SECONOMICS

D5.3 - General Methods for Security Risk Analysis

D. Ríos, J. Cano, F. Ortega, E. L. Cano, J. M. Moguerza, A. Alonso (URJC), A. Couce, S. Houmb (SNOK). A. Schmitz (ISST)

Pending of approval from the Research Executive Agency - EC

Document Number	D5.3
Document Title	General Methods for Security Risk Analysis
Version	2.0
Status	Final
Work Package	WP 5
Deliverable Type	Report
Contractual Date of Delivery	31.01.2015
Actual Date of Delivery	31.01.2015
Responsible Unit	URJC
Contributors	SNOK, ISST, Durham, UNITN
Keyword List	Adversarial risk; security risk; decision analysis; general methodology
Dissemination level	PU

Security Economics: Socio economics meets security

SECONOMICS Consortium

SECONOMICS “Socio-Economics meets Security” (Contract No. 285223) is a Collaborative project) within the 7th Framework Programme, theme SEC-2011.6.4-1 SEC-2011.7.5-2 ICT. The consortium members are:

1	UNIVERSITÀ DEGLI STUDI DI TRENTO	Università Degli Studi di Trento (UNITN) 38100 Trento, Italy http://www.unitn.it	Project Manager: Prof. Fabio Massacci fabio.massacci@unitn.it
2	DEEPBLUE	DEEP BLUE Srl (DBL) 00193 Roma, Italy http://www.dblue.it	Contact: Alessandra Tedeschi alessandra.tedeschi@dblue.it
3	Fraunhofer ISST	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., Hansastr. 27c, 80686 Munich, Germany http://www.isst.fraunhofer.de/en/	Contact: Prof. Jan Jürjens jan.juerjens@isst.fraunhofer.de
4	Universidad Rey Juan Carlos	UNIVERSIDAD REY JUAN CARLOS, Calle Tulipán s/n, 28933, Móstoles (Madrid), Spain. http://www.urjc.es	Contact: Prof. David Ríos Insua david.rios@urjc.es
5	1 4 9 5 UNIVERSITY OF ABERDEEN	THE UNIVERSITY COURT OF THE UNIVERSITY OF ABERDEEN, a Scottish charity (No. SC013683). King’s College Regent Walk, AB24 3FX, Aberdeen, United Kingdom http://www.abdn.ac.uk/	Contact: Dr Matthew Collinson matthew.collinson@abdn.ac.uk
6	TMB Transports Metropolitans de Barcelona	FERROCARRIL METROPOLITA DE BARCELONA SA, Carrer 60 Zona Franca, 21-23, 08040, Barcelona, Spain http://www.tmb.cat/ca/home	Contact: Michael Pellot mpellot@tmb.cat
7	AtoS	ATOS ORIGIN SOCIEDAD ANONIMA ESPANOLA, Calle Albarracin, 25, 28037, Madrid, Spain http://es.atos.net/es-es/	Contact: Alicia Garcia Medina alicia.garcia@atos.net
8	SECURENOK	SECURE-NOK AS, Professor Olav Hanssensvei, 7A, 4021, Stavanger, Norway Postadress: P.O. Box 8034, 4068, Stavanger, Norway http://www.securenok.com/	Contact: Siv Houmb sivhoumb@securenok.com
9	SOÚ Institute of Sociology AS CR	INSTITUTE OF SOCIOLOGY OF THE ACADEMY OF SCIENCES OF THE CZECH REPUBLIC PUBLIC RESEARCH INSTITUTION, Jilska 1, 11000, Praha 1, Czech Republic http://www.soc.cas.cz/	Contact: Dr. Zdenka Mansfeldová zdenka.mansfeldova@soc.cas.cz
10	nationalgrid THE POWER OF ACTION	NATIONAL GRID ELECTRICITY TRANSMISSION PLC, The Strand, 1-3, WC2N 5EH, London, United Kingdom http://www.nationalgrid.com/uk/	Contact: Dr. Ruprai Raminder raminder.ruprai@uk.ngrid.com
11	ANADOLU ÜNİVERSİTESİ	ANADOLU UNIVERSITY, SCHOOL OF CIVIL AVIATION Iki Eylul Kampusu, 26470, Eskisehir, Turkey http://www.anadolu.edu.tr/akademik/yo_svlhvc/	Contact: Nalan Ergun nergun@anadolu.edu.tr
12	Durham University	The Palatine Centre, Stockton Road, Durham, DH1 3LE, UK https://www.dur.ac.uk/	Contact: Prof. Julian Williams julian.williams@abdn.ac.uk

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	24/03/2014	Draft	D. Ríos, J. Cano (URJC)	ToC
0.2	28/05/2014	Draft	D. Ríos, J. Cano (URJC)	ToC + draft of sections
0.3	09/06/2014	Draft	D. Ríos, J. Cano (URJC), A. Couce, S. Houmb (SNOK)	ToC + full skeleton of sections
0.4	09/09/2014	Draft	D. Ríos, J. Cano (URJC)	Draft document body + annexes
1.0	21/11/2014	Draft	D. Ríos, J. Cano (URJC)	First complete version of document body
1.1	11/12/2014	Draft	D. Ríos, J. Cano, F. Ortega, E. López, J. M. Moguerza, A. Alonso (URJC)	Restructured main body
1.2	16/12/2014	Draft	D. Ríos, J. Cano, F. Ortega, E. López (URJC)	Consolidated main body
1.3	23/12/2014	Draft	E. Chiarani (UNITN)	Quality check completed. Minor changes
1.4	10/01/2015	Draft	J. Williams (Durham)	Scientific review completed. Minor changes
1.5	15/01/2015	Draft	A. Schmitz (ISST)	Input concerning computational integration within the tool
1.6	21/01/2015	Draft	A. Caranti (UNITN)	Second quality check completed
1.7	26/01/2015	Draft	F. Massacci (UNITN)	Final scientific review
2.0	29/01/2015	Final	D. Ríos, J. Cano, F. Ortega, E. López (URJC)	Final version ready to be submitted

Index

Executive summary

The main goal of this deliverable is to describe the general Adversarial Risk Analysis (ARA) methodology used in the SECONOMICS toolkit. This deliverable also addresses the issues and characteristics that are needed to model operational security problems for Critical Infrastructure Protection (CIP) in the real-world scenarios addressed in the project.

Some of these issues were already identified in the various case studies studied in *D5.2—Case Studies in Security Risk Analysis*, derived from the outcomes of SECONOMICS WP1, WP2 and WP3. The basic models used to solve these case studies needed to be expanded with *ad-hoc* modifications to accommodate the complexities posed by these new scenarios, as e.g. the presence of multiple risks simultaneously affecting several locations, among other advanced requirements. The general methodology proposed in this deliverable overcomes these shortcomings, presenting a rich framework to integrate additional dimensions that help us in reflecting the nuances of the underlying CIP problems in greater detail.

Specifically, this deliverable includes:

- A complete specification of a methodology to design general models based on ARA and its application to solve CIP problems.
- Design requirements for the development of tools implementing this methodology in different domains, which serves as an input for WP8-Tool Support.
- Two new case studies illustrating the application of this general methodology, along with its main advantages to identify and address future and emerging threats.

The main body of this document provides a high-level overview of the different aspects and factors that can be considered in this general ARA methodology to solve CIP problems. Besides, this document also includes several Annexes providing a more detailed and technical description of the core elements that enable such generalised approach (ANNEX1, ANNEX2 and ANNEX3), along with two case studies illustrating the application of the proposed methodology (ANNEX4 and ANNEX5). As a result, the main sections of this document body try to minimise as much as possible the use of mathematical, statistical and technical terminology and concepts, in order to provide an accessible description of the main scientific and technical contributions. In any case, multiple pointers to the Annexes are also provided in the corresponding sections for those readers interested in further theoretical details and practical technicalities.

1. Deliverable Scope and Structure

This deliverable describes a general methodology to design and implement Adversarial Risk Analysis (ARA) models to address Critical Infrastructure Protection (CIP) problems. It also provides an outline of requirements to devise tools implementing this model in practical settings.

The proposed solution described in this document builds on requirements and knowledge produced in WPs 1, 2 and 3. Likewise, this methodology stems from the outcomes of deliverables *D5.1—Basic Models for Security Risk Analysis* and *D5.2—Case Studies in Security Risk Analysis*. Such documents considered basic application settings and suggested the need for the general approach presented here to tackle more complex scenarios.

This document is organised as follows. Section 2 introduces the proposed methodology to build general ARA models and to address complex scenarios related to CIP problems. First, we briefly describe typical layouts and conditions defining these problems. They include structural aspects (like the existence of single or multiple objectives) as well as behavioural aspects of participants (such as coordination among attackers or defenders). Then, we outline the main steps to implement the proposed methodology, highlighting novel contributions and their impact to address emerging and future threats in different contexts. In Section 3 we summarise the main results and assessments obtained from the application of ARA models to address CIP problems in specific application domains, incorporating more advanced requirements.

Besides, we include several Annexes presenting in full detail the core elements of this general ARA methodology and its application to two case studies: ANNEX1 describes the modelling of CIP problems with multiple defenders and multiple attackers. ANNEX2 introduces the elements to model advanced strategic approaches that can be adopted by attackers. ANNEX3 provides additional details about Biagent Influence Diagrams (BAID), a graphical approach that help us formalised this type of problems. Finally, the two case studies concerning a cross-domain problem (Oil & Gas sector and cybersecurity) and a railway network in south Spain are developed in ANNEX4 and ANNEX5, respectively.

2. General Adversarial Risk Analysis Models

D5.3 describes a risk analysis methodology for security resource allocation in general CIP problems, involving complex interactions between participants. In these scenarios, one or multiple *defenders* will try to protect the utility of components of the critical infrastructure from multiple threats created by one or several adversaries, called *attackers*. The general risk analysis models derived from this methodology will let us formalise the characteristics of the critical infrastructure and the agents involved in the problem, as well as a variety of conditions that may exert a direct impact on their decision-making processes to achieve their respective goals.

The rationale for the design of this generalised risk analysis methodology stems from the previous work conducted in this WP, which was described in deliverables D5.1 and D5.2. In D5.1, five template models were presented to illustrate the application of simple risk analysis models to the resolution of security policy making problems. These models were founded on the ARA theoretical framework. For each model, we included a simple motivating example and a basic numerical illustration.

In turn, these basic models can be adopted as baseline building blocks to formalise and solve more general CIP risk analysis problems in different application domains, as it was illustrated in D5.2. In it, problems related to the airport (*D1.3—Airport Requirements final version*, from WP1) and metro (*D3.3—Urban Public Transport Requirements final version*, from WP3) case studies were formulated. Likewise, we also outlined the solution for the grid case study (*D2.3—National Grid Requirements final version*, from WP2). We used the ARA framework, adapting the basic templates as required to deal with the specific features and the inherent complexity of such case studies.

The proposed models could be applicable to other CIP problems with similar features or underlying structures. However, the case studies considered in D5.2 evidenced a number of modelling issues that would always imply the adaptation of this basic methodology to deal with the particular traits or structural features of new case studies. As a result, this caveat suggested us the need to create a generalised methodology, also founded on the ARA framework, that could be flexible enough as to accommodate more complex requirements in real-world scenarios. Additionally, this general methodology will be better suited to deal with any future emerging threat that might not have been initially considered in the definition and characterisation of the various problem scenarios.

Therefore, we briefly describe in this section our work to define a general ARA methodology to formalise and solve complex CIP problems. This methodology allows for more general conditions, such as more complex structures, advanced strategies adopted by participants or coordinated actions of attackers and/or defenders, that were not previously considered in our previous work. First, we break down the additional assumptions and conditions that can be incorporated in this generalised methodology for the resolution and assessment of CIP problems. Further details about this work can be found in ANNEX1, ANNEX2 and ANNEX3. Then, we detail the necessary steps to follow this methodology for the creation of models that can be applied to particular case studies.

2.1 Critical Infrastructure Protection problems

The goal of the taxonomy presented in this section is to offer a general picture of the different options and dimensions that can be considered in the formalisation of CIP problems. These options can exert an influence on how the different components of the CIP problem are modelled and interpreted in the ARA framework.

Topology and structure of the CIP

The first dimension to consider in the definition and formalisation of the CIP problem is the existence of any topological or structural characteristics of the underlying infrastructure that may be worth considering to create a model that better reflects the real situation to be assessed. In this case, the different options are:

- **No spatial or network structure**: The targets or components of the infrastructure neither have relevant physical or functional relationships among them nor they rely structurally on each other. Then, two possible alternatives can be considered:
  - *Single location*: We model the CI as a single location threaten by attackers. For instance, this corresponds to the case of the ATC Tower presented in Section 2 of D5.2.
  - *Multiple locations*: The CI is modelled as multiple locations which are not linked or related in any particular way among each other. For example, this pertains to the metro case study summarised in Section 3 of D5.2.
- **Spatial structure**: We considered the CI as composed of several targets which rely structurally on each other, so that an attack on one of such locations may increase the vulnerability of others depending on it. This alternative corresponds to the case of an urban space divided in neighbourhoods outlined in ANNEX5 of D5.1.
- **Network structure**: In this case, the underlying infrastructure is modelled as a network composed of nodes and links connecting them. Then, different approaches can be adopted, depending on where the value to be defended resides:
  - *Network with values at nodes only*: In some network problems, value is only attained at nodes. A typical example could be an underground transport system, where the value for different types of attackers is located only at the stations, as explained in Section 3 of D5.2.
  - *Network with values at nodes and links*: This is a generalisation of the previous case, in which value can be also encountered at links. Therefore, any attack or threat on one element may get transferred further to one or several other elements in the system, due to existing linkages. The work to address this advanced case is developed in complete detail in ANNEX4 of this deliverable.

- **Parallel systems**: The entire system is threatened only when all of its elements have been compromised. For instance, parallel computing systems typically consist of multiple nodes that perform similar operations. Among other benefits (like reducing execution time or load-balancing), these topologies can support failures in one or several nodes without compromising the entire system. ANNEX5 develops a complete case study on cybersecurity, in the context of the Oil & Gas business sector, considering this alternative.

Characterisation of defenders in CIP

- **Single defender**: In this case, the CI is protected by a single defender, as it occurs in all cases described in deliverables D5.1 and D5.2.

- **Multiple defenders**: ANNEX1 develops the case of multiple defenders who protect the CI, with two possible alternatives regarding the cooperation among them:

  - *Multiple uncoordinated defenders*: In this case, there is no coordination at all among defenders to protect the CI. It may be the case that each defender protects their own premises or locations, or that they all defend the same site or component. In ANNEX1, we provide details of such approach focusing on the case, in which the defenders first implement their actions and, then, the attacker, having observed them, performs his attack, which we called Sequential Defend-Attack problems.
  - *Multiple coordinated defenders*: Here, there exists a certain degree of coordination among all the defenders, usually in the form of resource sharing (technical or human resources, tactics, intelligence, etc.). As an example, we can consider the protection of an underground transport system against several threats, as specified in WP3, in which several private and (local, regional or national) government security bodies could be involved, sharing duties and responsibilities. An interesting issue here is to discern which benefits (if any) may the defenders obtain when coordinating their actions and sharing their resources, compared to the case in which they act uncoordinately. Further details about this work can be found in ANNEX1.

Besides the number and coordinated action of defenders, another dimension that can be considered in ARA models is the adoption of distinct **defensive strategies**, including: preventive strategies to minimise threats; recovery plans to minimise the impact of potential attacks; insurance plans to recover from potential losses, the use of false targets to confound attackers; separation of underlying system elements to reduce coupling and dependencies; addition of redundant components to improve resiliency and availability; the adoption of multilevel defence strategies, or the launch of preventive strikes to undermine the capacity and resources of potential attackers. Some of these strategies have been considered in examples and case studies previously presented in deliverables D5.1 and D5.2.

Characterisation of attackers in CIP

- **Single attacker**: In this case, the CI is threatened by a single attacker, as it occurs in most of the cases described in deliverables D5.1 and D5.2.
- **Multiple attackers**: In this type of problems, an organisation needs to protect from multiple threats.
  - *Multiple uncoordinated attackers*: We assume that the relevant multiple threats are uncoordinated, in the sense that different attackers do not make a common cause, although the outcome of different types of attacks might affect each other (see ANNEX1 for additional details).
  - *Multiple coordinated attackers*: The main difference with respect to the previous uncoordinated case is that the attackers will coordinately make their attacking decisions, as they are interested in common targets (again, see ANNEX1 for additional details).

Rationality level of agents

Inside the ARA framework, risks are derived from intentional actions of adversaries. Then, the analysis supports one of the decision-makers, who must forecast the actions of other agents. Typically, this forecast takes into account random consequences resulting from the set of selected actions. Therefore, to solve the problem we must model the behaviour of opponents, which entails strategic thinking.

ANNEX2 provides additional details about the work carried out to identify relevant options to model the strategic thinking of opponents in the proposed ARA (specially, in the generalised version which is the main focus of this deliverable). The different available options can be summarised as follows. Of course, combinations of different opponent models in the same CIP problem can also be considered.

- **Non-strategic participants**: The defender considers that the attacker does not follow a strategy, and thus acts randomly. Based on past data and/or expert opinion, the defender will elicit beliefs about the decision made by the attacker and deploy preventive measures, consequently.
- **Participants seeking Nash equilibrium**: In this case, the attacker and the defender are considered to have confronted each other many times before, and consequently they can anticipate their preferences, as well as the probabilities that the opponent selects specific actions according to them. We then compute the corresponding *Nash equilibrium* for each possible random scenario, which leads us to obtain the optimal actions that would be chosen by the attacker and their associated uncertainty. This result can be used to inform the decision-making process of the defender.
- **Participants with level-k thinking capacity**: In this alternative, the defender assumes that the attacker will select his action based upon a chain of reasoning of the form “I know that she knows that I know...”. The depth of this chain of reasoning may be of *k* levels, depending on how sophisticated the defender believes the attacker to be. For

example, if the defender is non-strategic, then she is a level-0 thinker and chooses randomly. If she chooses her action by assuming that the attacker is non-strategic, then she is a level-1 thinker. A level-2 defender assumes that the attacker is a level-1 thinker, who assumes she is a level-0 thinker, and so forth.

- **Participants seeking mirror equilibria**: As implied above, level-k thinking can lead to an infinite regress. However, we can overcome this problem if we assume that the defender has some information (represented by probability distributions) to model her own beliefs about the attacker intentions, along with information about the attacker beliefs regarding the defensive strategies. In that situation, the defender is able to develop a probabilistic model to predict the attacker's actions (further details about this approach can be found in ANNEX2).

- **Prospect maximising opponents**: There is abundant evidence showing that humans often make choices that do not maximise expected utility, but other type of individual or group prospects. In ANNEX2, we explain how to perform an ARA when the opponent maximises prospect functions, using prospect theory Wakker (2010).

2.2 Methodology outline

The main contribution reported in this document is the design and specification of a general ARA methodology suitable for capturing complex structural traits and behavioural patterns of agents involved in CIP problems.

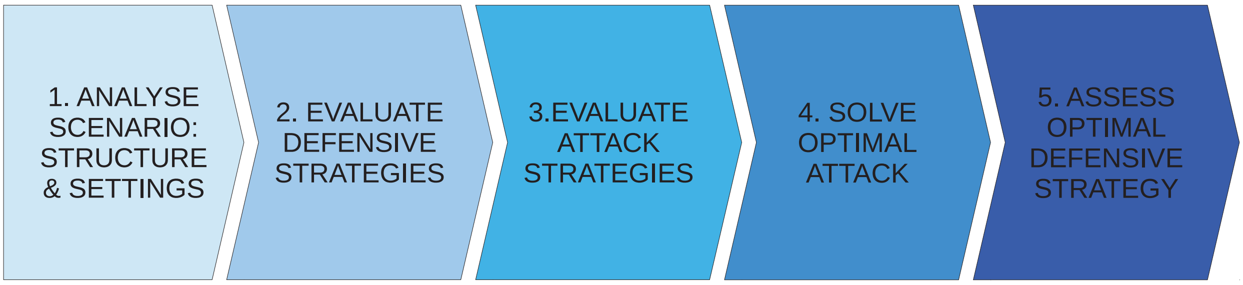

1. ANALYSE SCENARIO: STRUCTURE & SETTINGS
2. EVALUATE DEFENSIVE STRATEGIES
3.EVALUATE ATTACK STRATEGIES
4. SOLVE OPTIMAL ATTACK
5. ASSESS OPTIMAL DEFENSIVE STRATEGY

Figure 1: Overview of the main steps in the general ARA methodology for CIP.

Figure 1 shows the sequential procedure to follow this general ARA methodology to model complex CIP problems, comprising the following steps:

1. *Analyse the structure and characteristics of the underlying infrastructure*: The first step in our general methodology is to identify any relevant structural patterns and attributes describing the underlying infrastructure to be protected. As we have seen in the previous section, this includes identifying a single or multiple locations to be defended, any spatial or structural dependencies (such as a network structure) and defining which structural elements (e.g. nodes or links in a network) are valuable.

2. *Evaluate and model defensive agents and strategies*: After modelling the underlying infrastructure, the next step is to identify the number of defensive agents involved in the problem (single or multiple), as well as their likely defensive strategies. This includes the identification of any coordinated strategies (for example, to defend a common valuable asset), as well as defining a probabilistic model that reflects the defender's choices regarding each possible protective measure to be adopted. In the same way, the model can also accommodate more advanced rationality behaviours of defenders for whom we may assume the capacity of anticipating their attack intentions (as we already introduced in Section 2.1 above).

3. *Assess and model attackers and their likely strategies*: Like in the previous step, in this case we model the number of attacking agents (single or multiple), the options for coordinated actions against common targets, along with a probabilistic model to reflect the likely preferences of attackers to choose among their available strategies. Likewise, we can also assign more advance rationality levels to an attacker that may anticipate preventive measures deployed by the defender.

4. *Obtain the optimal attack options for opponents*: Once the initial set up of our model is completed, it is time to solve the ARA problem. To this aim, we select the most adequate template to model each attacker and site, include any additional uncertainties to better reflect the complexities of the real world scenario to be addressed and define any resource allocation constraints for the defender and the attacker. Finally, we assign the objectives and utilities for the defender and the attacker, which define their respective goals and steer their behaviour in the model.

   With these inputs, we can now solve the associated probabilistic model using Monte Carlo simulation to forecast the likely actions that will be pursued by the attacker (*random optimal attack*).

5. *Recommend best defensive options to counteract the attack*: The results from the previous sections will serve as an input for this final step. Once we have calculated the most likely actions that the attacker will perform, considering the characteristics and restrictions included in the ARA model, we can now assess the best strategies to be adopted by the defender to countermeasure the attack and attain optimal resource allocation.

   Once we obtain this initial result, it is advisable to perform a sensitivity analysis to better reflect the uncertainty associated to such results when reporting to the decision-makers. As well, we can also consider the alternative of sharing risks among defenders to further optimise the utilisation of available resources.

Thus, the main contributions of this general ARA model to solve more complex CIP problems can be summarised as follows.

- **Coordination among multiple attackers and multiple defenders**: As previously presented, one of the main advantages of this general ARA methodology is to allow for the explicit consideration of coordinated actions among defenders or attackers, who may decide to joint efforts on common goals or targets. ANNEX1 summarises the different available options and outlines algorithms to simulate these kind of problems in probabilistic computational models.

- **Considering advanced rationality types for attackers and defenders**: A contribution of key importance to address complex CIP problems is to integrate advanced strategic approaches that may be adopted by agents involved in the scenario, who try to anticipate the opponents decisions and their impact on their own plans, as shown above. In particular, the defender must account for some level of uncertainty about the decisions made by the attacker and also model the attacker's beliefs about defensive measures and strategies. Further details about this approach can be found in ANNEX2.

- **Allowing for more general interactions among different agents**: ARA is a decision-making methodology derived from *influence diagrams* Shachter (1986), a graphic representation to formalise the problem to be solved, in which we depict the different elements involved in the scenario and their relationships: agents, decisions, utilities and uncertain outcomes (e.g. the result of an hypothetical attack). The basic ARA models described in D5.1 and D5.2 can be generalised to allow for multiple dependencies between decisions taken by agents in the model and their consequences, using a more elaborated version of influence diagrams known as Biagent Influence Diagrams (BAID). ANNEX3 provides further details about this approach, as well as an example to illustrate its application.

2.3 Implementation guidelines

In order to be able to use the ARA methodology, present and discuss their inputs and output results with the stakeholders, as well as enable them to interact with the models, a more friendly interface seemed advisable. Therefore, the SECONOMICS Tool was implemented in WP8, whose integrated Tool framework was described in D8.4. The ARA models developed for the case studies in WPs 1 and 3 were implemented in Matlab and integrated within the SECONOMICS Tool. The Tool enables the user to enter the model parameters in a very intuitive way, presenting the obtained results in a graphical way together with a brief text description. In spite of the fact that modifying the model parameters is a straightforward task, adapting the models to arbitrary scenarios is a tougher process which would require a deep understanding of such models. The Tool, together with the integrated models, were evaluated at the SECONOMICS summit.

To conclude this section, we suggest a computational architecture to implement ARA models, that are structured in the project's Toolkit developed in WP8, following the general methodology proposed in this document. The Matlab code implementing all calculations underpinning this tool is also provided. The proposed architecture is composed of five main modules, see Figure 2, featuring the following functionalities:

- *Problem space*: We need to specify the relevant characteristics of the adversarial problem. Specifically, we need to define: (1) The problem topology, typically one among those discussed in Section 2.1; (2) Which ARA model will be used to address

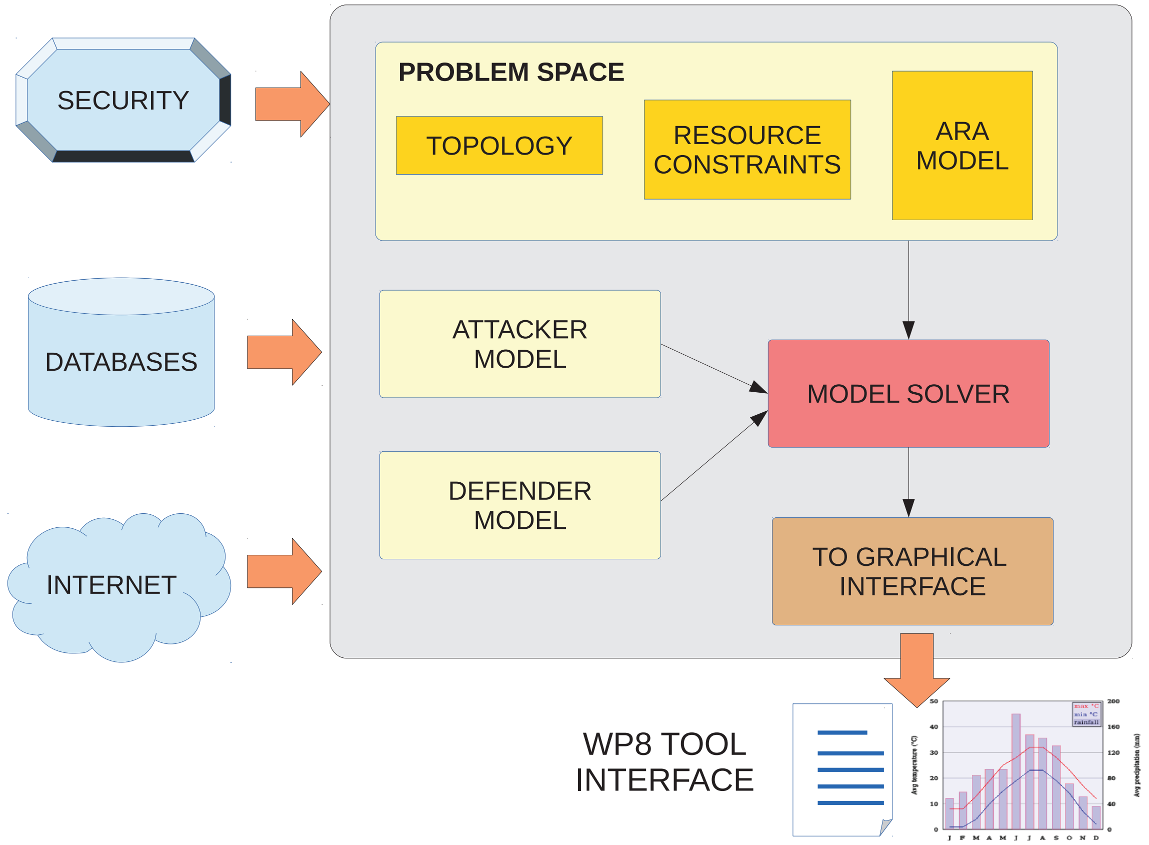


Figure 2: Proposed computational architecture to implement ARA models following the proposed general methodology.

the adversarial problem, either one of the basic templates or a more general model; and (3) The type and number of resources available to the defender, including their associated costs.

- *Attacker model*: In this module, we define the relevant quantities associated with the attacker's problem, such as objectives, probabilities, preferences and possible attacking strategies.
- *Defender model*: Here, we specify all quantities defining the defender's problem, such as probabilities, costs, consequences, preferences and utilities. Some of these inputs will come from the simulation of the attacker's problem, which must be specified, as well. Different defence strategies should also be selected and specified at this stage.
- *Model solver*. This is the central module, which will take the models for the underlying infrastructure, the attacker's problem and the defender's problem from the three previous modules, and will carry out the computational simulations to solve the probabilistic model. The results will be the optimal solution for the attacker's problem, and the optimal strategies and actions for the associated defender's problem, to counterbalance the attack.

- *Link to Graphical Interface (Toolkit)*. This module will provide the output to the Graphical Interface in the Toolkit developed in WP8, summarising the results and findings of the model, including numerical, graphical and textual information. This will facilitate the interpretation of these results by stakeholders and decision-makers.

3. Case Studies

In this section we offer a brief summary of the results and assessments derived from the application of ARA models to the resolution of different case studies representing a variety of CIP problems:

- The first two cases were solved in D5.2, using the basic ARA template models with ad-hoc modifications to address their peculiarities. As a result, the need for a general ARA model allowing the consideration of more general characteristics and problem conditions was clearly identified.
- The last two cases present the application of the generalised ARA model to examples in a cross-domain application and cybersecurity. These results illustrate implications of future and emerging threats that can be considered and modelled with this generalised approach. See ANNEX4 and ANNEX5 for further technical details about the work carried out in this regard.

Due to available time and resource constraints in the project, and given that the WP2 case study is strongly oriented to public policy aspects, the ARA case studies focus on those of WPs 1 and 3. Besides, the cross-domain case study features a unique combination of modelling requirements that makes it specially suitable to illustrate the capabilities of this general ARA methodology to address complex scenarios.

3.1 Airport case study

The first study presented in D5.2 illustrated the problem of protecting the ATC Tower of a small international airport from the assault of attackers who aim at hijacking the officers. The detailed requirements for this scenario were provided in D1.3 and *D1.4—Model Validation*.

The resolution of this problem was modelled using the Sequential Defend-Attack-Defend template introduced in Section 3.4 of D5.1. In this case, this template was applicable because we are considering the case of protecting the ATC Tower with defensive measures and then, in the event of a successful attack, we assume that a special unit of national authorities will be immediately called to take over the situation and try to reestablish the control of premises and liberating the personnel as soon as possible.

In this case, it was possible to find an optimal solution with the defender's problem within the available budget using modest computing resources (a complete summary of numerical results is available in Section 2.3 of D5.2). In addition, the model provides valuable insights about the likely strategies that could be followed by attackers, especially that they may tend to be cautious when additional protective measures are deployed, thus refusing to launch and attack or sending at most a single agent. However, we had to expand the basic Sequential Defend-Attack-Defend template to accommodate the additional complexities of this example, in particular:

- Modelling additional sources of uncertainty, as it would not be credible that the defender has precise information in advance about relevant factors such as the number of attackers, possible casualties or the total costs on the defender's side after an attack.

- The problem has a multiattribute nature, in that multiple relevant consequences of various types had to be considered, including economic, social (in extreme cases, even loss of human lives) and image consequences.
- In this example, the defender's decision to deploy additional protective measures was considered as fixed (deterministic) to simulate possible prospective scenarios. The model could be expanded to consider some uncertainty on the attacker's side when evaluating this decision.

The general ARA framework presented in Section 2.2 can integrate all these additional features in a more elaborated model.

3.2 Transport case study

The metro case study was specified in D3.3 and *D3.4—Model Validation*. This is an interesting example with two additional sources of complexity beyond the basic models: authorities must deal with more than one threat, and several premises can be affect by such threats. In this case the approach to solve this problem was to follow an incremental modelling:

- First solve the problem for one threat (unorganised fare evaders) and one location (single metro station).
- Then, repeat the model with a group of organised fare evaders in a single location.
- We join these two models, considering the case of both unorganised and organised fare evaders in one station.
- Finally, we expand the model to consider more than one threat (fare evaders and pickpockets) in a single station, and then extend this case for multiple stations.

Section 3 in D5.2 presents a complete summary of the setup of all aforementioned models, along with some numerical results. Further details for the final complex model can also be found in ANNEX3 of D5.2. In this example, the basic Sequential Defend-Attack template introduced in D5.1 was utilised to solve the problem. However, like in the previous case it had to be expanded to accommodate additional sources of complexity in this real case:

- The single uncertainty considered in the basic template is the result of the attack. Nonetheless, in real-world scenarios like this we must take into account additional factors, such as the fine imposed for fare evasion, the proportion of organised groups of fare evaders, and the potential dissuasive effect of protective measures deployed by the defender.
- Moreover, the basic template only allows us to consider a single threat, whereas in this case the ideal approach is to combine the two possible threats (fare evaders and pickpockets) in a single model.
- Finally, instead of considered a single location at risk, it is more desirable (and realistic) to take into account the simultaneous risk posed at different locations.

All these limitations can be addressed in a straightforward manner with the general ARA methodology proposed in this document. As a result, we can directly apply a model that can manage all these intricacies from the start, avoiding the need to implement ad-hoc expansions of the basic templates to tackle more advanced scenarios.

3.3 Cross-domain case study

In this example, we illustrate the applicability of the proposed general ARA methodology to directly solve complex CIP problems. The cross-domain case study presented in this example involves two apparently unrelated domains: the fossil fuel industry and cybersecurity.

Over the past years, the oil and gas (O&G) industry has progressively incorporated operational technology (OT) solutions, especially for the automation and control of offshore drilling premises. The benefits from the integration of OT and IT infrastructures in this business domain are clear, including the centralisation of oversight and decision-making processes, automating former manual mechanical activities, improving monitoring and telecommand with better, near real-time sensors and, as a result, global performance optimisation in their activities.

However, due to the gradual introduction of computational and networking resources these infrastructures has also become a very attractive target for cyber-attackers Shauk (2013), motivated by important economical and strategic interests that are at stake in this domain. ANNEX5 provides a detailed description of the many challenges faced by cybersecurity systems, as well as different methodologies to address them from distinct disciplines. In this example, we illustrate the application of our general ARA methodology to tackle this specific problem, putting special emphasis on the innovative contributions of this approach to guide evidence-based decision-making processes.

We present here an overview of the five different steps for the application of the general ARA methodology, introduced in Section 2.2. More complete details about this particular case study can be found in ANNEX5. In this example, the scope of the model is an assessment activity previous to the attack, providing assessment to underpinning incident handling plans.

In the first step, we model the spatial or structural characteristics defining the CI to be protected. In this case, the infrastructure is considered to be a single offshore drilling installation, in which OT and IT have already been deployed. Regarding the second step, we assume a single defender who will always be capable of detecting the attack, and will always respond to it. On the other side (step 3 in our methodology), the attacker do not represent a specific individual, but a generalisation of potential criminal organisations that represent business-oriented threats, guided mostly by monetary incentives. It is assumed that the attacker can commit a single attack, with several direct consequences for the CI to be defended, as well as several subsidiary consequences for the defender's goals depending on the risk treatment strategy that is finally selected. The combination of all these factors define the defender and attacker utility functions, that they will seek to maximise.

The flexibility of the proposed general ARA methodology allows us to consider, from the defender's point of view, all decisions that will be taken by the attacker as uncertain choices, which we can represent using a probabilistic model (e.g. based on information about previous attacks or input from experts in cybersecurity). The model is also general enough as to consider a longer sequence of attack and defensive actions intertwined among each other, should that be a case of interest in the future. We refer readers interested in the complete details describing the definition of the ARA model to consult ANNEX5 for additional information.

Finally, once the set up of the model is finished, we can proceed with steps 4 and 5 in the proposed methodology. We then solve the defender's problem by first simulating the attacker's problem to evaluate the likely attack actions that may be selected and their possible consequences. This input is then used to conclude the best recommendations to optimise the defender's protective plans. Again, the complete list of tables summarising the numerical results for this case study can be found in ANNEX5. The resulting assessments concerning the defender's strategy can be summarised as follows:

- The attacker decision will be strongly influenced by the defender's strategy. The attackers are more likely to commit an attack if they think the defender will accept the risk (assuming all possible consequences of the attack) instead of sharing the risk (buying insurance) or avoiding the risk at all (stop drilling operations). In general, perpetrating an attack is more attractive in case the attacker strongly believes that the defender is going to accept the risk or is going to continue drilling.

- If the defender thinks an attack will happen, then she would prefer sharing the risk (with an insurance service) and stop drilling after the incident. In case she believes that there will be no attack, she should accept the risk and continue drilling to maximise her utility. Accepting the risk in case of no attack is better than sharing the risk, but accepting the risk in case of attack is worse for the defender's interests.

In the near future, it is very likely that cyber-attackers will soon target several fuel production premises simultaneously. The motivations for these attacks may vary from undermining production capacity to accessing sensitive data or strategic information that other malicious agents with interest in this market could use in their own benefit. In response to these new emerging risks, the generalised ARA methodology presented here can provide a invaluable tool to address this challenge, as it can seamlessly incorporate complex structural features and dependencies characterising the underlying CI, as we briefly show in the following section.

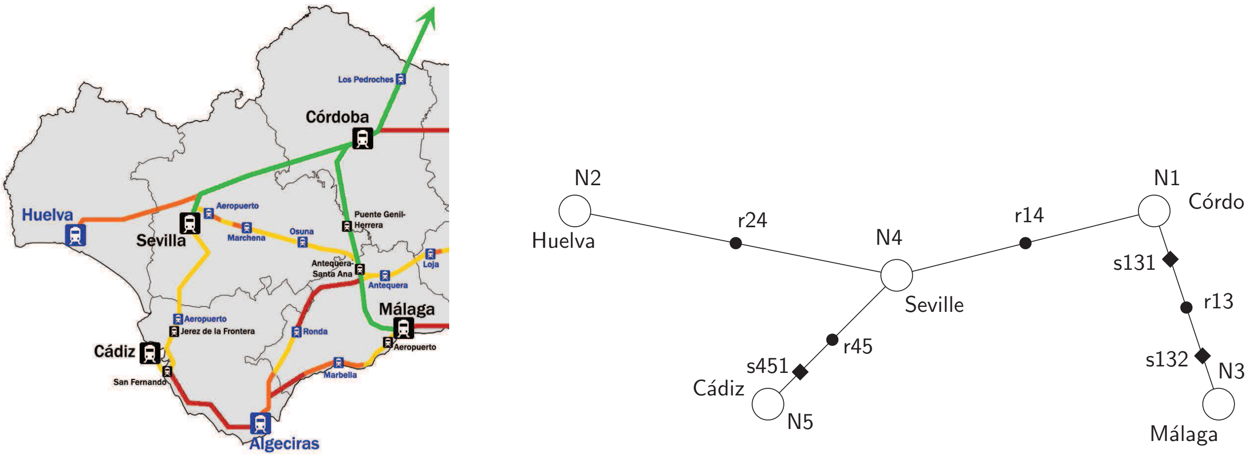
3.4 Addressing future and emerging threats

National security, and more specifically the oversight of critical infrastructures such as transport, power grids or telecommunication networks has been a key concern for governments and organisations around the world. Among all possible risks that may affect these infrastructures, terrorist attacks constitute one of the most worrisome threats for national and federal authorities. An important consequence derived from this serious concern has been to improve national security plans, including significant investments in protective responses Haberfeld and von Hassell (2009).

This CIP problem represents a paradigmatic example of scenarios in which new risks and threats may emerge in due course. Therefore, authorities and organisations in charge of national security must be prepared to address these challenges. The general ARA methodology that we have introduced in Section 2.2 offers an adaptive and flexible tool to model this kind of situations, providing valuable assessment to optimise the allocation of resources to prevent possible threats while minimising costs for organisations implicated in the deployment of the defensive plan.

In this case study, we consider the protection of the southwest section of the Spanish railway system against terrorist threats. Recent intelligence reports have alerted about the activation of a dormant cell established in Seville, integrated within the city for years without raising suspicion. In this regard, several Al-Qaeda members have been arrested in southern Spanish towns over the last years (BBC News Europe, 2011, 2012; New York Post, 2014). In this example, we consider the case in which terrorists intend to launch an attack in summer against the railway system and its users, taking advantage of large population flows during the vacation period along the Andalusian coast.

We summarise here the application of the general ARA methodology for this case study. As we showed in Figure 1, the first step is to study the spatial and structural characteristics of our problem scenario. Figure 3 depicts the section of the Spanish railway system considered for this case study. Figure 3a represents the actual map of the railway subnetwork that will be the setting for our CIP problem, whereas Figure 3b provides a more schematic representation for it. In this scheme, stations are represented as *N*, routes as *r* and critical points to be protected as *s*. More precisely, the following sensitive areas are considered: *s*131 represents Puente Genil's viaduct in the Córdoba-Málaga route; *s*132 represents a tunnel in Antequera within the Córdoba-Málaga route, and *s*451 stands for another tunnel in Jerez de la Frontera, along the Seville-Cádiz route.



(a) Network map.

(b) Network scheme.

Figure 3: Railway network for the case study.

The second step is to proceed with the evaluation of possible defensive strategies. The railway operator is a public company attached to the Spanish Ministry of Public Works, and it is responsible for the management and security of the infrastructure. It must follow a security budget that can be assigned to the allocation of defensive resources along the network, according to certain constraints. Table 1 summarises the main characteristics of available security measures, along with their estimated unit costs. For the security staff, we have provided their unit monthly gross salaries. Further details in this regard can be found in ANNEX4.

Table 1: Deployment and type of security measures

	Measure	Station	Track	Critical points	Type	Unit costs (€)
x_1	Metal detector	Yes	No	No	Static	6,500
x_2	CCTV camera	Yes	No	Yes	Static	650
x_3	Lamp posts	No	No	Yes	Static	3,000
x_4	Fence units	No	No	Yes	Static	4,200
x_5	Security guard	Yes	No	Yes	Mobile	2,600
x_6	Detection dog	Yes	No	Yes	Mobile	800
x_7	Helicopter	No	Yes	No	Mobile	90,000

Mobile defensive measures can be used for recovery purposes in the event of a successful attack. By ’recovery purposes’ we mean solely the detention of terrorists, thus not considering those protocols that the government and the railway operator should carry out after an attack. We shall typically use Monte Carlo methods to carry out the computations needed to solve the defender’s problem.

The third step is to characterise possible attack strategies. Initially, terrorists could attack any point in the network, trying to generate chaos, damage network elements and cause the largest number of casualties. Attacks against railway targets entailing large number of casualties have taken place all over the world since the beginning of the 21st century, as shown in Table 2.

Table 2: Terrorist attacks on rail transport. Source: Haberfeld and von Hassell (2009).

Date	Country	Casualties
2001/8/1	Angola	> 250
2004/3/11	Madrid, Spain	> 190
2005/7/7	London, UK	> 50
2006/7/11	Mumbai, India	> 180
2007/2/17	Pakistan	> 60

However, to simplify this example we consider that the attacker’s strategy can only comprise several options, summarised in Table 3.

Table 3: Type of attacks and their consequences

Attack	Description	Lives	Fixed assets	Station	Train	Reputation
a_1	Bomb in station	Yes	Yes	Yes	—	Yes
a_2	Bomb in train (station)	Yes	Yes	Yes	Yes	Yes
a_3	Bomb en route	Yes	Yes	—	Yes	Yes
a_4	WMD in train	Yes	—	—	—	Yes

Finally, we can solve the adversarial problem dealing with the defender’s problem, moving to the attacker’s problem when required to calculate the information about his probabilities and preferences that will be used as an input for the defender’s problem model. This corresponds to steps 4 and 5 in our general ARA methodology. To perform these simulations, we must consider all pairs of feasible combinations—those fulfilling the resource allocation constraints—and the attack strategies described in full detail in ANNEX4. Should too many combinations need to be considered, we could rely on alternative implementations, as e.g. genetic algorithms, see Goldberg (1989).

In this way, we have demonstrated how this general ARA methodology can address future and emerging threats applied to the specific domain of national security, choosing the best protective strategies against terrorist attacks who target a railway network. In this case, the defender's aims are: (a) Deter terrorists; (b) Minimise their chances of succeeding in their attack; and (c) Reduce as much as possible the impact of a hypothetically successful attack.

Likewise, additional complications could also be incorporated to this procedure, in order to consider additional threatening dimensions. For instance, in this example we have explicitly disregarded possible cascading effects resulting from terrorist actions, since we have considered that the impact on one target will not propagate along the network. However, in other types of CIP protection scenarios (e.g. communication or energy networks), these conditions may not hold and we could apply more general approaches to address this complexity, as suggested in Salmeron et al. (2004) or Holmgren (2006).

4. Conclusions

In this deliverable we have presented a generalised ARA methodology to create models that better reflect the characteristics and traits found in real-world CIP problems. This general methodology is derived from conclusions of the application of the basic ARA templates, introduced in D5.1, to solve the case studies presented in D5.2. The complex requirements exhibited by these case studies, compiled in WP1, WP2 and WP3, demanded an ad-hoc expansion of the basic ARA templates to accommodate certain advanced conditions (like multiple uncertain factors, multiple threats to be considered simultaneously or multiple premises to be protected from threats).

As a consequence, the development of this general methodology offers a richer and more flexible framework to undertake the analysis of CIP problems entailing advanced features, such as:

- Integrating spatial and structural characteristics of the underlying critical infrastructure to be modelled, as well as the identification of valuable spots and premises, or redundancy elements to improve resiliency against attacks.
- Consideration of multiple defenders and multiple attackers, who share interest in protecting or damaging the same targets and are willing to coordinate their actions and share risks.
- Modelling advanced strategic capacities of defenders and attackers, including the capacity to try and anticipate the opponent's decisions.

Two new case studies introduced in Sections 3.3 and 3.4 show the applicability of this general methodology to solve CIP problems involving complex scenarios. In the same way, we illustrate the capabilities of this methodology to address additional emerging threats that might be identified in future analyses, without implying any disruption or additional modification in the modelling process.

Finally, implementation guidelines are also provided in Section 2.3 to facilitate the development and integration of the proposed model in the software tools created in WP8. Therefore, this general methodology constitutes the main outcome of WP5 for the SECONOMICS project, paving the way for the assessment of decision-makers facing the challenge of optimising the allocation of available resources for the protection of critical infrastructures in a wide variety of circumstances.

Table of acronyms

Acronym	Description
ARA	Adversarial Risk Analysis
APT	Advanced Persistent Threat
ATC	Air Traffic Control
BAID	Biagent Influence Diagrams
CCTV	Close Circuit Television
CI	Critical Infrastructure
CIP	Critical Infrastructure Protection
CNI	Critical Network Infrastructure
CS	Control System
IT	Information Technology
MAID	Multi-Agent Influence Diagrams
O&G	Oil & Gas
OT	Operational Technology
WMD	Weapon of Mass Destruction

Glossary

Attacker (He): A participant willing to perform disruptive actions to damage the utility of components of a CI.

Critical Network Infrastructure: A Critical Infrastructure composed of nodes and links connecting the nodes.

Defender (She): A participant adopting protective measures to preserve the utility of components of a CI, usually constrained by resource allocation restrictions.

Influence diagram: A formal graphical representation of a decision-making process in a compact schema, illustrating the agents involved in the process, decisions made by participants, potential outcomes derived from such decisions and utility functions.

Level-k thinking opponent: A participant (attacker or defender) who is assumed to anticipate the strategy that her adversary will adopt to try to fulfil her goals. The *k* level refers to the depth of the strategic plan, i.e. number of attack/defence actions anticipated by such participant.

Link: In the context of CNI, a link is any physical or functional connection between two nodes in the network. These connections may or may not have a utility for the defender.

Monte Carlo simulation: A simulation technique to solve probabilistic models, generating random samples from the target probability distribution that is the outcome of the model in a computationally efficient way.

Nash equilibrium: A formal rule to describe the behaviour of two or more players in a noncooperative game. In such a context, participants reach *Nash equilibrium* when each one knows the optimal strategy of all other players, and none of them can benefit from changing their strategy while the other players keep their strategies unaltered.

Node: In the context of CNI, a node is any physical or functional asset connected to other assets in the infrastructure layout by means of one or more links.

Non-strategic player: A participant in a game who does not perform rational actions based on a strategic plan. This corresponds to a *level-0 thinking opponent*.

Optimal CIP: Best strategic combination of defensive decisions that a defender can make to minimise the damage on the utility of the CI assets, according to resource allocation restrictions.

Random optimal attack: Best offensive action that an attacker could perpetrate on a CI, assuming that he does not follow a strategic plan.

Resource allocation constraints: Set of limitations and restrictions that attackers and defenders must fulfil when using available resources to meet their corresponding goals.

Sequential model: In ARA, a model that considers a strict order and number of actions performed by participants in a specific scenario.

Strategic player: A participant in a game who performs rational actions according to a preconceived strategic plan. This corresponds to a *level-k thinking opponent*.

Threat: Any risk represented for the utility of any component in a CI, generated as a result of disrupting actions that may be undertaken by an attacker.

Utility (function): The value of components of a CI as perceived by participants in an ARA model.

BIBLIOGRAPHY

BBC News Europe. Spain arrests al-Qaeda in Islamic Maghreb suspect. Press Release, http://www.bbc.com/news/world-europe-14563948, August 2011.

BBC News Europe. ’Al-Qaeda trio’ arrested in southern Spanish towns. Press Release, http://www.bbc.com/news/world-europe-19091753, August 2012.

D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

R. Haberfeld and A. von Hassell. *A New Understanding of Terrorism: Case Studies, Trajectories and Lessons Learned*. Humanities, Social Sciences and Law. Springer, 2009.

Å. J. Holmgren. Using graph models to analyze the vulnerability of electric power networks. *Risk Analysis*, 26(4):955–969, 2006.

New York Post. Spain, Morocco arrest 9 in ISIS terror cell. Press Release, http://nypost.com/2014/09/26/spain-morocco-arrest-9-in-isis-terror-cell/, September 2014.

J. Salmeron, K. Wood, and R. Baldick. Analysis of electric grid security under terrorist threat. *IEEE Transactions on Power Systems*, 19(2):905–912, 2004.

R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

Z. Shauk. Hackers hit energy companies more than others. `http://fuelfix.com/blog/2013/03/25/electronic-attacks-hit-two-thirds-of-energy-companies-in-study/`, 2013.

P. P. Wakker. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press Cambridge, 2010.

WP5 METHODOLOGY PAPERS

This document presents a compendium of methodology papers produced by URJC as a complement to the work accomplished in the deliverable *D5.3—General Methods for Security Risk Analysis.* The proposed methodologies are aimed at providing a general and flexible framework that could be applicable to generic CIP problems, also addressing future and emerging threats.

TABLE OF CONTENTS

Introduction to the General Methods for Security Risk Analysis Papers

In D5.3[1] we provide a risk analysis methodology for security resource allocation in general critical infrastructure protection (CIP) problems, with complex interactions between the intervening participants, who aim at adversarial objectives. We provide security prescriptions for one or more of the participants, generally called (She) the Defender, in their effort to defend themselves against multiple threats created by the adversaries called (He) the Attacker. Part of these attackers may coordinate their actions to attain better results. We consider that valuable targets may be distributed across various locations. Both the defenders and the attackers may dynamically reallocate, if necessary, part of their resources over different targets. We further assume that all participants have limited resources and are subject to other general constraints as e.g. political, economical, logistic, legal, etc.

The methodology developed in D5.3 can be regarded as a generalisation of that introduced in D5.1 and D5.2. D5.1 provided five Adversarial Risk Analysis (ARA) templates for security policy making. For each model, we included a simple motivating example and a basic numerical illustration. Those models were suggested as basic building blocks for general CIP risk analysis problems, as illustrated in D5.2. We formulated there several incumbent problems of the airport and metro case studies, from WP1 and WP3, respectively, also outlying the solution for the grid case study in WP2. We used the ARA framework, adapting the basic templates as required to deal with the specific features and the inherent complexity of such case studies. The proposed methodologies could be applicable to other CIP problems with similar features and/or underlying structures. However, such case studies suggested a number of issues that require generalisations. Furthermore, the models developed are flexible enough to allow their extendability to cope with future and emerging threats as e.g., the occurrence of cyberattacks in the airport domain or in the Gas & Oil sector. Such extensions would possibly require additional modelling advancements to support potentially novel features in the model as, e.g., the assessment of certain parameters, probabilities, preferences or defence and attack strategies.

In essence, as described in the deliverable body, the general approach proposed here consists of deploying one of the previous models over each of the targets within the critical infrastructure, relating them through resource constraints and value aggregation. From a procedural point of view, the methodology depicted graphically in the deliverable body could be implemented according to the following steps:

1. Problem structure.

   i. Choose the underlying CIP topological structure.

   ii. Determine the number of defenders and their eventual coordination.

   iii. Determine the defensive strategies available.

   iv. Determine the number of attackers, their rationality and eventual coordination.

   v. Determine the relevant ARA template model for each attacker and site.

   vi. Expand each of the templates for additional uncertainties.

   vii. Define resource constraints for the Defender(s) and the Attacker(s).

[1] Contributions provided by D. Ríos, J. Cano, F. Ortega, E. L. Cano, J. M. Moguerza, A. Alonso (URJC), A. Couce, S. Houmb (SNOK). A. Schmitz (ISST) provided input concerning computational integration within the tool. J. Williams (Durham) revised the document technically. Discussions with J. Ríos (IBM Research) and D. Banks (Duke) are gratefully acknowledged.

2. Problem assessments.

    i. Assess the Defender’s objectives, probabilities and utilities.

    ii. Assess the Attacker’s objectives, (random) probabilities and (random) utilities.

3. Problem evaluation

    i. Simulate from the Attacker’s problem to forecast his actions.

    ii. Optimise the Defender’s problem to obtain the optimal resource allocation.

    iii. Perform sensitivity analysis.

    iv. If necessary, share risks among defenders.

The general methodology required the following developments outlined in the first three appendices:

- Annex 1. *Enhancement of Sequential Defend-Attack Models*. Considers the possibility that several Attackers (with various degrees of coordination) and/or several Defenders (also with various degrees of coordination) and/or several targets (with various topological layouts) are present in the CIP problem at hand. We deal with each of these issues one at a time, stemming from the basic sequential Defend-Attack template. Combinations of the three themes are based on the previous ideas.

- Annex 2. *Modelling Opponents in Adversarial Risk Analysis*. Considers that the Attackers may have a different rationality to that entailed by the expected utility paradigm. We consider random attackers, Nash equilibria seeking attackers, level-*k* thinkers, mirror equilibria seeking attackers and prospect maximisers. We also consider uncertainty about such paradigms through a model mixture approach. We use the basic simultaneous Defend-Attack template as starting point.

- Annex 3. *Adversarial Risk Analysis for Biagent Influence Diagrams*. Considers that the Attackers and Defenders engage in much more involved interactions than those in the five basic templates, possibly across several time periods. We define a class of general interactions through Biagent Influence Diagrams and describe how to handle such problems, using relevance concepts and showing that, indeed as forecasted in Deliverables 5.1 and 5.2, we may use our five templates as basic bricks for general security risk models.

The outlined methodology is then tested in two additional annexes which complete the range of case studies from Deliverable 5.2.

- Annex 4. *Optimal CIP with Network Structure* The general methodology is applied to a problem with network structure, specifically to railway counterterrorism. The elements to be protected are nodes, links and what we call link hotspots. Different types of resources and ARA models are used at different elements. Beyond illustrating the feasibility of the general methodology, we deal with a major threat for an essential public infrastructure and we show how to deal with networked infrastructures as in WP2.

- Annex 5. *A Graphical Adversarial Risk Analysis Model for Oil & Gas Cybersecurity*. The general methodology is applied to a cybercontrolled critical infrastructure. Several defense stages (including insurance) and several attackers are considered. Beyond illustrating the feasibility of the general methodology, we address important threats over critical infrastructures for the Oil & Gas energy sector, which entails unusual settings and requirements, and illustrate the multistage and multiattacker aspects of our approach.

With these, we have provided the suggested general methodology for CIP risk analysis, focusing on the defence resource allocation strategic-tactical problem. The methodology essentially goes through three stages of problem structuring, problem assessment and problem evaluation, and aims at supporting a group of defenders in deciding optimal defences when protecting a CIP against one or more attackers.

All in all, we have provided an innovative, rigorous and powerful methodology for security resource allocation in CIP. The methodology nicely integrate with the rest of SECONOMICS technical WPs as follows:

- WP4 would provide advice about the objectives of Attackers and Defenders in their interaction, as well as about the Defender concerns and risk perceptions. These would be crucial elements in the (random) utility model of the Attacker and the utility model of the Defender.

- WP6, of a more strategic nature, would be used to decide the security budget available as well as an orientation about the effectiveness of various countermeasures.

- WP8 would implement the models developed here, as it already does with the most basic ones. As described in the deliverable body, we suggest a computational architecture to implement ARA models, that are structured in the project's Toolkit developed in WP8, following the general methodology proposed in this document. The proposed architecture is composed of five main modules: (1) *Problem space*, specifying the relevant characteristics of the adversarial problem; (2) *Attacker model*, defining the relevant quantities associated with the attacker's problem; (3) *Defender model*, similarly for the defender's problem; (4) *Model solver*, the central module, carrying out the computational simulations to solve the probabilistic model; and (5) *Link to Graphical Interface (Toolkit)*, providing the output to the Graphical Interface in the Toolkit developed in WP8, summarising the results and findings of the model. WP5 also contributed with the tool tuning, specifying and explaining the meaning and suggested values of the model parameters implemented. WP5 implemented Matlab code for the various models developed throughout the SECONOMICS project, as e.g. the basic template models introduced in D5.1; a model for the protection of the ATC Tower in the airport case study; or models for the simultaneous protection against pickpocketing and fare evasion in the metro case study in D5.2.

The cases performed within WP1, WP2 and WP3, and others beyond, show that ARA indeed provide a powerful and well founded approach for security resource allocation.

Enhancements of Sequential Defend Attack Models

Adversarial Risk Analysis has been introduced as a framework to deal with risks derived from the intentional actions of adversaries. The typical use is in security resource allocation. The analysis supports one of the decision makers, which we designate the Defender, who must forecast the actions of the other agents, taking account of random consequences resulting from the set of selected actions. Stemming from the basic Sequential Defend-Attack template model, we provide several variations which bring further realism to it. First we consider the case in which the Defender needs to face several attackers, which might be coordinated or not. We then consider the case in which there are several Defenders facing a single Attacker. Multiple attacker vs multiple defender problems may be seen in the light of earlier approaches. Finally, we consider the case in which several targets need to be protected.

1 Introduction

Recent applications in counterterrorism, cybersecurity, auctions and competitive marketing are driving renewed interest in developing practical tools and theory for analysing the strategic calculation of intelligent opponents who must act in scenarios with random outcomes. We use the term Adversarial Risk Analysis (ARA) to describe approaches in which the solution is based upon an explicit Bayesian model of the capabilities, probabilities and utilities used by the opponent in his analysis. For various concepts, methods and applications see Ríos Insua et al. (2009), Banks et al. (2011), Ríos and Ríos Insua (2012), and Rázuri et al. (2013).

The aim is to support one of the players who will use a decision analytic approach to solve her decision-making problem. To this end, she needs to forecast the actions of the other agents and, based on her own choice, the outcomes which she and her opponents will receive. This can be viewed as a Bayesian approach to game theory, and was proposed, non-constructively, by Kadane and Larkey (1982), Raiffa (1982) and Raiffa et al. (2002). The approach has been criticised by Harsanyi (1982) and Myerson (1991), among others. From a practical standpoint, the main obstacle in implementing this approach to conflict situations has been the lack of explicit mechanisms which allow the supported decision maker to encode her subjective probabilities about all components in her opponents' decision making. In earlier work, we have focused on relatively simple models, which serve as templates for complex models. Stemming from them, we explore here a number of variations which may bring further model realism: the presence of several attackers and/or several and/or defenders and/or several targets to be protected.

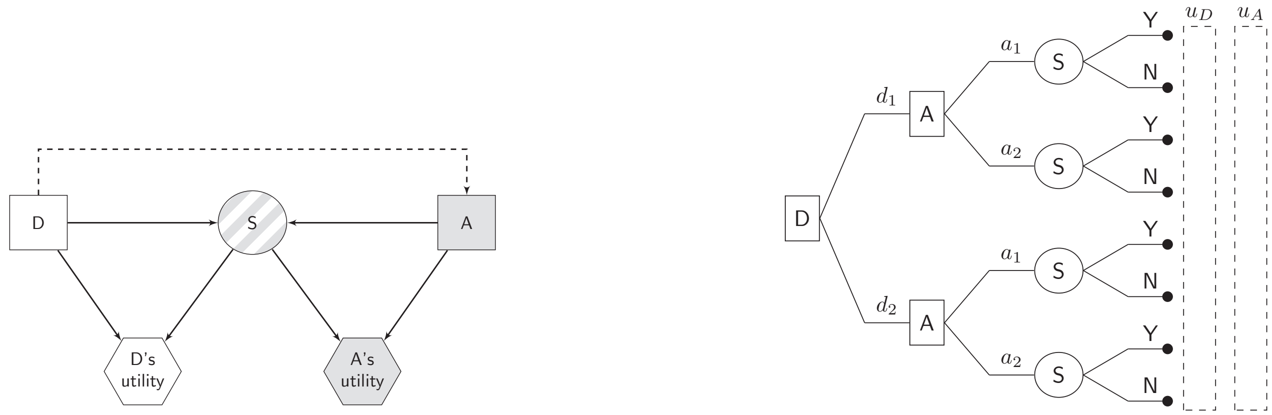
We choose the Sequential Defend-Attack Model as our initial template, which we outline in Section 2. Then, we address in Section 3 the case in which we need to protect one defender from multiple attackers. We distinguish between the cases in which the attackers are coordinated or not. We then consider in Section 4 the case in which several defenders face a single attacker, again distinguishing between coordinated and uncoordinated defenders. Multiple attacker *vs* multiple defender problems may be seen in the light of earlier approaches. We finally describe in Section 5 the case in which several targets need to be protected. We end up with a discussion.

2 The Sequential Defend-Attack Model

We start by considering the Sequential Defend-Attack model, to which we shall add complexities in stages. The Defender first chooses a defense and, then, having observed it, the Attacker chooses an attack. This corresponds to a Stackelberg game, see Aliprantis and Chakrabarti (2000), and have been studied in detail in the security domain from a classical game-theoretic perspective by Bier and

Azaiez (2009), and Brown et al. (2006). To simplify the discussion, we assume that the Defender (she) has a discrete set of possible defenses $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$ from which she must choose one. Similarly, the Attacker (he) has his set of possible attacks $\mathcal{A} = \{a_1, a_2, \dots, a_k\}$ to choose one from. We shall also simplify the problem by assuming that the only uncertainty deemed relevant is a binary outcome $S \in \{0, 1\}$ representing the failure or success of the attack. Finally, for both adversaries, the consequences depend on the success of this attack and their own action.

Figure 1 depicts the problem graphically. On one hand it shows a coupled influence diagram, an influence diagram for each participant with a shared uncertain node and a linking arrow. The influence diagram shows explicitly that the uncertainty associated with the success S of an attack is probabilistically dependent on the actions of both the Attacker and the Defender: $S|d, a$. Recall that arcs into a utility node represent functional dependence, see Shachter (1986). Thus, the utility functions over the consequences for the Defender and the Attacker are, respectively, $u_D(d, S)$ and $u_A(a, S)$. The arc in the influence diagram from the Defender's decision node to the Attacker's reflects that the Defender's choice is observed by the Attacker. We also show a game tree (with only two actions per adversary: $m = k = 2$) for the problem, reflecting its sequential nature. Note that there are two utility values, for the Attacker and the Defender, at the tree terminal nodes.



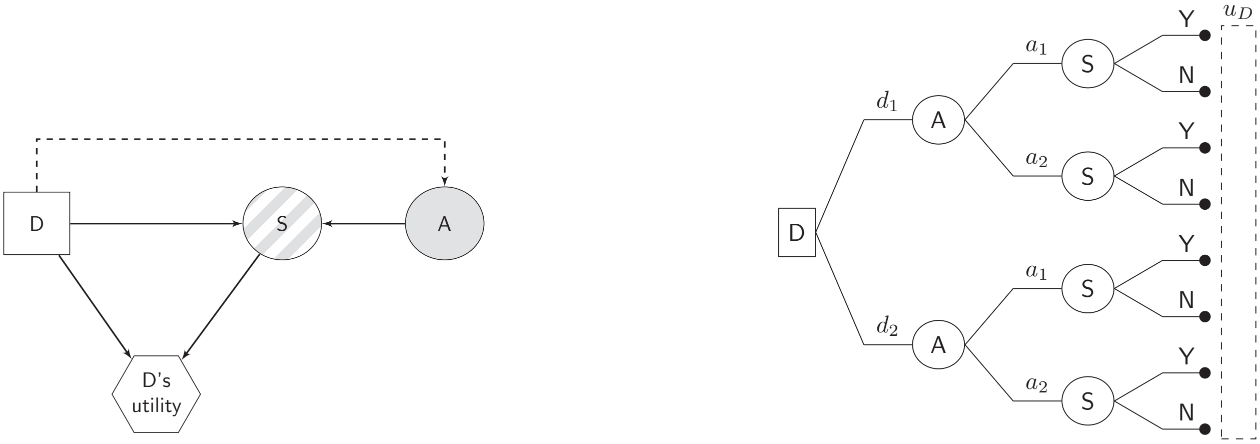
(a) Influence diagram

(b) Game tree ($m = n = 2$)

Figure 1: The Sequential Defend-Attack model

We weaken the common knowledge assumption required by the standard game theoretic approach: the Defender does not actually know (p_A, u_A), which model the beliefs and preferences of the Attacker. We consider the Defender's problem as a standard decision analysis problem: the Defender's influence diagram in Figure 2, no longer has the hexagonal utility node with the Attacker's information and his decision node is perceived as random variable. Similarly, her decision tree denotes uncertainty about the Attacker's decision by replacing $\boxed{\text{A}}$ with Ⓐ and including a reference only to the Defender's utility function. By looking at the influence diagram, note that in order to solve her decision problem, suppose the Defender has already assessed $p_D(S|d, a)$ and $u_D(d, S)$. She also needs $p_D(A|d)$, which is her assessment of the probability that the Attacker will choose attack a, after observing that the Defender has chosen defense d. This assessment requires the Defender to analyze the problem from the Attacker's perspective, possibly as we describe.

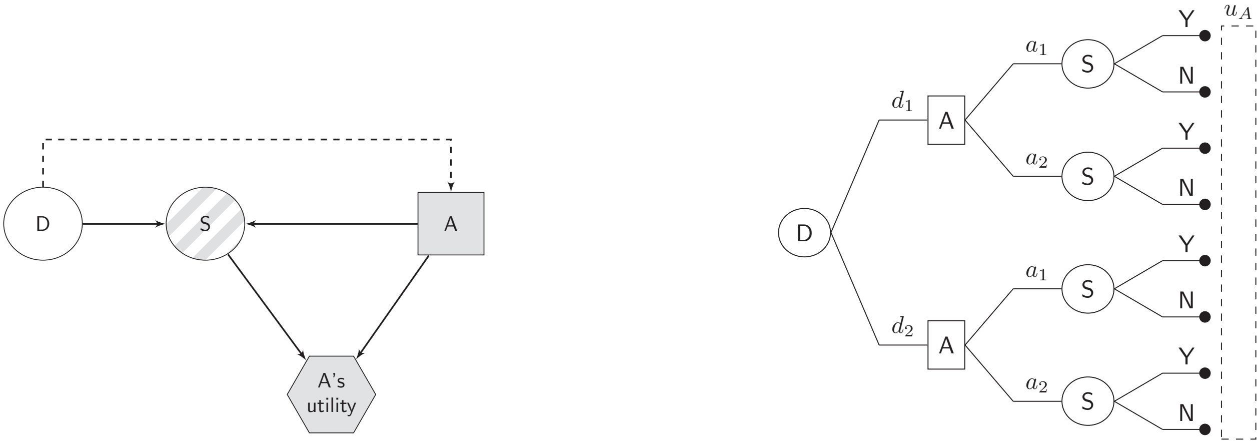
First, the Defender must place herself in the Attacker's shoes, and consider his decision problem. Figure 3 represents the Attacker's problem, as seen by the Defender. We assume that the Defender analyzes the Attacker's problem considering that he is an expected utility maximizer. Thus, she will use all the information and judgment available she can about the Attacker's utilities and probabilities.



(a) Influence diagram

(b) Decision tree ($m = n = 2$)

Figure 2: The Defender's decision problem



(a) Influence diagram

(b) Decision tree ($m = n = 2$)

Figure 3: The Defender's analysis of the Attacker's problem

To find $p_D(A|d)$, she should first estimate the Attacker's utility function and his probabilities about success S, conditional on (d, a), and consequently compute the required probability. However, instead of using point estimates for p_A and u_A to find the Attacker's optimal decision $a^*(d)$ as in the standard game-theoretic approach, the Defender's uncertainty about the Attacker's decision should derive from her uncertainty about the Attacker's (p_A, u_A), which we describe through a distribution F. This, in turn, will induce a distribution on the Attacker's expected utility $\psi_A(a, d)$. Thus, assuming the Attacker is an expected utility maximizer, the Defender's distribution about the Attacker's choice, given her defense choice d, is

$$p_D(A = a|d) = \mathbb{P}_F[a = \arg\max_{x \in \mathcal{A}} \Psi_A(d, x)], \quad \forall a \in \mathcal{A},$$

where

$$\Psi_A(d, a) = P_A(S = 0|d, a)U_A(a, S = 0) + P_A(S = 1|d, a)U_A(a, S = 1)$$

for $(P_A, U_A) \sim F$. She can use Monte Carlo simulation to approximate $p_D(A|d)$ by drawing n samples

$\left\{\left(p_A^i, u_A^i\right)\right\}_{i=1}^n$ from F, which produce $\{\psi_A^i\}_{i=1}^n \sim \Psi_A$, and approximating $p_D(A = a|d)$ by

$$\widehat{p}_D(A = a|d) = \frac{\#\{a = \arg\max_{x \in \mathcal{A}} \ \psi_A^i(d, x)\}}{n}, \quad \forall a \in \mathcal{A}.$$

Once the Defender has completed these assessments, she can solve her problem. Her expected utilities at node $\textcircled{S}$ in Figure 2 for each $(d, a) \in \mathcal{D} \times \mathcal{A}$ are

$$\psi_D(d, a) = p_D(S = 0|d, a)u_D(d, S = 0) + p_D(S = 1|d, a)u_D(d, S = 1).$$

Then, her estimated expected utilities at node $\textcircled{A}$ for each $d \in \mathcal{D}$ are

$$\widehat{\psi}_D(d) = \sum_{i=1}^{k} \psi_D(d, a_i)\widehat{p}_D(A = a_i|d).$$

Finally, her optimal decision is $d^* = \arg\max_{d \in \mathcal{D}} \ \widehat{\psi}_D(d)$.

Note that, in terms of classic game theory, the solution d^* for the sequential game need not correspond to a Nash equilibrium. Assume there would be a third party who knows the Defender's true (p_D, u_D) and her beliefs F about the Attacker's utilities and probabilities, as well as the Attacker's true (p_A, u_A) and his beliefs G about the Defender's. That party would then be able to predict the game, identifying the decisions chosen by each player. However, this omniscient prediction would not be the Nash equilibrium computed based on the true (p_D, u_D) and (p_A, u_A). Since the players lack full and common knowledge, their choices are unlikely to coincide with those made in the traditional game theory formulation.

This approach requires the assessment of $\left(U_A, P_A(s|d, a)\right)$. With respect to the random probabilities, we could base them on the corresponding assessments for the Defender, $p_D(s|d, a)$, possibly as follows:

- If S is discrete, $P_A(s|d, a)$ could be modeled as a Dirichlet distribution with mean $p_D(s|d, a)$ and variance accounting for the incumbent uncertainty. In particular, when S is binary, $P_A(s|d, a)$ could be modeled as a beta distribution.

- If S is continuous, then $P_A(s|d, a)$ could be a Dirichlet process with base distribution $p_D(s|d, a)$ and concentration parameter δ, expressing our uncertainty about such base, see Ferguson (1973).

In both cases, when lacking information, we could set a sufficiently large value for the variance or concentration parameter, respectively.

For the random utility model, the Defender must study whatever information she has about the aims of the attackers, see Keeney (2007), Keeney and von Winterfeldt (2011) or Keeney and von Winterfeldt (2010) for a detailed treatment of interests, values and objectives of terrorists. Based on their suggestion, we could view as reasonable a model based on a weighted measurable value function, as in Dyer and Sarin (1979). To take into account risk attitudes, we could appeal to the relative risk aversion concept, see Dyer and Sarin (1982), and assume risk proneness on the attackers. Finally, the uncertainty would be reflected by distributions over the weights and risk proneness coefficients. Wang and Bier (2013) provide another approach for assessing adversary preferences using ordinal judgments and the probabilistic inversion method, see Kraan and Bedford (2005).

3 Multiple attackers vs one defender

In the Sequential Defend-Attack model, and, more generally, in any adversarial situation, it is entirely plausible to face more than one Attacker, and these opponents may have different sets of resources, different goals and different degrees of cooperation. For example, governments must simultaneously defend against state-sponsored terrorism, franchise terrorists and solitary actors; similarly, police must defend against vandals, gangs, and organised crime. Within a corporate competition environment, a company may enter into a bidding against two or more competitors, or an organisation enter a marketing campaign to improve its market share. We distinguish between the cases in which the attackers are coordinated or not.

3.1 Uncoordinated attackers

We study first the case of a defender which faces several uncoordinated attackers. As an example, consider the urban police in a city which needs to face drug dealers, pickpockets, car thieves, house thieves and so on, and suppose that each class of delinquents operates in a manner uncoordinated with the others.

Although there are several variants of the problem, that shall be outlined below, to fix ideas we shall use the version illustrated by the multiagent influence diagram in Figure 4, see Koller and Milch (2003) for further details on MAIDs. It is a case in which various attacks have detrimental effect over the results for the defender. For example, in the motivating case, the police would need to use its limited human resources to face simultaneously all types of delinquents, with the ensuing detrimental effect.

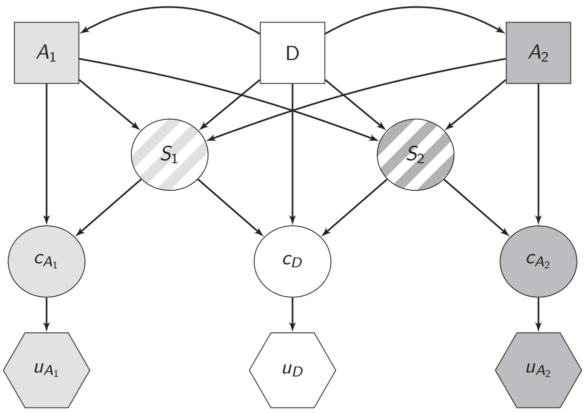


Figure 4: Multiagent influence diagram for a bithreat problem.

We, thus, consider a Defender, D, who needs to deploy defensive resources $d \in \mathcal{D}$ to face m uncoordinated attackers $A_1, \dots, A_m$. These observe her decision, and, respectively, make attacking decisions $a_i \in \mathcal{A}_i$, $i = 1, \dots, m$. The interaction between D and the A_i's through their respective decisions d and a_i, leads to random results $S_i \in \mathcal{S}_i$, which depend on all decisions. The Defender faces multiattribute consequences c_D, which depend on her defense effort d and the results $s_1, \dots, s_m$. She then gets her utility u_D. Each attacker will get his multiattribute consequences c_{A_i}, which depend on his attack effort a_i and the result s_i. He then gets his utility u_{A_i}. Note that, in this formulation, her probability for the success of attack a_k does not depend on whether the other attacks were successful, but only upon which choices the other Attackers made. This is a reasonable approximation when the Defender is highly resourced and the Attackers do not coordinate. For example, the outcome of one burglary attempt is probably not affected by whether or not other burglars are successful, but it may be affected by the fact that other people choose to burgle; i.e., if a neighborhood sees a rash of attempted robberies, successful or not, then police increase their surveillance, lowering the chance of success. But if the Defender is not sufficiently well resourced, this version of the problem is less plausible: an understaffed police department means that a successful attack diverts resources, increasing the chance that other attacks will be successful. Similarly, if Attackers coordinate, so that multiple burglaries occur simultaneously, this may increase the chance of success burglary for all.

The Defender aims at finding her optimal defense strategy d^*. The consequences for the Defender are evaluated through her utility $u_D(d, s_1, \dots, s_m)$. Assuming conditional independence between the outcomes S_i of different attacks, given the defensive resources d and the implemented attacks a_i, she needs to assess the probability models $p_D(s_i|d, a_1, \dots, a_n)$, $i = 1, \dots, m$, reflecting which outcomes are more likely when attackers A_i launch attack a_i and defensive resources d have been deployed. She gets her expected utility given the attacks, integrating out the uncertainty over the outcomes of the attacks:

$$\psi_D(d|a_1, \dots, a_m) = \int \cdots \int u_D(d, s_1, \dots, s_m)\, p_D(s_1|d, a_1, \dots, a_m) \cdots p_D(s_m|d, a_1, \dots, a_m)\, \mathrm{d}s_1 \dots \mathrm{d}s_m. \quad (1)$$

Suppose that the Defender is able to build the models $p_D(a_i|d)$, $i = 1, \dots, m$, expressing her beliefs about which attack will be chosen by the i-th attacker after having observed the defense d. Our assumption of uncoordinated attacks is reflected on the conditional independence of $a_1, \dots, a_m$ given d. Then, D may compute

$$\psi_D(d) = \int \cdots \int \psi_D(d|a_1, \dots, a_m)\, p_D(a_1|d) \cdots p_D(a_m|d)\, \mathrm{d}a_1 \dots \mathrm{d}a_m,$$

and solve

$$\max_{d \in \mathcal{D}} \quad \psi_D(d)$$

to find her optimal defense resource allocation d^*.

In order to solve her problem, the Defender needs to assess $u_D(d, s_1, \dots, s_m)$, the distributions $p_D(s_i|d, a_1, \dots, a_m)$ and the distributions $p_D(a_i|d)$, $i = 1, \dots, m$, which are the only nonstandard assessments in her formulation. To obtain them, the Defender needs to put herself into the shoes of each attacker, and solve their corresponding problem separately, as they are uncoordinated. For instance, for the problem faced by attacker A_1, assuming that he is an expected utility maximizer, she would need his utility $u_{A_1}(a_1, s_1)$ and probabilities $p_{A_1}(s_1|d, a_1)$. Then, she would solve

$$a_1^*(d) = \arg\max_{a_1 \in \mathcal{A}_1} \int u_{A_1}(a_1, s_1)\, p_{A_1}(s_1|d, a_1)\, \mathrm{d}s_1.$$

However, the Defender lacks knowledge about u_{A_1} and p_{A_1}. Suppose we may model her uncertainty about them through random utilities and probabilities $\left(U_{A_1}, P_{A_1}\right)$. Then, we could propagate such uncertainty to obtain the random optimal attack, given her defense d

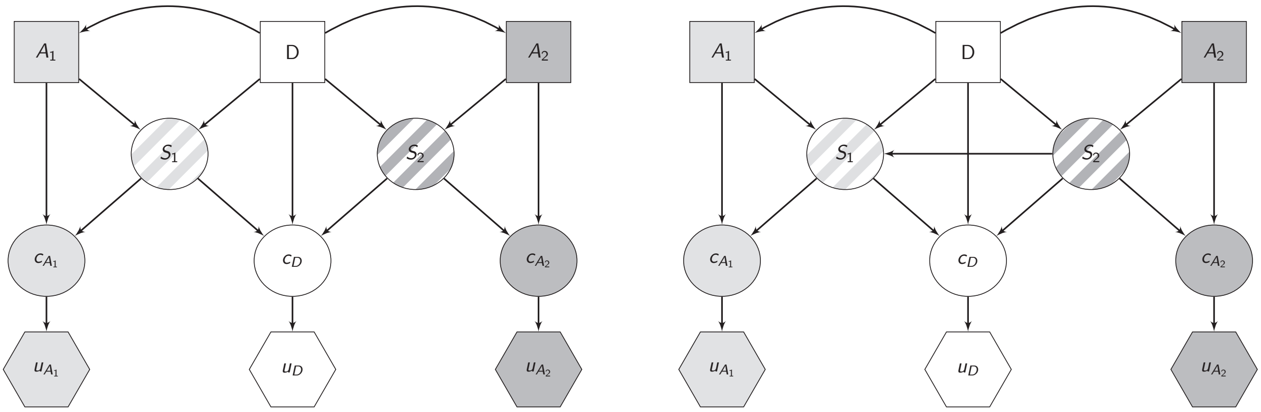
$$A_1^*(d) = \arg\max_{a_1 \in \mathcal{A}_1} \int U_{A_1}(a_1, s_1)\, P_{A_1}(s_1|d, a_1)\, \mathrm{d}s_1,$$

and, consequently, obtain $p_D(a_1|d) = \Pr(A_1^*(d) \leq a_1)$, which may be approximated through simulation by sampling from the random utilities and probabilities, finding the corresponding optimal attacks and using the Monte Carlo fraction of samples with the relevant optimal attacks as in Section 2. A similar scheme would be implemented, in parallel, for the other attackers, $A_2, \dots, A_m$, leading to estimates $\widehat{p_D}(a_i|d)$, $i = 2, \dots, m$, of the required probabilities.

As we said, the approach may be generalized in several ways. Sometimes it is reasonable to suppose that uncoordinated attacks have independent effects. For example, this would occur if one Attacker were a murderer and the other Attacker were a burglar, and if the police department had different divisions to handle those crimes, with little interaction between various sections. The outcome for the attempted murder would not affect the outcome for the attempted burglary. This is shown in Figure 5a. Then, we could rewrite the probability model $p_D(s_1|d, a_1, \ldots, a_m) \cdots p_D(s_m|d, a_1, \ldots, a_m)$ in (3.2) as

$$p_D(s_1|d, a_1) \cdots p_D(s_m|d, a_m),$$

and proceed in a similar fashion. Here, the Defender's action would be to decide how much money to allocate to the Homicide Unit and to the Burglary Unit.



(a)

(b)

Figure 5: Some generalizations for the bithreat problem.

Alternatively, it could be the case that there is some cascading effect between the results of the attacks, see Figure 5b. This pattern is apparent in franchise terrorism, suicide bombings, school shootings, and other contemporary threats, and corresponds to very weak coordination. A slightly stronger form of coordination occurs if the Attackers agree to order their decisions, so that A_1 chooses first, A_2 chooses second, and so forth, For example, assuming that $m = 2$, it could happen that s_2 affects s_1, so that $p_D(s_1|d, a_1)\, p_D(s_2|d, a_2)$ in (3.2) becomes

$$p_D(s_1|d, a_1, s_2)\, p_D(s_2|d, a_2).$$

Under this assumption, the general scheme required to estimate $\widehat{p_D}(a_i|d)$, $i = 1, \ldots, m$ cannot be implemented in parallel, but requires some sequentiality, as easily generalised.

As a final variation, it could be the case that there is just one random outcome S which depends on the decisions d of the Defender and $a_1, \ldots, a_m$ of the various Attackers. A typical example would be within an auction in which D is the Auctioneer designing the auction mechanism and A_i places his bid in the designed auction, with S being the result of such auction.

3.2 Coordinated attackers

When multiple opponents coordinate their attacks, the ARA for the Sequential Defend-Attack game must take account of the kind of cooperation that exists. We have seen examples of partial coordination in attack cascades inspired by previous attacks and in turn-taking in certain games. But often attacks are strongly coordinated and explicitly strategic. Important examples include:

- The Spanish government must defend its people against the joint operations of the Basque terrorist organisation ETA, the 'Ndrangheta and the Colombian narco mafia (http://www.terra.com.mx/articulo.aspx?articuloid=578597).
- Iran is the subject of coordinated economic sanctions imposed by many other (mostly Western) nations.
- Exxon must compete in a world market in which critical price controls are imposed by the OPEC.

Combating cooperative opponents is common, and arises in terrorism, international relations, business, and many other circumstances.

Our defender needs now to protect from the coordinated actions of attackers. We thus consider now the case of a Defender, D, who needs to deploy defensive resources $d \in \mathcal{D}$ to face m coordinated attackers $A_1, \ldots, A_m$. These observe her decision, and, coordinately, make attacking decisions $a_i \in \mathcal{A}_i,\; i = 1, \ldots, m$. The interaction between D and the attackers A_i, through their corresponding decisions d and a_i, leads to a random result $S_i \in \mathcal{S}_i$. The Defender faces multiattribute consequences c_D, which depend on her defense effort d and the results $s_1, \ldots, s_m$. She then gets her utility u_D.

When there is strong coordination, one can view the problem as one in which there is a single Attacker. The challenge is to determine the group utility function that represents the shared interest of the Attackers. Group utilities combine the individual utility functions of each Attacker. Each attacker will get his multiattribute consequences c_{A_i}, which depend on his attack effort a_i and his result s_i. Then the group of attackers get their (group) utility u_G, which somehow combine their individual utilities. Keeney and Raiffa (1993) and Ríos Insua et al. (2008) provide discussions concerning group utilities. Note that many of the bargaining algorithms, see Thomson (1994) may be seen in the light of maximising a group value function, aggregating individual value functions.

As before, the Defender aims at finding her optimal defense strategy d^*. She evaluates consequences through her utility $u_D(d, s_1, \ldots, s_m)$. She needs to assess the probability model $p_D(s_1, \ldots, s_m | d, a_1, \ldots, a_m),\; i = 1, \ldots, m$, reflecting which outcomes are more likely when the attackers launch their coordinated attacks a_i and defensive resources d have been deployed. She gets her expected utility, given the attacks, integrating out the uncertainty over the outcomes:

$$\psi_D(d|a_1, \ldots, a_m) = \int \cdots \int u_D(d, s_1, \ldots, s_m)\, p_D(s_1, \ldots, s_m | d, (a_1, \ldots, a_m))\, \mathrm{d}s_1 \ldots \mathrm{d}s_m.$$

Suppose now that the Defender is able to build the model $p_D(a_1, \ldots, a_m|d)$, expressing her beliefs about which (coordinated) attacks will be chosen by the attackers after having observed d. Then, she may compute

$$\psi_D(d) = \int \cdots \int \psi_D(d|a_1, \ldots, a_m)\, p_D(a_1, \ldots, a_m|d)\, \mathrm{d}a_1 \ldots \mathrm{d}a_m,$$

and solve

$$\max_{d \in \mathcal{D}} \quad \psi_D(d)$$

to find her optimal defense resource allocation d^*.

Again, the assessment of $p_D(a_1, \ldots, a_m|d)$ is nonstandard. We assume that the attackers are a (group) expected utility maximizer, see Keeney and Raiffa (1993). Assuming that their group utility is $u_G((a_1, \ldots, a_m), (s_1, \ldots, s_m))$ and their group probabilities are $p_G(s_1, \ldots, s_m|d, a_1, \ldots, a_m)$, then the Defender would solve for

$$a^*(d) = \operatorname*{arg\,max}_{(a_1, \ldots, a_m) \in \mathcal{A}_1 \times \mathcal{A}_m} \int u_G((a_1, \ldots, a_m), (s_1, \ldots, s_m))\, p_G(s_1, \ldots, s_m|d, a_1, \ldots, a_m)\, \mathrm{d}s_1 \ldots \mathrm{d}s_m.$$

However, the Defender lacks knowledge about u_G and p_G. Suppose we may model her uncertainty about them, through random utilities and probabilities $\left(U_G, P_G\right)$, and propagate that uncertainty to obtain the random optimal attack, given her defense d

$$A^*(d) = \operatorname*{arg\,max}_{(a_1,\ldots,a_m)\in\mathcal{A}_1\times\mathcal{A}_m} \int U_G(a_1,\ldots,a_m)\, P_G(s_1,\ldots,s_m|d,a_1,\ldots,a_m)\, \mathrm{d}s_1\ldots\mathrm{d}s_m.$$

Then, we would get $p_D(a_1,\ldots,a_m|d) = \Pr(A^*(d) \leq (a_1,\ldots,a_m))$, which may be approximated by simulation as earlier.

We discuss now the assessments of P_G and U_G. For P_G we may use the same approach as described above, based on Dirichlet distributions and/or processes. This means that, e.g., in the continuous case we could use a Dirichlet process centered around the Defender's assessment p_D with concentration δ_i for each attacker. Then, based on the expert judgment aggregation literature, see Cooke (1991), O'Hagan et al. (2006) and Clemen and Winkler (1990), we could aggregate through various rules, including the arithmetic mean. In this case, the random P_g would be $1/n \sum P_{A_i}$. Alternatively, we could have a distribution over the weights of the various P_{A_i} terms.

For U_G we could use the following approach. Suppose that $u_1,\ldots,u_m$ are our point estimate utility functions for each of the attackers, and $d_1,\ldots,d_m$ is a disagreement point, obtained e.g. through uncoordinated ARA analysis. Then, we could define $u_G = ((u_1 - d_1)^q + \cdots + (u_m - d_m)^q))^{1/q}$, as the group utility function. Finally we would just need to consider a density $f(q)$ over the exponent q of the group utility function. See Esteban and Ríos Insua (2014) for a justification.

4 Multiple defenders vs one attacker

Besides multiple attackers, it is also common to have multiple defenders. Examples include:

- Different banks, which share information to create credit scores for customers in order to reduce default rates and fraud.
- Airline companies,which each run baggage screening systems to prevent the introduction of bombs on planes, but must generally trust the security systems of each other when transferring luggage between carriers, see Kunreuther and Heal (2003).
- The multinational military effort in Afghanistan,where the United States, Great Britain, Canada and others seek to protect the government from overthrow by the Taliban.

Note that each defender may each be protecting its own targets (as in the case of several countries, each defending its territory against al Qaeda), or they may defend a common target (as when several companies jointly invest in computer security to protect a database that they all use). These defenders may act independently, or with weak or strong coordination.

We thus consider now the case in which several defenders face a single attacker.

4.1 Uncoordinated defenders

We start by the case of n uncoordinated defenders D_i, $i = 1,\ldots,n$ that need to face an attacker A. Examples include neighbours who decide whether or to protect their homes from burglars with an alarm, irrespective of what the other neighbours make; companies which decide to protect their intranets from internet attacks, irrespective of what other companies do; or countries which decide to protect their port entries from bioterrorist attackers, irrespective of what other countries decide to do.

We focus on supporting defender D_1, who needs to deploy defensive resources $d_1 \in \mathcal{D}_1$ to face an attacker A, in company of the other $n-1$ defenders, who make their corresponding decisions

$d_i, i = 2, \ldots, n$. The attacker observes their decisions, and make his attacking decision $a \in \mathcal{A}$. The interaction between D_i and A through their respective decisions d_i and a, leads to a random result $S_i \in \mathcal{S}_i$. The i-th Defender faces multiattribute consequences c_D, which depend on her defense effort d and the results s_i. She then gets her utility u_i. The attacker will get his multiattribute consequences c_A, which depend on his attack effort a and his results $s_i, i = 1, \ldots, n$. He then gets his utility u_A.

We thus assume we are supporting the first Defender, who aims at finding her optimal defense strategy d_1^*. The consequences for the Defender are evaluated through her utility $u_1(d, s_1)$. For this, she computes her expected utility, conditional on what the other defenders and the attacker will do,

$$\psi_1(d_1|d_2, \ldots, d_n, a) = \int u_1(d, s_1) p_1(s_1|d_1, d_2, \ldots, d_n, a) ds_1.$$

Then, based on her forecast of what the other defenders will do, and what the attacker will do given the deployed defences, she will compute

$$\psi_1(d_1) = \int \cdots \int \psi_1(d_1|d_2, \ldots, d_n, a) p_1(a|d_1, d_2, \ldots, d_n) p_1(d_2, \ldots, d_n) \, \mathrm{d}d_2 \ldots \mathrm{d}d_n \, \mathrm{d}a,$$

and solve for

$$\max_{d \in \mathcal{D}_1} \quad \psi_1(d),$$

to find her optimal defense resource allocation.

As before, we need to assess $p_1(a|d_1, \ldots, d_n)$, as well as $p_1(d_1, \ldots, d_n)$. For the first distribution, we use the earlier argument that

$$a^*(d_1, \ldots, d_n) = \arg\max_{a \in \mathcal{A}} \int \cdots \int u_A(a, s_1, \ldots, s_n) \, p_A(s_1, \ldots, s_n|a, d_1, \ldots, d_n) \, \mathrm{d}s_1 \ldots \mathrm{d}s_n.$$

However, the Defender lacks knowledge about u_A and p_A. We may model her uncertainty about them, through random utilities and probabilities $\left(U_A, P_A\right)$, and propagate that uncertainty to obtain the random optimal attack, given their defenses $d_1, d_2, \ldots, d_n$

$$A^*(d_1, \ldots, d_n) = \arg\max_{a \in \mathcal{A}} \int \cdots \int U_A(a, s_1, \ldots, s_n) \, P_A(s_1, \ldots, s_n|a, d_1, \ldots, d_n) \, \mathrm{d}s_1 \ldots \mathrm{d}s_n.$$

Note that a typical structural assumption would be

$$p_A(s_1, \ldots, s_n|a, d_1, \ldots, d_n) = \Pi \;\; p_A(s_i|a, d_i).$$

We need to assess also $p_1(d_2, \ldots, d_n)$, that is what defender D_1 believes how the other defenders will perform. Because of the simultaneous and uncoordinated nature of the defenders, we may assume that they are independent so that $p_1(d_2, \ldots, d_n) = \Pi_{i=2}^n p_1(d_i)$. We thus describe how to assess $p_1(d_2)$, through a similar argument as above. Indeed, D_2 would aim at solving, similarly to what D_1 does,

$$\psi_2(d_2|d_1, d_3, \ldots, d_n, a) = \int u_2(d_2, s_2) p_2(s_2|d_1, d_2, \ldots, d_n, a) ds_2,$$

$$\psi_2(d_2) = \int \cdots \int \psi_2(d_2|d_1, d_3, \ldots, d_n, a) p_2(a|d_1, d_2, \ldots, d_n) p_2(d_2, \ldots, d_n) \, \mathrm{d}d_1 \ldots \mathrm{d}d_n \, \mathrm{d}a,$$

$$d_2^* = \arg\max_{d \in \mathcal{D}_2} \psi_2(d),$$

to find her optimal defense resource allocation. However, we lack knowledge about u_2 and the p_2's. If we model our uncertainty through U_2 and the corresponding P_2's we would get the desired distribution as

$$D_2^* = \arg\max_{d \in \mathcal{D}_2} \int \cdots \int U_2(d_2, s_2) P_2(s_2 | d_1, d_2, \ldots, d_n, a) P_2(a | d_1, d_2, \ldots, d_n) P_2(d_2, \ldots, d_n) \, \mathrm{d}d_1 \ldots \mathrm{d}d_n \, \mathrm{d}a,$$

which, again, may be approximated by simulation. The same argument would be applied for the other defenders.

Note that in the previous argument we could start a recursion based on what the other defenders think of what a given defender is doing. This reminds us of the level-k thinking model in Stahl and Wilson (1995).

4.2 Coordinated defenders

When the Defenders coordinate, one must distinguish complete cooperation from partial cooperation. Examples of complete cooperation include soldiers in a combat squadron, or a neighborhood association that requires all households to contribute a fixed amount to provide security. In these situations, there is centralized decision making and so one can view the problem as a two-person Sequential Defend-Attack game, as analyzed in Section 2.

Partial cooperation is typically more complicated. A prominent example is the international military alliance between the United States, the United Kingdom, Australia, and Poland which led to the invasion of Iraq in 2003. Each nation had different interests and made different contributions, but there was sufficient common ground and negotiated structure that mutual choices were made. A looser level of coordination occurs in, say, a Neighborhood Watch program, for which different individuals volunteer different amounts of time and financial support, and coverage may vary widely, depending upon personal circumstances.

In the ideal cooperative case, each Defender elicits her probabilities about the Attacker's choice, conditional on all sets of defensive actions, and also her probabilities for the outcomes given the attack and the defenses. If each Defender finds that her expected utility is maximized by the same set of set of joint decisions, then the problem is solved. But such agreement is rare.

An alternative, when there is disagreement, is to compromise. The Defenders could accept any set of decisions for which each Defender receives an increase in her expected utility. If there is more than one set of defense choices that has that property, then the Defenders negotiate, and perhaps decide to use the set that maximizes the minimum gain in expected utility, or which maximizes the average gain in expected utility.

Regrettably, it may often happen that there is no set of decisions that improves all the expected utility of all Defenders. In that case hard negotiation is required, and Defenders who anticipate large gains in expected utility must find ways to compensate those who expect a loss. The chosen solution depends sensitively upon the resources and relationships between the Defenders.

There are other options. If none of the Defenders feels confident in her elicited probabilities for the Attacker's choice and/or her elicited probabilities for the outcomes, conditional on all sets of defense choices, then they might regard their probabilities as a draw from a common distribution. By pooling their beliefs, the Defenders could broker agreement on a common set of probabilities, enabling a unified solution as if there were a single Defender. This approach entails combination of subjective beliefs, which is notoriously problematic, but also often necessary.

Formally, first, we could perform our ARA based single defender analysis for each defender, thus obtaining a disagreement point $(d_1^*, \ldots, d_n^*)$, which would be associated with the corresponding optimal expected utilities $(\Psi_1^*, \ldots, \Psi_n^*)$. Note though that such values may not be jointly attainable, since they are obtained individually. Now given that the defenders jointly implement $(d_1, \ldots, d_n)$ and the attacker implements a, the results of the interaction will be $(s_1, \dots, s_n)$, which will lead to a utility $u_i(d_i, s_i)$ for each defender D_i. Define, then, the overachievement for defender D_i as

$$g_i(s_i) = \max((u_i(d_i, s_i) - \Psi_i^*), 0).$$

Finally, we could solve the problem

$$\max_{d_1,\dots,d_n} \int \cdots \int \left(\sum g_i(s_i)^p\right)^{1/p} p(s_1, \dots, s_n | a, d_1, \dots, d_n)\, \mathrm{d}s_1 \dots \mathrm{d}s_n.$$

5 Multitarget Protection

We, finally, consider Sequential Defend-Attack games in which the Defender must protect multiple targets from a threat. Examples of Defenders in such games include:

- A mayor, who must allocate police resources across multiple precincts, to control several kinds of criminal activity.
- A CEO, who must develop a budget that funds different departments within the organization, where each department (target) faces competition from a competitor (threats).
- A government, which assigns security personnel to embassies in other countries, where its interests may be threatened by bombs, mobs, or espionage.

The game can be seen as an application of portfolio theory in which an opponent observes the Defender's investments and seeks to minimize her return.

Suppose there are I targets and J kinds of resources. The Defender may deploy an amount d_{ij} of the jth resource to protect the ith target, so the entire decision is represented by the matrix $\boldsymbol{D} = \left\{d_{ij}\right\}$. These allocations must satisfy two standard constraints:

$$d_{ij} \geq 0, \qquad \sum_{i=1}^{I} d_{ij} \leq T_j.$$

The first ensures that negative investment is impossible, and the second implies that there is a ceiling, T_j, on the amount of the jth resource that is available. In general, there are additional constraints, such as the requirement that a bomb detecting dog must always be accompanied by a security officer, or a directive that some targets receive a minimum level of protection. The feasible choice set for the Defender is denoted by $\mathcal{D}$.

In this Sequential Defend-Attack game, the Attacker has K resources that may be used for attack. He observes the initial set of investments $\boldsymbol{D}$ and decides to allocate to the ith target an amount a_{ik} of the kth attack resources. Similarly to the Defender, his full decision is represented by a matrix $\boldsymbol{A}$ such that entry $a_{ik} \geq 0$ and $\sum_i a_{ik} \leq T_k$. The Attacker may also have additional constraints, such as a policy of not using more than three bombs on a single target. The feasible choice set for the Attacker is denoted by $\mathcal{A}$.

The interaction between the Defender and the Attacker at the ith target produces a random outcome S_i which takes a value $s_i \in \mathcal{S}_i$. A specific set of outcomes across all targets is denoted by $\boldsymbol{s} = (s_1, \dots, s_I)$, where $\boldsymbol{s} \in \mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_I$. The Defender's realized utility is $u_D(\boldsymbol{D}, \boldsymbol{s})$ and the Attacker's realized utility is $u_A(\boldsymbol{A}, \boldsymbol{s})$. In some applications the utility might also depend upon the actions chosen by the opponent, and then the utility functions for the Defender and Attacker would be written as $u_D(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{s})$ and $u_A(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{s})$, respectively. The ARA is a bit more complicated, but conceptually it is straightforward.

The Defender seeks the optimal investment $\boldsymbol{D}^* \in \mathcal{D}$. Often she can make plausible conditional independence assumptions that assert that the outcome at the ith target only depends on the total investments by both opponents for that target, and not upon the outcomes at other targets. That assumption would fail if, e.g., one target was a power plant and another target was protected by an electric fence powered by that plant. The conditional independence assumption could be relaxed, at the cost of having to elicit full joint distributions over outcomes at multiple targets, which imposes significant cognitive burden. So this ARA assumes conditional independence, and the Defender need only assess the probabilities $p_D(s_i|\boldsymbol{d}_i, \boldsymbol{a}_i)$, where $\boldsymbol{d}_i = (d_{i1}, \dots, d_{iJ})$ and $\boldsymbol{a}_i = (a_{i1}, \dots, a_{iK})$, for $i = 1, \dots, I$.

The Defender calculates her expected utility for each feasible allocation $\boldsymbol{D}$, conditional on each feasible allocation of the attack, $\boldsymbol{A}$:

$$\psi_D(\boldsymbol{D}|\boldsymbol{A}) = \int_{\mathcal{S}} u_D(\boldsymbol{D}, \boldsymbol{s}) \prod_{i=1}^{I} p_D(s_i|\boldsymbol{d}_i, \boldsymbol{a}_i) d\boldsymbol{s}.$$

If the Defender is also able to assess probabilities $p_D(\boldsymbol{A}|\boldsymbol{D})$, reflecting her belief about which attack will be chosen when she selects allocation $\boldsymbol{D}$, then she can compute her unconditional expected utility for each possible defense:

$$\psi_D(\boldsymbol{D}) = \int_{\mathcal{A}} \psi_D(\boldsymbol{D}|\boldsymbol{A})\, p_D(\boldsymbol{A}|\boldsymbol{D})\, d\boldsymbol{A}, \tag{2}$$

and solve for the optimal defense, $\boldsymbol{D}^* = \arg\max_{\boldsymbol{D}\in\mathcal{D}} \ \psi_D(\boldsymbol{D})$.

As usual, the trick is to assess $p_D(\boldsymbol{A}|\boldsymbol{D})$. Following the method in Section 2, the Defender attempts to solve the problem faced by the attacker A. If she knew his utility function $u_A(\boldsymbol{A}, \boldsymbol{s})$ and his probbilities for outcomes conditional on defenses and attacks on each target, or $p_A(s_i|\boldsymbol{d}_i, \boldsymbol{a}_i)$, then she would calculate his expected utility as

$$\psi_A(\boldsymbol{A}|\boldsymbol{D}) = \int_{\mathcal{S}} u_A(\boldsymbol{A}, \boldsymbol{s}) \prod_{i=1}^{I} p_A(s_i|\boldsymbol{d}_i, \boldsymbol{a}_i) d\boldsymbol{s}.$$

Of course, she does not know his true utilities and probabilities, but she can place subjective distributions over both, and thus can generate random (U_A, P_A). Thus she can repeatedly sample and solve

$$\boldsymbol{A}^*(\boldsymbol{D}) = \arg\max_{\boldsymbol{A}\in\mathcal{A}} \int_{\mathcal{S}} U_A(\boldsymbol{A}, \boldsymbol{s}) \prod_{i=1}^{I} P_A(s_i|\boldsymbol{d}_i, \boldsymbol{a}_i) d\boldsymbol{s}$$

to find $\widehat{p}_D(\boldsymbol{A}|\boldsymbol{D})$, her estimate for the probability of the attack which maximizes the Attacker's expected utility. She uses this distribution in (2) to find her best feasible allocation.

6 Discussion

We have provided approaches to generalisations of the Sequential Defend-Attack model. We first dealt with cases in which several defenders need to face several attackers. The standard approach would combine ideas from noncooperative and cooperative game theory. We have focused here in an approach based on the ARA framework, distinguishing the cases in which the defenders and/or the attackers are coordinated or not. We have analysed the case of one attacker *vs* several defenders and one defender *vs* several attackers. We have also analysed cases in which several targets need to be protected.

It is possible to generalize the previous discussion to include cases with multiple attackers, multiple defenders and multiple targets, all in the same context of an Sequential Defend-Attack game, by combining the earlier principles. The ideas extend to other ARA templates like the simultaneous Defend-Attack or the Sequential Defend-Attack-Defend models.

Acknowledgments

This work has been partly funded by the European Union's 7th Framework Programme under grant agreement no 285223 - SECONOMICS.

BIBLIOGRAPHY

C. D. Aliprantis and S. K. Chakrabarti. *Games and Decision Making*. Oxford University Press, 2000.

D. Banks, F. Petralia, and S. Wang. Adversarial risk analysis: Borel games. *Applied Stochastic Models in Business and Industry*, 27(2):72–86, 2011.

V. M. Bier and M. N. Azaiez. *Game Theoretic Risk Analysis of Security Threats*. Springer, 2009.

G. Brown, M. Carlyle, J. Salmerón, and K. Wood. Defending critical infrastructure. *Interfaces*, 36(6): 530–544, 2006.

R. T. Clemen and R. L. Winkler. Unanimity and compromise among probability forecasters. *Management Science*, 36(7):767–779, 1990.

R. M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, New York, 1991.

J. S. Dyer and R. K. Sarin. Measurable multiattribute value functions. *Operations Research*, 27(4): 810–822, 1979.

J. S. Dyer and R. K. Sarin. Relative risk aversion. *Management Science*, 28(8):875–886, 1982.

P. G. Esteban and D. Ríos Insua. Supporting an autonomous social agent within a competitive environment. *Cybernetics and Systems*, 45(3):241–253, 2014.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.

J. C. Harsanyi. Subjective probability and the theory of games: Comments on Kadane and Larkey's paper. *Management Science*, 28(2):120–124, 1982.

J. B. Kadane and P. D. Larkey. Subjective probability and the theory of games. *Management Science*, 28(2):113–120, 1982.

G. L. Keeney and D. von Winterfeldt. Identifying and structuring the objectives of terrorists. *Risk Analysis*, 30(12):1803–1816, 2010.

R. L. Keeney. Modeling values for anti-terrorism analysis. *Risk Analysis*, 27(3):585–596, 2007.

R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, Cambridge, 1993.

R. L. Keeney and D. von Winterfeldt. A value model for evaluating homeland security decisions. *Risk Analysis*, 31(9):1470–1487, 2011.

D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003.

B. Kraan and T. Bedford. Probabilistic inversion of expert judgments in the quantification of model uncertainty. *Management Science*, 51(6):995–1006, 2005.

H. Kunreuther and G. Heal. Interdependent security. *Journal of Risk and Uncertainty*, 26(2/3):231–249, 2003.

R. B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, MA, 1991.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Chichester, West Sussex, 2006.

H. Raiffa. *The Art and Science of Negotiation*. Harvard University Press, 1982.

H. Raiffa, J. Richardson, and D. Metcalfe. *Negotiation Analysis: the Science and Art of Collaborative Decision Making*. Harvard University Press, Cambridge, MA, 2002.

J. G. Rázuri, P. G. Esteban, and D. Ríos Insua. An adversarial risk analysis model for an autonomous imperfect decision agent. In T. V. Guy, M. Karny, and D. Wolpert, editors, *Decision Making and Imperfection*, volume 474 of *Studies in Computational Intelligence*, pages 163–187. Springer Berlin Heidelberg, 2013.

J. Ríos and D. Ríos Insua. Adversarial risk analysis for counterterrorism modeling. *Risk Analysis*, 32 (5):894–915, 2012.

D. Ríos Insua, J. Ríos, and D. Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854, 2009.

D. Ríos Insua, G. E. Kersten, J. Ríos, and C. Grima. Towards decision support for participatory democracy. *Information Systems and e-Business Management*, 6(2):161–191, 2008.

R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

D. O. Stahl and P. W. Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

W. Thomson. Cooperative models of bargaining. In Aumann R. J. and S. Hart, editors, *Handbook of Game Theory with Economic Applications*, volume 2, chapter 35, pages 1237–1284. Elsevier, 1 edition, 1994.

C. Wang and V. M. Bier. Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, 61(2):372–385, 2013.

Modelling Opponents in Adversarial Risk Analysis

Adversarial Risk Analysis has been introduced as a framework to deal with risks derived from the intentional actions of adversaries. The analysis supports one of the decision makers, who must forecast the actions of the other agents, and typically this forecast must take account of random consequences resulting from the set of selected actions. The solution requires one to model the behavior of the opponents, which entails strategic thinking. The supported agent may face different kinds of opponents, who may use different rationality paradigms; e.g., the opponent may behave randomly, or seek a Nash equilibrium, or perform level-*k* thinking, or use mirroring, or employ prospect theory, among many other possibilities. We describe the appropriate analysis for these situations, and also show how to model the uncertainty about the rationality paradigm used by the opponent through a Bayesian model averaging approach, suggesting a way to validate the opponent models. We focus on simultaneous decision-making by two agents.

1 Introduction

Recent applications in counterterrorism, cybersecurity and competitive corporate decision making have driven renewed interest in developing practical tools and theory for analyzing the strategic calculation of intelligent opponents who must act in scenarios with random outcomes. We use the term Adversarial Risk Analysis (ARA) to describe approaches in which the solution is based upon an explicit Bayesian model of the capabilities, probabilities and utilities used by the opponent in his analysis. For various concepts, methods and applications see Ríos Insua et al. (2009), Wang and Banks (2011), Banks et al. (2011), Ríos and Ríos Insua (2012), Sevillano et al. (2012), and Rázuri et al. (2013).

In ARA, the aim is to support one of the players who will use a decision analytic approach to solve her decision-making problem. To this end, she needs to forecast the actions of the other agents and, based on her own choice, the outcomes which she and her opponents will receive. This is not a new problem; it can be viewed as a Bayesian approach to game theory, and was proposed, non-constructively, by Kadane and Larkey (1982) and Raiffa (1982) and Raiffa et al. (2002). The approach has been criticised by Harsanyi (1982) and Myerson (1997), among others. From a practical standpoint, the main obstacle in implementing the decision analytic approach has been the lack of explicit mechanisms which allow the supported decision maker to encode her subjective probabilities about all the components in her opponents' decision making.

ARA deals with this problem within the framework of a Bayesian model for the supported decision-maker's uncertainty. She may face various kinds of opponents, who may use different rationality paradigms. However, she herself is a rational expected utility maximiser, in accordance with the Bayesian decision theory developed in Savage (1954).

Specifically, this paper treats opponents who may act at random, or be Nash equilibria seeking, level-*k* thinking, mirror equilibria seeking, or prospect maximising, but other kinds are possible. We describe how to model each of these opponents, and then use Bayesian model averaging to incorporate uncertainty about the rationality paradigm used by the opponent. Details on Bayesian model averaging can be found in Hoeting et al. (1999) and Clyde and George (2004); a longer treatment is given in Chipman et al. (2001). As discussed, this may be used to validate the supported decision-maker's model for her opponent.

The structure of the paper is as follows. We first present the problem of discrete simultaneous games, and briefly compare the game-theoretic and ARA approaches. Next we will describe models for the various kinds of rationality used by opponents, and then how those models can be combined in order to reflect uncertainty about the kind of rationality being used by the opponents in the game. We will also examine the cognitive burden of the analyses that pertain to different rationality paradigms. Although this paper focuses on two-person discrete simultaneous games, the methodology may be extended to more complex cases.

2 Basics

The two-person discrete simultaneous game is described by the Multi-Agent Influence Diagram in Figure 6. These diagrams were proposed by Koller and Milch (2003), and use rectangles to indicate decisions, ovals to indicate probability distributions, and hexagons to indicate preferences.

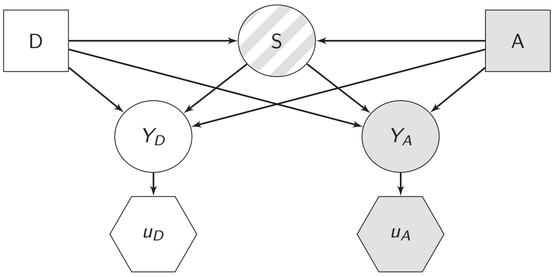


Figure 6: Two person simultaneous game

In this example there are two agents, the Defender and the Attacker. the Defender chooses from a finite set of actions $\mathcal{D} = \{d_1, \ldots, d_m\}$, while the Attacker simultaneously chooses from the finite set $\mathcal{A} = \{a_1, \ldots, a_n\}$. These choices correspond to the rectangles indicated by the corresponding initial. the Defender and the Attacker receive payoffs Y_D and Y_A, respectively, which depend upon both of their actions (d, a) and (in general) a random outcome ω; e.g., in a Prisoner's Dilemma game, this randomness might correspond to the chance of getting a strict or lenient judge. The hexagons in the diagram indicate the utilities u_D and u_A received by the Defender and the Attacker, respectively, from the payoffs Y_D and Y_A.

the Defender's expected utility associated with the pair of actions $(d, a) \in \mathcal{D} \times \mathcal{A}$ is

$$\psi_D(d, a) = \int u_D(a, d, \omega) p_D(\omega | a, d)\, \mathrm{d}\omega,$$

where $u_D(a, d, \omega)$ represents the utility she gets from a payoff $Y_D(a, d, \omega)$ and $p_D(\omega | a, d)$ represents her beliefs about the chance of obtaining the outcome ω, given the chosen pair of actions a and d. Similarly, the Attacker's expected utility is

$$\psi_A(d, a) = \int u_A(a, d, \omega) p_A(\omega | a, d)\, \mathrm{d}\omega.$$

Under the strong assumption of common knowledge (Gibbons, 1992 or Myerson, 1997), the Defender and the Attacker are expected utility maximisers who know the other's probabilities, utilities, and choice set, and both know that all of this is commonly known. In that case, the game may be described in bimatrix form:

$$\begin{array}{c|c} & a \\ \hline d & \psi_D(d, a), \psi_A(d, a) \end{array}$$

When the common knowledge assumption holds, players can predict with certitude their opponent's best responses to each of their actions by solving the opponent's decision problem. the Attacker's best response, known by the Defender, is then

$$R_A(d) = \arg\max_{a \in \mathcal{A}} \psi_A(d, a)$$

and, similarly, the Defender's best response, known by the Attacker, is

$$R_D(a) = \arg\max_{d \in \mathcal{D}} \psi_D(d, a).$$

These predictive models of best response R_A and R_D are used to find a Nash equilibrium solution for the game, consisting of a pair of actions (d^*, a^*) that are best responses to each other, so that $d^* = R_D(a^*)$ and $a^* = R_A(d^*)$. Since a Nash equilibrium solution is not guaranteed to exist for discrete games, often the action sets are extended to include randomised strategies, consisting of probability distributions over the original set of actions, so that a solution can always be found for the extended game (Nash, 1951). However, the common knowledge assumption is implausible in many applications. Raiffa et al. (2002), Rothkopf (2007), and Lippman and McCardle (2012) discuss its failure in detail. ARA avoids this problematic assumption through Bayesian modelling.

Specifically, the Defender must acknowledge her uncertainty about how her opponent solves his decision problem. Depending upon the kind of rationality paradigm she believes the Attacker uses, she may have to place subjective distributions over the utilities and probabilities held by the Attacker. She may also have to model what the Attacker thinks about her decision behavior. Things can quickly become complicated. To start with a very simple example, suppose the Defender believes that the Attacker has some value $v(a)$ associated with each of his possible actions, and that he will select the action $a^* = \arg\max_{a \in \mathcal{A}} v(a)$. This model would predict the Attacker's action if the Defender knew the values $v(a_i) = v_i$, for $i = 1, \ldots, n$. But without telepathy, the Defender must proceed as a Bayesian and describe her uncertainty about $(v_1, \ldots, v_n)$ through a joint distribution $(V_1, \ldots, V_n) \sim F$, where F represents her probabilistic beliefs about the values the Attacker holds. Thus, the Defender believes that the Attacker will choose action a_i with probability $p_D(a_i) = \mathbb{P}_F(V_i = \max\{V_1, \ldots, V_n\})$. Now, in order to maximize her own expected utility, the Defender solves

$$\max_d \psi_D(d) = \sum_{i=1}^{n} \psi_D(d, a_i)\, p_D(a_i) \tag{3}$$

to find the action d that maximizes her expected utility. The point of this simple example is to show how the Defender can replace common knowledge with subjective belief, and then proceed to select the action that is optimal under traditional Bayesian decision theory.

The next section extends this simple example by showing how the Defender can obtain the probabilities $p_D(a)$ that she needs for her solution under various models for the kind of rationality that the Attacker uses. At various points, we shall mention random probabilities and utilities, which, when invoked, will refer to a common probability space $(\Theta, \mathcal{F}, \mathcal{P})$ with atomic elements θ.

3 Probabilistic models of opponent behavior

In general, the Defender does not know what kind of rationality (or solution concept) the Attacker will use in choosing his decision. There are many standard kinds of rationality described in the literature, and this section shows how the Defender should use ARA with respect to some of them. Later, we combine them into a mixture model, so that the Defender can incorporate her personal uncertainty about the Attacker's solution concept, both to make her decision and to learn about the Attacker's rationality, eventually validating the models she uses.

3.1 Non-strategic opponents

First, assume that the Defender believes that the Attacker is non-strategic. In that case, it is as if she were playing Nature, in the standard decision-theoretic parlance (French and Ríos Insua, 2000), since the Attacker makes his decision at random, without regard to the Defender's action. Based on past data and/or expert opinion, the Defender will self-elicit her distribution $p_D(a)$.

If the contest is repeated and the Attacker lacks memory of previous moves, then he chooses randomly and independently each time. Then, the Defender would learn her distribution over his action space through a multinomial distribution, perhaps starting with a Dirichlet prior. Here, the Defender initially assumes a Dirichlet distribution $(p_1, \dots, p_n) \sim \mathcal{D}(\alpha_1, \dots, \alpha_n)$, where p_i is her initial subjective probability that the Attacker chooses action a_i. Then, after T iterations of the game, suppose the Attacker has selected action a_i exactly h_i times, so that the counts for his actions are $(h_1, \dots, h_n)$, with $\sum_i h_i = T$. Bayesian inference shows that the Defender's posterior distribution for the Attacker's choice is $\mathcal{D}(\alpha_1 + h_1, \dots, \alpha_n + h_n)$.

the Defender can now make point probability forecasts for the Attacker's non-strategic choice. Under the Dirichlet-Multinomial model, one possibility is that

$$p_D^{NS}(a_i) = E(p_i|\text{data}) = \frac{\alpha_i + h_i}{T + \sum \alpha_i}, \quad i = 1, \dots, n.$$

Now the Defender can solve

$$d^*_{NS} = \arg\max_d \sum_{i=1}^{n} \psi_D(d, a_i)\, p_D^{NS}(a_i)$$

to determine her best choice in the non-strategic case.

In a slightly more general vein, suppose that the Attacker has memory, and can recall previous games $\{(a_i, d_j, \omega)\}_{t=1}^{T}$. To simplify the exegesis, we assume he can recall only the preceding game, and we also limit the outcome of that game to discrete levels indexed by ω, reflecting, say, the degree of success from his action. Both of these simplifying assumptions can be easily weakened.

In this case, through the Markov property, the Defender can use a matrix beta prior, as in Ríos Insua et al. (2012), to learn the corresponding parameters through

$$(p_1, \dots, p_n)|a_i, d_j, \omega \sim \mathcal{D}(\alpha_1^{ij\omega}, \dots, \alpha_n^{ij\omega}).$$

If, after an (a_i, d_j, ω) game, the Attacker has selected action a_k exactly $h_k^{ij\omega}$ times, then

$$(p_1, \dots, p_n)|a_i, d_j, \omega, \text{data} \sim \mathcal{D}(\alpha_1^{ij\omega} + h_1^{ij\omega}, \dots, \alpha_n^{ij\omega} + h_n^{ij\omega}),$$

and the Defender could use the probability mass function

$$p_D^{NS}(a_k|a_i, d_j, \omega, \text{data}) = E(p_k|a_i, d_j, \omega, \text{data}) = \frac{\alpha_k^{ij\omega} + h_k^{ij\omega}}{A^{ij\omega} + T^{ij\omega}},$$

where $A^{ij\omega} = \sum_k \alpha_k^{ij\omega}$ and $T^{ij\omega} = \sum_k h_k^{ij\omega}$.

A shortcoming of this approach is that the size of the conditioning set grows according to the product of the cardinalities of the sets $\mathcal{D}$, $\mathcal{A}$ and Ω. However, by using the concept of mixtures of Markov chains (Raftery, 1985), one can linearly control the size of the conditioning set by writing

$$p_D(a|a_i, d_j, \omega) = w_1 p_D(a|a_i) + w_2 p_D(a|d_j) + w_3 p_D(a|\omega).$$

In order to make inference about the transition probabilities and weights, one can use Gibbs sampling, as described in Ríos Insua et al. (2012). Note that the inference made on the weights may be used to check the influence of various elements a_i, d_j or ω over the decisions made by the Attacker, through the posterior probabilities $p(w_i|\text{data}), i = 1, 2, 3$. This approach extends to the case in which the Attacker can recall a larger number of previous games.

This Bayesian analysis is related to learning and the fictitious play approach in games (Ozdaglar and Menache, 2011). If all opponents play in this way, then under certain conditions, they converge to a Nash equilibrium, if they interact for a sufficiently long time.

3.2 A Nash equilibrium seeking opponent

Suppose now that the Defender believes that the Attacker will compute a Nash equilibrium in order to select his action. This could be because she believes that he has studied game theory or he is long-memoried non-strategic player, as in Section 3.1, and they have played many games. the Defender has a subjective distribution for (U_A, P_A), the random variables that represent her beliefs about the Attacker's utility and probability functions (u_A, p_A). She also has a subjective distribution for (U_D, P_D), which is what she thinks the Attacker believes are her utility and probability functions (u_D, p_D).

For the probability space $(\Theta, \mathcal{F}, \mathcal{P})$ with atomic elements $\theta \in \Theta$, the random quantities ARA needs are $(u_A^\theta, p_A^\theta) \sim (U_A, P_A)$ and $(u_D^\theta, p_D^\theta) \sim (U_D, P_D)$. Also, the Defender believes the expected utilities are $(\psi_D^\theta(d,a), \psi_A^\theta(d,a))$, with, for example,

$$\psi_D^\theta(d,a) = \int u_D^\theta(a,d,\omega) p_D^\theta(\omega|a,d)\,\mathrm{d}\omega.$$

For each θ, she can compute the corresponding Nash equilibrium $(d^{NE}(\theta), a^{NE}(\theta))$, which will typically be a distribution over the d's and a's, respectively. Then, she would calculate

$$p_D^{NE}(a) = \mathcal{P}(\{\theta : a^{NE}(\theta) = a\}).$$

Ultimately, she will choose the action d_{NE}^* that solves

$$d_{NE}^* = \arg\max_d \sum_{i=1}^n \psi_D(d, a_i)\, p_D^{NE}(a_i).$$

3.3 A level-k thinking opponent

When the Attacker's rationality entails level-k thinking (Stahl and Wilson, 1994, 1995), the Defender knows that the Attacker selects his action based upon a chain of reasoning of the form "I know that she knows that I know ..." which will go k levels deep, depending on how sophisticated she believes the Attacker to be. Thus, if the Defender is non-strategic, then she is a level-0 thinker and chooses randomly. If she chooses her action by assuming that the Attacker is non-strategic, then she is a level-1 thinker, and so forth. In this situation, the Defender will maximize her expected utility by reasoning one level further in the chain than the Attacker.

Concretely, the Defender needs to solve (3), so that her optimal decision is

$$d^* = \arg\max_d \sum_a \psi_D(d,a) p_D(a).$$

In thinking about $p_D(a)$, she considers the problem faced by the Attacker and assumes he is an expected utility maximizer, so that his decision can be predicted by solving

$$\begin{aligned} a^* &= \arg\max_a \sum_d \psi_A(d,a) p_A(d) \\ &= \arg\max_a \sum_d \left[\int u_A(d,a,\omega) p_A(\omega|a,d)\,\mathrm{d}\omega \right] p_A(d). \end{aligned} \tag{4}$$

the Defender does not know the elements $(u_A, p_A(\cdot\,|\,\cdot), p_A)$ required to solve (4). As a Bayesian, she expresses her uncertainty about these through random utilities and probabilities $(U_A, P_A(\cdot\,|\,\cdot),\ P_A)$.

Substituting these in (4), she obtains a predictive distribution for the random action that the Attacker will choose through

$$A = \arg\max_a \sum_d \left[\int U_A(d, a, \omega) P_A(\omega|a, d)\, \mathrm{d}\omega \right] P_A(d), \tag{5}$$

which provides the $p_D(a) = \mathbb{P}(A = a)$ required in (3). Here, the distribution of A may be estimated through Monte Carlo simulation by sampling from the random utilities and probabilities and computing (and accumulating) the corresponding optimal alternatives as follows:

Algorithm 1 Simulating from the Attacker's problem

```
For k = 1 to N
    Sample $(u_A^k, p_A^k(\cdot \,|\, \cdot), p_A^k) \sim (U_A, P_A(\cdot \,|\, \cdot), P_A)$
    Compute $a_k^* = \mathrm{argmax}_a \sum_d \left[\int u_A^k(d, a, \omega) p_A^k(\omega|a, d) d\omega\right] p_A^k(d)$
Approximate $p_D(a_i) \approx \mathrm{card}\{1 \leq k \leq N : a_k^* = a_i\}/N$, for $i = 1, \ldots, n$
```

In the above triplet $(U_A, P_A(\cdot \,|\, \cdot), P_A)$, the first two elements are relatively easy for the Defender to elicit, since they represent what the Defender believes are the Attacker's utility function and beliefs about the outcome of the game, conditional on their decisions. For example, $P_A(\omega|a, d)$ could be centered around her own $p_D(\omega|a, d)$ with some additional uncertainty. As far as U_A is concerned, typically she shall have information about the interests of the Attacker, which she would aggregate with a weighted measurable utility function. Using the relative risk aversion concept, as in Dyer and Sarin (1979), she could model the risk attitude of the Attacker that determines the functional form of his utility function. Finally, her uncertainty would be reflected by distributions over the weights and the risk coefficient.

In contrast, the third element, $P_A(d)$, often requires higher-level strategic thinking. She must model what the Attacker thinks is the Defender's decision analysis. Thus, if the Defender supposes that the Attacker is a level-1 thinker and that he is modeling her as an expected utility maximizer, then from his perspective the Defender is modeled as solving the optimization problem in (3) where the required $(u_D, p_D(\cdot \,|\, \cdot), p_D)$ are unknown to him and therefore must be represented through random utilities and probabilities $(U_D, P_D(\cdot \,|\, \cdot), P_D)$. This allows the Attacker to elicit his predictive probability distribution over her possible actions through

$$p_A(D = d) = \mathbb{P}\left(\arg\max_{x \in \mathcal{D}} \sum_a \left[\int U_D(x, a, \omega) P_D(\omega|x, a)\, \mathrm{d}\omega \right] P_D(a) = d \right), \tag{6}$$

which he needs to solve (4).

Now, the Defender's uncertainty about the Attacker's distribution F_D for $(U_D, P_D(\cdot \,|\, \cdot), P_D)$ can be modeled though $F_D \sim \mathcal{F}_D$, her probabilistic beliefs about the distributions F_D used by the Attacker to solve her decision problem and compute his $p_A(d)$. This gives the $P_A(d)$ required in (5) through

$$P_A(D = d) = \mathbb{P}\left(\arg\max_{x \in \mathcal{D}} \sum_a \left[\int U_D(x, a, \omega) P_D(\omega|x, a)\, \mathrm{d}\omega \right] P_D(a) = d \right) \tag{7}$$

with $(U_D, P_D(\cdot \,|\, \cdot), P_D) \sim F_D$ and now $F_D \sim \mathcal{F}_D$.

Once the Defender obtains $P_A(d)$, she plugs it into (5) to obtain the $p_D(a)$ required in (3), thus making the Defender a level-2 thinker. This level-k thinking process would continue, to the level that the Defender deems necessary, as in the following loop, which constitutes a hierarchy of nested decision models:

Algorithm 2 Simulating the level-k thinking process.

```
For i = 2 until necessary
    Sample $\left(U_A^i, P_A^i(\cdot|\cdot), P_A^i\right) \sim F_A^i$, with $F_A^i \sim \mathcal{F}_A^i$
    Compute $P_D^i(A^i = a) = \mathbb{P}\left(\text{argmax}_{x \in \mathcal{A}} \sum_{d \in \mathcal{D}} \left[\int U_A^i(d, x, \omega) P_A^i(\omega | d, x)\, d\omega\right] P_A^i(D^i = d) = a\right)$
    Sample $\left(U_D^i, P_D^i(\cdot|\cdot), P_D^i\right) \sim F_D^i$, with $F_D^i \sim \mathcal{F}_D^i$
    Compute $P_A^i(D^i = d) = \mathbb{P}\left(\text{argmax}_{x \in \mathcal{D}} \sum_{a \in \mathcal{A}} \left[\int U_D^i(x, a, \omega) P_D^i(\omega | x, a)\, d\omega\right] P_D^{i+1}(A^{i+1} = a) = d\right)$
i = i + 1
```

Note that $p_D(a)$ and $P_A(d)$ defined, respectively, by (5) and (7) would correspond to a step $i = 1$ in the above loop. To sum up the levels of thinking in level-k rationality, in terms of the notation we have used,

- A level-0 the Defender acts at random (non-strategically).
- A level-1 the Defender chooses her alternative optimally, but assumes that the Attacker acts randomly, since he is level-0, as in Section 3.1.
- A level-2 the Defender assumes that the Attacker is a level-1 thinker, who assumes she is a level-0 thinker. the Defender stops at $i = 1$ in the hierarchy, with the elicitation of $F_D \sim \mathcal{F}_D$, which determines $P_A(d)$ representing her beliefs about the probability model used by the Attacker to predict her action.
- A level-3 the Defender assumes that she faces a level-2 adversary: the Attacker's calculation assumes she is a level-1 thinker, who thinks about his decision problem. the Defender stops at $i = 2$ in the hierarchy, with the elicitation of $F_A^2 \sim \mathcal{F}_A^2$, which determines $P_D^2(A^2 = a)$.
- A level-4 the Defender assumes she is facing a level-3 adversary: the Attacker takes strategic account of what he thinks she thinks he thinks that she thinks. the Defender stops at $i = 2$ in the hierarchy, with the elicitation of $F_D^2 \sim \mathcal{F}_D^2$, which determines $P_A^2(D^2 = d)$.
- And so forth.

Rothschild et al. (2012) use this framework to provide an algorithmic approach to level-k thinking. the Defender first selects the value k ($k > 1$) that she believes is the depth of the Attacker's analysis. Then she places a uniform distribution over the Attacker's actions and supposes that the Attacker has a uniform distribution over her action space. the Defender climbs up one level at a time in the hierarchy by simulating from these distributions and solving to find the new distribution for the Attacker's optimal action, and consequently inferring his corresponding new distribution over her own action set. She repeats until she reaches the selected value of k.

We believe it is more natural to think in terms of an alternative approach, as suggested in Ríos Insua et al. (2009) and Ríos and Ríos Insua (2012), which proceeds by climbing up in the hierarchy until the Defender finds it difficult to reason meaningfully. Indeed, at higher levels of thinking, the Defender will probably lack the information necessary to assess the distributions $\mathcal{F}_A^i$ or $\mathcal{F}_D^i$ associated with the decision analysis of A^i or D^i, respectively. At this point, the Defender might assign a probability distribution over A^i or D^i, without going deeper in the hierarchy, thus summarizing all remaining information she might have through the direct assessment of $P_D^i(A^i = a)$ or $P_A^i(D^i = d)$, as appropriate. At this stage, one reasonable possibility is to assign a noninformative distribution. Lee and Wolpert (2012) describe experiments in behavioral game theory which suggest that opponents rarely go further than levels $k = 2$ or 3. So, in most cases, the Defender need not to go beyond $k = 3$ or 4 in order to be one level deeper than the Attacker.

3.4 Mirror equilibria seeking opponents

As implied above, level-k thinking can lead to an infinite regress. Classical game theory avoids this through the common knowledge assumption, which allows players to use deterministic predictive models of their opponents decisions. Another way to preclude the infinite regress is through the mirroring equilibria concept (Banks et al., 2011), which we formalize here.

Assume the Defender has distributions for the random quantities $(U_A, P_A(\cdot \mid \cdot), P_A(\cdot))$ which describe the Defender's beliefs about the Attacker's utilities and probabilities, as in (5), and she also has $(U_D, P_D(\cdot \mid \cdot), P_D(\cdot))$ which describe the Defender's beliefs about the Attacker's beliefs regarding her own utilities and probabilities, as in (6). This completes step $i = 1$ in the hierarchy for level-k thinking.

Suppose for a moment that the Defender has a point mass in θ, in the generic probability space $(\Theta, \mathcal{F}, \mathcal{P})$ introduced previously. In this case, she believes that the Attacker will solve for his optimal decision

$$a^*(\theta) = \arg\max_a \sum_d \left[\int u_A^\theta(d, a, \omega) p_A^\theta(\omega \mid a, d)\, \mathrm{d}\omega \right] p_A^\theta(d).$$

Next, by assuming non-point mass support $\mathcal{P}$, she deduces her predictive distribution over the Attacker's choice in $\mathcal{A}$,

$$p_D^{ME}(a) = \mathcal{P}(\{\theta \in \Theta : a^*(\theta) = a\}).$$

Note that this may be written as

$$p_D^{ME}(a) = \mathbb{P}\left(\arg\max_{x \in \mathcal{A}} \sum_d \left[\int U_A(d, x, \omega) P_A(\omega \mid d, x)\, \mathrm{d}\omega \right] P_A(d) = a \right).$$

Symmetrically, knowing θ and $p_D^{ME}(a)$, the Defender thinks that the Attacker believes that she is trying to solve

$$d^*(\theta) = \arg\max_d \left[\sum_d \int u_D^\theta(d, a, \omega) p_D^\theta(\omega | a, d)\, \mathrm{d}\omega \right] p_D^{ME}(a),$$

which yields a random optimal decision $d(\theta)$ on the underlying probability space $(\Theta, \mathcal{F}, \mathcal{P})$ with distribution

$$p_A^{ME}(d) = \mathcal{P}(\{\theta \in \Theta : d^*(\theta) = d\}).$$

We say that the distributions $p_D^{ME}(a)$ and $p_A^{ME}(d)$ are consistent and constitute a *mirroring equilibrium* if they jointly satisfy

$$p_A^{ME}(d) = \mathcal{P}\left(\arg\max_x \sum_a \left[\int U_D(x, a, \omega) P_D(\omega \mid x, a)\, \mathrm{d}\omega \right] p_D^{ME}(a) = d \right), \tag{8a}$$

$$p_D^{ME}(a) = \mathcal{P}\left(\arg\max_x \sum_d \left[\int U_A(d, x, \omega) P_A(\omega \mid d, x)\, \mathrm{d}\omega \right] p_A^{ME}(d) = a \right). \tag{8b}$$

When such a pair of consistent distributions is found, this provides the Defender with a probabilistic model to predict the Attacker's actions, in which it is assumed that he uses the mirroring equilibria as the solution concept. At this point, the Defender steps out of the mirror-equilibrium paradigm and uses her utility and probability functions to select the action that maximizes her expected utility; i.e.,

$$\max_d \sum_{i=1}^n \psi_D(d, a_i) p_D^{ME}(a_i),$$

with $p_D^{ME}(a)$ obtained as the fixed point solution of the mirroring analysis (8). The mirroring solution concept may be viewed as a way to enforce coherence in the Defender's probability judgments.

The existence of fixed points solutions to the mirroring distribution equations in (8) is a complex question and only partial solutions are generally available. The problem is closely related to the existence of Bayes Nash equilibria solutions. The most complete theory exists in the context of two-person asymmetric auctions. Lebrun (1996, 1999) shows that when bidders' valuations of the item on auction have continuous densities, then the solution to the mirror equilibrium exists, is unique, and the cumulative distribution functions over the bidders decisions are continuous. However, except in some special cases (Kaplan and Zamir, 2012), no closed form solution exists. Various algorithms for solving the asymmetric two-person auction have been proposed (Li and Riley, 2007; Gayle and Richard, 2008), but none has been proven to converge. In fact, the most popular approaches, based on the back-shooting algorithm, are provably non-convergent in an open ball around 0 (Fibich and Gavish, 2011). This is an important open problem.

4 A cognitive comparison

We compare now the four previous models for opponent rationality, and the standard game theoretic approach, in terms of the cognitive load that the analysis imposes upon the Defender. The following table summarizes the elements which must be assessed according to the hierarchy in level-k thinking introduced in Section 3.3.

	1	2	3	4	5	6
0	u_D	$p_D(\cdot\,\vert\,a,d)$	$p_D(a)$	u_A	$p_A(\cdot\,\vert\,a,d)$	$p_A(d)$
1	U_D	$P_D(\cdot\,\vert\,a,d)$	$P_D(a)$	U_A	$P_A(\cdot\,\vert\,a,d)$	$P_A(d)$
2	U_D^2	$P_D^2(\cdot\,\vert\,a,d)$	$P_D^2(a)$	U_A^2	$P_A^2(\cdot\,\vert\,a,d)$	$P_A^2(d)$
...	...	...	...	...	...	...

Row 0 corresponds to the utilities and beliefs of the Defender and the Attacker, as perceived by themselves. The elements in row 1 all correspond to random utilities and probabilities perceived by the Defender, including those she believes are the Attacker's utilities and probabilities, subscripted by A, and those she believes are the Attacker's beliefs about her own utilities and probabilities, subscripted by D. Row 2 includes what she believes the Attacker believes are her beliefs about the Attacker's utilities and probabilities, subscripted by A, and so forth.

To summarize the cognitive demands of the various approaches, we shall refer to element (i, j) in the matrix as the cognitive element in row i and column j. All the approaches we have discussed require the elements (0,1) and (0,2). Beyond that:

- The standard game theoretic approach requires also elements (0,4), and (0,5), and these must be common knowledge. This limits the scope in realistic applications.
- The non-strategic opponent model requires also (0,3), but uses no specific strategic principles in assessing it; as sketched in Section 3.1, the Defender needs only data and/or expert opinion.
- As sketched in Section 3.2, the Nash equilibrium opponent model requires elements (1,1), (1,2), (1,4) and (1,5), which are used to calculate (0,3).
- The level-k opponent model requires elements:
  - (1,4), (1,5) and (1,6), which are used to produce (0,3), or,
  - (1,1), (1,2) and (1,3), which are used to produce (1,6), and this in turn, with (1,4) and (1,5) produces (0,3), or,

- additional layers are needed to handle $k > 2$.

- The mirror equilibrium opponent model requires elements (1,4), (1,5) and (1,1), (1,2), and uses the consistency condition to produce (0,3), as sketched in Section 3.4.

4.1 Prospect maximising opponents

One possible criticism of the opponent models presented so far is that they assume that both players essentially seek to maximise expected utility. The assumption is arguably reasonable for the Defender—ARA is designed to support her decision-making, and can prescribe that perspective for her. But it is more tenuous for the Attacker, and there is much evidence that humans often make choices that do not maximise expected utility (Camerer, 2003). In the context of counterterrorism, both terrorism psychology (see, e.g., English, 2010) and logistics (see, e.g., Brown et al., 2006) suggest that adversaries tend to invest their attack resources in order to produce the best outcomes in a certain sense. A well reported example is the corporate-like organisation of piracy in Somalia (Sevillano et al., 2012), referred to as cutthroat capitalism (Carney et al., 2009). Thus, attackers optimise, but they may not be optimising expected utility.

An alternative descriptive model of decision making under uncertainty is prospect theory (Wakker, 2010); here one optimises some criterion other than expected utility. To illustrate how the expected utility solutions may be extended to other optimization criteria, suppose that the Attacker's choice is modeled as the solution of

$$a^* = \arg\max_a \left[\sum_d \int v_A(d, a, \omega) \pi_A(p_A(\omega|a, d))\, \mathrm{d}\omega \right] \pi_A(p_A(d)),$$

where v_A is a value function and π_A is a weighting function over the Attacker's probabilities p_A, as defined in Kahneman and Tversky (1979). Note the formal similarity with equation (4).

the Defender does not know these probability distributions and value and weighting functions. She must elicit personal distributions over them, which we model through $(V_A, \Pi_A, P_A(\cdot\,|\,\cdot), P_A)$. By propagating the uncertainty quantified in these distributions, she induces a probability distribution over the Attacker's action space, as in (5), through:

$$A^* = \arg\max_a \left[\sum_d \int V_A(d, a, \omega) \Pi_A(P_A(\omega\,|\,a, d))\, \mathrm{d}\omega \right] \Pi_A(P_A(d)).$$

This distribution is the $p_D^{PT}(a) = \mathbb{P}(A^* = a)$ required for her to maximize her expected utility in (3). Also as before, the Defender could examine a hierarchy of nested decision analysis, in the more complex level-k framework, in order to ultimately elicit $p_D^{PT}(a)$.

5 Combining opponent models

The five procedures that we have discussed correspond to different opponent models. These lead to different adversarial forecasting models and, consequently, potentially different optimal actions for the Defender. But in many situations she will not know which model among these possibilities (and others) correctly describes the Attacker's behaviour. However, as a Bayesian, she will have a subjective probability about the relevance of each model she considers. These subjective probabilities allow her to combine her solutions through a Bayesian mixture model, as described in Clyde and George (2004) or Hoeting et al. (1999).

Let $p(M_i)$ be the probability that the Defender has for each of the k opponent models that she thinks might describe the Attacker, with $\sum_{i=1}^k p(M_i) = 1$ and $p(M_i) \geq 0, i = 1, \ldots, k$. Let $p_D^i(a)$ be the probability distribution induced by each opponent model over the Attacker's action set, through the kind of analysis described in Section 3. the Defender then combines all these distributions into a single distribution $p_D(a) = \sum_{i=1}^{k} p(M_i)\, p_D^i(a)$, a weighted average for which the weights are her beliefs about the type of reasoning the Attacker will use. She should then solve

$$d^* = \max_d \sum_a \psi_D(d,a) \left(\sum_{i=1}^{k} p(M_i) p_D^i(a) \right) = \max_d \sum_{i=1}^{k} p(M_i) \left[\sum_a \psi_D(d,a) p_D^i(a) \right]$$

to determine her best possible choice.

When there is repeated play, the Defender can learn what kind of rationality the Attacker employs, and thus validate her model for the Attacker's decision making. By appropriately embedding these as parametric models, she can build in a model selection strategy based on $p(M_i|data)$, as data about games accumulate. At a given time she could decide to use in her mixture only opponent models which have high posterior probability, or, if one model emerges as a winner, she could use it solely. But, as argued in Draper (1995), this underestimates her uncertainty, which may be disastrous in certain applications, such as national security. The model selection strategy may be seen as a model validation approach. As data accumulates, $p(M_i|data)$ may be viewed as a measure of how valid she believes model M_i to be. For additional discussion of this perspective, see Berger and Ríos Insua (1998).

6 Discussion

We have provided a description of a family of opponent models for adversarial risk analysis, in the context of discrete, simultaneous two-person games. The ideas extend easily to continuous games. The extension to other forms of interaction (like sequential defend-attack, sequential defend-attack-defend, or general coupled influence diagrams) will become complicated, but the approach is clear. When extending the methods to multiple adversaries, one needs to consider whether to allow for the possibility that opponents coordinate their decisions. Beyond that, there are also technical issues, such as determining when mirroring equilibria exist, or determining whether it is necessary to climb higher in the level-k hierarchy, taking into account value-of-information concepts. Additionally, although we discussed some aspects of learning from repeated play, much more could be done.

There is no simple solution for a serious strategic analysis. But this is exactly the conclusion one should expect. the Defender's success depends critically upon the accuracy of the information that she has about her opponent, and, in particular, it is often sensitive to what she believes the Attacker believes about her.

In general, the Defender will not have precise knowledge about what kind of rationality (or solution concept) the Attacker is using to select his action. But, at the cost of additional computation, the Defender can use a mixture model to combine the results from different opponent models, then use this distribution to choose the action that maximizes her expected utility. In addition, if there is data on the opponent's decisions in similar situations, one can compute the posterior probabilities of each opponent model and the parameters within each model (e.g., the opponent's random utilities and probabilities), thus validating them. From this perspective, ARA provides a flexible and attractive framework for making strategic decisions.

Acknowledgments

This work has been partly funded by the European Union's 7th Framework Programme under grant agreement no 285223 - SECONOMICS.

BIBLIOGRAPHY

D. Banks, F. Petralia, and S. Wang. Adversarial risk analysis: Borel games. *Applied Stochastic Models in Business and Industry*, 27(2):72–86, 2011.

J. O. Berger and D. Ríos Insua. *Statistical and Bayesian Methods in Hydrological Sciences*, chapter Recent developments in Bayesian inference with applications in hydrology, pages 56–80. UNESCO Press, 1998.

G. Brown, M. Carlyle, J. Salmerón, and K. Wood. Defending critical infrastructure. *Interfaces*, 36(6): 530–544, 2006.

C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

S. Carney, S. Eggertsson, and M. Doret. Cutthroat capitalism: An economic analysis of the Somali pirate business model. *Wired Magazine*, 17(07), 2009.

H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, 38:65–134, 2001.

M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97, 1995.

J. S. Dyer and R. K. Sarin. Measurable multiattribute value functions. *Operations Research*, 27(4): 810–822, 1979.

R. English. *Terrorism: How to Respond*. Oxford University Press, 2010.

G. Fibich and N. Gavish. Numerical simulations of asymmetric first-price auctions. *Games and Economic Behavior*, 73(2):479–495, 2011.

S. French and D. Ríos Insua. *Statistical Decision Theory*. Arnold, 2000.

W. R. Gayle and J. F. Richard. Numerical solutions of asymmetric, first-price, independent private values auctions. *Computational Economics*, 32(3):245–278, 2008.

R. Gibbons. *A Primer in Game Theory*. Pearson Education Ltd., 1992.

J. C. Harsanyi. Subjective probability and the theory of games: Comments on Kadane and Larkey's paper. *Management Science*, 28(2):120–124, 1982.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999.

J. B. Kadane and P. D. Larkey. Subjective probability and the theory of games. *Management Science*, 28(2):113–120, 1982.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47 (2):263–291, 1979.

T. R. Kaplan and S. Zamir. Asymmetric first-price auctions with uniform distributions: Analytic solutions to the general case. *Economic Theory*, 50(2):269–302, 2012.

D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003.

B. Lebrun. Existence of an equilibrium in first price auctions. *Economic Theory*, 7(3):421–443, 1996.

B. Lebrun. First price auctions in the asymmetric *n* bidder case. *International Economic Review*, 40 (1):125–142, 1999.

R. Lee and D. Wolpert. Game theoretic modeling of pilot behavior during mid-air encounters. In T. V. Guy, M. Kárný, and D. H. Wolpert, editors, *Decision Making with Imperfect Decision Makers*, volume 28 of *Intelligent Systems Reference Library*, pages 75–111. Springer Berlin Heidelberg, 2012.

H. Li and J. G. Riley. Auction choice. *International Journal of Industrial Organization*, 25(6):1269–1298, 2007.

S. A Lippman and K. F. McCardle. Embedded Nash bargaining: Risk aversion and impatience. *Decision Analysis*, 9(1):31–40, 2012.

R. B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1997.

J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

A. Ozdaglar and I. Menache. *Network Games: Theory, Models, and Dynamics*. Synthesis Lectures on Communication Networks. Morgan & Claypool Publishers, 2011.

A. E. Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):528–539, 1985.

H. Raiffa. *The Art and Science of Negotiation*. Harvard University Press, 1982.

H. Raiffa, J. Richardson, and D. Metcalfe. *Negotiation Analysis: the Science and Art of Collaborative Decision Making*. Harvard University Press, 2002.

J. G. Rázuri, P. G. Esteban, and D. Ríos Insua. An adversarial risk analysis model for an autonomous imperfect decision agent. In T. V. Guy, M. Karny, and D. Wolpert, editors, *Decision Making and Imperfection*, volume 474 of *Studies in Computational Intelligence*, pages 163–187. Springer Berlin Heidelberg, 2013.

J. Ríos and D. Ríos Insua. Adversarial risk analysis for counterterrorism modeling. *Risk Analysis*, 32 (5):894–915, 2012.

D. Ríos Insua, J. Ríos, and D. Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854, 2009.

D. Ríos Insua, F. Ruggeri, and M. Wiper. *Bayesian analysis of stochastic process models*, volume 978. John Wiley & Sons, 2012.

M. H. Rothkopf. Decision analysis: The right tool for auctions. *Decision Analysis*, 4(3):167–172, 2007.

C. Rothschild, L. McLay, and S. Guikema. Adversarial risk analysis with incomplete information: A level-k approach. *Risk Analysis*, 32(7):1219–1231, 2012.

L. J. Savage. *The Foundations of Statistics*. Courier Dover Publications, 1954.

J. C. Sevillano, D. Ríos Insua, and J. Ríos. Adversarial risk analysis: The Somali pirates case. *Decision Analysis*, 9(2):86–95, 2012.

D. O. Stahl and P. W. Wilson. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3):309–327, 1994.

D. O. Stahl and P. W. Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

P. P. Wakker. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press Cambridge, 2010.

S. Wang and D. Banks. Network routing for insurgency: An adversarial risk analysis framework. *Naval Research Logistics*, 58(6):595–607, 2011.

Adversarial Risk Analysis for Biagent Influence Diagrams

We describe how to support an agent's decision making when facing an adversary such that their joint problem is modelled as a biagent influence diagram. General interactions entailing sequences of defence and attack movements, possibly repeated across time, are explored. We describe an adversarial risk analysis methodology to support the agent, avoiding strong common knowledge assumptions.

1 Introduction

In his landmark paper, Shachter (1986) suggested extending influence diagrams to multiagent cases as an important problem. This suggestion has been faced mainly from a game theoretic perspective, stemming from Koller and Milch (2003) who introduced Multiagent influence diagrams (MAID) to find Nash equilibrium solutions in general non-cooperative game-theoretic problems, discussing how they provide equivalent solutions to game trees. A main drawback of such game-theory based methodology is its common knowledge assumption, criticised in e.g. Raiffa et al. (2002) or Lippman and McCardle (2012). Most versions of non-cooperative game theory assume that the adversaries not only know their own payoffs, preferences, beliefs, and possible actions, but also those of their opponents. When there is uncertainty in the game, it is assumed that players have common probabilities as in games of incomplete information, see Harsanyi (1967). These strong common knowledge assumptions allow for a symmetric joint normative analysis in which players try to maximise their expected utilities (and expect other players to do the same). Their decisions can be anticipated and dominated by Nash equilibria concepts. However, in many contexts, including counterterrorism, cybersecurity or competitive marketing, players will not typically have full knowledge of their opponent's objectives, beliefs and possible moves. This may be aggravated as participants try to conceal information.

At this point, Adversarial Risk Analysis (ARA) provides a solution, as common knowledge is no longer required. In supporting one of the participants, which we call the Defender, ARA views the Defender problem as a decision analytic one, but procedures which somehow employ the game-theoretical structure and other information available are used to estimate the probabilities of the opponents' actions. ARA is a methodology that combines statistical risk analysis and game theory from a Bayesian perspective, see e.g. Ríos Insua et al. (2009). A main motivation for ARA developments arises from security and counterterrorism studies. Specific ARA related work presenting case studies dealing with protection from intelligent threats include defending an airport against terrorist attacks (Cano et al., 2014); preventing ships from piracy risks (Sevillano et al., 2012); or anti-IED defence in routing problems, see Wang and Banks (2011).

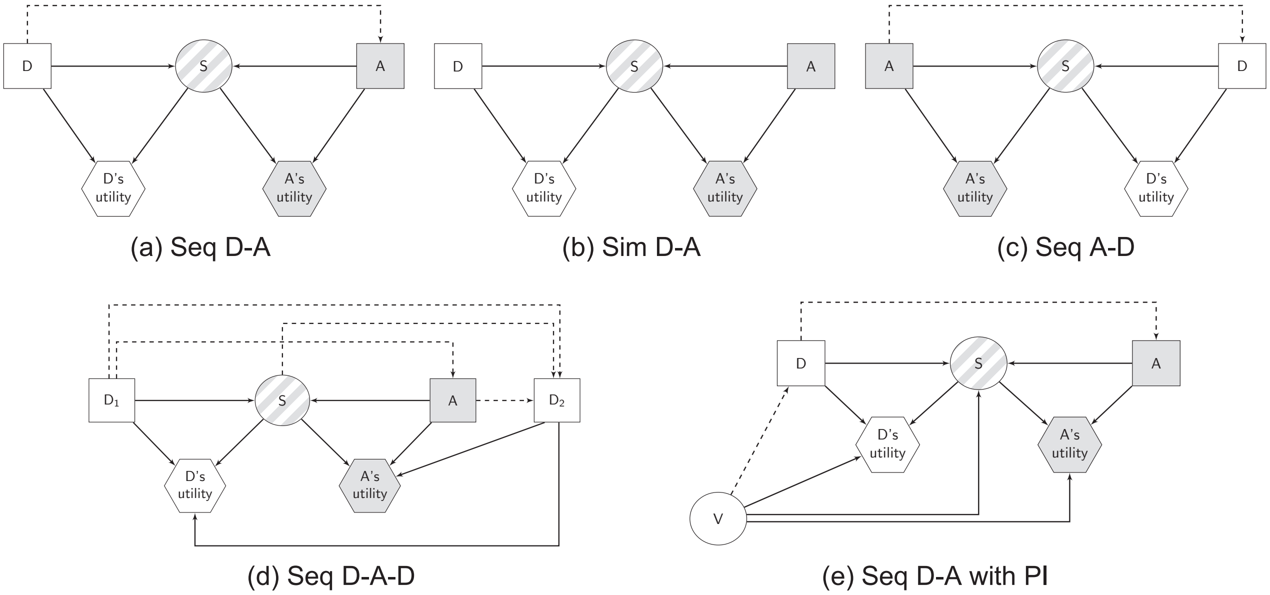
These and other applications have been modelled using relatively simple ARA models, with basic sequences of defence and attack movements. Indeed, we can identify a number of templates that can be viewed as basic building blocks for general security risk analysis problems, see Brown et al. (2006), Brown et al. (2008) or Ríos and Ríos Insua (2012) for additional details. They differ from each other in the way and order in which attack and defence movements take place within the global sequence of decisions and events, as well as in the information revealed. Five of the most basic templates, with self-explanatory names, are: the Sequential Defend-Attack model; the Simultaneous Defend-Attack model; the Sequential Attack-Defend model; the Sequential Defend-Attack-Defend model; and, finally, the Sequential Defend-Attack with Private Information model.

Beyond these templates, we consider here general adversarial problems in which we allow for more complex interactions between the intervening agents, typically consisting of intermingled sequential and simultaneous movements, possibly spanning across different planning periods. Our aim is to support one of the agents, the Defender, in her security resource allocation decision making problem. For that, she needs to forecast the Attacker's intentions. Assuming that the Attacker is an expected utility maximiser, we can predict his actions by finding his maximum expected utility action. The uncertainty in our assessments about the Attacker's probabilities and utilities is propagated over to his random optimal decision. Sometimes, such assessments may lead to a hierarchy of nested decision problems, as described in Ríos Insua et al. (2009), close to the concept of level-*k* thinking, see Stahl and Wilson (1995). Thus, we solve general adversarial problems using MAIDs ability to model complex interaction problems, taking advantage of the concept of strategic relevance (Koller and Milch, 2003), but relaxing the common knowledge assumption by using the ARA methodology.

The paper is structured as follows. In Section 2, we present the biagent influence diagrams we shall be dealing with, illustrating them with a driving example. Section 3 presents the main features of our proposal applied to the example. We generalise the methodology in Section 4. Section 5 provides further examples. We end up with some discussion.

2 Biagent Influence Diagrams

We describe the basic structures that we shall be dealing with. We shall essentially face coupled influence diagrams, one for the Defender and one for the Attacker, possibly with shared uncertainty nodes and some links between the Attacker and the Defender decision nodes. We designate them Biagent influence diagrams (BAID), see Koller and Milch (2003) who introduce Multiagent influence diagrams. Figure 7 presents the BAIDs for the five template models sketched in the introduction.



(a) Seq D-A

(b) Sim D-A

(c) Seq A-D

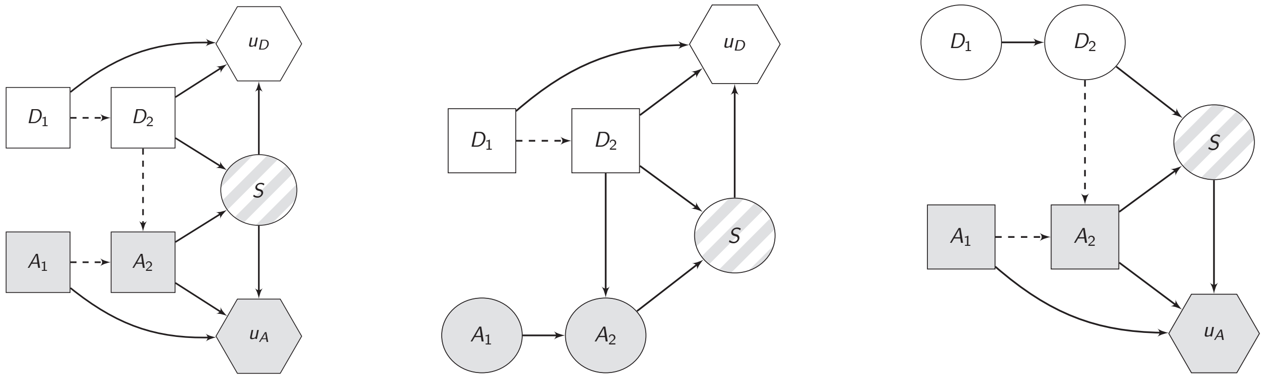
(d) Seq D-A-D

(e) Seq D-A with PI

Figure 7: Influence diagrams for the five template models.

In them, we can observe several decision (rectangle), chance (circle) and utility (hexagon) nodes, corresponding to the Defender (white) and the Attacker's (grey) problems, respectively. Stripped nodes *S* represent common uncertainty nodes. Besides, when an agent's action is observed by his/her adversary prior to her/his own decision, there is an arrow pointing to the decision node of the observing agent, see e.g. the arrow between decision nodes *D* and *A* in Figure 7a. The remaining features are shared with standard influence diagrams, as in Shachter (1986). Schemes for implementing ARA within such stylised settings may be seen in Ríos and Ríos Insua (2012) and Ríos Insua et al. (2013). However, such templates may not be sufficient to cope with complexities in many real problems.

We thus consider general adversarial problems for two agents interacting with each other over time. As an illustration, consider the influence diagram in Figure 8a, which we shall use to outline the methodology. To fix ideas, we could associate it with a scenario related with the protection of a critical infrastructure. At any given time, the incumbent authorities might have to decide whether or not to increase the infrastructure's protection against terrorist attacks. To this aim, they could order, decision D_1, an internal audit to obtain a reliable diagnosis of the compliance with security standards and protocols. Meanwhile, the terrorists, decision A_1, might be pondering about infiltrating within the infrastructure to gain intelligence for future attacks. D_1 and A_1 are simultaneous, yet unknown to each other, actions. Depending on the results of the audit, the Defender might deploy additional measures, D_2, by e.g. reinforcing security controls on people and items. This time, the Attacker observes the Defender action, and feels that his time might have come, fearing that new countermeasures could be added in the near future. He then might choose his attack A_2, possibly consisting of a direct attempt to cause damages to people and/or assets. The interaction between actions D_2 and A_2 would yield a random outcome S, which describes the result of the attack. This or similar sequences of defence-attack movements could be repeated across time, possibly spanning along different planning periods. For each agent and planning period, there is a utility node which aggregates all their costs and consequences, as e.g. represented by u_D and u_A in Figure 8a.



(a) Influence diagram (b) The Defender problem (c) The Attacker problem

Figure 8: ARA modelling of a general multistage biagent problem.

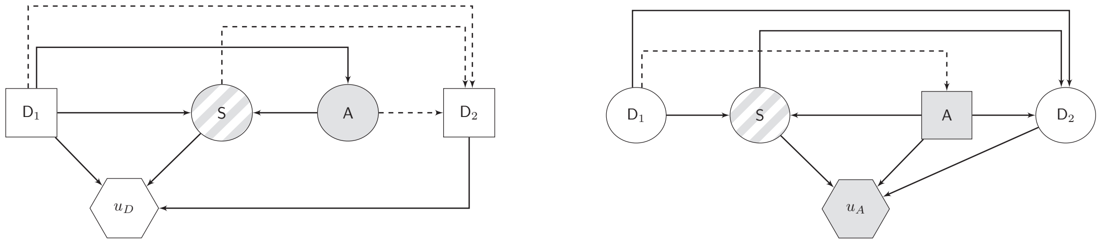
Note that, within the general layout of the influence diagram shown in Figure 8a, we may identify specific sequences of defence-attack moves from those described as basic templates in Section 1. In particular, nodes D_1–A_1, and D_2–A_1, correspond to a Simultaneous Defend-Attack template, in which both agents decide their movement simultaneously, without knowing the action chosen by each other. Similarly, nodes D_2–A_2 reproduce the backbone structure of a Sequential Defend-Attack template, in which the Defender first chooses her action and, then, having observed it, the Attacker decides his own move. As we discuss in the next sections, detecting this type of patterns within the global layout of an influence diagram will be of great help when dealing with complex adversarial problems.

We shall not be able to deal with all BAIDs, conceivable as the mere superposition of two IDs (one for the Defender and one for the Attacker) with some shared chance nodes and some arrows linking their nodes. To ensure consistency between the informational structure and the ordering of the decision makers' analysis, we follow Shachter's (1986) terminology and algorithm. Essentially, we require that if any two decisions are simultaneous, then there is no directed path between them. Thus, for our purposes, a BAID is an acyclic directed graph over decision, chance and utility nodes, where chance nodes can be shared, such that, from each decision maker's perspective is a proper ID.

To check whether a BAID satisfies this constraint, one must test whether the diagram generated from the BAID for each player is a proper ID. For each player, an ID is obtained by deleting the other decision maker's preference nodes in the BAID, and converting his/her decision nodes into chance nodes. Chance nodes that are not shared with the chosen player are eliminated. Also, decision nodes owned by the other decision maker may become barren nodes and, thus, be eliminated. Each player's ID must then define a total order of decisions and a corresponding partial order of chance nodes.

The total order for the Defender is associated with the order of decisions made by her: $D_1 \to D_2 \to \cdots \to D_m$ for m decision nodes. This implies the existence of a time or precedence-based ordering and induces a partition of the set $\mathcal{C}$ of chance nodes relevant for her into the subsets: (1) $\mathcal{C}_0$, consisting of those random events that are known when she makes her first decision D_1; (2) $\mathcal{C}_i$, composed of those chance nodes whose values are observed between decisions D_{i-1} and D_i, for $i = 1, \dots, m$; and (3) $\mathcal{C}_m$, including those chance nodes that are not observed—and therefore are unknown—before any of the decisions is made. Some of these sets $\mathcal{C}_i$ might be empty. This partition defines a partial order over decision and chance nodes $\mathcal{C}_0 \prec D_1 \prec \cdots \prec \mathcal{C}_{m-1} \prec D_m \prec \mathcal{C}_m$, which induces an information structure within the temporal order of decisions that specifies the information known at the time each decision is made. Similarly, we may define a partial order $\mathcal{B}_0 \prec A_1 \prec \cdots \prec \mathcal{B}_{n-1} \prec A_n \prec \mathcal{B}_n$, for the n decisions $A_1 \to A_2 \to \cdots \to A_n$ to be made by the Attacker and the incumbent set of chance nodes $\mathcal{B}$. Note that $\{D_1, \dots, D_m\} \subset \mathcal{B}$ and $\{A_1, \dots, A_n\} \subset \mathcal{C}$, and it could be the case that $\mathcal{C} \cap \mathcal{B} \neq \emptyset$.

Given a BAID, we thus associate two influence diagrams with it, representing the Defender and the Attacker's perception of their problems. As an example, consider the Sequential Defend-Attack-Defend model in Figure 7d. When we address the Defender's problem, we treat the Attacker's decision at node A as uncertain from the Defender's viewpoint and model such uncertainty. This is reflected in the influence diagram in Figure 9a, where the Attacker's decision node has been converted into a chance node, by replacing $\boxed{\text{A}}$ with Ⓐ. On the other hand, when the Defender tries to solve her own problem, she needs to analyse the problem from the Attacker's perspective, represented in Figure 9b. As we can observe, decision nodes D_1 and D_2 have turned now into chance nodes for the Attacker, as he does not know about the Defender's intentions. Note, however, that the Attacker will observe the outcome of (for him) chance node D_1 prior to his own decision A_1. The Attacker's influence diagram is built following a similar reasoning.



(a) The Defender problem

(b) The Attacker problem

Figure 9: The Sequential Defend-Attack-Defend model.

3 Computational Strategy with the Example

As a motivation, we describe how to deal with the security example in Section 2. We first solve the Defender's problem, depicted in Figure 8b. The Attacker's decision nodes have been converted into chance nodes, since they are uncertain to the Defender.

We need to assess the Defender's beliefs about which attacks will be chosen by the Attacker at decision nodes A_1 ($p_D(A_1 = a_1)$)and A_2, conditional on attack a_1 and defence d_2 ($p_D(A_2 = a_2|a_1, d_2)$), besides the (more standard) assessments about her own utilities $u_D(d_1, d_2, s)$ and probabilities $p_D(S = s|d_2, a_2)$. Given these, the Defender could solve her decision problem working backwards the influence diagram in Figure 8b, following the standard approach in Shachter (1986).

$\mathcal{D}1$. At chance node S, compute the expected utilities

$$(d_1, d_2, a_2) \to \psi_D(d_1, d_2, a_2) = \int u_D(d_1, d_2, s)\, p_D(S = s|d_2, a_2)\, \mathrm{d}s.$$

$\mathcal{D}2$. At chance (for the Defender) node A_2, compute the expected utilities

$$(d_1, a_1, d_2) \to \psi_D(d_1, a_1, d_2) = \int \psi_D(d_1, d_2, a_2)\, p_D(A_2 = a_2|a_1, d_2)\, \mathrm{d}a_2.$$

$\mathcal{D}3$. At chance (for the Defender) node A_1 obtain the expected utilities

$$(d_1, d_2) \to \psi_D(d_1, d_2) = \int \psi_D(d_1, a_1, d_2)\, p_D(A_1 = a_1)\, \mathrm{d}a_1.$$

$\mathcal{D}4$. At decision node D_2, compute the optimal action $d_2^*(d_1)$, given d_1,

$$d_1 \to d_2^*(d_1) = \underset{d_2 \in \mathcal{D}_2}{\arg\max}\ \psi_D(d_1, d_2).$$

$\mathcal{D}5$. Finally, at node D_1, find the maximum expected utility decision

$$d_1^* = \underset{d_1 \in \mathcal{D}_1}{\arg\max}\ \psi_D(d_1, d_2^*(d_1)).$$

Then, the Defender's optimal strategy would be to choose d_1^* at node D_1 and, later, choose $d_2^*(d_1^*)$ at node D_2.

Note that while the assessment of the utility $u_D(d_1, d_2, s)$ and the probability $p_D(S = s|d_2, a_2)$ is relatively standard, those of $p_D(A_2 = a_2|a_1, d_2)$ and $p_D(A_1 = a_1)$ require strategic thinking. In order to anticipate her adversary's movements, the Defender may analyse the Attacker's problem, shown in Figure 8c.

Our proposal is to model the Defender's uncertainty about the Attacker's decisions, which stems from her uncertainty about the Attacker's probabilities and utility, assuming he is an expected utility maximiser. For other attacker rationalities, see Rios Insua et al (2015). Therefore, the Defender needs to assess $u_A(a_1, a_2, s)$ and $p_A(S = s|d_2, a_2)$, as well as $p_A(D_1 = d_1)$ and $p_A(D_2 = d_2|d_1)$. In general, the Defender will not know the Attacker's probabilities and utility, but she may acknowledge her uncertainty about them through random probabilities and utilities, which we describe through $F \sim \big(P_A(S = s|d_2, a_2), U_A(a_1, a_2, s), P_A(D_2 = d_2|d_1), P_A(D_1 = d_1)\big)$. Then, for each (d_1, d_2), we would propagate the uncertainty in F to obtain the random optimal alternatives as follows:

$\mathcal{A}1$. At chance node S, compute the Attacker's (random) expected utilities

$$(a_1, d_2, a_2) \to \Psi_A(a_1, d_2, a_2) = \int U_A(a_1, a_2, s)\, P_A(S = s|d_2, a_2)\, \mathrm{d}s.$$

$\mathcal{A}2$. At decision node A_2, compute the Attacker's (random) optimal decisions

$$(a_1, d_2) \to A_2^*(a_1, d_2) = \arg\max_{a_2 \in \mathcal{A}_2} \mathit{\Psi}_A(a_1, d_2, a_2).$$

and record the (random) optimal expected utilities

$$\Psi(a_1, d_2) = \max_{a_2 \in \mathcal{A}_2} \mathit{\Psi}_A(a_1, d_2, a_2).$$

$\mathcal{A}3$. At chance (for the Attacker) node D_2, compute the Attacker's (random) expected utilities

$$(a_1, d_1) \to \mathit{\Psi}_A(a_1, d_1) = \int \mathit{\Psi}_A(a_1, d_2)\, P_A(D_2 = d_2 | d_1)\, \mathrm{d}d_2.$$

$\mathcal{A}4$. At chance node (for the Attacker) node D_1, compute the Attacker's (random) expected utilities

$$a_1 \to \mathit{\Psi}_A(a_1) = \int \mathit{\Psi}_A(a_1, d_1)\, P_A(D_1 = d_1)\, \mathrm{d}d_1.$$

$\mathcal{A}5$. At decision node A_1, compute the Attacker's (random) optimal decision

$$A_1^* = \arg\max_{a_1 \in \mathcal{A}_1} \mathit{\Psi}_A(a_1).$$

Then, the optimal (random) attacks would be A_1^* and $A_2^*(A_1^*, d_2)$. Finally, the Defender's predictive density over attack A_2, conditional on her second defence decision d_2 and the Attacker's first move a_1, would be given by

$$\int_{-\infty}^{a_2} p_D(A_2 = y | a_1, d_2)\, \mathrm{d}y = \Pr(A_2^*(a_1, d_2) \leq a_2),$$

and, similarly, her predictive density $p_D(A_1 = a_1)$ over attack A_1, is given by

$$\int_{-\infty}^{a_1} p_D(A_1 = x)\, \mathrm{d}x = \Pr(A_1^* \leq a_1).$$

This information would be incorporated back into the Defender's problem to obtain her optimal defence. Both probabilities can be approximated through Monte Carlo simulation.

Note that of the four elements in F, the first two are relatively standard probabilistic assessments, but the last two entail strategic thinking and may lead to recursions as discussed in Ríos and Ríos Insua (2012), close to level-k thinking, see Stahl and Wilson (1995).

Given the sequentiality of the problem, with alternation of simultaneous and sequential decisions, we shall need to solve it stepwise, scheduling optimising stages for the Defender's problem and simulating stages from those of the Attacker. Indeed, both schemes $\mathcal{D}1$–$\mathcal{D}5$ and $\mathcal{A}1$–$\mathcal{A}5$ can be combined into a single one which jumps from steps $\mathcal{D}$ to steps $\mathcal{A}$, and backwards, as follows:

$\mathcal{S}1$. Perform $\mathcal{D}1$. The quantities $p_D(S = s | d_2, a_2)$ and $u_D(d_1, d_2, s)$ are assessed for the Defender. Then compute the expected utilities in $\mathcal{D}1$.

$\mathcal{S}2$. At step $\mathcal{D}2$, the Defender needs $p_D(A_2 = a_2 | a_1, d_2)$. Since she lacks it, she switches to the Attacker's problem. Using her assessments $P_A(S = s | d_2, a_2)$ and $U_A(a_1, a_2, s)$, she performs steps $\mathcal{A}1$ and, then, $\mathcal{A}2$, getting $p_D(a_2 | a_1, d_2)$ from $\Pr(A_2^*(a_1, d_2) \leq a_2)$. She then completes $\mathcal{D}2$.

$\mathcal{S}3$. At step $\mathcal{D}3$, she needs $p_D(A_1 = a_1)$. Since she lacks it, she switches again to the Attacker's problem,. Using $P_A(D_2 = d_2|d_1)$ she performs $\mathcal{A}3$. Using $P_A(D_1 = d_1)$ she performs $\mathcal{A}4$. Then, she performs $\mathcal{A}5$, getting $p_D(a_1)$ from $\Pr(A_1^* \leq a_1)$, and completes $\mathcal{D}3$. She completes also $\mathcal{D}4$ to eliminate D_2.

$\mathcal{S}4$. She finally completes step $\mathcal{D}5$.

Essentially, the approach solves as many $\mathcal{D}$ steps as possible with standard ID reduction operations, until some assessment from the Attacker is required to solve another $\mathcal{D}$ step. Then, as few steps from the $\mathcal{A}$ problem are solved, with ID reduction operations modified to take into account the uncertainty about the Attacker's utilities and probabilities, until the required Attacker assessment is obtained. At this point we jump back to the Defender problem.

Deciding when to jump from problem $\mathcal{D}$ to problem $\mathcal{A}$ and backwards is relatively simple to perform by hand, but it can be messy from an algorithmic point of view. We may achieve this through the use of the relevance and component graphs, described in Koller and Milch (2003). Basically, the relevance graph for a BAID is a directed graph whose nodes are the decision nodes of the BAID, and which contains an edge $N_j \rightarrow N_i$ if and only if N_j is strategically relevant for N_i. The relevance graph for our example in Section 2 is shown in Figure 10, together with the so-called maximal strongly connected components (SCCs, encircled by dotted lines) which, in turn, define the component graph.

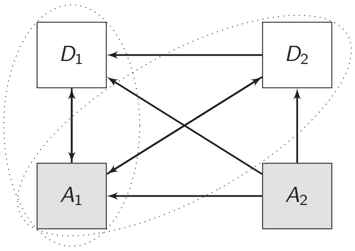


Figure 10: Relevance graph for the general multistage problem.

The relevance graph induces a topological ordering among the attacker and the defender decisions, which is

$$A_2 \longrightarrow (A_1, D_2) \longrightarrow (D_1, A_1),$$

in consonance with the ordering established in steps $\mathcal{S}1$–$\mathcal{S}4$ above. To wit,

- We first reduce all possible nodes in problem $\mathcal{D}$ prior to dealing with A_2, step $\mathcal{S}1$.
- We switch to problem $\mathcal{A}$ and perform as few reductions as possible until A_2 is eliminated, reducing as many nodes as possible in problem $\mathcal{D}$, and use them before dealing with (A_1, D_2), step $\mathcal{S}2$.
- Since it is an (A, D) simultaneous pair, we perform reductions in the $\mathcal{A}$ and $\mathcal{D}$ problems, before dealing with (A_1, D_1), step $\mathcal{S}3$.
- Finally, we deal with (A_1, D_1). A_1 was already eliminated, so we just need to eliminate D_1, step $\mathcal{S}4$.

In the next section, we generalise this to provide a general ARA strategy to solve complex adversarial biagent problems.

4 General Computational Strategy

We describe now the general methodology for solving complex adversarial biagent problems. We assume that two agents, the Defender and the Attacker, deal with a problem which may be modelled with a BAID as defined in Section 2. We first describe the basic BAID reduction operations; then, we provide a brief review of relevance concepts, see further details in Koller and Milch (2003). Finally, we provide an algorithm for ARA support to an agent in a problem modeled as a BAID.

4.1 BAID reduction operations

As shown in our running example, we may distinguish between reduction operations referring to $\mathcal{D}$ steps and reduction operations referring to $\mathcal{A}$ steps. Those in relation with $\mathcal{D}$ steps correspond to standard ID reduction operations as in Shachter (1986), which we just enumerate here:

- Barren node removal. Barren nodes do not affect the problem value and thus can be eliminated from the ID.
- Arc inversion. This corresponds to applying Bayes' formula.
- Chance node removal. This corresponds to computing expected utilities with respect to the corresponding probability distribution.
- Decision node removal. This corresponds to computing (and storing) the maximum expected utility alternative.

Those in relation with $\mathcal{A}$ steps must take into account the uncertainty about the probabilities and utilities of the Attacker and thus need to be conveniently modified, except for the barren node removal which coincides. With no loss of generality, we may assume that all involved random utilities and probabilities are defined over the same basic probability space:

- A-Arc inversion. With the same preconditions and postconditions as in Shachter (1986), we condition on the base measure for the random probabilities and apply Bayes' formula, to obtain new random probabilities over the nodes.
- A-Chance node removal. With the same preconditions and postconditions as in Shachter (1986), we condition on the base measure for the random probabilities and utilities, and compute expected utilities, to obtain random expected utilities.
- A-Decision node removal. With the same preconditions and postconditions as in Shachter (1986), we condition on the base measure for the random probabilities and utilities, and compute optimal alternatives, to obtain random optimal alternatives. We also store the maximal random expected utility.

4.2 Relevance concepts

The key concept is that of strategic relevance. A decision node N_j is strategically relevant for decision node N_i if to make the decision N_i we need to know the decision made at N_j. The relevance graph for a BAID is a directed graph whose nodes are the decision nodes of the BAID, which contains an arc $N_j \rightarrow N_i$ if and only if N_j is strategically relevant for N_i. If two decision variables N_i and N_j rely on each other, we say that the relevance graph is cyclic. On the other hand, a set $\mathcal{N}$ of nodes is said to be a strongly connected component (SCC) if for every pair of nodes $N_i, N_j \in \mathcal{N}$, there exists a directed path from N_i to N_j. A maximal SCC is an SCC which is not a strict subset of any other SCC. The component graph, which can be shown to be always acyclic, is composed of the maximal SCCs.

For any BAID, we can construct the relevance graph following the strategy in Koller and Milch (2003). If all the decisions made by the Defender and the Attacker are accomplished sequentially, the relevance graph is acyclic, and we then obtain a complete topological ordering of the decision nodes, $N_1 \rightarrow N_2 \rightarrow \cdots \rightarrow N_{m+n}$, such that if N_j strategically relies on N_i, then N_i precedes N_j. Such ordering corresponds to the particular sequence of movements that both adversaries carry out during their interaction, corresponding to our sequential Defend-Attack or Attack-Defend blocks.

However, it could happen that certain pairs of decisions are made simultaneously—one by each agent—without knowledge to each other, corresponding to our simultaneous Defend-Attack blocks. Such paired decisions rely on each other, preventing the establishment of a precedence ordering between them and, consequently, making the relevance graph cyclic. Should that be the case, we would build the component graph from the SCCs and proceed in a similar manner as in the acyclic case.

4.3 The global approach

We describe now the global approach in Algorithm 3. To simplify the discussion, we assume that the relevance graph is acyclic.

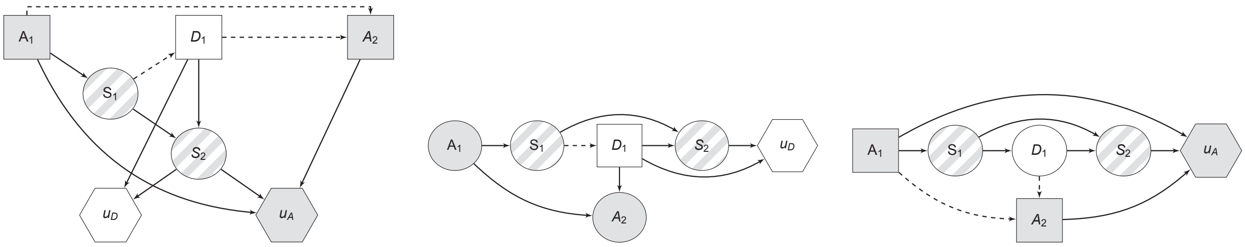
Algorithm 3 General computational strategy

```
Build D’s problem, check if proper.  If not, STOP
Build A’s problem, check if proper.  If not, STOP
Repeat no decision node antecessors of value node left in D’s problem
    While no A assessment required
        Apply as many standard ID reduction operations (invert D arcs, remove D
        decision nodes, remove D chance nodes, remove barren nodes) to D problem
        (in Shachter’s order).
    Repeat requested A assessment obtained
        Apply A-ID reduction operations (A-arc inversion, A-chance node removal,
        A-decision node removal, barren node removal) to A problem (in Shachter’s
        order).
```

5 Revisiting the Tree Killer Problem

We end up illustrating the general methodology with the sequence of moves in the tree killer example from Koller and Milch (2003), which we adapt to the following CIP scenario described in Figure 11. A terrorist group is planning to attack the control centre of a critical infrastructure, killing or taking as hostages the personnel working at the premises. This corresponds to decision node A_2. The critical infrastructure is protected with a sophisticated computerised surveillance system, which poses considerable difficulties to any unauthorised person trying to access the premises. The Attacker is considering the possibility of hacking the security system, represented by the decision node A_1. The preparation and execution of the cyber-attack entails some costs for the Attacker, which are incorporated into the utility node u_A. The result of the cyber-attack is uncertain, as described by chance node S_1, affecting the security system in case of being successful. We assume that the Defender will not be able to ascertain whether a cyber-attack has been launched, but she will be able to detect any unusual performance in the HW/SW components conforming the security system, i.e., she observes the outcome of chance node S_1. Should she perceive anomalies, she could decide to accomplish a deep inspection of all security protocols. This corresponds to decision node D_1, which would entail some costs to the Defender, whose consequences are encompassed in the utility node

u_D. In return for making such decision, the vulnerability of the security system would be reduced to some extent, guaranteeing a better protection than if no action is made. However, there is still a possibility that the whole security system collapses, as reflected by the chance node S_2. Meanwhile, the Attacker must make a decision about attacking physically the control centre, as he fears that new protective measures could eventually be deployed in a near future. When he makes his decision A_2, he knows whether the Defender has accomplished the inspection D_1, but he does not know the real impact of his cyber-attack on the security system. Finally, depending on the outcome of S_2, this would entail consequences both for the Defender and the Attacker, which will be also aggregated in utility nodes u_D and u_A, respectively. Figure 11 provides the BAID, which coincides with that in Koller and Milch (2003), except that we have aggregated the utility nodes. We also present the Defender and the Attacker problems as influence diagrams.



(a) BAID (b) Defender problem (c) Attacker problem

Figure 11: Adapted tree killer problem.

The relevance graph is shown in Figure 12. The topological ordering induced is $D_1 \to A_2 \to A_1$.

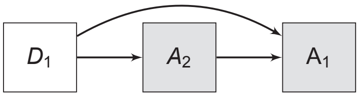


Figure 12: Relevance graph for the BAID example.

5.1 Supporting the Defender

Since the only D decision node precedes all A decisions, the Defender essentially solves a standard decision analysis problem. Indeed, based on Figure 11b, we assess from the Defender her utility $u_D(d_1, s_2)$ and probabilities $p_D(s_2|s_1, d_1)$. In theory, we should also assess $p_D(a_2|a_1, d_1)$, $p_D(s_1|a_1)$ and $p_D(a_1)$, but we shall see that we do not actually need them. Then, we proceed through the following steps:

$\mathcal{D}$1. Eliminate the chance (for the Defender) node A_2, as it is a barren node.

$\mathcal{D}$2. Eliminate chance node S_2 by computing the expected utilities

$$\psi_D(s_1, d_1) = \int u_D(d_1, s_2) p_D(s_2|s_1, d_1)\, \mathrm{d}s_2.$$

$\mathcal{D}$3. Eliminate decision node D_1 by computing $d_1^*(s_1) = \arg\max_{d_1} \psi_D(s_1, d_1)$.

Once the Defender observes s_1, then her optimal decision is $d_1^*(s_1)$. Thus, note that the Defender does not need to switch to the $\mathcal{A}$ problem in this particular case. A more complete response, would accompany the above answer with the assessment of $p_D(s_1|a_1)$, which requires assessing $p_D(a_1)$, for which we would need to switch to problem $\mathcal{A}$, but we shall not refer to this here.

5.2 Supporting the Attacker

Suppose now that we support the Attacker in his decision making. We thus invert the roles of Defender and the Attacker. Based on Figure 11c, we need to assess his utilities $u_A(a_1, a_2, s_2)$ and probabilities $p_A(s_2|s_1, d_1)$, $p_A(d_1|s_1)$ and $p_A(s_1|a_1)$. Once with them, he may solve the diagram through these steps:

$\mathcal{A}$1. Eliminate chance node S_2 by computing the expected utilities

$$\psi_A(a_1, s_1, d_1, a_2) = \int u_A(a_1, a_2, s_2) p_A(s_2|s_1, d_1)\,\mathrm{d}s_2.$$

$\mathcal{A}$2. Eliminate decision node A_2 by maximising expected utilities

$$\psi_A(a_1, s_1, d_1) = \max_{a_2} \psi_A(a_1, s_1, d_1, a_2)$$

and recording

$$a_2^*(a_1, s_1, d_1) = \arg\max_{a_2} \psi_A(a_1, s_1, d_1, a_2).$$

$\mathcal{A}$3. Eliminate chance (for the attacker) node D_1 by computing the expected utilities

$$\psi_A(a_1, s_1) = \int \psi_A(a_1, s_1, d_1) p_A(d_1|s_1)\,\mathrm{d}d_1.$$

$\mathcal{A}$4. Eliminate chance node S_1 by computing the expected utilities

$$\psi_A(a_1) = \int \psi_A(a_1, s_1) p_A(s_1|a_1)\,\mathrm{d}s_1.$$

$\mathcal{A}$5. Eliminate decision node A_1 by computing

$$a_1^* = \arg\max\ \psi_A(a_1).$$

Then, a_1^* and $a_2^*(a_1^*, s_1, d_1)$ constitute an optimal decision for the Attacker.

Note though that the Attacker would need to assess $p_A(d_1|s_1)$ which has an intrinsic strategic component. One way to obtain it is by spreading the Attacker's uncertainty over the Defender's probabilities and utilities onto the Defender's problem. Thus, if the corresponding Attacker beliefs are modelled through the random utility $U_D(d_1, s_2)$ and probabilities $P_D(s_2|s_1, d_1)$, we propagate them through:

$\mathcal{D}$1. Eliminate A_2, which is a barren node.

$\mathcal{D}$2. Eliminate chance node S_2 by computing the random expected utilities

$$\Psi_D(s_1, d_1) = \int U_D(d_1, s_2) P_D(s_2|s_1, d_1)\,\mathrm{d}s_2.$$

$\mathcal{D}3$. Eliminate decision node D_1 by computing the random optimal alternatives $D_1^*(s_1) = \arg\max_d \Psi_D(s_1, d_1)$.

Then, assuming that node D_1 has a discrete decision space, we have that $p_A(d_1|s_1) = \Pr\big(D_1^*(s_1) = d_1\big)$. In this case, because of the acyclic nature of the relevance graph and there being just one change from D decisions to A decisions, the combined scheme is simple to state.

$\mathcal{S}1$. Obtain $u_A(a_1, a_2, s_2)$ and $p_A(s_2|s_1, d_1)$. Eliminate chance node S_2 and decision node A_2, thus performing $\mathcal{A}1$–$\mathcal{A}2$.

$\mathcal{S}2$. At step $\mathcal{A}3$, we need $p_A(d_1|s_1)$. We switch to $\mathcal{D}$ problem to perform steps $\mathcal{D}1$–$\mathcal{D}3$ to get it. Then, perform $\mathcal{A}3$.

$\mathcal{S}3$. Perform steps $\mathcal{A}4$–$\mathcal{A}5$.

6 Discussion

We have provided an adversarial risk analysis approach to dealing with BAIDs. We assume we were supporting one of the agents, who essentially forecasts the actions of the adversary and then optimises. This needs to be performed sequentially based on the informational needs of the supported agent, which takes advantage of the strategic relevance concept. This leads to a strategy which combines simulation and optimisation stages.

As with standard the IDs the approach has exponential complexity and an effort should be made to improving computational response times. One possibility would be to try a single stage process based on the augmented simulation approach to ID reduction see Bielza et al. (1999). It might be interesting also to produce a computational environment supporting this methodology possibly linked with GeNIe (2014).

Acknowledgments

This work has been partly funded by the European Union's 7th Framework Programme under grant agreement no 285223 - SECONOMICS.

BIBLIOGRAPHY

C. Bielza, P. Müller, and D. Ríos Insua. Decision analysis by augmented probability simulation. *Management Science*, 45(7):995–1007, 1999.

G. Brown, M. Carlyle, J. Salmerón, and K. Wood. Defending critical infrastructure. *Interfaces*, 36(6): 530–544, 2006.

G. Brown, W. M. Carlyle, and R. Wood. Optimizing department of homeland security defense investments: Applying defender-attacker(-defender) optimization to terror risk assessment and mitigation. Technical report, National Academies Press, Appendix E, 2008.

J. Cano, D. Ríos Insua, A. Tedeschi, and U. Turhan. Security economics: An adversarial risk analysis approach to airport protection. *Annals of Operational Research*, Accepted for publication, 2014.

GeNIe. `GeNIe` (Graphical Network Interface) Software Package. http://genie.sis.pitt.edu, 2014.

J. C. Harsanyi. Games with Incomplete Information Played by “Bayesian” Players, I-III. Part I. The Basic Model. *Management Science*, 14(3):159–182, 1967.

D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003.

S. A. Lippman and K. F. McCardle. Embedded Nash bargaining: Risk aversion and impatience. *Decision Analysis*, 9(1):31–40, 2012.

H. Raiffa, J. Richardson, and D. Metcalfe. *Negotiation Analysis: the Science and Art of Collaborative Decision Making*. Harvard University Press, 2002.

J. Ríos and D. Ríos Insua. Adversarial risk analysis for counterterrorism modeling. *Risk Analysis*, 32 (5):894–915, 2012.

D. Ríos Insua, J. Ríos, and D. Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854, 2009.

D. Ríos Insua, J. Cano, W. Shim, F. Massacci, and A. Schmitz. SECONOMICS “Socio-Economics meets Security”. Deliverable 5.1. Basic Models for Security Risk Analysis. Technical report, European Union, 2013.

J. C. Sevillano, D. Ríos Insua, and J. Ríos. Adversarial risk analysis: The Somali pirates case. *Decision Analysis*, 9(2):86–95, 2012.

R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

D. O. Stahl and P. W. Wilson. On players’ models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

S. Wang and D. Banks. Network routing for insurgency: An adversarial risk analysis framework. *Naval Research Logistics*, 58(6):595–607, 2011.

Optimal CIP with Network Structure

An escalation in terrorist attacks in recent years has stirred the development of new methodological solutions to support governments in their fight against such threats. In this paper, we use the adversarial risk analysis (ARA) framework to deal with the protection of critical networked infrastructures. We deploy an ARA model for each relevant element (node, link, hotspot in link) in the network, taking the so-called Sequential Defend-Attack-Defend model as a reference. We assume that there will be just one attack against the whole infrastructure over the relevant planning period. The ARA models over the network elements are related by resource constraints and aggregation of results over various sites for the Defender. As an illustration, we consider the case of a potential terrorist attack over a section of the Spanish rail network.

1 Introduction

Security is one of the main concerns of governments and organisations worldwide, see e.g. the World Economic Forum (2013, 2014) Global Risks reports over the last few years. Indeed, large-scale terrorist attacks constitute one of the most worrisome threats faced by authorities. As a consequence, there has been an intense debate in recent years about how to best defend critical infrastructures against terrorist attacks. Large-scale terrorist events like 9/11 or the Madrid train bombings have led to significant national investments in protective responses, see Haberfeld and von Hassell (2009). Such expenditures have received ample criticism from broad sectors of public opinion, who have seen them sometimes as disproportionate, see e.g. Sunstein (2007).

The complexity of the current phenomenon of terrorism and its associated challenges reinforce the need to implement novel analysis tools to support decision makers, see e.g. Ezell et al. (2010) or Wein (2009) for recent accounts of various methodologies and applications. Lewis (2006) proposes several quantitative procedures for evaluating the vulnerability of critical infrastructures against terrorist threats and establishing optimal policies for reducing such weaknesses. In the same line, Brown et al. (2006) apply game-theoretic based optimisation models to make critical infrastructures more resilient against terrorism. From a more qualitative perspective, Haimes and Longstaff (2002) ponder on the usefulness of risk analysis as a valid tool for CIP. Other relevant works include Parnell et al. (2008) and Enders and Sandler (2011), who provide overviews on strategies, models, and research issues in terrorism risk analysis.

Most of the previous approaches to this problem have a game-theoretic flavour. As such, they can take account of the intelligent and adaptive nature of the attackers. However, some of the assumptions on which they are based are not realistic, mainly that of common knowledge about each other's preferences and probabilities. Alternatively, we propose the use of Adversarial Risk Analysis (ARA) to address the problem of how to best protect from intelligent terrorist threats. ARA provides one-sided prescriptive support to one of the intervening agents (she, the Defender), based on a subjective expected utility model, treating the decisions of the adversary (he, the Attacker) as uncertainties.

We shall consider here problems in which an organisation needs to protect a critical networked infrastructure (CNI) from terrorist threats. For a recent review on security analysis of general critical network systems under terrorist threats, see Dziubiński and Goyal (2013). Networked infrastructures are systems composed of two types of elements: nodes and links. Some links may also contain critical points with a particularly important strategic, economic and/or functional value. There are numerous examples of CNIs, including water, oil or natural gas pipelines, transportation routes and facilities, power grids, and telecommunication networks, see Lewis (2006). Hausken and Levitin (2012) provide a classification of systems defence and attack models. Within such classification, we shall be facing a case of protecting a network from attacks against a single element with incomplete information.

Specifically, our case study will refer to deciding the security resource allocation for a railway network whose operator fears the onslaught of a terrorist group. Railway systems constitute a preferred terrorist target. In our problem, we assume that the relevant authorities will face a single strike. In principle, the terrorists may attack any point in the network, trying to wreak havoc, damage network elements and cause the largest number of casualties. Indeed, attacks against rail targets entailing large number of casualties have taken place all over the world since the beginning of the 21st century, see Table 1. Following these events, various studies on the vulnerability of railway transport have been conducted, see e.g. Haberfeld and von Hassell (2009). They highlight that despite all the improvements in security systems, it is difficult to control and protect stations and tracks against terrorist attacks 24/7. Therefore, railway networks still constitute very vulnerable targets.

Table 1: Terrorist attacks on rail transport. Source: Haberfeld and von Hassell (2009)

Date	Country	Casualties
2001/8/1	Angola	> 250
2004/3/11	Madrid, Spain	> 190
2005/7/7	London, UK	> 50
2006/7/11	Mumbai, India	> 180
2007/2/17	Pakistan	> 60

To address the protection of the railway network, we will adopt an ARA strategy, based on the Sequential Defend-Attack-Defend model, see Brown et al. (2006), Parnell et al. (2010) or Ríos and Ríos Insua (2012). In this model, the Defender first deploys defensive resources. Then, the Attacker, having observed such decision, performs an attack. Finally, the Defender tries to recover from the attack as best as she can. We will deploy one of such models for each relevant element in the network: nodes, links and critical points, with possibly different types of resources for each element class. Models will be related by resource constraints and by aggregation of results over various targets for both the Defender and the Attacker. We will populate the model with a case study referring to a section of a national railway system.

The paper is structured as follows. Section 2 outlines the ARA methodology we shall use to address adversarial problems over networks. In Section 3, we describe the main features of the motivating case study. Section 4 adapts the ARA methodology for networks to deal with the case study, whereas Section 5 discusses relevant modelling issues. We remain at a conceptual level not focusing on computational aspects, which may be seen in Ríos Insua et al. (2013). We end up with some discussion.

2 Methodology

We present the ARA methodology that we shall use to deal with CNI protection problems. A network operator (she, the Defender) needs to protect from terrorist threats (he, the Attacker). As mentioned, networked infrastructures are composed of two types of elements: nodes and links. Besides them, we shall also consider that some links may include critical points, which are specific locations of particularly high-value for both terrorists and the operator. For our case of railway networks, nodes correspond to stations, whereas links refer to the tracks connecting them. Critical points could be, e.g., viaducts, tunnels or bridges along certain routes.

We thus consider that we have, in general, ℓ targets, $g_1, \dots, g_\ell$, of which ℓ_s correspond to nodes, ℓ_r to links, and ℓ_c to critical points, with $\ell = \ell_s + \ell_r + \ell_c$. We assume that there is a unique directed path connecting any two nodes. For example, if there is a path joining $A \rightarrow B \rightarrow C$, we rule out the possibility of having a directed link $A \rightarrow C$.

For the protection of each element, we shall deploy a particular ARA model. The models will be related by resource constraints and aggregation of results over various sites for the Defender and the Attacker. As an illustration, consider a generic element in the network, and assume we use a Sequential Defend-Attack-Defend model for it, shown in Figure 13. Note though that other models could be used and, even, different models for different types of elements. Node D_1 corresponds to the Defender's initial decision about the deployment of countermeasures to protect herself against the terrorist threat. After observing such deployment, the Attacker decides his attack A. S_1 represents the outcome of the attack. The Defender, then, makes a decision D_2, trying to recover as best as she can from the attack. S_2 expresses the result of the Defender's attempt to recover from it. The adversaries aggregate their consequences in nodes c_D and c_A, respectively. The consequences for the Defender depend on (d_1, s_1, s_2), i.e., on the effort in implementing her protective measures and the results of the initial attack and her recovery action. Similarly, the consequences for the Attacker depend on (a, s_1, s_2), i.e., on the costs of launching his attack, its outcome, and the result after the Defender's subsequent move. The obtained results would then be aggregated by the Defender over the various sites. The consequences are then evaluated through their respective utility nodes, u_D and u_A.

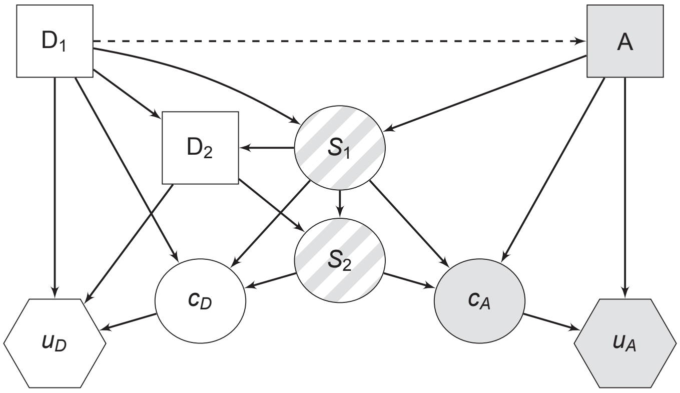


Figure 13: Influence diagram for the railway network case study.

The Defender aims at finding her optimal defence strategy $(d_1^*, d_2^*(d_1^*, s_1))$. For this, she needs to assess the probability models $p_D(s_1|d_1, a)$ and $p_D(s_2|s_1, d_2)$, reflecting her beliefs about: (i) The initial attack outcome, S_1, when defensive resources d_1 have been deployed and the Attacker launches attack a; and (ii) The final attack outcome, S_2, when the result of the attack is s_1 and the recovery action d_2 has been performed. The Defender would also need to assess $p_D(a|d_1)$, expressing her beliefs about what attack will the Attacker choose once he has observed the deployed countermeasures. Assuming we obtain such assessments, we would solve the Defender's problem in Figure 14, based on the standard influence diagram reduction algorithm, see Shachter (1986), proceeding through the steps below.

D0. Aggregate results over the various sites

D1. At node c_D, obtain the expected utilities

$$(d_1, s_1, s_2) \to \psi_D(d_1, s_1, s_2) = \int u_D(c_D) p_D(c_D|d_1, s_1, s_2)\,\mathrm{d}c_D.$$

D2. At node S_2, obtain the expected utilities

$$(d_1, s_1, d_2) \to \psi_D(d_1, s_1, d_2) = \int \psi_D(d_1, s_1, s_2) p_D(s_2|s_1, d_2)\,\mathrm{d}s_2.$$

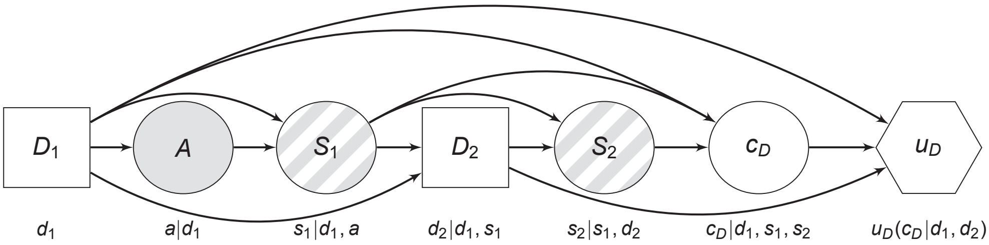


Figure 14: Defender's influence diagram.

D3. At node D_2, compute the optimal decisions

$$(d_1, s_1) \to d_2^*(d_1, s_1) = \arg\max_{d_2 \in \mathcal{D}_2} \ \psi_D(d_1, s_1, d_2).$$

D4. At node S_1, obtain the expected utilities

$$(d_1, a) \to \psi_D(d_1, a) = \int \psi_D(d_1, s_1, d_2^*(d_1, s_1)) p_D(s_1|d_1, a) \, \mathrm{d}s_1.$$

D5. At node A, obtain the expected utilities

$$d_1 \to \psi_D(d_1) = \int \psi_D(d_1, a) p_D(a|d_1) \, \mathrm{d}a.$$

D6. Finally, at node D_1, compute the optimal decisions

$$d_1^* = \arg\max_{d_1 \in \mathcal{D}_1} \ \psi_D(d_1).$$

Provided that the number of portfolios is not too large, this can be accomplished by enumerating and evaluating all $d_1 \in \mathcal{D}_1$. Should that number be large, we could proceed by simulating ψ_D at a few d_1 values, fitting a regression metamodel $\widehat{\psi}_D(d_1)$, see e.g. Kleijnen and Sargent (2000), and solving for

$$\max_{d_1 \in \mathcal{D}_1} \quad \widehat{\psi}_D(d_1).$$

We shall typically need Monte Carlo simulation to perform the above computations.

The only nonstandard assessment in the above formulation is that of $p_D(a|d_1)$. To assess it, the Defender may consider the Attacker's decision problem, see Figure 15, possibly as described below. Assuming that the Attacker is an expected utility maximiser, see French and Ríos Insua (2000), the Defender would need his probabilities $p_A(c_A|a, s_1, s_2)$, $p_A(s_2|s_1, d_2)$, $p_A(d_2|d_1, s_1)$ and $p_A(s_1|d_1, a)$ and utility $u_A(c_A)$. However, the Defender lacks knowledge about them. Suppose we may model her uncertainty through random utilities and probabilities $\big(U_A(c_A), P_A(c_A|\cdot), P_A(s_2|\cdot), P_A(d_2|\cdot), P_A(s_1|\cdot)\big)$. Then, for each d_1, she would solve

A0. Aggregate results over the various sites

A1. At node c_A, obtain the (random) expected utilities

$$\Psi_A(a, s_1, s_2) = \int U_A(c_A) P_A(c_A|a, s_1, s_2) \, \mathrm{d}c_A.$$

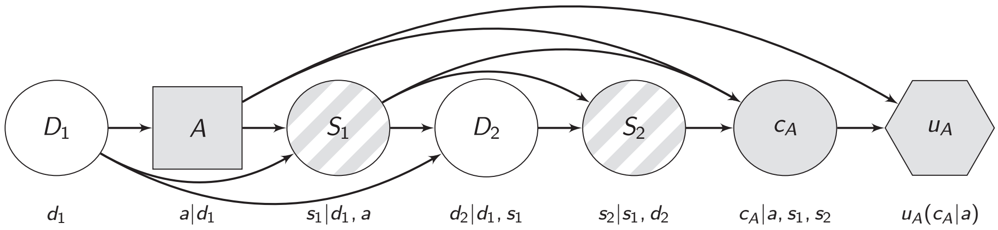


Figure 15: Attacker's influence diagram.

A2. At node S_2, obtain the (random) expected utilities

$$\Psi_A(a, s_1, d_2) = \int \Psi_A(a, s_1, s_2) P_A(s_2|s_1, d_2)\, \mathrm{d}s_2.$$

A3. At node D_2, obtain the (random) expected utilities

$$\Psi_A(a, s_1) = \int \Psi_A(a, s_1, d_2) P_A(d_2|d_1, s_1)\, \mathrm{d}d_2.$$

A4. At node S_1, obtain the (random) expected utilities

$$\Psi_A(d_1, a) = \int \Psi_A(a, s_1) P_A(s_1|d_1, a)\, \mathrm{d}s_1.$$

A5. At node A, obtain the (random) optimal decisions, given d_1

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

Then, we would get $\int_{-\infty}^{a} p_D(\xi|d_1)\, \mathrm{d}\xi = \Pr(A^*(d_1) \leq a)$, which may be approximated through

Algorithm 4 Simulating the Attacker's problem

```
For $d_1 \in \mathcal{D}_1$
    For $k = 1$ to $K$
        For $a \in \mathcal{A}$
            Sample $\left(U_A^k(c_A), P_A^k(c_A|\cdot), P_A^k(s_2|\cdot), P_A^k(d_2|\cdot), P_A^k(s_1|\cdot)\right)$
            Compute $\Psi_A^k(a, s_1, s_2) = \int U_A^k(c_A) P_A^k(c_A|a, s_1, s_2)\, dc_A$
            Compute $\Psi_A^k(a, s_1, d_2) = \int \Psi_A^k(a, s_1, s_2) P_A^k(s_2|s_1, d_2)\, ds_2$
            Compute $\Psi_A^k(a, s_1) = \int \Psi_A^k(a, s_1, d_2) P_A^k(d_2|d_1, s_1)\, dd_2$
            Compute $\Psi_A^k(d_1, a) = \int \Psi_A^k(a, s_1) P_A^k(s_1|d_1, a)\, ds_1$
            Compute $a^k(d_1) = \arg\max_{a \in \mathcal{A}} \Psi_A^k(d_1, a)$
    Approximate $\int_{-\infty}^{a} \widehat{p_D}(\xi|d_1) d\xi \approx \#\{1 \leq k \leq K : a^k(d_1) \leq a\}/K$
```

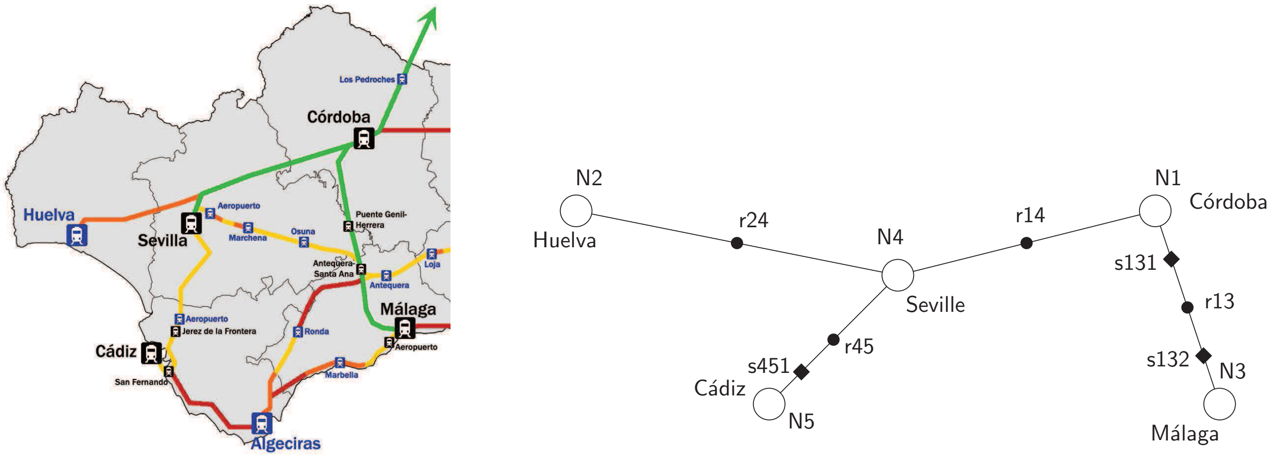
Note that different models could be used for different targets to accommodate their own specificities. Similarly, the nature and number of deployable resources may depend on the type of target.

3 Case Description

We analyse the protection of the southwest section of the Spanish railway system against terrorist threats. Recent intelligence reports have alerted about the activation of a dormant cell established in Seville, integrated within the city for years without raising suspicion. In this regard, several Al-Qaeda members have been arrested in Southern Spanish towns in the last years, see BBC News Europe (2011, 2012); New York Post (2014). The terrorists intend to launch an attack in summer against the railway system and its users, taking advantage of large population flows during the vacation period along the Andalusian coast.

3.1 Layout

The section of the Spanish railway system under consideration is shown in Figure 16a. We have represented the network more schematically in Figure 16b, indicating the stations (N), routes (r) and critical points (s). The latter correspond to the following sensitive areas: s131 represents Puente Genil's viaduct in the Córdoba-Málaga route, whereas s132 and s451 stand for tunnels in Antequera and Jerez de la Frontera, in the Córdoba-Málaga and Seville-Cádiz routes, respectively.



(a) Network map.

(b) Network scheme.

Figure 16: Railway network for the case study.

Table 2 displays the main features of the stations (associated population), routes (length) and critical points within the incumbent network. In what follows, we shall use the notation $g_1, \ldots, g_{12}$ for the twelve potential targets.

Table 2: Targets and their main features

Station (city population)					Route (km)				Critical points (distance from head)		
g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}
Córdoba	Huelva	Málaga	Seville	Cádiz	Co-Ma	Co-Se	Hu-Se	Se-Ca	s131 (v)	s132 (t)	s451 (t)
328.488	148.568	567.433	702.355	123.948	159	142	92	123	Co-Ma (71)	Co-Ma (107.5)	Se-Ca (84.8)

The type of train circulating across the network is composed of three units, each of which consists of five coaches. The train maximum capacity is 780 seats, which is typically full at this time of the year.

3.2 Countermeasures deployable by the railway operator

The railway operator is a public company attached to the Ministry of Public Works, responsible of security, in coordination with Homeland Security and Defense. It has a security budget to be distributed among various countermeasures across the network, subject to constraints specified below. Countermeasures can be static or mobile; the latter may be used for recovery purposes in the event of a successful attack. Note that by ’recovery’ we mean solely the detention of terrorists, thus not considering those protocols that the government and railway operator should carry out after an attack to: (1) Ensure an adequate response to potential victims; and (2) Restore the service as soon as possible. These aspects pertain to the domain of contingency plans, with a special unit attached to the Ministry of Defence, trained to intervene in emergencies. However, the special unit is not responsible of capturing terrorists, which belongs to Homeland Security, so we will not include its eventual deployment in our model.

The countermeasures that the operator considers include:

- Walk-through detectors. Guarantee all passengers to be checked. Placed at passenger entrance of platform areas, trying to detect metal objects that could be part of an explosive. May provoke long queues, especially at peak hours.
- Cameras. Part of a CCTV system to monitor sensitive areas in stations, trains and tracks.
- Additional illumination. Long-range self-powered directional spotlights in specific areas, improving visibility at night.
- Fences. Prevent persons and animals from accessing tracks in sensitive areas.
- Security staff. Patrol along stations, checking anyone suspicious. Periodically, will inspect critical points along the routes accompanied by detection dogs.
- Detection dogs. Trained to detect explosives. Must be always accompanied by security personnel.
- Helicopter patrols. Each helicopter is assigned to two crews of one pilot and one co-pilot. Will randomly patrol along the network, communicating anomalies to mobile units at stations or critical points.

All the above measures will have a deterrent effect over the terrorist intentions. Besides, security personnel and detection dogs have also a recovery role, as they will help in trying to capture the terrorists in case of an attack. Table 3 summarises the features of security measures, together with their associated unit costs. For the security staff, we provide their unit monthly gross salaries. Agents will be outsourced to a private security company for two months—July and August, the period in which the terrorist alert is active. Similarly, we have provided the monthly cost of hiring a detection dog from a specialised company. The costs entailed by the helicopter patrols include crew salaries and those associated with equipment operation and maintenance. We have included the overall costs of installing a walking-through detector, a CCTV camera, a lamp post and a fence unit, respectively, over the planning period, including their eventual maintenance and repair, and taking into account their typical lifetime.

3.3 Possible terrorist strategies

In order to assess our uncertainty about the Attacker’s decision, we must analyse the motivations of this type of terrorist groups, see Keeney and von Winterfeldt (2010). In this regard, terrorists are

Table 3: Features of security measures and associated costs

	Measure	Station	Track	Critical points	Type	Unit costs (€)
x_1	Metal detector	Yes	No	No	Static	6,500
x_2	CCTV camera	Yes	No	Yes	Static	650
x_3	Lamp posts	No	No	Yes	Static	3,000
x_4	Fence units	No	No	Yes	Static	4,200
x_5	Security guard	Yes	No	Yes	Mobile	2,600
x_6	Detection dog	Yes	No	Yes	Mobile	800
x_7	Helicopter	No	Yes	No	Mobile	90,000

usually assumed to be risk seekers, being their main aims: (1) Inflict the largest number of casualties and damage to their enemies; (2) Have large impact on the media; and (3) Cause panic and chaos within the civilian population. In our scenario, this turns into the following facts:

- The places where the largest number of casualties can be caused are stations, due to their higher concentration of people.

- The most devastating attack would be the detonation of an explosive inside a train as it enters a station. This would cause victims within the train and at the station platform, besides damaging the train, the station and other assets.

- An attack using weapons of mass destruction (WMD) would cause the greatest panic among civilians, as with the sarin gas released at the Tokyo subway attack in 1995, see Haberfeld and von Hassell (2009, chap. 12) and Fellman et al. (2011).

We consider that the terrorist cell has resources to launch a single attack over a single target within the network over the incumbent period. We further assume that the attack strategies being considered by the terrorists are:

- a_1. Place bomb in station.

- a_2. Place bomb inside a train and detonate when approaching a station.

- a_3. Place bomb along the route, and make it explode as a train passes by.

- a_4. Place WMD of chemical or biological nature inside a train and detonate it either when approaching a station.

Note that for attack type a_3, terrorists would ideally aim at putting the bomb as close as possible to a critical point, so as to cause the largest damage. However, critical points are, normally, particularly well protected spots, thus becoming a more perilous target for terrorists. We address this issue in Section 5.2.

In deciding their strategy, the terrorists should take into account the following inherent risks: (1) Manipulation and transportation of bombs or weapons intended to be used in the attack, since, for instance, WMDs require more caution and care than explosives; (2) The possibility of being detected by security forces (aided by technological resources) before putting the explosive in the station, train or tracks; and (3) Introducing the explosive inside the train and not being able of getting off it before detonation.

4 The Model

We discuss now in detail the model elements for Section 2, particularised to our case study.

- D_1 represents the decision about the countermeasures deployed by the Defender at each of the targets in Table 2. Let x_{ik} be the number of units of countermeasure x_i deployed over objective g_k, $i = 1, \ldots, 7$, $k = 1, \ldots, 12$, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i12})$, and $\overline{x_i} = \sum_{k=1}^{12} x_{ik}$. Their unit costs q_i, $i = 1, \ldots, 7$ were defined in Table 3. We denote the set of possible decisions by $\mathcal{D}_1$. The countermeasures must fulfil the following constraints, globally displayed in (9):
  - Political. All stations and critical points must have a minimum level of protection, with at least three security guards and two detection dogs patrolling 24/7 (9h).
  - Economic. Due to their high costs, a maximum of two metal detectors can be installed at each station (9g). Similarly, at most two helicopters may be used for patrolling (9i).
  - Logistic. Only mobile elements can be shifted from the nearest post to the point in which the attack has occurred. Walk-through scanners will be used only at stations (9c). Cameras and mobile resources will be used only at stations and around critical points (9d). Lamp posts and fences will be used only around critical points in the routes (9e). Helicopters will only patrol over unprotected track stretches in the network (9f). Detection dogs must be at any time accompanied by a security guard (9h).

  Let B_D be the maximum budget available for new countermeasures (9a), on top of the existing ones, for the relevant two-month planning period. Then, the feasible security portfolios $d_1 = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7)$ will satisfy the above mentioned constraints expressed through

$$c_{\text{inv}} = \sum_{i=1}^{7} q_i \overline{x_i} \leq B_D, \tag{9a}$$

$$x_{ik} \text{ integer}, \ i = 1, \ldots, 7, \ k = 1, \ldots, 12, \tag{9b}$$

$$x_{1k} = 0, \ k = 6, \ldots, 12, \tag{9c}$$

$$x_{2k} = x_{5k} = x_{6k} = 0, \ k = 6, \ldots, 9, \tag{9d}$$

$$x_{3k} = x_{4k} = 0, \ k = 1, \ldots, 9, \tag{9e}$$

$$x_{7k} = 0, \ k = 1, \ldots, 5, 10, \ldots, 12, \tag{9f}$$

$$x_{1k} \leq 2, \ k = 1, \ldots, 5, \tag{9g}$$

$$x_{5k} \geq 3, x_{6k} \geq 2, x_{5k} \geq x_{6k}, \ k = 1, \ldots, 5, 10, \ldots, 12, \tag{9h}$$

$$\overline{x_7} \leq 2. \tag{9i}$$

  Besides the above constraints, the operator also imposes additional lower and upper bounds on the number of deployable units over each target for the rest of the countermeasures, see Table 4. Then, e.g., the constraint referring to upper and lower bounds about x_1 is expressed like $1 \leq x_1 \leq 2$, meaning that, at each station, there must be mandatorily one or, at most, two metal detectors.

  Table 4: Bounds over deployable units of countermeasures

Measure	x_1	x_2	x_3	x_4	x_5	x_6	$\overline{x_7}$
Min	1	0	0	0	3	2	1
Max	2	4	2	2	4	3	2

- A stands for the attack chosen by the terrorists. We denote the set of possible decisions by $\mathcal{A}$. We define the binary decision variables $a_{jk} \in \{0, 1\}$, $j = 1, \ldots, 4$, $k = 1, \ldots, 12$, with $a_{jk} = 1$

representing that terrorists decide to use attack type a_j against target g_k, and $a_{jk} = 0$, otherwise. Since we assume that the terrorists will launch, at most, one attack, this implies that

$$\sum_{j=1}^{4}\sum_{k=1}^{12} a_{jk} \leq 1.$$

The sum will equal zero only if the terrorists decide not to attack at all. Additional logistic constraints apply to the Attacker's decision:

- Strategies a_1, a_2 and a_4 can only be carried out close to the stations, thus $a_{1k} = a_{2k} = 0,\ k \in \{6, \dots, 12\}$.
- Similarly, strategy a_3 can only be implemented when the train is en route. Therefore, $a_{3k} = 0,\ k \in \{1, \dots, 5\}$.

- S_1 reflects the outcome of an eventual attack against a certain target. We shall assume that it takes values in the set $\mathcal{S}_1 = \{0, 1\}$. We define binary variables $s_{jk}^{(1)} \in \mathcal{S}_1$, $j = 1, \dots, 4$, $k = 1, \dots, 12$, with $s_{jk}^{(1)} = 1$ meaning a successful attack of type a_j against target g_k, and $s_{jk}^{(1)} = 0$, otherwise.

- D_2. In case of a successful attack against a station or a critical point, the Defender would use the available mobile resources to try to capture the terrorists. If the attack occurs at some noncritical point along the route, the Defender will shift the available mobile resources from the nearest location, whether a station or a critical point. D_2 takes values in $\mathcal{D}_2 = \{0, 1\}$, with $d_2 = 1$ meaning that the recovery measures are deployed, and $d_2 = 0$, otherwise.

- S_2 represents the result of the Defender's recovery action. It takes values in the set $\mathcal{S}_2 = \{0, 1\}$. We define binary variables $s_{jk}^{(2)} \in \mathcal{S}_2$, $j = 1, \dots, 4$, $k = 1, \dots, 12$, with $s_{jk}^{(2)} = 1$ meaning that all the terrorists—those still operational after an attack of type a_j against target g_k—will be captured, and $s_{jk}^{(2)} = 0$, otherwise.

4.1 The Defender problem

The dynamics for the Defender are:

- She deploys her countermeasures $d_1 = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7)$ over the network elements.

- She observes whether an attack is launched by the terrorists. In case of a successful attack, she would use available security units from the nearest location, trying to capture the terrorists.

- She faces the multiple consequences in relation with the eventual terrorist attack, and her recovery action, as specified below.

- She attains her utility.

The repercussion of a successful attack may be tragic and devastating in terms of casualties, damages and recovery costs, as already experienced in the real cases presented in Table 1. The most relevant consequences for the Defender are listed below:

- Casualties (killed or wounded people), with an associated cost of c_{life}.

- Damages in fixed infrastructure assets (tracks, signals, viaducts, tunnels, stations,...), whose costs will be aggregated into a single variable c_{fixed}.

- Damages in mobile elements (trains), denoted by c_{train}.
- Fear in the population, and a negative perception of security, particularly of the railway transportation system. A negative vision of the country's image overseas. We encompass all these issues in a single variable c_{image}.

We provide details about the involved costs c_{life}, c_{fixed}, c_{train} and c_{image} in Section 5.1. We summarise in Table 5 the eventual consequences of successful attack strategies. In case of no attack, there are no consequences for the Defender, other than the security investment incurred.

Table 5: Type of attacks and consequences

Attack	Description	Lives	Fixed assets	Station	Train	Image
a_1	Bomb in station	Yes	Yes	Yes	—	Yes
a_2	Bomb in train (station)	Yes	Yes	Yes	Yes	Yes
a_3	Bomb en route	Yes	Yes	—	Yes	Yes
a_4	WMD in train	Yes	—	—	—	Yes

We use a measurable multiattribute value function together with relative risk aversion as in Dyer and Sarin (1979, 1982) to model the Defender's utility function. First, the multiattribute value function for the Defender will be described as a linear additive value function, see Keeney and von Winterfeldt (2011), through

$$c_D(d_1) = c_{\text{inv}} + c_{\text{life}} + c_{\text{fixed}} + c_{\text{train}} + c_{\text{image}}, \tag{10}$$

aggregating consequences over all targets. Note that, effectively, we are monetising consequences. The Defender's utility function is, then, $u_D(c_D)$.

4.2 The Attacker problem

The dynamics of the Attacker are:

- They see the preventive measures d_1 deployed by the Defender over the targets.
- They decide about the type of attack and its target.
- In case of attacking, they observe the result of the attack and face its operational consequences.
- In case of a successful attack, they tackle the recovery measures deployed by the Defender and face the consequences.
- They get the corresponding utility.

The Defender considers that the relevant multiple consequences for the Attacker are:

- Preparation costs, c_{prep};
- The costs associated with the number of terrorists killed or detained over the operation, to which the Attacker puts a value c'_{life}; and
- Whether they are able to wreak havoc among population, since terrorists also focus on image consequences, c'_{image}.

We provide details about the involved costs c_{prep}, c'_{life} and c'_{image} in Section 5.2. As before, we first aggregate the attributes in a linear additive value function

$$c_A(a) = -c_{\text{prep}} - c'_{\text{life}} + c'_{\text{image}}, \tag{11}$$

and then build the utility function $u_A(c_A)$.

5 Case Study

We start with the assessment of the quantities related with the Defender's problem. We assume that while the terrorist threat is active, trains operate at their maximum capacity with all seats occupied. The available security budget is $B_D = 270,000$ €.

5.1 Defender's assessments

We assess now the relevant consequences, utilities and probabilities in the Defender's problem.

Human losses c_{life} We first discuss the costs associated with the number of casualties on the Defender's side. As for the number of victims within the station when terrorists are using conventional bombs—attack types a_1 and a_2—we define a random variable z_s accounting for the number of casualties that would occur at the target station. We assume a binomial distribution

$$z_s|g_k \sim \mathcal{B}in(n_s, p_s), \; k = 1, \dots, 5,$$

where n_s depends on the particular station, and represents the theoretical maximum number of people that could be potentially affected by the blast if present at station g_k. Based on expert opinion, we assume n_s to be equal to, approximately, 0.5% of the corresponding population size for each station, see Table 2. In turn, p_s corresponds to the probability of a random person being killed if present at the incumbent station when the attack occurs. The operator assumes the same probability for all stations, although she has uncertainty about its value, placing a beta distribution $p_s \sim \mathcal{B}e(\alpha_s, \beta_s)$. To assess the expected probability of being killed or badly injured in a station, $E[p_s]$, we rely on previous similar terrorist attacks, see Haberfeld and von Hassell (2009, chap. 10): we have estimated the expected number of casualties at 50 for g_5 ($E[z_s|g_5] = 50$), from which we obtain the expected probability, equal to 0.081. The operator also assessed a value for the variance $\text{Var}[p_s] = 0.001$, which will determine, in turn, the values of $\alpha_s = 5.90$ and $\beta_s = 67.27$. Finally, we have obtained the expected number of casualties for the other stations, using their population size ratios, see Table 6. The associated expected costs are obtained by multiplying the expected number of casualties by the statistical value of life for the Defender q_{svl}, see Riera Font et al. (2007). We estimate $q_{\text{svl}} = 2.04$ M€ for killed or badly injured people, disregarding the associated costs of mildly injured people.

Table 6: Expected number of victims after a bomb attack

Target	g_1	g_2	g_3	g_4	g_5	Train in station	En route train
Casualties	133	60	229	283	50	49.4	156

We have also estimated the number of casualties that would occur inside the train, either when it is entering a station, z_e, or en route, z_r, corresponding to attack types a_2 and a_3, respectively. For the first case, we assume that just the coach containing the bomb would be affected, while a whole unit (five coaches) will be damaged in the second one. We use binomial distributions for $z_e \sim \mathcal{B}in(n_e, p_e)$ and $z_r \sim \mathcal{B}in(n_r, p_r)$, being $n_e = 52$ and $n_r = 260$ the maximum capacity of a coach and a unit, respectively. To account for the inherent uncertainty over the probabilities, we use beta distributions for $p_e \sim \mathcal{B}e(\alpha_e, \beta_e)$ and $p_r \sim \mathcal{B}e(\alpha_r, \beta_r)$, with expected values $E[p_e] = 0.95$ and $E[p_r] = 0.6$ and moderate variances 0.02 and 0.01, respectively, as assessed by the operator. The expected numbers of victims are shown in Table 6.

In an attack using WMDs (attack type a_4), according to expert opinion, approximately one half of the train's passengers is expected to be contaminated. We use a binomial distribution $z_w \sim$

$\mathcal{B}in(n_w, p_w)$ for the number of victims, being $n_w = 780$ the train maximum capacity, and a beta distribution for the incumbent probability $p_w \sim \mathcal{B}e(\alpha_w, \beta_w)$ with expected value $E[p_w] = 0.5$ and variance $\text{Var}[p_w] = 0.008$. In addition, all the people present at the station at the time of the attack within a certain distance from the WMD focus would be affected as well. For simplicity, we consider the range of the WMD similar to that of a conventional explosive. Subsuming both contributions, we obtain the estimations in Table 7.

Table 7: Expected number of victims after a WMD attack

Target	g_1	g_2	g_3	g_4	g_5
Casualties	523	450	619	673	440

We aggregate now the costs due to the occurrence of casualties

$$c_{\text{life}} = q_{\text{svl}} \left[\sum_{k=1}^{5} a_{1k} z_s + \sum_{k=1}^{5} a_{2k}(z_s + z_e) + \sum_{k=6}^{12} a_{3k} z_r + \sum_{k=1}^{5} a_{4k}(z_s + z_w) \right],$$

accounting for attack types a_1–a_4. Note that c_{life} would inherit the uncertainty in z_s, z_e, z_r and z_w, just described.

Material losses c_{fixed} For the costs associated with infrastructure damages, we have consulted the construction value of the different elements, assuming an affected area of 500 m around the blast. We denote the costs in tracks and signals, tunnels and viaducts by q_{tas}, q_{tun} and q_{via}, respectively. For damages at stations, denoted by q_{sta}, we have used as a reference the cost estimation of Atocha station damages, one of the targets in the Madrid train bombings, see Buesa Blanco et al. (2005). The estimated values (in millions of euros) of the above costs are displayed in Table 8.

To discriminate whether a bomb has been put sufficiently close to a critical point (i.e. less than 500 m from it) under attack type a_3, we define a binary variable $h \in \{0, 1\}$, such that $h = 1$ if the shock wave affects the critical point, and $h = 0$, otherwise. Note that if $h = 1$, then one of the variables a_{3k}, $k = 10, 11, 12$ must be equal to 1, whereas if $h = 0$, all of them are zero. We describe how to determine h in Section 5.2.

Table 8: Associated costs (M€) with damages in fixed infrastructure assets

	Tracks & signals	Tunnel	Viaduct	Station
Cost	0.700	4.620	2.000	0.218

Since material costs are relatively low compared with human costs, see Tables 6 and 7, we do not take into account their associated uncertainty. We express the overall material costs for the Defender as

$$c_{\text{fixed}} = (q_{\text{tas}} + q_{\text{sta}}) \sum_{j=1}^{2} \sum_{k=1}^{5} a_{jk} + q_{\text{tas}} \sum_{k=6}^{9} a_{3k} + (q_{\text{tas}} + h \cdot q_{\text{via}}) a_{310} + h(q_{\text{tas}} + q_{\text{tun}}) \sum_{k=11}^{12} a_{3k}.$$

The first term corresponds to attacks a_1 and a_2. The others, to attack a_3, distinguishing whether the bomb has been placed at a critical point or not.

We have also estimated the costs associated with damages in the train, differentiating whether it is entering a station (when, according to experts, just the coach carrying the bomb would be affected) or en route (when five coaches would suffer the consequences). Each unit has a cost of $q_{\text{unit}} = 5.77$

M€. Again, we disregard the associated uncertainty due to the comparatively much smaller costs. Then, the costs caused by damages in trains are

$$c_{\text{train}} = q_{\text{unit}} \left[\frac{1}{5} \sum_{k=1}^{5} a_{2k} + \sum_{k=6}^{12} a_{3k} \right],$$

corresponding to attacks a_2 and a_3, respectively.

Image costs c_{image} Image costs will be modelled through a random variable, c_{image}, which depends on the type of strategy chosen by the Attacker and on whether the Defender's recovery action is effective or not. Should that be the case—when all terrorists are detained—the impact on the country's image will be lower than otherwise. We use a truncated normal model with truncation point zero

$$p_D(c_{\text{image}} | a, s_1 = 1, s_2) \sim \mathcal{TN}(\mu_a, \sigma_a^2),$$

whose expected values and variances, shown in Table 9, have been assessed by experts from the operator.

Table 9: Expected image costs and associated standard deviation (M€)

	a_1	a_2	a_3	a_4
$s_2 = 0$	29(14)	29(14)	20(9)	35(17)
$s_2 = 1$	15(7)	15(7)	10(6)	18(8)

Assessment of the Defender's probabilities We discuss now the assessment of the probabilities in the Defender's problem. We start with the probability that an attack of type j against target g_k is successful, conditional on the deployed countermeasures, $p_D\left(s_{jk}^{(1)} = 1 | d_1, a_{jk} = 1\right) = p_{jk}^{(1)}$, $j = 1, \ldots, 4$, $k = 1, \ldots, 12$. The operator is quite confident about her assessment over such probabilities, so we explicitly disregard the associated uncertainty. The Defender believes that the expected probability depends on the risk factors mentioned in Section 3.3. We use

$$E\left[p_{jk}^{(1)}\right] = 1 - (1 - \phi_1) \cdot (1 - \phi_2),$$

where ϕ_1 represents the chance that the bomb is ruined during the manipulation and/or transportation phases, estimated at $\phi_1 = 0.3$ for a conventional bomb and $\phi_1 = 0.1$ for a WMD; and ϕ_2 is the probability that the terrorists are detected, thus being forced to abort the mission. ϕ_2 depends on the type and number of deployed countermeasures. We use an exponential model

$$\phi_2(d_1) = 1 - \phi_d \cdot \exp\left(- \sum_{i=1}^{7} \gamma_i x_{ik} \right)$$

to account for the fact that each countermeasure is expected to decrease the chances of a successful attack. $(1 - \phi_d)$ represents the probability of detection if no additional measures are deployed. Table 10 shows the estimated values of the γ's and ϕ_d, as assessed by an expert, depending on whether the attack requires the terrorists entering a station or not. For the γ's, we asked the experts from the operator about the expected deterrent effect of each countermeasure when considered separately, fitting the expression for ϕ_2 and obtaining the corresponding coefficient.

We assess now the probability that the recovery action—provided the Defender carries it out as a response to a successful attack—is also successful, i.e., $p_D\left(s_{jk}^{(2)} = 1 | s_{jk}^{(1)} = 1, d_2 = 1\right) = p_{jk}^{(2)}$,

Table 10: Probability of detection and deterrence parameters

	ϕ_d	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
a_1, a_2, a_4	0.4	0.28	0.13	—	—	0.69	0.69	—
a_3	0.7	—	—	0.05	0.02	—	—	1.54

$j = 1, \ldots, 4$, $k = 1, \ldots, 12$. Again, due to the small variance assessed by the operator, we explicitly disregard the associated uncertainty. The Defender believes that $p_{jk}^{(2)}$ depends on the number and proximity of the nearest available mobile resources, as we detail. First, we choose an exponential model

$$E\left[p_{jk}^{(2)}\right] = 1 - p_0 \cdot \exp\left(-\mu_5 \tilde{x}_{5k} - \mu_6 \tilde{x}_{6k}\right),$$

to express the fact that the presence of security members and detection dogs would increase the probability of capturing the terrorists. $(1 - p_0)$ represents the probability of a successful recovery if no additional countermeasures are deployed. $\tilde{x}_{5k}$ and $\tilde{x}_{6k}$ stand for, respectively, the number of security members and detection dogs shifted towards target g_k from the nearest locations. For an attack at a station or a critical point, there is no need to transfer resources, since they have their own assignment, see (9h), therefore yielding $\tilde{x}_{5k} = x_{5k}$ and $\tilde{x}_{6k} = x_{6k}$, $k = \{1, \ldots, 5, 10, \ldots, 12\}$. On the contrary, for an en route attack, mobile resources will be transferred from the nearest points (either a station or a critical point). μ_5 and μ_6 represent the efficiency of such countermeasures, attenuated by a remoteness factor: intervention speed is crucial, which is influenced by the distance between the attack point and the post from which resources are mobilised. Denoting such distance (in km) by d, we determine the following expressions:

$$\mu_5 = \mu_{50} \cdot \exp(-\lambda_5 \cdot d), \quad \mu_6 = \mu_{60} \cdot \exp(-\lambda_6 \cdot d).$$

The operator has estimated the parameters $\mu_{50} = 0.28$, $\mu_{60} = 0.13$ and $\lambda_5 = \lambda_6 = 0.03$. She also estimates that p_0 depends on the type of attack, with $p_0 = 0.4$ for attack types a_1, a_2 and a_4, and $p_0 = 0.8$ for a_3.

Finally, we found a good fit with a constant risk averse, with respect to c_D, utility function, thus being (strategically equivalent to) $u_D(c_D) = -\exp(k_D \cdot c_D)$, with $k_D > 0$. We have assessed, with the aid of experts, a value $k_D = 1.4 \cdot 10^{-3}$ for the risk aversion parameter, using standard utility assessment techniques, see Farquhar (1984).

5.2 Defender's assessments of the Attacker's probabilities and utilities

We discuss now the Defender's assessment of quantities corresponding to the Attacker's problem. Note first that the costs associated with the preparation of an attack, c_{prep}, will depend on the type of attack chosen by the Attacker, and are the result of aggregating the following expenses: (1) Fabrication of the explosive; (2) Train ticket; (3) Transportation to the target location. However, we shall disregard all the above costs due to their extremely low relative value compared to other relevant costs in this case study.

Life costs c'_{life} As a consequence of the attack, we take into account the number v of terrorists captured during the preparation of the attack or killed in action, when manipulating the explosive or because they could not get off the train after depositing the bomb. The Defender uses a binomial distribution to model v

$$v | d_1 \sim \mathcal{B}in(t, p_t),$$

where t is the number of terrorists and p_t is the probability that a terrorist is killed or detained in that phase. The Defender believes that the terrorist cell will be composed of between three and five members with probabilities 0.2, 0.5 and 0.3, respectively. She also estimates that p_t can be modelled as $p_t = 1 - (1 - \phi_1)(1 - \phi_2)$, with ϕ_1 and ϕ_2 as in Section 5.1. Recall that p_t depends on d_1 through ϕ_2. We further assume that the terrorists put a value $q'_{\text{svl}} = 300,000$ € to the possibility of being killed or detained. Thus, the total costs faced by the terrorists in this concept would amount to

$$c'_{\text{life}} = q'_{\text{svl}} \cdot v.$$

Note also that, in case of a successful recovery action by the Defender, we have assumed that the rest of the terrorists, $t - v$, will be also killed or detained, so the expected global life costs for the Attacker would then be $c'_{\text{life}} = q'_{\text{svl}} \cdot t$. The uncertainty in v would be propagated to c'_{life}.

Image costs c'_{image} Concerning the image costs, the Defender believes that the terrorists have similar expectations to hers, although with a greater uncertainty. Then, she uses a truncated normal distribution, with the same expected value than $c_{\text{image}} | a, s_1 = 1, s_2$, although with variance ten times bigger, that is,

$$c'_{\text{image}} | a, s_1 = 1, s_2 \sim \mathcal{TN}\left(\mu_a, 10 \cdot \sigma_a^2\right).$$

Once we have assessed the associated random costs for the Attacker, we can aggregate the consequences for him as expressed in (11). The probability distribution describing the Defender's beliefs about such consequences, $p_A(c_A | a, s_1, s_2)$, would inherit the uncertainty in c'_{life} and c'_{image}.

Assessment of the Attacker's probabilities For the Attacker's beliefs over the probability of a successful attack of type a_j against target g_k, we model $P_A\left(s_{jk}^{(1)} = 1 | d_1, a_{jk} = 1\right)$ as a beta distribution, with mean $p_{jk}^{(1)}$ and variance δ_1. The operator has little uncertainty about such assessment of the Attacker's probability. Thus, she sets $\delta_1 = 0.1$.

Mimicking the reasoning about the probability of a successful attack, we assess the Defender's beliefs about the probability that the Attacker gives to a successful recovery action, $P_A\left(s_{jk}^{(2)} = 1 | s_{jk}^{(1)} = 1, d_2 = 1\right)$, as a beta distribution with mean $p_{jk}^{(2)}$ and moderate variance, estimated by the operator at $\delta_2 = 0.15$.

We consider now the case in which the terrorists decide to use attack type a_3, placing a bomb somewhere along the route. If the chosen route does not contain critical points, the operator considers that the terrorists will put the bomb at a random point, θ, sufficiently far from protected places. We model the value of θ through a triangular distribution. Let us consider a railway section $\overline{ab}$ of length d_{ab}, and let (x_{5a}, x_{6a}) and (x_{5b}, x_{6b}) be the number of deployed mobile resources in the extremes (a, b) of such section. Then, the distribution of θ given d_1 is

$$\mathcal{T}ri(a, b, m),$$

where $m = d_{ab}(x_{5a} + x_{6a} + 1)/(x_{5a} + x_{6a} + x_{5b} + x_{6b} + 2)$ is the mode of the triangular distribution, reflecting that the terrorists would tend to put the bomb as far as possible from the most protected places. Note that the shortest of the distances, $d = \min(\overline{\theta a}, \overline{\theta b})$ will determine from which post will the mobile resources be transferred, see Section 5.1.

For a route containing one or more critical points, we will use a mixture model of triangular distributions. Let us discuss the case in which we have one critical point at a given point c. If we denote by (x_{5c}, x_{6c}) the number of deployed mobile resources at the critical point, then, θ given d_1 will follow the mixture

$$w_1 \cdot \mathcal{T}ri(a, c, m_1) + w_2 \cdot \mathcal{T}ri(c, b, m_2),$$

where $w_1 = 1 - w_2 = (x_{5b} + x_{6b} + 1)/(x_{5a} + x_{6a} + x_{5b} + x_{6b} + 2)$ are the weights of the mixture, and $m_1 = d_{ac}(x_{5a} + x_{6a} + 1)/(x_{5a} + x_{6a} + x_{5c} + x_{6c} + 2)$ and $m_2 = d_{cb}(x_{5c} + x_{6c} + 1)/(x_{5b} + x_{6b} + x_{5c} + x_{6c} + 2)$

are the modes of the corresponding mixture component. Again, the shortest of the distances $d = \min(\overline{\theta a}, \overline{\theta b}, \overline{\theta c})$ will determine from which post will the mobile resources be transferred. For a route containing one or more critical points, we will use a similar model, but including as many additional mixture terms as necessary to account for the mobile resources assigned to the incumbent critical points. Finally, note that should θ be a point within 500 m of a tunnel or a viaduct, then such critical point would be affected by the blast and, in consequence, we would set $h = 1$, see Section 5.1.

Assessment of the Attacker's utility We assume that the Attacker is constant risk prone in benefits. Therefore, his utility function will be strategically equivalent to

$$u_A(c_A) = \exp(k_A \cdot c_A), \; k_A > 0.$$

We consider the random utility model for the Attacker

$$U_A(c_A) = \exp(k_A \cdot c_A), \; k_A \sim \mathcal{U}(0, K_A).$$

The Defender thinks that the parameter k_A takes a maximum value $K_A = 2.5 \cdot 10^{-3}$.

6 Discussion

We have analysed the protection of critical networked infrastructures from terrorist attacks. We have considered a generic network in which value can be located at nodes, links and critical points, the latter regarded as particularly important locations for both the Defender and the Attacker. We have deployed an ARA model over each relevant target in the network, relating the models by resource constraints and aggregation of results over various sites for the Defender. The Sequential Defend-Attack-Defend model has been used as a reference template, although different ARA models could be accommodated for each target.

We have illustrated our methodology with a case study dealing with a section of the Spanish railway system. We have remained at a conceptual level not focusing on computational aspects. Through it, we have shown how ARA can support the railway operator in choosing her best protection strategy against terrorist attacks, trying to: (1) Deter terrorists; (2) Minimise their chances of succeeding; and (3) In case of a successful attack, reduce its impact and consequences as much as possible. Should the attack succeed, we have also considered the recovery decision that the Defender could make in her aim to capture the perpetrators.

The model chosen is dynamic, in the sense that we have allowed for mobility of resources for the Defender in case of a successful attack. The model could be further sophisticated at an operational level, by deciding appropriate patrolling schedules for the mobile resources, with models as in Alpern et al. (2011), Brown et al. (2014), or Zoroa et al. (2012).

Note that we have explicitly disregarded possible cascading effects resulting from terrorist actions, in the sense that we assume that the impact on one target will not propagate along the network. This could be relevant for e.g. communication or energy networks, see Salmerón et al. (2004) or Holmgren (2006) for related applications.

Acknowledgments

This work has been partly funded by the European Union's 7th Framework Programme under grant agreement no 285223 - SECONOMICS.

BIBLIOGRAPHY

S. Alpern, A. Morton, and K. Papadaki. Patrolling games. *Operations Research*, 59(5):1246–1257, 2011.

BBC News Europe. Spain arrests al-Qaeda in Islamic Maghreb suspect. Press Release, http://www.bbc.com/news/world-europe-14563948, August 2011.

BBC News Europe. ’Al-Qaeda trio’ arrested in southern Spanish towns. Press Release, http://www.bbc.com/news/world-europe-19091753, August 2012.

G. Brown, M. Carlyle, J. Salmerón, and K. Wood. Defending critical infrastructure. *Interfaces*, 36(6): 530–544, 2006.

M. Brown, S. Saisubramanian, P. R. Varakantham, and M. Tambe. STREETS: Game-theoretic traffic patrolling with exploration and exploitation. In *Innovative Applications in Artificial Intelligence (IAAI)*. Twenty Eighth AAAI Conference on Artificial Intelligence, AAAI-14, Research Collection School Of Information Systems, 2014.

M. Buesa Blanco, A. Valiño Castro, J. Heijs, T. Baumert, and J. González Gómez. Evaluación del coste directo de los atentados terroristas del 11-M para la economía de la Comunidad de Madrid. Documento de trabajo no. 51. Technical report, *Instituto de Análisis Industrial y Financiero*, UCM, 2005.

J. S. Dyer and R. K. Sarin. Measurable multiattribute value functions. *Operations Research*, 27(4): 810–822, 1979.

J. S. Dyer and R. K. Sarin. Relative risk aversion. *Management Science*, 28(8):875–886, 1982.

M. Dziubiński and S. Goyal. Network design and defence. *Games and Economic Behavior*, 79:30–43, 2013.

W. Enders and T. Sandler. *The Political Economy of Terrorism, 2nd edition*. Cambridge University Press, New York, 2011.

B. C. Ezell, S. P. Bennett, D. von Winterfeldt, J. Sokolowski, and A. J. Collins. Probabilistic risk analysis and terrorism risk. *Risk Analysis*, 30(4):575–589, 2010.

P. H. Farquhar. State of the art—utility assessment methods. *Management Science*, 30(11):1283–1300, 1984.

P. V. Fellman, G. S. Parnell, and K. M. Carley. Biowar and bioterrorism risk assessment. In *Unifying Themes in Complex Systems Volume VIII: Eighth International Conference on Complex Systems*, pages 1382–1396, 2011.

S. French and D. Ríos Insua. *Statistical Decision Theory*. Arnold, London, 2000.

M. R. Haberfeld and A. von Hassell. *A New Understanding of Terrorism: Case Studies, Trajectories and Lessons Learned*. Humanities, Social Sciences and Law. Springer, New York, 2009.

Y.Y. Haimes and T. Longstaff. The role of risk analysis in the protection of critical infrastructures against terrorism. *Risk Analysis*, 22(3):439–444, 2002.

K. Hausken and G. Levitin. Review of systems defense and attack models. *International Journal of Performability Engineering*, 8(4):355–366, 2012.

Å. J. Holmgren. Using graph models to analyze the vulnerability of electric power networks. *Risk Analysis*, 26(4):955–969, 2006.

G. L. Keeney and D. von Winterfeldt. Identifying and structuring the objectives of terrorists. *Risk Analysis*, 30(12):1803–1816, 2010.

R. L. Keeney and D. von Winterfeldt. A value model for evaluating homeland security decisions. *Risk Analysis*, 31(9):1470–1487, 2011.

J. P. C. Kleijnen and R. G. Sargent. A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120(1):14–29, 2000.

T. G. Lewis. *Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation*. John Wiley & Sons, New Jersey, 2006.

New York Post. Spain, Morocco arrest 9 in ISIS terror cell. Press Release, http://nypost.com/2014/09/26/spain-morocco-arrest-9-in-isis-terror-cell/, September 2014.

G. S. Parnell, D. Banks, L. Borio, G. Brown, L. A. T. Cox Jr, J. Gannon, E. Harvill, H. Kunreuther, S. Morse, M. Pappaioanou, S. Pollock, N. Singpurwalla, and A. Wilson. *Report on Methodological Improvements to the Department of Homeland Security's Biological Agent Risk Analysis*. National Academies Press, 2008.

G. S. Parnell, C. M. Smith, and F. I. Moxley. Intelligent adversary risk analysis: A bioterrorism risk management model. *Risk Analysis*, 30(1):32–48, 2010.

A. Riera Font, A. Ripoll Penalva, and J. Mateu Sbert. Estimación del valor estadístico de la vida en España: Una aplicación del método de salarios hedónicos. *Hacienda Pública Española*, 2(181): 29–48, 2007.

J. Ríos and D. Ríos Insua. Adversarial risk analysis for counterterrorism modeling. *Risk Analysis*, 32 (5):894–915, 2012.

D. Ríos Insua, J. Cano, A. Tedeschi, A. Pollini, U. Turhan, M. Pellot, R. Ortega, and R. Munné. SECONOMICS "Socio-Economics meets Security". Deliverable 5.2. Case Studies in Security Risk Analysis. Technical report, European Union, 2013.

J. Salmerón, K. Wood, and R. Baldick. Analysis of electric grid security under terrorist threat. *IEEE Transactions on Power Systems*, 19(2):905–912, 2004.

R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

C. R. Sunstein. *Worst-Case Scenarios*. Harvard University Press, Boston, 2007.

L. M. Wein. OR Forum—Homeland Security: From Mathematical Models to Policy Implementation. *Operations Research*, 57(4):801–811, 2009.

World Economic Forum. *World Economic Forum. Global Risks*. http://www.weforum.org/issues/global-risks, 2013.

World Economic Forum. *World Economic Forum. Global Risks*, 2014. URL http://www3.weforum.org/docs/WEF_GlobalRisks_Report_2014.pdf.

N. Zoroa, M. J. Fernández-Sáez, and P. Zoroa. Patrolling a perimeter. *European Journal of Operational Research*, 222(3):571–582, 2012.

Model for Oil and Gas Drilling Cybersecurity

Oil and gas drilling is based, increasingly, on operational technology, whose cybersecurity is complicated by several challenges. We propose a graphical model for cybersecurity risk assessment based on Adversarial Risk Analysis to face those challenges. We also provide an example of the model in the context of an offshore drilling rig. The proposed model provides a more formal and comprehensive analysis of risks, still using the standard business language based on decisions, risks, and value.

1 Introduction

Operational technology (OT) refers to “hardware and software that detects or causes a change through the direct monitoring and/or control of physical devices, processes and events in the enterprise”, see IT (2013). It includes technologies such as SCADA systems. Implementing OT and information technology (IT) typically leads to considerable improvements in industrial and business activities, through facilitating the mechanization, automation, and relocation of activities in remote control centers. These changes usually improve the safety of personnel, and both the cost-efficiency and overall effectiveness of operations.

The oil and gas industry (O&G) is increasingly adopting OT solutions, in particular offshore drilling, through drilling control systems (drilling CS) and automation, which have been key innovations over the last few years. The potential of OT is particularly relevant for these activities: centralizing decision-making and supervisory activities at safer places with more and better information; substituting manual mechanical activities by automation; improving data through better and near real-time sensors; and optimizing drilling processes. In turn, they will reduce rig crew and dangerous operations, and improve efficiency in operations, reducing operating costs (typically about $300,000 per day).

Since many of the involved OT employed in O&G are currently computerized, they have become a major potential target for cyber attacks, see Shauk (2013), given their economical relevance, with large stakes at play. Indeed, we may face the actual loss of large oil reserves because of delayed maneuvers, the death of platform personnel, or potential large spills with major environmental impact and potentially catastrophic consequences. Moreover, it is expected that security attacks will soon target several production installations simultaneously, with the purpose of sabotaging production, possibly taking advantage of extreme weather events, and attacks oriented towards manipulating or obtaining data or information. Cybersecurity poses several challenges, which are enhanced in the context of operational technology. Such challenges are sketched in the following section.

1.1 Cybersecurity Challenges in Operational Technology

Technical vulnerabilities in operational technology encompass most of those related with IT vulnerabilities Byres and Lowe (2004), complex software Board (2013), and integration with external networks Giani et al. (2009). There are also and specific OT vulnerabilities Zhu et al. (2011); Brenner (2013). However, OT has also strengths in comparison with typical IT systems employing simpler network dynamics.

Sound organizational cybersecurity is even more important with OT given the risks that these systems bring in. Uncertainties are considerable in both economical and technical sense Anderson and Fuloria (2010). Therefore better data about intrusion attempts are required for improving cybersecurity Pfleeger and Rue (2008), although gathering them is difficult since organizations are reluctant about disclosing such information Ten et al. (2008).

More formal approaches to controls and measures are needed to deal with advanced threat agents such as assessing their attack patterns and behavior Hutchins et al. (2011) or implementing intelligent sensor and control algorithms Cárdenas et al. (2008). An additional problem is that metrics used by technical cybersecurity to evaluate risks usually tell little to those evaluating or making-decisions at the organizational cybersecurity level. Understanding the consequences of a cyber attack to an OT system is difficult. They could lead to production losses or the inability to control a plant, multimillion financial losses, and even impact stock prices Byres and Lowe (2004). One of the key problems for understanding such consequences is that OT systems are also cyber-physical systems (CPS) encompassing both computational and complex physical elements Thomas et al. (2013).

Risk management is also difficult in this context Mulligan and Schneider (2011). Even risk standards differ on how to interpret risk: some of them assess the probabilities of risk, others focus on the vulnerability component Hutchins et al. (2011). Standards also tend to present oversimplifications that might alter the optimal decision or a proper understanding of the problem, such as the well-known shortcomings of the widely employed risk matrices Cox (2008).

Cyber attacks are the continuation of physical attacks by digital means. They are less risky, cheaper, easier to replicate and coordinate, unconstrained by distance Cárdenas et al. (2009), and they could be oriented towards causing high impact consequences Board (2013). It is also difficult to measure data related with attacks such as their rate and severity, or the cost of recovery Anderson and Fuloria (2010). Examples include Stuxnet Brenner (2013), Shamoon Brenner (2013), and others Cárdenas et al. (2008). Non targeted attacks could be a problem also.

Several kinds of highly skilled menaces of different nature (e.g., military, hacktivists, criminal organizations, insiders or even malware agents) can be found in the cyber environment Board (2013), all of them motivated and ware of the possibilities offered by OT Byres and Lowe (2004). Indeed, the concept Advanced Persistent Threat (APT) has arisen to name some of the threats Ltd (2011). The diversity of menaces could be classified according their attitude, skill and time constraints Dantu et al. (2007), or by their ability to exploit, discover or even create vulnerabilities on the system Board (2013). Consequently, a sound way to face them is profiling Atzeni et al. (2011) and treating Li et al. (2009) them as adversarial actors.

1.2 Related Work Addressing the Complexities of Cybersecurity Challenges

Several approaches have been proposed to model attackers and attacks, including stochastic modelling Muehrcke et al. (2010); Sallhammar (2007), attack graph models Kotenko and Stepashkin (2006) and attack trees Mauw and Oostdijk (2006), models of directed and intelligent attacks Ten et al. (2008); models based on the kill chain attack phases Hutchins et al. (2011), models of APT attack phases Ltd (2011), or even frameworks incorporating some aspects of intentionality or a more comprehensive approach to risk such as CORAS Lund et al. (2011) or ADVISE Conning (2013).

Game theory has provided insights concerning the behavior of several types of attackers—such as cyber criminal APTs—and how to deal with them. The concept of incentives can unify a large variety of agent intents, whereas the concept of utility can integrate incentives and costs in such a way that the agent objectives can be modeled in practice Liu et al. (2005). Important insights from game theory are that the defender with lowest protection level tends to be a target for rational attackers Johnson et al. (2012), that defenders tend to under-invest in cybersecurity Amin et al. (2011), and that the attacker's target selection is costly and hard, and thus it needs to be carefully carried on Florêncio and Herley (2013). In addition to such general findings, some game-theoretic models exist for cybersecurity or are applicable to it, modelling static and dynamic games in all information contexts Roy et al. (2010). However, game-theoretic models have their limitations Hamilton et al. (2002); Roy et al. (2010) such as limited data, the difficulty to identify the end goal of the attacker, the existence of a dynamic and continuous context, and that they are not scalable to the complexity of real cybersecurity problems in consideration. Moreover, from the conceptual point of view, they require common knowledge assumptions that are not tenable in this type of applications.

Additionally, several Bayesian models have been proposed for cybersecurity risk management such as a model for network security risk analysis Xie et al. (2010); a model representing nodes as events and arcs as successful attacks Dantu et al. (2007); a dynamic Bayesian model for continuously measuring network security Frigault et al. (2008); a model for Security Risk Management incorporating attacker capabilities and behavior Dantu et al. (2009): or models for intrusion detection systems (IDS) Balchanos (2012). However, these models require forecasting attack behavior which is hard to come by.

Adversarial Risk Analysis (ARA) Ríos Insua et al. (2009) combine ideas from Risk Analysis, Decision Analysis, Game-Theory, and Bayesian Networks to help characterizing the motivations and decisions of the attackers. ARA is emerging as a main methodological development in this area Merrick and Parnell (2011), providing a powerful framework to model risk analysis situations with adversaries ready to increase our threats. Applications in physical security may be seen in Sevillano et al. (2012).

1.3 Our Proposal

The challenges that face OT, cybersecurity and the O&G sector create a need of a practical, yet rigorous approach, to deal with them. Work related with such challenges provides interesting insights and tools for specific issues. However, more formal but understandable tools are needed to deal with such problems from a general point of view, without oversimplifying the complexity underlying the problem. We propose a model for cybersecurity risk decisions based on ARA, taking into account the attacker behavior. Additionally, an application of the model in drilling cybersecurity is presented, tailored to decision problems that may arise in offshore rigs employing drilling CS.

2 Model

2.1 Introduction to Adversarial Risk Analysis

ARA aims at providing one-sided prescriptive support to one of the intervening agents, the Defender (she), based on a subjective expected utility model, treating the decisions of the Attacker (he) as uncertainties. In order to predict the Attacker's actions, the Defender models her decision problem and tries to assess her probabilities and utilities but also those of the Attacker, assuming that the adversary is an expected utility maximizer. Since she typically has uncertainty about those, she models it through random probabilities and uncertainties. She propagates such uncertainty to obtain the Attacker's optimal random attack, which she then uses to find her optimal defense.

ARA enriches risk analysis in several ways. While traditional approaches provide information about risk to decision-making, ARA integrates decision-making within risk analysis. ARA assess intentionality thoroughly, enabling the anticipation and even the manipulation of the Attacker decisions. ARA incorporates stronger statistical and mathematical tools to risk analysis that permit a more formal approach of other elements involved in the risk analysis. It improves utility treatment and evaluation. Finally, an ARA graphical model improves the understandability of complex cases, through visualizing the causal relations between nodes.

The main structuring and graphical tool for decision problems are Multi-Agent Influence Diagrams (MAID), a generalization of Bayesian networks. ARA is a decision methodology derived from Influence Diagrams, and it could be structured with the following basic elements:

- *Decisions or Actions*. Set of alternatives which can be implemented by the decision makers. They represent what one can do. They are characterized as decision nodes (rectangles).
- *Uncertain States*. Set of uncontrollable scenarios. They represent what could happen. They are characterized as uncertainty nodes (ovals).
- *Utility and Value*. Set of preferences over the consequences. They represent how the previous elements would affect the agents. They are characterized as value nodes (rhombi).
- *Agents*. Set of people involved in the decision problem: decision makers, experts and affected people. In this context, there are several agents with opposed interests. They are represented through different colors.

We describe now the basic MAID that may serve as a template for cybersecurity problems in O&G drilling CS, developed using GeNIe Laboratory.

2.2 Graphical Model

Our model captures the Defender cybersecurity main decisions prior to an attack perpetrated by an APT, which is strongly "business-oriented". Such cyber criminal organization behavior suits utility-maximizing analysis, as it pursues monetary gains. A sabotage could also be performed by this type of agents, and they could be hired to make the dirty job for a foreign power or rival company. We make several assumptions in the Model, to make it more synthetic:

- We assume one Defender. The Attacker's nodes do not represent a specific attacker, but a generalization of potential criminal organizations that represent business-oriented APTs, guided mostly by monetary incentives.
- We assume an atomic attack (the attacker makes one action), with several consequences, as well as several residual consequences once the risk treatment strategy is selected.
- The Defender and Attacker costs are deterministic nodes.
- We avoid detection-related activities or uncertainties to simplify the Model. Thus, the attack is always detected and the Defender is always able to respond to it.
- The scope of the Model is an assessment activity prior to any attack, as a risk assessment exercise to support incident handling planning.
- The agents are expected utility maximizers.
- The Model is discrete.

By adapting the proposed template in Figure 17, we may generalize most of the above assumptions to the cases at hand.

Defender Decision and Utility Nodes The Defender nodes, in white, are:

- *Protect* (*DP*) decision node. The Defender selects among security measures portfolios to increase protection against an Attack, e.g., access control, encryption, secure design, firewalls, or personal training and awareness.
- *Forensic System* (*DF*) decision node. The Defender selects among different security measures portfolios that may harm the Attacker, e.g., forensic activities that enable prosecution of the Attacker.

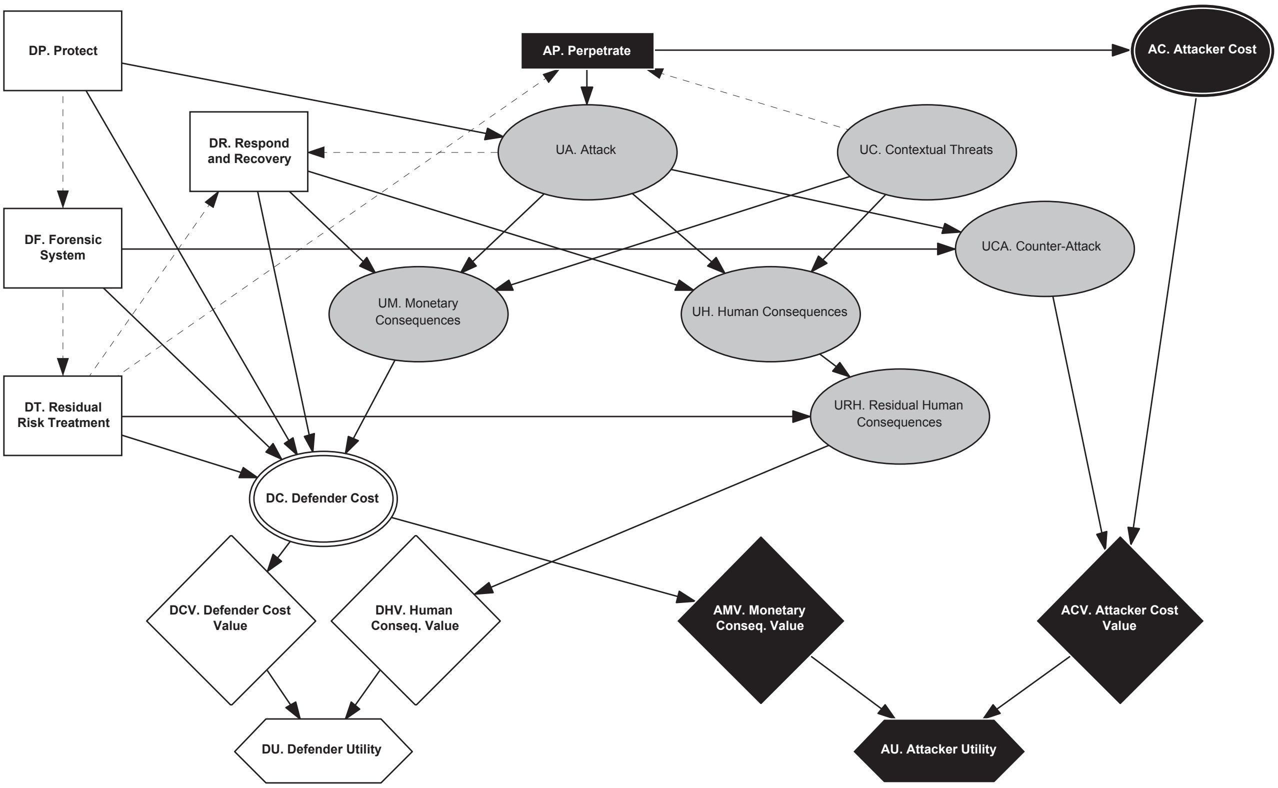


Figure 17: MAID of the ARA Model for O&G drilling cybersecurity.

- *Residual Risk Treatment* (*DT*) decision node. This node models Defender actions after the assessment of other decisions made by the Defender and the Attacker. They are based on the main risk treatment strategies excluding risk mitigation, as they are carried on through the Protect and the Respond and Recovery nodes: avoiding, sharing, or accepting risk. This node must be preceded by the Protect defender decision node, and it must precede the Attack uncertainty node (the residual risk assessment is made in advance).
- *Respond and Recovery* (*DR*) decision node. The Defender selects between different response and recovery actions after the materialization of the attack, trying to mitigate the attack consequences. This will depend on the attack uncertainty node.
- *Defender Cost* (*DC*) deterministic node. The costs of the decisions made by the Defender are deterministic, as well as the monetary consequences of the attack (the uncertainty about such consequences is solved in the Monetary Consequences node). In a more sophisticated model, most of the costs could be modeled as uncertain nodes. This node depends on all decision nodes of the Defender and the Monetary Consequences uncertainty node.
- *Value Nodes* (*DCV* and *DHV*). The Defender evaluates the consequences and costs, taking into account her risk attitude. They depend on the particular nodes evaluated at each Value node.
- *Utility Nodes* (*DU*). This node merges the Value nodes of the Defender. It depends on the Defender's Value nodes.

The Decision nodes are adapted to the typical risk management steps, incorporating ways of evaluating managing sound organizational cybersecurity strategy, which takes into account the business implications of security controls, and prepare the evaluation of risk consequences. Related work (Section 1.2) on security costs and investments could incorporate further complexities underlying the above nodes.

Attacker Decision and Utility Nodes The Attacker nodes, in black, are:

- *Perpetrate* (AP) decision node. The [generic] Attacker decides whether he attacks or not. It could be useful to have a set of options for a same type of attack (e.g., preparing a quick and cheap attack, or a more elaborated one with higher probabilities of success). It should be preceded by the Protect and Residual Risk Treatment decision nodes, and might be preceded by the Contextual Threat node (in case the Attacker observes it).
- *Attacker Cost* (*AC*) deterministic node. Cost of the Attacker decisions. Preceded by the Perpetrate decision node.
- *Value Nodes* (*AMV* and *ACV*). The Attacker evaluates the different consequences and costs, taking into account his risk attitude. They depend on the deterministic or uncertainty nodes evaluated at each Value node.
- *Utility Nodes* (*AU*). It merges the Value nodes of the Attacker to a final set of values. It must depend on the Attacker's Value nodes.

These nodes help in characterizing the Attacker, avoiding the oversimplification of other approaches. Additionally, the Defender has uncertainty about the Attacker probabilities and utilities. This is propagated over their nodes, affecting the Attacker expected utility and optimal alternatives, which are random. Such distribution over optimal alternatives is our forecast for the Attacker's actions.

Uncertainty Nodes The uncertainty nodes in grey are:

- *Contextual Threats* (*UC*) uncertainty node. Those threats (materialized or not) present during the Attack. The Attacker may carry out a selected opportunistic Attack (e.g. hurricanes or a critical moment during drilling).
- *Attack* (*UA*) uncertainty node. It represents the likelihood of the attack event, given its conditioning nodes. It depends on the Perpetrate decision node, and on the Protect decision node.
- *Consequences* (*UM* and *UV*) uncertainty node. It represents the likelihood of different consequence levels that a successful attack may lead to. They depend on the Attack and Contextual Threat uncertainty nodes, and on the Respond and Recovery decision node.
- *Residual Consequences* (*URH*) uncertainty node. It represents the likelihood of different consequence levels after applying residual risk treatment actions. They depend on the Consequence node modelling the same type of impact (e.g., human, environmental, or reputation).
- *Counter-Attack* (*UCA*) uncertainty node. Possibility, enabled by a forensic system, to counter-attack and cause harm to the Attacker. Most of the impacts may be monetized. It depends on the Forensic System decision node.

Dealing with the uncertainties and complexities and obtaining a probability distribution for these nodes could be hard. Some of the methodologies and findings proposed in the sections 1.1 and 1.2 are tailored to deal with some of these complexities. Using them, the Model proposed in this paper could lead to limit the uncertainties in cybersecurity elements such as vulnerabilities, controls, consequences, attacks, attacker behavior, and risks. This will enable achieving simplification, through the proposed Model, without limiting the understanding of the complexities involved, and a sounder organizational cybersecurity.

3 Example

We present a numerical example of the previous Model tailored to a generic decision problem prototypical of a cybersecurity case that may arise in O&G offshore rig using drilling CS. The model specifies a case in which the driller makes decisions to prevent and respond to a cyber attack perpetrated by a criminal organization with APT capabilities, in the context of offshore drilling and drilling CS. The data employed in this example are just plausible figures helpful to provide an overview of the problems that drilling cybersecurity faces. Carrying on the assessment that the Model enables may be helpful for feeding a threat knowledge base, incident management procedures or incident detection systems.

The context is that of an offshore drilling rig, a floating platform with equipment to drill a well through the seafloor, trying to achieve a hydrocarbon reservoir. Drilling operations are dangerous and several incidents may happen in the few months (usually between 2 or 4) that the entire operation may last. As OT, drilling CS may face most of the challenges presented in Section 1.1 (including being connected to Enterprise networks, an entry path for attackers) in the context of high-risk incidents that occur in offshore drilling.

3.1 Agent Decisions

Defender Decisions The Defender has to make three decisions in advance of the potential attack. In the Protect decision node (DP), the Defender must decide whether she invests in additional protection: if the Defender implements additional protective measures, the system will be less vulnerable to attacks. In the Forensic System decision node (DF), the Defender must decide whether she implements a forensic system or not. Implementing it enables the option of identifying the Attacker and pursuing legal or counter-hacking actions against him. The Residual Risk Treatment decision node (DT) represents additional risk treatment strategies that the Defender is able to implement: avoiding (aborting the entire drilling operation to elude the attack), sharing (buying insurance to cover the monetary losses of the attack), and accepting the risk (inheriting all the consequences of the attack, conditional on to the mitigation decisions of DP, FD, and DR).

Additionally, the Respond and Recovery decision node (DR) represents the Defender's decision between continuing and stopping the drilling operations as a reaction to the attack. Continuing the drilling may lead to worsen the consequences of the attack, whereas stopping the drilling will incur in higher costs due to holding operations. This is a major issue for drilling CS. In general, critical equipment should not be stopped, since core operations or even the safety of the equipment or the crew may be compromised.

Attacker Decisions For simplicity, in the Perpetrate decision node (AP) the Attacker decides whether he perpetrates the attack or not, although further attack options could be added. In this example, the attack aims at manipulating the devices directly under control of physical systems with the purpose of compromising drilling operations or harming equipment, the well, the reservoir, or even people.

3.2 Threat Outcomes and Uncertainty

Outcomes and Uncertainty during the Incident The Contextual Threats uncertainty node (UC) represents the existence of riskier conditions in the drilling operations (e.g., bad weather or one of the usual incidents during drilling), which can clearly worsen the consequences of the attack. In this scenario, the Attacker is able to know, to some extent, these contextual threats (e.g., a weather forecast, a previous hacking in the drilling CS that permits the attacker to read what is going on in the rig).

The Attack uncertainty node (UA) represents the chances of the Attacker of causing the incident. If the Attacker decides not to execute his action, no attack event will happen. However, in case of perpetration, the chances of a successful attack will be lower if the Defender invests in protective measures (DC node). An additional uncertainty arises in case of materialization of the attack: the possibility to identify and counter-attack the node, represented by the Counter-Attack uncertainty node (UCA).

If the attack happens, the Defender will have to deal with different consequence scenarios. The Monetary (UM) and Human Consequences (UH) nodes represent the chances of different consequences or impact levels that the Defender may face. The monetary consequences refer to all impacts that can be measured as monetary losses, whereas human consequences represent casualties that may occur during an incident or normal operations. However, the Defender has the option to react to the attack by deciding whether she continues or stops the drilling (DR node). If the Defender decides to stop, there will be lower chances of casualties and lower chances of worst monetary consequences (e.g., loss of assets or compensations for injuries or deaths), but she will have to assume the costs of keeping the rig held (one day in our example) to deal with the cyber threat.

Outcomes and Uncertainty in Risk Management Process The previous uncertainties appear after the Attacker's decision to attack or not. The Defender faces additional relevant uncertainties. She must make a decision between avoiding, sharing, or accepting the risk (DT node). Such decision will determine the final or residual consequences. The final monetary consequences are modeled through the Defender Cost deterministic node (DC node), whose outcome represents the cost of different Defender decisions (nodes DP, DF, DT, and DR). In case of accepting or sharing the risk, the outcome of the DC node will also inherit the monetary consequences of the attack (UM node). Similarly, the outcome of the Residual Human Consequences uncertainty node (URH) is conditioned by the risk treatment decisions (DC node) and, in case of accepting or sharing the risk, it will inherit the human consequences of the attack (UH node). If the Defender decides to avoid the risk, she will assume the cost of avoiding the entire drilling operations and will cause that the crew face a regular death risk rather than the higher death risk of offshore operations. If the Defender shares the risk, she will assume the same casualties as in UH and a fixed insurance payment, but she will avoid paying high monetary consequences. Finally, in case the Defender accepts the risk, she will inherit the consequences from the UM and UH nodes.

The Attacker Cost deterministic node (AC) provides the costs (non-uncertain by assumption) of the decision made by the Attacker. Since he only has two decisions (perpetrate or not), the node has only two outcomes: cost or not. This node could be eliminated, but we keep it to preserve the business semantics within the graphical model.

3.3 Agent Preferences

The Defender aims at maximizing her expected utility, with the utility function being additive, through the Defender Utility node (DU). The Defender key objective is minimizing casualties, but he also considers minimizing his costs (in this example we assume she is risk-neutral). Each objective has its own weight in the utility function.

The objective of the Attacker is to maximize his expected utility, represented by an additive utility function, through the Attacker Utility node (AU). The Attacker key objective is maximizing the monetary consequences for the Defender. We assume that he is risk-averse towards this monetary impact (he prefers ensuring a lower impact than risking the operations trying to get a higher impact). He also considers minimizing his costs (i.e., being identified and perpetrating the attack). Each of these objectives has its own weight in the utility function, and its own value function. The Attacker does not care about eventual victims.

3.4 Uncertainty about the Opponent Decisions

The Attacker is able to know to some extent the protective decisions of the defender (DP node), gathering information while he tries to gain access to the drilling CS. While knowing if the Defender avoided the risk (avoiding all the drilling operations) is easy, knowing if the Defender chose between sharing or accepting the risk is difficult. The most important factor, the decision between continue or stop drilling in case of an attack, could be assessed by observing the industry or company practices. The Defender may be able to assess also how frequent similar attacks are, or how attractive the drilling rig is for this kind of attacker. In ARA, and from the Defender perspective, the AP node would be an uncertainty node whose values should be provided by assessing the probabilities of the different attack actions, through analyzing the decision problem from the Attacker perspective and obtaining his random optimal alternative.

3.5 Example Values

An annex provides the probability tables of the different uncertainty nodes employed to simulate the example in Genie (Tables 11 to 17). It also provides the different parameters employed in the utility and value functions (Tables 18 to 20). Additionally, the "risk-averse" values for AMV are obtained with $AMV = \sqrt[3]{\frac{DC}{10^7}}$; the "risk-neutral" values for DCV are obtained with $DCV = 1 - \frac{DC}{10^7}$; and, the values for DHV are 0 in case of victims and 1 in case of no victims.

3.6 Evaluation of Decisions

Based on the solution of the example, we may say that the Attacker should not perpetrate his action in case he believes the Defender will avoid or share the risk. However, the Attacker may be interested in perpetrating his action in case he believes that the Defender is accepting the risk. Additionally, the less preventive measures the Defender implements (DP and DT nodes), the more motivated the Attacker would be (if he thinks the Defender is sharing the risk). The Attacker's expected utility is listed in Table 21 in the Annex. The Defender will choose in this example not to implement additional protection (DP node) without a forensic system (DF node). If the Defender believes that she is going to be attacked, then she would prefer sharing the risk (DT node) and stop drilling after the incident (DR node). In case she believes that there will be no attack, she should accept the risk and continue drilling. The Defender's expected utilities are listed in Table 22 in Annex.

Thus, the Defender optimal decisions create a situation in which the Attacker is more interested in perpetrating the attack. Therefore, to affect the Attacker's behavior, the Defender should provide the image that her organization is concerned with safety, and especially that it is going to share risks. On the other hand, if the Attacker perceives that the Defender pays no attention to safety or that she is going to accept the risk, he will try to carry on his attack. The ARA solution for the Defender is the following:

1. Assess the problem from the point of view of the Attacker. The DT and DR nodes are uncertainty nodes since that Defender decisions are uncertain for the Attacker. The Defender must model such nodes in the way that she thinks the Attacker models such uncertainties. In gen-

eral, perpetrating an attack is more attractive in case the Attacker strongly believes that the Defender is going to accept the risk or is going to continue drilling.

2. Once forecasted the Attacker's decision, the Defender should choose between sharing and accepting the risk. Accepting the risk in case of no attack is better than sharing the risk, but accepting the risk in case of attack is worse.

Thus, the key factor for optimizing the decision of the Defender are her estimations on the uncertainty nodes that represent the DT and DR nodes for the attacker. Such nodes will determine the Attacker best decision, and this decision the Defender best decision.

4 Conclusions and Further Work

We have presented the real problem and extreme consequences that OT cybersecurity in general, and drilling cybersecurity in particular, are facing. We also explained some of the questions that complicate cybersecurity, especially in OT systems. The proposed graphical model provides a more comprehensive, formal and rigorous risk analysis for cybersecurity. It is also a suitable tool, able of being fed by, or compatible with, other more specific models such as those explained in Section 1.

Multi-Agent Influence Diagrams provide a formal and understandable way of dealing with complex interactive issues. In particular, they have a high value as business tools, since its nodes translate the problem directly into business language: decisions, risks, and value. Typical tools employed in widely used risk standards, such as risk matrices, oversimplify the problem and limit understanding. The proposed ARA-based model provides a business-friendly interpretation of a risk management process without oversimplifying its underlying complexity.

The ARA approach permits us to include some of the findings of game theory applied to cybersecurity, and it also permits to achieve new findings. The model provides an easier way to understand the problem but it is still formal since the causes and consequences in the model are clearly presented, while avoiding common knowledge assumptions in game theory.

Our model presents a richer approach for assessing risk than risk matrices, but it still has the security and risk management language. In addition, it is more interactive and modular, nodes can be split into more specific ones. The proposed model can still seem quite formal to business users. However, data can be characterized using ordinal values (e.g., if we only know that one thing is more likely/valuable than other), using methods taken from traditional risk management, employing expert opinion, or using worst case figures considered realistic. The analysis would be poorer but much more operational.

Using the nodes of the proposed model as building blocks, the model could gain in comprehensiveness through adding more attackers or attacks, more specific decision nodes, more uncertainty nodes, or additional consequence nodes, such as environmental impact or reputation. Other operations with significant business interpretation can be done, such as sensitivity analysis (how much the decision-makers should trust a figure) or strength of the influence analysis (which are the key elements).

Its applicability is not exempt of difficulties and uncertainties, but in the same way than other approaches. Further work is needed to verify and validate the model and its procedures (in a similar way to the validation of other ARA-based modelsRíos Insua and Cano (2013)), and to identify the applicability and usability issues that may arise. The model could gain usability through mapping only the relevant information to decision-makers (roughly, decisions and consequences) rather than the entire model.

Appendix: Tables with Example Data

Table 11: Probability table for UC node

Riskier conditions	30%
Normal conditions	70%

Table 12: Probability table for UA node

Attacker's Perpetrate decision	Perpetrate		No perpetrate	
Defender's Protect decision	Additional protection	Non additional protection	Additional protection	Non additional protection
Attack event	5%	40%	0%	0%
No attack event	95%	60%	100%	100%

Table 13: Probability table for UM node

Attack event	Attack				No attack			
Contextual Threat event	Riskier conditions		Normal conditions		Riskier conditions		Normal conditions	
Defender's Respond and Recovery decision	Continue drilling	Stop drilling	Continue drilling	Stop drilling	Continue drilling	Stop drilling	Continue drilling	Stop drilling
Losing 0 $ event	3%	0%	10%	0%	92%	0%	96%	0%
Losing 0 - 1 Million $ event	12%	85%	20%	90%	7%	97%	4%	99%
Losing 1 - 5 Million $ event	85%	15%	70%	10%	1%	3%	0%	1%

Table 14: Probability table for UH node

Attack event	Attack				No attack			
Contextual Threat event	Riskier conditions		Normal conditions		Riskier conditions		Normal conditions	
Defender's Respond and Recovery decision	Continue drilling	Stop drilling	Continue drilling	Stop drilling	Continue drilling	Stop drilling	Continue drilling	Stop drilling
Non casualties event	96%	99.2%	99.4%	99.96%	99.6%	99.96%	99.9%	99.99%
Casualties event	4%	0.8%	0.6%	0.04%	0.4%	0.04%	0.1%	0.01%

Table 15: Probability table for URH node

Human Consequences event	No casualties			Casualties		
Defender's Residual Risk Treatment decision	Avoid	Share	Accept	Avoid	Share	Accept
No casualties event	99.95%	100%	100%	0%	0%	0%
casualties event	0.05%	0%	0%	100%	100%	100%

Table 16: Probability table for UCA node

Attack event	Attack		No attack	
Defender's Forensic System decision	Forensic	No forensic	Forensic	No forensic
No identification event	30%	90%	100%	100%
Identification event	70%	10%	0%	0%

Acknowledgments

This work has been partly funded by the European Union's 7th Framework Programme under grant agreement no 285223 - SECONOMICS.

Table 17: Probability table for DC node

Avoiding the risk		10,000,000 $		
Sharing the risk		500,000 $		
Accepting the risk	Monetary Consequences event	0 $	0 - 1,000,000 $	1,000,000 - 5,000,000 $
	Value assigned	0 $	500,000$	2,500,000 $
Additional protection		20,000 $		
Forensic system		10,000 $		
Stop drilling		300,000 $		

Table 18: Probability table for DU node

Importance of the Costs	5%
Importance of the Human Consequences	95%

Table 19: Probability table for ACV node

Attacker Cost event	Cost		No cost	
Counter Attack Consequences event	No identification	Identification	No identification	Identification
Value	0.75	0	1	0.25

Table 20: Probability table for AU node

Importance of the costs	3%
Importance of the Monetary Consequences on the Defender	97%

Table 21: Attacker expected utilities (in black the highest among the different Attacker's decisions)

DP node	DF node	DT node	UC node	Defender continues drilling		Defender stops drilling	
				Perpetrate decision	Non perpetrate decision	Perpetrate decision	Non perpetrate decision
Additional protection	Forensic	Avoid	Riskier conditions	**1**			
			Normal conditions	**1**			
		Share	Riskier conditions	0.56074	**0.56903**	0.61138	**0.61966**
			Normal conditions	0.56074	**0.56903**	0.61138	**0.61966**
		Accept	Riskier conditions	0.36484	0.35433	0.61728	0.62458
			Normal conditions	0.35170	0.34293	0.61375	0.62130
	No forensic	Avoid	Riskier conditions	**1**			
			Normal conditions	**1**			
		Share	Riskier conditions	0.55938	**0.56699**	0.61060	**0.61821**
			Normal conditions	0.55938	**0.56699**	0.61060	**0.61821**
		Accept	Riskier conditions	0.34461	0.33241	0.61653	0.62315
			Normal conditions	0.33055	0.32013	0.61299	0.61986
No additional protection	Forensic	Avoid	Riskier conditions	1			
			Normal conditions	1			
		Share	Riskier conditions	0.55116	**0.56496**	0.60295	**0.61675**
			Normal conditions	0.55116	**0.56496**	0.60295	**0.61675**
		Accept	Riskier conditions	0.45634	0.29898	0.61588	0.62173
			Normal conditions	0.42794	0.28532	0.61058	0.61841
	No forensic	Avoid	Riskier conditions	1			
			Normal conditions	1			
		Share	Riskier conditions	0.55442	**0.56282**	0.60690	**0.61530**
			Normal conditions	0.55442	**0.56282**	0.60690	**0.61530**
		Accept	Riskier conditions	0.32392	0.07465	0.61990	0.62030
			Normal conditions	0.28286	0.05131	0.61456	0.61696

BIBLIOGRAPHY

S. Amin, G. A. Schwartz, and S. S. Sastry. On the interdependence of reliability and security in networked control systems. In *Decision and Control and European Control Conference (CDC-*

Table 22: Defender expected utilities (in black the highest among the different Defender's decisions)

DP node	DF node	DT node	DR node	Possible events			
				Riskier conditions		Normal conditions	
				Attack event	Non attack event	Attack event	Non attack event
Additional protection	Forensic	Avoid	Continue drilling	0.91154	0.94573	0.94383	0.94858
			Stop drilling	0.94193	0.94915	0.94915	0.94943
		Share	Continue drilling	0.95935	0.99355	0.99165	0.99640
			Stop drilling	0.98825	0.99547	0.99547	0.99576
		Accept	Continue drilling	0.95092	0.99575	0.98490	0.99880
			Stop drilling	0.98675	0.99517	0.99447	0.99566
	No forensic	Avoid	Continue drilling	0.91154	0.94573	0.94383	0.94858
			Stop drilling	0.94193	0.94915	0.94915	0.94943
		Share	Continue drilling	0.95940	0.99360	0.99170	0.99645
			Stop drilling	0.98830	0.99552	0.99552	0.99581
		Accept	Continue drilling	0.95097	0.99580	0.98495	0.99885
			Stop drilling	0.98680	0.99522	0.99452	0.99571
No additional protection	Forensic	Avoid	Continue drilling	0.91154	0.94573	0.94383	0.94858
			Stop drilling	0.94193	0.94915	0.94915	0.94943
		Share	Continue drilling	0.95945	0.99365	0.99175	0.99650
			Stop drilling	0.98835	0.99557	0.99557	0.99586
		Accept	Continue drilling	0.95102	0.99585	0.98500	0.99890
			Stop drilling	0.98685	0.99527	0.99457	0.99576
	No forensic	Avoid	Continue drilling	0.91154	0.94573	0.94383	0.94858
			Stop drilling	0.94193	0.94915	0.94915	0.94943
		Share	Continue drilling	0.95950	0.99370	0.99180	0.99655
			Stop drilling	**0.98840**	0.99562	**0.99562**	0.99591
		Accept	Continue drilling	0.95107	**0.99590**	0.98505	**0.99895**
			Stop drilling	0.98690	0.99532	0.99462	0.99581

ECC), 2011 50th IEEE Conference on, pages 4078–4083. IEEE, 2011.

R. Anderson and S. Fuloria. Security economics and critical national infrastructure. In *Economics of Information Security and Privacy*, pages 55–66. Springer, 2010.

A. Atzeni, C. Cameroni, S. Faily, J. Lyle, and I. Fléchais. Here's Johnny: A methodology for developing attacker personas. In *Availability, Reliability and Security (ARES), 2011 Sixth International Conference on*, pages 722–727. IEEE, 2011.

M. G. Balchanos. *A probabilistic technique for the assessment of complex dynamic system resilience*. PhD thesis, Georgia Institute of Technology, 2012. URL https://smartech.gatech.edu/bitstream/handle/1853/43730/balchanos_michael_g_201205_phd.pdf.

Defense Science Board. *Task Force report: Resilient military systems and the advanced cyber threat*. Department of Defense, 2013. URL http://www.acq.osd.mil/dsb/reports/ResilientMilitarySystems.CyberThreat.pdf.

J. F. Brenner. Eyes wide shut: The growing threat of cyber attacks on industrial control systems. *Bulletin of the Atomic Scientists (1974)*, 69(5):15–20, 2013.

E. Byres and J. Lowe. The myths and facts behind cyber security risks for industrial control systems. In *Proceedings of the VDE Kongress*, volume 116, 2004. URL http://www.isa.org/CustomSource/ISA/Div_PDFs/PDF_News/Glss_2.pdf.

A. A. Cárdenas, S. Amin, and S. Sastry. Research challenges for the security of control systems. In *HotSec*, 2008. URL http://robotics.eecs.berkeley.edu/~sastry/pubs/Pdfs%20of%202008/CardenasResearch2008.pdf.

A. A. Cárdenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry. Challenges for securing cyber physical systems. In *Workshop on Future Directions in Cyber-Physical Systems Security*, 2009. URL http://cimic.rutgers.edu/positionPapers/cps-security-challenges-Cardenas.pdf.

Conning. ADVISE Enterprise Risk Modeler, 2013. URL https://www.conning.com/risk-and-capital-management/software/advise.html. Retrieved: 12/13/2013.

L. A. T. Cox. What's wrong with risk matrices? *Risk Analysis*, 28(2):497–512, 2008.

R. Dantu, P. Kolan, R. Akl, and K. Loper. Classification of attributes and behavior in risk management using Bayesian networks. In *Intelligence and Security Informatics, 2007 IEEE*, pages 71–74. IEEE, 2007.

R. Dantu, P. Kolan, and J. Cangussu. Network risk management using attacker profiling. *Security and Communication Networks*, 2(1):83–96, 2009.

D. Florêncio and C. Herley. Where do all the attacks go? In *Economics of Information Security and Privacy III*, pages 13–33. Springer, 2013.

M. Frigault, L. Wang, A. Singhal, and S. Jajodia. Measuring network security using dynamic Bayesian network. In *Proceedings of the 4th ACM workshop on Quality of protection*, pages 23–30. ACM, 2008.

A. Giani, S. Sastry, K. H. Johansson, and H. Sandberg. The VIKING project: An initiative on resilient control of power networks. In *Resilient Control Systems, 2009. ISRCS'09. 2nd International Symposium on*, pages 31–35. IEEE, 2009.

S. N. Hamilton, W. L. Miller, A. Ott, and O. Saydjari. Challenges in applying game theory to the domain of information warfare. In *4th Information Survivability Workshop (ISW-2001/2002), Vancouver, Canada*, 2002. URL http://www.au.af.mil/au/awc/awcgate/afrl/hamilton-31-08-a.pdf.

E. M. Hutchins, M. J. Cloppert, and R. M. Amin. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, 1:80, 2011. URL http://www.f35team.com/content/dam/lockheed/data/corporate/documents/LM-White-Paper-Intel-Driven-Defense.pdf.

Gartner IT. Gartner it glossary, 2013. URL http://www.gartner.com/it-glossary/operational-technology-ot. Retrieved: 12/13/2013.

B. Johnson, J. Grossklags, N. Christin, and J. Chuang. Nash equilibria for weakest target security games with heterogeneous agents. In *Game Theory for Networks*, pages 444–458. Springer Berlin Heidelberg, 2012.

I. Kotenko and M. Stepashkin. Attack graph based evaluation of network security. In *Communications and Multimedia Security*, pages 216–227. Springer, 2006.

Decision Systems Laboratory. Genie. URL http://genie.sis.pitt.edu/.

Z. Li, Q. Liao, and A. Striegel. Botnet economics: Uncertainty matters. In *Managing Information Risk and the Economics of Security*, pages 245–267. Springer, 2009.

P. Liu, W. Zang, and M. Yu. Incentive-based modeling and inference of attacker intent, objectives, and strategies. *ACM Transactions on Information and System Security (TISSEC)*, 8(1):78–118, 2005.

Command Five Pty Ltd. Advanced persistent threats: A decade in review, 2011. URL `http://www.commandfive.com/papers/C5_APT_ADecadeInReview.pdf`. Retrieved: 12/13/2013.

M. S. Lund, B. Solhaug, and K. Stolen. *Model-driven risk analysis: the CORAS approach*. Springer, 2011.

S. Mauw and M. Oostdijk. Foundations of attack trees. In *Information Security and Cryptology-ICISC 2005*, pages 186–198. Springer, 2006.

J. Merrick and G. S. Parnell. A comparative analysis of PRA and intelligent adversary methods for counterterrorism risk management. *Risk Analysis*, 31(9):1488–1510, 2011.

C. Muehrcke, E. V. Ruitenbeek, K. Keefe, and W. H. Sanders. Characterizing the behavior of cyber adversaries: The means, motive, and opportunity of cyberattacks. In *2010 International Conference on Dependable Systems and Networks Supplemental*. IEEE/IFIP, 2010. URL `https://www.perform.illinois.edu/Papers/USAN_papers/10VAN01.pdf`.

D. K. Mulligan and F. B. Schneider. Doctrine for cybersecurity. *Daedalus*, 140(4):70–92, 2011.

S. L. Pfleeger and R. Rue. Cybersecurity economic issues: Clearing the path to good practice. *Software, IEEE*, 25(1):35–42, 2008.

D. Ríos Insua and J. Cano. Basic Models for Security Risk Analysis (SECONOMICS D5.1). Technical report, SECONOMICS Project, 2013. URL `http://seconomicsproject.eu/content/d051-basic-models-security-risk-analysis`.

D. Ríos Insua, J. Ríos, and D. Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854, 2009.

S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu. A survey of game theory as applied to network security. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

K. Sallhammar. *Stochastic models for combined security and dependability evaluation*. PhD thesis, Norwegian University of Science and Technology, 2007. URL `http://www.diva-portal.org/smash/get/diva2:123582/FULLTEXT01`.

J. C. Sevillano, D Ríos Insua, and J. Ríos. Adversarial risk analysis: The Somali pirates case. *Decision Analysis*, 9(2):86–95, 2012.

Z. Shauk. Hackers hit energy companies more than others, March 2013. URL `http://fuelfix.com/blog/2013/03/25/electronic-attacks-hit-two-thirds-of-energy-companies-in-study/`. Retrieved: 12/13/2013.

C. W. Ten, C. C. Liu, and G. Manimaran. Vulnerability assessment of cybersecurity for SCADA systems. *IEEE Transactions on Power Systems*, 23(4):1836–1846, 2008.

R. C. Thomas, M. Antkiewicz, P. Florer, S. Widup, and M. Woodyard. How bad is it?—A branching activity model to estimate the impact of information security breaches. 2013.

P. Xie, J. H. Li, X. Ou, P. Liu, and R. Levy. Using Bayesian networks for cyber security analysis. In *Dependable Systems and Networks (DSN), 2010 IEEE/IFIP International Conference on*, pages 211–220. IEEE, 2010. URL 10.1109/DSN.2010.5544924.

B. Zhu, A. Joseph, and S. Sastry. A taxonomy of cyber attacks on SCADA systems. In *Internet of Things (iThings/CPSCom), 4th International Conference on Cyber, Physical and Social Computing*, pages 380–388. IEEE, 2011.