analysis: The Somali Pirates Case," *Decision Analysis*, vol. 9, no. 2, pp. 86–95, June 2012.
[24] J. Rios and D. R. Insua, "Adversarial risk analysis for counterterrorism modeling," *Risk Analysis*, vol. 32, no. 5, pp. 894–915, 2012.
[25] P. G. Esteban and D. R. Insua, "Supporting an autonomous social agent within a competitive environment," *Cybernetics and Systems*, vol. 45, no. 3, pp. 241–253, 2014.
[26] C. Gil, D. Rios Insua, and J. Rios, "Adversarial risk analysis for urban security resource allocation," *Risk Analysis*, vol. 36, no. 4, pp. 727–741, 2016.
[27] D. Banks, J. Rios, and D. R. Insua, *Adversarial Risk Analysis*, 1st ed. CRC Press, 2015.
[28] B. Ezell, S. Bennett, D. Winterfeldt, J. Sokolowski, and A. Collins, "Probabilistic risk analysis and terrorism risk," *Risk Analysis*, vol. 30, no. 4, 2010.
[29] C. Wang and V. Bier, "Expert elicitation of adversary preferences using ordinal judgments," *Operations Research*, vol. 61, no. 2, pp. 372–385, 2013.
[30] A. Moore, W. Novak, M. Collins, R. Trzeciak, and M. Theis, *Effective Insider Threat Programs: Understanding and Avoiding Potential Pitfalls*, 2015, white Paper.
[31] D. Liu, X. Wang, and J. Camp, "Mitigating inadvertent insider threats with incentives," *BUSCAR!!*, vol. ??, p. ??, 2008.
[32] I. Martinez-Moyano, E. Rich, S. Conrad, D. Andersen, and T. Stewart, "A behavioral theory of insider-threat risks: a system dynamic approach," *ACM Transactions on Modeling and Computer Simulation*, vol. 18, no. 2, 2008.
[33] D. Ríos Insua, D. Banks, J. Ros, and J. Ortega, *Adversarial Risk Analysis as an Expert Judgement Methodology*. Springer International Publishing, 2019, pp. –.
[34] D. Draper, "Assessment and propagation of model uncertainty," *Journal Royal Statistical Society*, vol. 57, no. 1, pp. 45 – 97, 1995.
[35] D. Ríos Insua, F. Ruggeri, C. Alfaro, and J. Gomez, *Robustness for Adversarial Risk Analysis*. Springer International Publishing, 2016, pp. 39–58.

PLACE PHOTO HERE

Chaitanya Joshi is a senior lecturer at the Dept of Mathematics and Statistics at the University of Waikato. He received is his Ph.D in Statistics from Trinity College Dublin and M.Sc. in Statistics from Indian Institute of Technology Kanpur. His area of expertise are Bayesian methods, specifically, computational Bayesian methods, Bayesian robustness and Statistical modeling.
Prior to his Ph.D, he has worked in statistical positions at various corporations including Novartis, Bristol-Myers Squibb and Dun and Bradstreet.

David Rios Insua is AXA-ICMAT Chair in Adversarial Risk Analysis and Member of the Spanish Royal Academy of Sciences. He has formerly held teaching and research positions at SAMSI, Duke, Purdue, Paris-Dauphine, Leeds, Manchester, IIASA, CNR-IMATI and Madrid Technical University. He is Professor of Statistics and Operations Research at Rey Juan Carlos University (on leave). He has written 13 books, edited 7 special issues and published more than 100 refereed papers in his areas of interest, which include decision analysis, risk analysis, negotiation analysis and Bayesian statistics, and their applications to robotics, aviation safety, critical infrastrtucture protection and wter resources maangement, among others.

He has been PI of over 50 sponsored projects and supervised 17 PhD theses. Has received research awards from INFORMS, IFORS, IIASA, SRA, Wirsbo, the Everis Foundation and SEIO, among others.

Jesus Rios is a Research Staff Member at the IBM Research Division. He joined IBM Research in 2010. Prior to joining IBM, he has held several research positions at the University of Manchester, University of Luxembourg, Aalborg University, SAMSI (Statistical and Applied Mathematical Sciences Institute), and Concordia University.