CYBECO

Supporting Cyberinsurance from a Behavioural Choice Perspective

D3.2: Improved modelling framework for cyber risk management

Due date: M24

Abstract:

This deliverable presents the improved CYBECO modelling framework for cybersecurity risk management. It draws and expands upon deliverable *3.1 Modelling framework for cyber risk management*. Suggestions from improvements have come from diverse sources. Improvements are essentially oriented towards being implementable within the CYBECO toolbox and consist of: an improved definition of the basic underlying model, better adapted also to cybersecurity terminologies; improved computational algorithms adapted to the new model; improved preference models for the cyber defender and the cyber attacker; the consideration of issues concerning insider threats and third party risks; the consideration of multiple attackers. On the whole, we move beyond standard cybersecurity frameworks which do not take into account the intentionality of certain threats, properly considering these; we provide models that overcome ordinal scales used in risk matrices; allow for repeated interactions between defenders and attackers; include behavioural elements in relation with risk aversion and, very importantly, reflect the decision of adopting cyber insurance. Our model aims at supporting an organization which needs to decide its optimal cybersecurity resource allocation, including what security controls and insurance product to acquire, if any. We also include a description of how to implement the models in the CYBECO Toolbox as well as a case study for illustration purposes. The core of the document describes in an accessible manner the above developments, after motivating the proposed improvements. We then include technical appendices referring to the new papers/reports developed, and additional support material. The core activities refer to Task 3.4 which took place during the second year of the CYBECO project.

Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document Status

Document Title	Modelling framework for cyber risk management
Version	3.0
Work Package	3
Deliverable #	3.2
Prepared by	David Rios Insua, Aitor Couce Vieira (CSIC)
Contributors	Wolter Pieters, Kate Labunets, Pieter Van Gelder (TUDELFT), Ernesto Nungesser, Manuel de León, Jorge G. Ortega, Alberto Torres, Roi Naveiro, Alex Kosgodogan (CSIC), José Vila, Yolanda Gómez (DEVSTAT)
Checked by	TUDELFT
Approved by	TREK
Date	April 30th 2019
Confidentiality	PU

Document Change Log

Each change or set of changes made to this document will result in an increment to the version number of the document. This change log records the process and identifies for each version number of the document the modification(s) which caused the version number to be incremented.

Change Log	Version	Date
First draft	1.0	March 16, 2019
Second draft after peer-review	2.0	April 28, 2019
Final version after quality check	3.0	April 30, 2019

Table of Contents

List of Figures

1 Introduction

Deliverable objectives. In this deliverable we aim at improving the framework for cybersecurity risk management available at D3.1. Such framework was implemented in a first version of the CYBECO Toolbox and was presented to a group of experts at AXA, two focus groups with experts in cybersecurity risk management and was subject to two behavioural experiments. It was also assessed by the members of the Advisory Board as well as by the project reviewers and participants at the Lorentz workshop.

As a consequence of such process a number of weak and strong points emerged. In this Deliverable, we aim at mitigating the weak ones and promote the strong ones always oriented to its final implementation of the framework in the CYBECO toolbox prototype. We emphasize the first model from D3.1, which introduces an integrated cybersecurity risk analysis approach to facilitate decision-making regarding ICT systems security as proposed in the CYBECO DoW.

Deliverable structure. The rest of D3.2 is structured as follows. After providing some background in the rest of Section 1, Section 2 summarises the key information collected from various agents concerning our initial models and implementation. Section 3 describes the new version of our core model. Section 4 provides some modelling enhancements suggested referring to the consideration of cyber insiders, third party cyber risk management and belief formation. Section 5 refers to computational enhancements in relation with faster computations with our model and modelling with attack trees. Section 6 provides generic preference models for cyber defenders and attackers. Section 7 describes the transition of the ideas towards the CYBECO tool. Section 8 refers to other cyberinsurance modelling efforts undertaken. Section 9 provides a final discussion.

The core of this document aims at describing in an accessible non-technical manner our developments, with more technical details in the eight annexes corresponding to papers and technical reports. These are also compiled in D8.2, with three other related papers from the first year already published or accepted. A ninth annex contains a sketch of the R routines developed which form the underlying core of the CYBECO toolbox. A tenth annex contains suggestions for improvement from focus groups, the advisory board and reviewers. Finally, we include information about a final experiment performed concerning belief formation.

We provide now some background information.

Cybersecurity. A defining feature of our society is its almost pervasive digitalization. All kinds of organizations, from corporations to governments to SMEs, may be critically impacted by cyber-attacks (Andress and Winterfeld, 2013). Indeed, the economic impact of cyber attacks is outstanding and, consequently, cybersecurity has become an issue of major importance, both technically and financially. Furthermore, attacks, espionage, insiders and breaches appear to increase in frequency, impact and sophistication (Lloyd's, 2017). For instance, the industry estimates that attacks costed as much as 450 billion globally in 2016, causing an impact over the global GDP (0.8% in 2014) of a similar magnitude to drug trade (0.9%) or international crime (1.2%) (McAfee, 2017).

Cybersecurity is emerging as one of the major global concerns (WEF, 2019). Although some experts criticize an excessive hype about the potential disruptive capability of large-scale cyber-attacks, cybersecurity is a truly relevant problem due to the persistence, frequency and variety of threats. Such diversity may be classified according to their motivation, skill and constraints (Dantu et al., 2007), and their ability to exploit or create vulnerabilities on the targeted systems (DSB, 2013). Important cyber threat sources include the military units maintained by global powers; 'hacktivists'; insiders; and, profit-oriented cyber-criminal groups. When it comes to malware, they are usually developed with a goal-oriented behaviour (Li et al., 2009) and, consequently, a sound approach is to treat them as adversarial actors and counter-attack them with behavioural approaches (Li et al., 2009). This is a very important motivating point for CYBECO, as with few exceptions like IS1 (CESG, 2012), standard methodologies do not explicitly take into account the intentionality of certain threats, in contrast with the relevance that organisations like the Information Security Forum (ISF, 2016) or Trend Micro (2015) and legislation (Gob. de España, 2018) start to give to targeted threats. Relevant cases (Command Five, 2011) include the 2007 Aurora attacks against Google to obtain confidential data about their algorithms and Chinese dissidents; the 2012 Shamoon attack that disabled 30.000 computers of Aramco (Brenner, 2013); and the 2013 credit card breach of 40 million customers of the US retailer Target (DeNardis, 2015). Attacks with physical consequences are also emerging, including the 2010 Stuxnet attack against an Iranian nuclear plant that disabled a fifth of its nuclear centrifuges (Brenner, 2013) or the attack on a German steelworks in 2014 that stopped their operations (Lee et al., 2014). Another notorious trend over the last years have been the indiscriminate ransomware attacks such as the 2017 Wannacry case (Yaqoob et al., 2017) that affected thousands of large and small organizations across the globe for several hours. Similarly, the NotPetya malware (Greenberg, 2018) affected thousands of organisations worldwide with an estimated cost of more than 8 billion EUR.

Cyber risk analysis. Risk analysis is a fundamental tool to help manage these problems (Cooke and Bedford, 2001). With it, organizations can assess the risks affecting their assets and what safeguards should be implemented to reduce the likelihood of such threats and/or their impacts, in case they are produced. Numerous frameworks have been developed to screen cyber risks and support risk management resource allocation, including CRAMM (CCTA, 2003), ISO 27005 (ISO, 2011), or SP 800-30 (NIST, 2012). Similarly, diverse compliance and control assessment frameworks, like ISO 27001 (2013), Common Criteria (2012), or CCM (CSA, 2016) provide guidance on the implementation of cybersecurity best practices. These standards and frameworks cover detailed catalogues of security measures suggested for protecting an organization's assets. Although these proposals have virtues, particularly their extensive catalogues of threats and assets, much remains left to be done regarding cyber risk analysis from a methodological point of view.

As an example, we reproduce some ideas about MAGERIT (Min. Hacienda, 2012). This methodology provides a very detailed catalogue of assets, threats and impacts. However, one of its more relevant weaknesses is the use of qualitative methods and risk matrices for risk analysis purposes. For instance, the treatment of threat occurrence is weak and based on a qualitative likelihood approach, as sketched in Table 1a, and analysed in detail in D3.1. Impacts associated with threats are treated in a similarly ambiguous fashion, as shown in Table 1b. Again, qualitative values are used as analysed in D3.1. Risks present a similar issue, as shown in Table 1c. Ultimately, MAGERIT sheds ambiguous results by using risk matrices.

VH	100	Very frequent	Daily
H	10	Frequent	Daily
M	1	Normal	Monthly
L	1/10	Infrequent	Every few years
VL	1/100	Very infrequent	Every century

(a) Probabilities

Impact		**Degradation**		
		1%	**10%**	**100%**
Value	**VH**	M	H	VH
	H	L	M	H
	M	VL	L	M
	L	VL	VL	L
	VL	VL	VL	VL

(b) Impacts

Risk		**Probability**				
		VL	**L**	**M**	**H**	**VH**
Impact	**VH**	H	VH	VH	VH	VH
	H	M	H	H	VH	VH
	M	L	M	M	H	H
	L	VL	L	L	M	M
	VL	VL	VL	VL	L	L

(c) Risk analysis

Figure 1: Qualitative vision of threats in MAGERIT

As indicated in Cox (2008) and Thomas et al (2014), these methods suffer from numerous shortcomings including a poor resolution to compare threats, the introduction of errors while assigning qualitative values leading to risk acceptance inconsistencies; the promotion of several biases like those of centering and category-definition; or, more importantly for our cybersecurity problem field, the induction of potentially suboptimal resource allocations, originated from range compressions, possibly rank reversals and instabilities. We emphasize also Hubbard and Seiersen (2015) and Allodi and Massacci (2017) who focus in incorrect applications of risk matrices in ICT. It is also important to stress the absence of consideration of the threat behaviour in the elicitation, which is a major component in the occurrence of targeted cyber threats.

Indeed, with counted exceptions like IS1 (NTAIA, 2012), standards do not explicitly take into account the intentionality of some of the cyber threats, a key factor to forecast what threats would target the system and what their strategic behaviour would be. Thus, ICT owners may obtain unsatisfactory results about the proper prioritization of risks and the security controls that should be implemented. Another critical issue, unlike other risky domains, is that it is difficult to obtain and analyse data, since organizations are reluctant to disclose information about intrusion attempts or consequences of attacks (Balchanos, 2012), for reputational reasons.

Cyberinsurance. Regarding this, it is important to highlight how in recent years new cyber insurance products have been introduced, of very different nature and not in every country, by companies like AXA, Generali, Allianz, or Zurich. However, cyber insurance has yet to take off (Marotta et al., 2017; Low, 2017), in spite that organizations are increasingly aware of their dependence on new technologies and on how information is a critical asset that must be secured so as to not incur in loss of customers, reputational damage and sanctions by regulators. Obstacles for researching and developing cyber insurance (Marotta et al., 2017) include information asymmetry between agents that undermines trust, lack of data due to sensitivity concerns, the evolutionary nature of attacks and attackers, and the difficulty of specifying rates of occurrence or damages.

2 Input from initial model

We revise here the main comments from various sources (referees, focus groups with experts, advisory board, reviewers) concerning the initial framework and first version of the tool prepared. We provide first a comparison of use cases, scenarios, models and Toolbox v1 and then sketch suggestions. Annex 10 contain the literals of the suggestions and a collation of improvement directions.

2.1 Comparison with use cases, scenarios and Toolbox v1

The table below provides a matching of the use cases and scenarios (D4.1) (columns 1 and 2) with the models in the papers (D3.1) (columns 3 and 4) and the CYBECO Toolbox v1 (D5.1) (column 5) and proposals for improvements (column 6). Paper 1 refers to appendix 1 in D3.1 entitled “An Adversarial Risk Analysis Framework for Cybersecurity” (to appear in *Risk Analysis*); Paper 2 refers to appendix 2 in D3.1 entitled “Some Risk Analysis Problems in Cyber Insurance Economics” (appeared in *Estudios de Economía Aplicada*); and Toolv1 refers to Toolbox version 1 (Risk case 1 to a fixed demo case, Risk case 3 to a variable demo case). It is important to note that the risk cases in the tool refer to SMEs, but several of the uses cases/scenarios do not belong to the realm of SMEs (although they coincide conceptually).

Use case	Scenario	Paper 1	Paper 2	Toolv1	Proposal
Use case 1: Cyber-insurance selection process for an SME		Model in paper	Model 1 in paper	Risk case 1 Risk case 3	Include more details in tool and improve model (Annex 1)
Use case 2: Loss of personally identifiable data for a large company in the financial sector	Scenario 1: Loss of personally identifiable data for a large company in the financial sector	Model in paper	Model 1 in paper	Risk case 1 Risk case 3	Include details in tool and improve model (Annex 1 and 2)
Use case 3: Insurance fraud for an SME in the professional services sector	Scenario 2: Insurance fraud for an SME in the professional services sector		Model 2 in paper		Structurally similar to the insider problem dealt with in Annex 3
Use case 4: Products / Services Manipulation for a large company in the manufacturing sector	Scenario 3: Manipulation of Products/Services for a large company in manufacturing	Model in paper	Model 1 in paper	Risk case 1 Risk case 3	Include more details in tool and improve model (Annex 1)
Use case 5: Insufficient insurance coverage for an SME operating in the IT industry sector		Model in paper	Model 1 in paper	Risk case 1 Risk case 3	Treatment available in Del 3.1. Some other ideas in Annex 1.

Use case 6: Accumulation of cyber-incidents following a single large-scale attack with involvement of reinsurance in the claim process			Model 3 in paper	☒	

The following conclusions emerge from the above comparison and drive the developments of the deliverable:

- The scenarios and use cases applying to the tool refer only to risk cases 1 and 3.
- Risk case 1 will be left as a demo case with minimal involvement from the user. However, given the comments in Annex 10, we shall need to improve language, documentation and interface concerning it.
- We shall freeze risk case 2 in the tool for these reasons:
  - Not required by use cases or scenarios proposed.
  - It would be part more of a back office tool for an insurance company, whereas CYBECO would be more of a front office tool for insurance companies, cybersecurity consultants and their customers
  - However, the design of cyber insurance products problem is mentioned in both papers and we describe below how, and in Appendix 1, the design of cyber insurance products may take place based on what we have developed.
- We should enhance as much as possible risk case 3 with more threats, security controls, insurance products, type of company detailed in the use cases and scenarios.
- For use case 3, scenario 2, the problem is conceptually close to the insider threat model developed below.
- We shall prepare a paper describing the CYBECO tool and final model developed (Annex 1).
- For the crossed cells, additional tools/products could be developed, but are out of scope of this project (as they were not in the proposal).

2.2 Recommendations from focus groups, advisory board and reviewers

We also collected comments and recommendations from focus groups including experts in AXA Group security, the project advisory board, the EC’s reviewers of the project, referees of the papers submitted, participants at the Lorentz seminar and internal discussions. Annex 10 specifies their comments and specific answers for each of the relevant questions, issues or input provided by these parties. In the rest of this section, we highlight the main drivers emerging from such recommendations.

- There is a need to better adapt terminology and wording to standards like NIST or ISF.
- Even though these are prototype models and tools, there is a need to increase the number of elements (threats, controls, impacts,…) in them. Include also an option 'Other ….'. Take into account the Cyber Essentials. In that sense, we should include also constraints referring to not issuing cyber insurance unless minimal controls are implemented.
- There is a need to better link with standard IT measures of availability, integrity and confidentiality. More generally, an improved preference model for cyber defenders seems in order, as well as better for attackers. Objectives of defenders and attackers do not need to be aligned.
- There is a need to consider issues concerning insiders.
- There is a need to consider third party cyber risk issues.
- There is a need to stress the combination of adversarial and non-adversarial cyber threats as well as cyber and environmental threat, and human errors.
- There is a need to further split assets (e.g., servers and computers).
- There is a need to split between operational and maintenance costs.
- There is a need to consider the impact of controls over insurance prices.
- There is a need to better inform about how likely are attacks.
- There is a need to include various types of attackers and multiple attackers.
- There is a need to better account for available cyber insurance products and how the models may be used for designing such products.

3 Updated model

In D3.1 and Rios et al. (2018 and 2019, both annexes 1 and 2 of D3.1), we have presented an approach to support cybersecurity resource allocation. When compared with standard cybersecurity frameworks, our proposal provides a more comprehensive method leading to a more detailed modelling of risk problems, yet, no doubt, more demanding in terms of analysis. We believe though that the stakes at play are so high at many organisations, especially, in critical infrastructures and sectors, that this additional work should be worth the effort.

We start by providing an increasingly detailed view of the cybersecurity risk management problem. It will lead to the definition of our basic underlying model below. For additional motivation, recall our Introduction to risk analysis in Deliverable 3.1

Overview of our model

In a nutshell, as reflected in Figure 2, the cybersecurity risk management problem refers to an organisation which faces potential threats that may have impacts on it. We could choose an appropriate cybersecurity portfolio to manage such risks as best as possible. The portfolio might include a cyber insurance contract to eventually transfer risks.

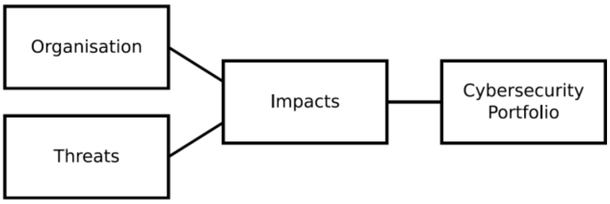


Figure 2: High-level categories of the cybersecurity risk management problem

These elements can be further specified as reflected in Figure 3 as follows:

- The *organisation* is basically described in terms of its *profile* and *assets*, together with what we shall call *other organizational features*.
- Wie distinguish between the following *threats*, reflecting the ISF classification:
    - *Environmental threats*: Incidents in external systems outside the control of the organisation.
    - *Accidental threats*: Failures or human errors in systems within the control of the organisation.
    - *Non-targeted cyber threats*: Cyber attacks that target organisations opportunistically, e.g., through the most vulnerable target, or randomly.
    - *Targeted cyber threats*: These are differentiated from non-targeted ones because another organisation, called attacker, devotes resources to specifically harm the defending organisation.

- Impacts will be separated according to:
    - *Insurable impacts*, which may be partly covered by an eventual insurance contract.
    - *Non-insurable impacts*, which will not typically be covered by insurance.
- Finally, the cybersecurity portfolio will include:
    - *Security controls*, put in place by the organisation covering protection and prevention activities as well as detection of and response to threats, partly reducing the likelihood of the threats, partly mitigating their impact.
    - *Recovery controls*, typically implemented to respond and recover from the attacks, reducing their eventual impact.
    - *Insurance* aimed at reducing the burden of the attacks, partly serving to transfer risks.

As in ISF's classification, we further separate the security controls between *procedural* (referring to practices and procedures to enhance security), *technical* (concerning digital protection technologies) and *physical* (dealing with physical protection means). The above instruments will typically have to satisfy certain constraints in relation with available cyber security budgets, compliance requirements and others.

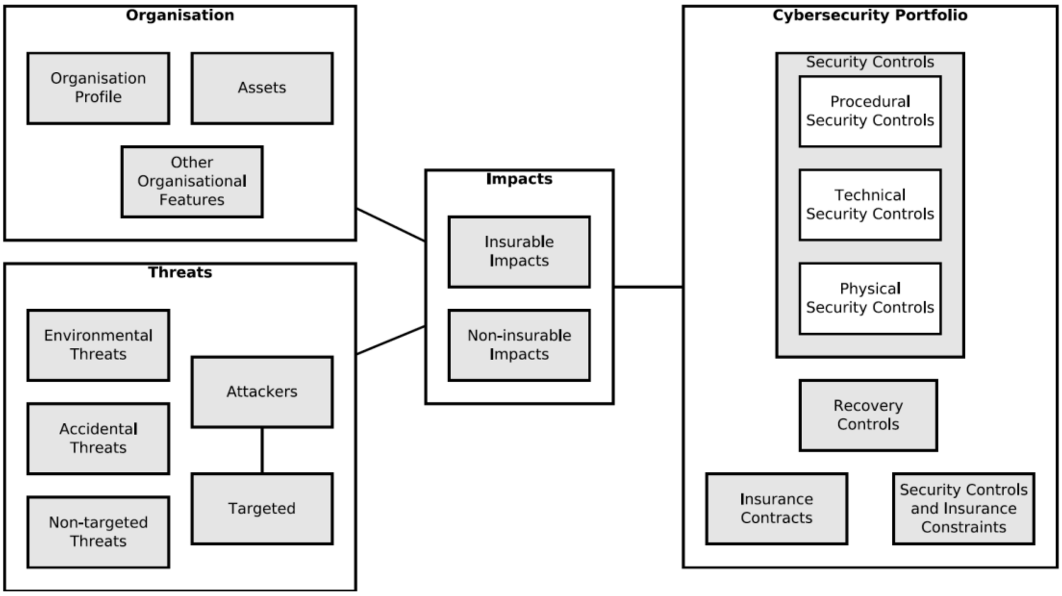


Figure 3: Subcategories of the cybersecurity risk management problem

Once we have found the relevant categories and subcategories, we may consider the elements within each of such subcategories. For this, we may take into account various catalogues from e.g. the methodologies mentioned above, as exemplified in Figure 4 in which we have included just a couple of instances per category.

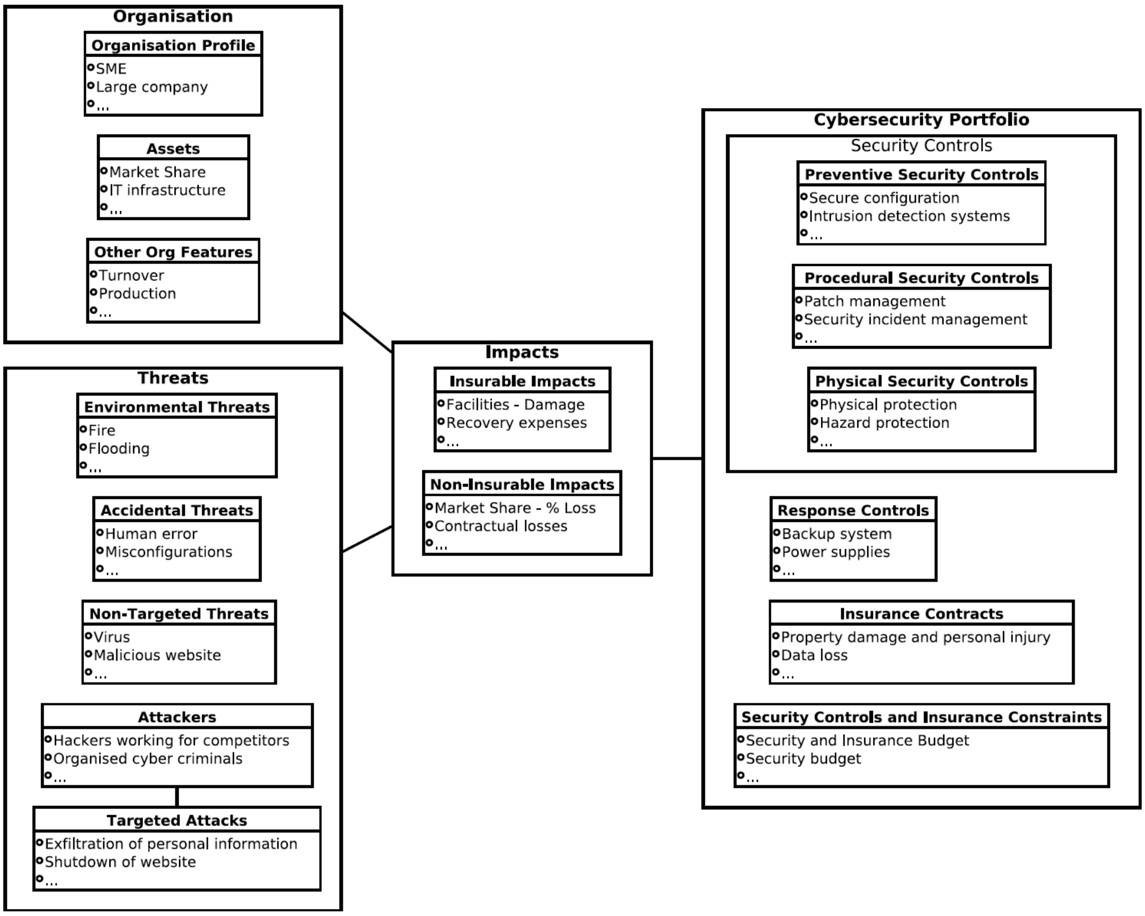


Figure 4: Subcategories with examples

A final important defining issue refers to how do we link the corresponding assets with the relevant impacts which will characterise cybersecurity risk management from the perspective of the incumbent organisation. Conceptually, we just need to link the identified relevant assets with the corresponding identified relevant impacts as reflected in Figure 5.

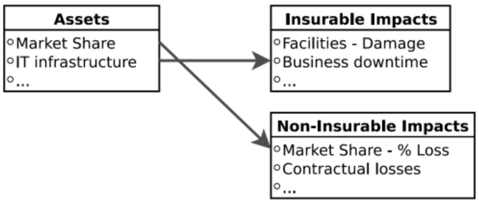


Figure 5: Linking assets and impacts

Sketch of model

Based on the schematic description above, we formulate the basic qualitative model underlying our DSS and then sketch how it may be solved once we have provided the required quantitative inputs. We start by describing our generic qualitative model for cybersecurity risk management. The aim is to support an organisation in its cybersecurity resource allocation process. We distinguish between a Defender to which our DSS will support in her allocation and, initially, an Attacker which will try to perpetrate attacks to the Defender in pursue of certain goals.

We represent the problem as a bi-agent influence diagram (BAID) in Figure 6, with the terminology used in Banks et al. (2015): the diagram includes oval nodes which represent uncertainties relevant for the problem; hexagonal utility nodes which model preferences over consequences; rectangle nodes portraying decisions modelled through the set of relevant alternatives at such point; and, finally, double oval nodes representing deterministic factors relating the values of the predecessors of the node. The diagram also includes arrows pointing to decision (meaning that such decisions are undertaken knowing the values of the predecessors) and chance and value nodes (meaning that the corresponding events or consequences are influenced by the predecessors). Light coloured nodes designate nodes belonging just to the Defender problem; darker ones to the Attacker; stripped ones are relevant to both agents' decisions.

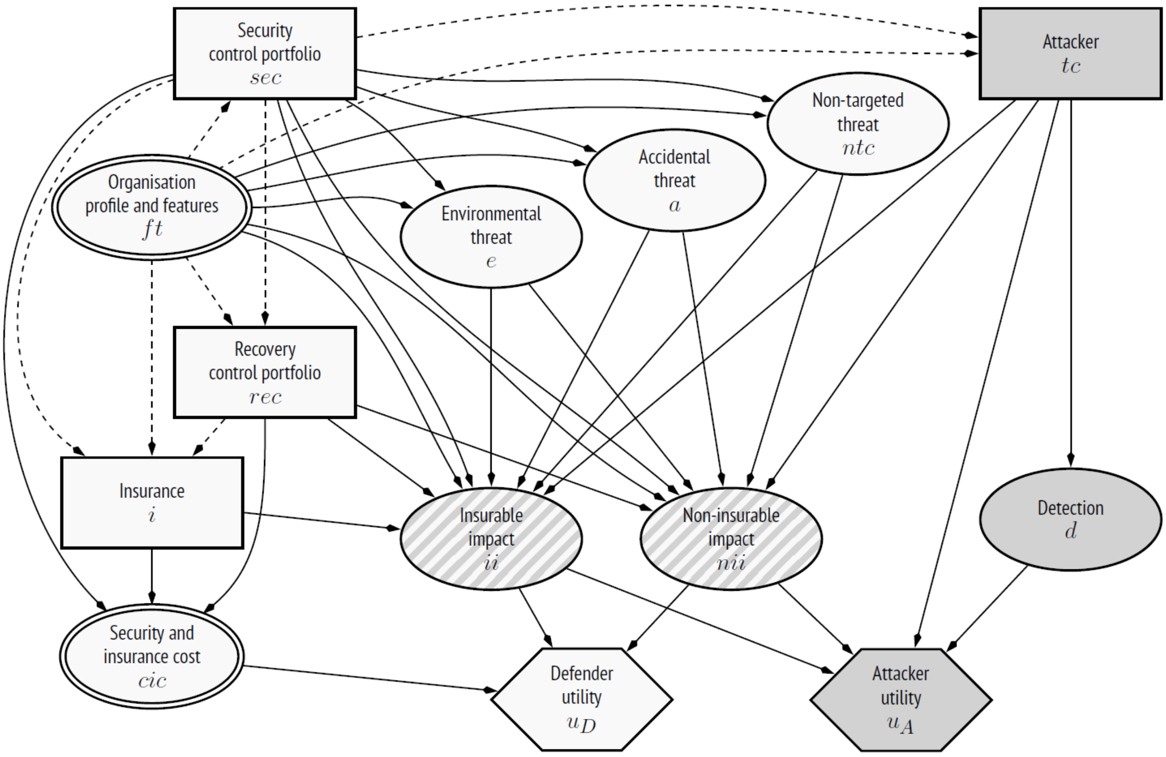

Security control portfolio sec
Attacker tc
Non-targeted threat ntc
Accidental threat a
Organisation profile and features ft
Environmental threat e
Recovery control portfolio rec
Insurance i
Insurable impact ii
Non-insurable impact nii
Detection d
Security and insurance cost cic
Defender utility u_D
Attacker utility u_A

Figure 6: Bi-agent influence diagram for the basic underlying model

We detail now the various nodes available which correspond with the problem elements described in the earlier overview.

We start with a description of the organization profile and features, which will be deterministic; we designate them with ft. We then identify the threats relevant to the organisation. As mentioned, we distinguish between: *environmental*, designated through e; *accidental*, designated through a; and *non-targeted cyber threats*, designated ntc, all of them modelled through uncertain nodes. Besides, we shall also consider, *targeted cyber threats* designated tc; we model them as decisions, but associated with a different agent, the Attacker, who perpetrates its attacks with information concerning ft and sec. Having identified the threats and relevant assets, we may identify the impacts which we separate between *insurable,* called ii, and *non-insurable* ones, referred to by nii. Then, the organisation may identify the actions that may be undertaken to mitigate the likelihood and/or impact of the threats. The three types of instruments, *security controls* sec, *recovery controls* rec and *insurance* i are modelled with decision nodes. They will have (security and insurance) costs, which will typically be deterministic; we designate them by cic. The above instruments may have to satisfy certain constraints. With all the relevant elements in place, we may then prepare the preference model for the Defender through her utility uD, based on a value node.

We turn now to the remaining elements faced by the Attacker mainly referring to the detection and identification of the attacker, which we designate by d. Note that detection could depend on the controls implemented, but we may ignore such fact for most attackers because they typically undertake the attack remotely and their identification, in the sense of being able to prosecute them, depends more on their failures as attackers (e.g., the language they use in the code) than on the defender forensic capabilities. The exception are insiders, who are more exposed to being detected (in this case we will link the security control portfolio with the insider detection node). Finally, with all his relevant elements in place, we consider the preference model for the Attacker through the utility uA of the Attacker through a value node.

The above may be seen as a basic template which may be simplified, for example, by eliminating the rec node, if it is not deemed relevant, or by including several Attacker blocks. An example in Annex 1 showcases this idea.

Assessment and Computation

Once structured, we first assess the non-strategic elements (utilities and non-adversarial probabilities) in the Defender problem. We then assess the random utilities and probabilities in the Attacker problem and simulate from it to estimate the adversarial probabilities in the Defender problem. In section 6 we describe the computational refinements introduced. Annex 6 contains all these developments with an example.

The first stage when solving the Defender problem in Figure 6 is the provision of the quantitative models required at various nodes. We need to provide the following elements:

- At chance nodes, we must define the relevant events and the corresponding conditional probability distributions. For example, at the environmental threat node we need to elicit the relevant events e that may affect the organization and the distribution $p(e|sec, ft)$, which represents the probability that environmental threat e happens, assuming that the preventive portfolio sec has been implemented by an organization with features ft. The threats would be chosen from a catalogue of relevant threats and the distributions from one of relevant parametrised distributions as illustrated in Annex 1.
- At deterministic nodes, we must define the functions relating the values of the node with the predecessors, if they exist. For example, at the security and insurance costs node we would have a function $cic = g(sec, rec, i)$, which represents the costs associated with the portfolios *sec*, *rec* and insurance *i*, typically aggregated additively.
- At decision nodes, we must define the alternatives available at the corresponding time point. For example, at node *Security control portfolio* we must define the sec portfolios relevant to the organization, characterized by its features ft. The features induce which controls from an available catalogue are relevant, from which we choose those considered incumbent. From them, we would generate the portfolios satisfying the corresponding constraints.
- At the utility node, we must define the relevant preference model which, in our case, would be a function $u_D(cic, ii, nii)$, adopting the general form we detail in Annex 2.

The above assessments are standard and may be based on data and/or expert judgement, except for $p(tc|sec)$ which models the beliefs concerning the targeted cyber threats given the security control portfolio implemented, as it entails strategic thinking.

When dealing with the strategic problem of the Attacker, we have to take into account the fact that the Attacker has means to obtain information about the organisation; namely, some of their security controls and organisation features (sec, ft). We also use random probability and utility models to model our uncertainty about the Attacker beliefs. So, given the portfolio (sec, ft) - observable by the Attacker - we associate with each attack tc their random expected utility to determine the random optimal attack $TC^*|sec, ft$ maximising the Attacker random utility and, subsequently, estimating the probability distribution of the attacks $p(tc|sec, ft)$.

Based on the above assessments, we associate with each feasible combination of security controls and insurance products (sec, rec, ii) their expected utility to find the portfolio with maximum expected utility, defined as $(\mathrm{sec}^*, rec^*, ii^*)$.

As mentioned, Annex 1 contains a full description and an illustration. Below we provide an illustrative example of the information that the model generates, basically what the model provides to the CYBECO Toolbox for its output.

Example - Solution (and other relevant results)

Solving the defender problem, we find the best portfolio that would consist of installing all the security controls and insurance. Additionally, the model provides the different combinations of security controls and insurance from best to worst, as well as other associated indicators such as probabilities or expected monetary impacts. The table below provides an example of this.

Ranking	1	2	3	4	5
Firewalls and Gateways	YES	YES	YES	YES	YES
Secure configuration	YES	YES	YES	YES	YES
Access control	YES	YES	YES	YES	NO
Malware protection	YES	YES	NO	NO	NO
Patch and vulnerability mgmt.	YES	YES	YES	YES	YES
Hazard protection	YES	YES	YES	YES	YES
Security controls cost	€3990	€3990	€3390	€3390	€3060
Physical and personal damage insurance	YES	YES	YES	YES	YES
Data loss insurance	YES	NO	YES	NO	NO
Insurance cost	€500	€300	€500	€300	€200
Expected** asset losses	€68,716.67	€68,716.67	€69,472.73	€69,472.73	€70,724.67
Expected insurance payment	€21,386.51	€20,441.51	€21,521.51	€20,441,51	€20,441,51
Expected damages to facilities and property	€22,712.79	€22,712.79	€22,712.79	€22,712.79	€22,712.79
Expected impact of business downtime	€44,953.88	€44,953.88	€45.559.94	€45.559.94	€46.661,88
Expected number of PII* records lost	7	7	8	8	9
Expected impact of PII liability	€1050	€1050	€1200	€1200	€1350
Probability of fire	2%	2%	2%	2%	2%
Expected number of employee errors per year	0.31	0.31	0.31	0.31	0.31
Expected number of malware attacks per year	0.25	0.25	0.25	0.25	0.34
Probability of targeted exfiltration***	32.76%	32.76%	33.74%	33.74%	35.35%
Probability of targeted data manipulation	48.16%	48.16%	49.20%	49.20%	50.70%
Probability of targeted denial of service	52.78%	52.78%	54.01%	54.01%	55.24%

** PII: Personal identifiable information*

*** Note that expected values should be interpreted in statistical terms. For instance, the €22,712.79 expected damages to facilities come from the fact that there is a 2% of fire every year that, if happens, would cause hundreds of thousands of damages in euros. It should not be interpreted as the most frequent value to expect - which in the case of damages due to fire is 0 per year.*

**** Note that in this example case the defender is being targeted by hacktivists and cybercriminals and, thus, the probability of targeted attacks is very high.*

4 Modelling enhancements

In this section we describe the main modelling developments enhancing the basic model described in Section 3. They were based on comments received from various sources as outlined in Section 2 as well as by the referees of papers submitted in Year 1. They include three topics: insider threats, supply chain cyber risk management and belief formation.

Insider threats

Insider threats are encountered in many risk analysis areas including international security, geo-politics, business, and, specially, cyber security. They are not only widely perceived to be significant, but also often considered to be more damaging and more likely than outsider attacks. Moreover, it is feared that the impact of the insider threat problem actually known is only the tip of an iceberg as many organizations are choosing not to report such incidents unless required to do so by law: it is a field in which little data is available, specially in the cyber security domain. Protection from insider threats is challenging as the perpetrators might have access to sensitive resources and privileged system accounts. Finally, solutions to insider threat problems are considered to be complex: technical solutions do not suffice since insider threats are fundamentally a people issue. In its simplest form, it is natural to view the insider threat problem as a two player game. We may call the first player the organization and the second one, the employee. A typical scenario would be as follows: since insider threats are a well-known phenomena, it will frequently be the case that several measures would have already been implemented by the organization to prevent or deter an insider attack. The employee will typically be aware of the measures in place and plans an attack accordingly. Once the attack has been carried out and detected, the organization will undertake actions to end the attack and mitigate any damage caused, possibly based on the resources deployed at the first stage. This type of interactions have been named sequential Defend-Attack-Defend games.

During CYBECO we have provided two ARA models to deal with the problem of insiders. First, we use an ARA Defend-Attack-Defend model between the organization and the employee. We then segment the employees in three classes (good, inadvertent and malicious insiders) considering more sophisticated ARA models and illustrate the concepts with a cyber security example. We could include the insider Attacker node as another instance of Attacker node as in Figure 6, and consequently in the Toolbox.

This material is covered in detail in Annex 3.

Supply chain cyber risk management

Supply chain risk management (SCRM) has come into place to implement strategies to manage risks in a supply chain with the goal of reducing vulnerabilities and avoid service and product disruptions. As in other risk analysis application areas, SCRM usually involves four processes: identification, assessment, controlling and monitoring of risks. Tang and

Tomlin (2008) define the field as the management of supply chain risks through coordination or collaboration among supply chain partners to ensure profitability and continuity and consider four basic areas to mitigate their impact: supply, demand, product and information management. Due to the proliferation of cyber attacks and the increasing interconnectedness of organizations, a major feature of recent interest refers to new cyber threats affecting supply chain operations in what we shall call Supply Chain Cyber Risk Management (SCCRM). We could also call it third party cyber risk management. Recall the Target attack described in Section 1. Another relevant attack was Wannacry which took over, among many others, Telefonica and the UK NHS producing the unavailability of numerous services, which entailed costs estimated to have reached $4 billion, Berr (2016).

During CYBECO, we have presented a general framework for SCCRM. We provide a general description, covering models to forecast attacks and their impacts, integrating such information to provide relevant risk indicators. Due to lack of data typical of cyber security contexts, we need to rely on expert judgment to assess the involved parameters. We also outline its implementation and a numerical example. This may be used for third party impacts if we decide to include them in our cyber risk analysis.

This material is covered in detail in Annexes 4 (general framework) and 5 (expert judgement assessment). Moreover, this one provides further illustrations on the use of SEJ for cybersecurity risk assessment. Third party cyber risk is also mentioned in Annex 2 when referring to cyber security preference models.

Belief formation

Structured Expert Judgement (SEJ) elicitation has a long history of successes, both in methodology and application. Hence, it has become a major ingredient within the risk and decision analysis practice (Bedford and Cooke, 2001). A significant feature of these disciplines is their emphasis in decomposing complex problems into smaller pieces that are easier to handle and then recombining the piecewise solutions to tackle the global problem. One example refers to belief assessment which benefits from decomposition, typically through the argument of extending the conversation. Rather than directly assessing the probability of an outcome, one finds a conditioning partition and assesses the probabilities of the outcome given the conditioning events. From these, and the probabilities of the conditioning events, the law of total probabilities enables calculation of the unconditional probability of the outcome. Tetlock and Gardner (2015) call this approach Fermitisation and present it as a key strategy for the success of their super-forecasters.

Various forms of decomposition pervade risk and decision analysis. They simplify the complex cognitive tasks and mitigate expert reliance on heuristics that can introduce bias, ensuring that experts and decision makers actually analyze their decision making problems. The decomposition typically entails more assessments, though these tend to be simpler and more meaningful, leading to better decisions.

During CYBECO we have presented and studied Adversarial Risk Analysis (ARA) as a decomposition strategy for game theoretic problems from a Bayesian perspective. ARA can be framed as a tool for SEJ elicitation when we need to deal with probabilities referring to actions by opponents. As an example, in Chen et al. (2016) nearly 30% of the questions posed to experts somehow involved adversaries (e.g. Will Syria use chemical or biological weapons before January 2013?, asked in 2011). We do show how this strategy can actually improve non-structured expert assessment of the opponent's actions as well as propose several ways to implement ARA in practice.

We illustrate the relevance of such approach as a decomposition technique to forecast adversarial actions in game theoretic contexts, which could be added to the SEJ toolkit. We show how the ARA decomposition strategy breaks down an attack probability assessment into multi-attribute utility and probability assessments for the adversary. For the ARA approach to be worthwhile, it is expected that the resulting probabilities are more accurate than the ones that would have been directly obtained and, also, that the corresponding increased number of necessary judgements are cognitively easier. Theoretical simulations have shown the relevance of the approach. Initial experiments have been conducted to validate these ideas. Thus, we have shown that ARA may improve the results of direct SEJ.

Annex 8 reflects a theoretical justification for ARA as a SEJ tool. Annex 11 reflects the initial experimental setup to validate such ideas currently under processing.

5 Preference models

In this section we cover generic families of preference models relevant for both the Defender and the Attacker. The first issue was suggested by the reviewers and the advisory board, although it was sketched in Deliverable 3.1; the second one is a natural extension.

Defender preference model

We have consolidated and expanded the work in this topic undertaken in Deliverable 3.1 now presented as Annex 2. Relevant aspects in our risk analysis approach is (a) forecasting the potential consequences of the different risk identified in the risk analysis and (b) building the evaluation of them regarding stakeholder preferences and risk attitudes. To facilitate the identification and assessment of objectives and preferences, we propose three templates:

- A generic tree of potential cybersecurity objectives for ICT owners. We provide several relevant attributes for their subobjectives.
- Forecasting models for such objectives.
- A generic multi-attribute utility function to assess the previous cybersecurity objectives regarding ICT owners' preferences and risk attitudes.

Tree of Cybersecurity objectives. We provide a generic tree of objectives (the performance measures that we want to optimise) for ICT owners in a cybersecurity context. Ideally this could be shown to cybersecurity stakeholders who would pick from it the relevant objectives for their problem at hand. For each objective, we identify the corresponding attribute in which we assess it. There are several requirements that the objectives in a decision-making problem should meet (Keeney and Gregory, 2005): Comprehensive; Measurable; Relevant; Unambiguous; Understandable and communicable. We distinguish between **natural attributes** that provide a direct measure of the objective involved (e.g., repair costs in €), **proxy attributes** that have a relationship to the objective (e.g., website downtime – in an online store – regarding the income generation objective) and used when no natural attributes are suitable, and **constructed scales**.

Some cybersecurity frameworks provide catalogues of concepts analogue to our cybersecurity objectives, mostly those addressing business impact analysis in cybersecurity or business impact analysis in general including ETSI GS ISI 002 v1.2.1 (2015), ISO 22317 (2015b), OWASP business impacts (2017), OECD types of cyber losses (2017) the ENISA Information Package for SMEs (2007), the ENISA report on ICT business continuity management for SMEs (2010), and CYBECO deliverable on the definition of cyber insurance scenarios (CYBECO D4.2, 2018). In general, they depict a few general categories of impacts (legal and regulatory, productivity, financial, reputation and loss of customers) with some examples or subcategories. However they do not meet the requirements for objectives we mentioned earlier. Most of them provide a list of recurrent business impacts rather than a comprehensive list that encompasses less typical impacts (e.g., physical impacts).

Similarly, they provide types of objectives that somehow overlap: most of the impacts affect monetary objectives and, thus, some categorisation among them is recommended. For instance, some costs affect specific assets (e.g., asset degradation or activity interruption) whereas others are more general (e.g., competitive advantage, reputation).

Besides the existing lists of cybersecurity impacts, the main conceptual influences on our final list come from asset management and law. First, asset management – e.g., ISO 55000 (ISO, 2014) for assets in general or ISO 19770 for ICT assets (2015a) – it helps to conceptualise the different status that an asset could attain is important, so that engineers could characterise how an asset affects a system or the organisation in terms, for instance, of reliability or predictability. Another conceptual influence comes from law, in particular, the distinction between damages on property (a.k.a. economic or pecuniary damages) and damage on persons (a.k.a. general or non-pecuniary damages). This facilitates the distinction between objectives that can be measured in monetary terms (directly or through estimation) and others that are of non-monetary nature and, thus, need special considerations when it comes to their evaluation (e.g., through the value of statistical life). It also helps on the distinction between the owners of the objectives (i.e., health and environmental damages are suffered always by third parties besides the monetary, legal or reputational consequences that these damages could cause to the organisation). We have also made an effort concerning relating the usual availability, integrity and confidentiality criteria in business terms.

Based on such catalogues and the described approach, we have developed a generic tree of cybersecurity objectives for a generic organisation, which we summarise in Figure 7. The general categories are the following:

- **Minimize operational costs**. We refer here to the assets and activities that constitute the inventory and operations of an organisation. All of them measurable in monetary terms, i.e. the corresponding attribute would be euros.

- **Minimize income reduction**: We refer here to impacts that reduce the income obtained by the organisation. All of them are measurable in monetary terms. All of them measurable in monetary terms.

- **Minimize other costs**: These refer to other impacts that affect an organisation. It include some strategic, compliance and financial costs. All of them measurable in monetary terms.

- **Minimize reputation impact**: We refer here to impacts over reputation that affect the trustworthiness of the organisation as an institution, rather than those more directly measurable in monetary terms that impact brand value or minimise income/service. In principle, these impacts cannot be fully represented with monetary attributes.

- **Cybersecurity costs**: It is practical to separate the costs related with managing cybersecurity, since this is the activity we aim to support in our decision-making.

- **Impact to other organizations**: A cybersecurity incident in our organisation might affect other organisations and, thus, the organisation objectives also involve minimising damage to them. This is related with the supply chain cyber risks mentioned above.

- **Harm to people**: A cybersecurity incident might also affect people such as employees, customers, by-standers or local communities. Therefore, the organisation objectives could also involve minimising harm to people. Some of the sub-objectives entail impacts with have been very rare, so far, in cybersecurity, but the emergence of industrial systems, smart infrastructures and mass surveillance bring these risks to the fore (e.g., Stuxnet).

- **Environmental damage**: Similar to damage inflicted to people, the environment might be affected by cyber attacks against systems with physical operations.

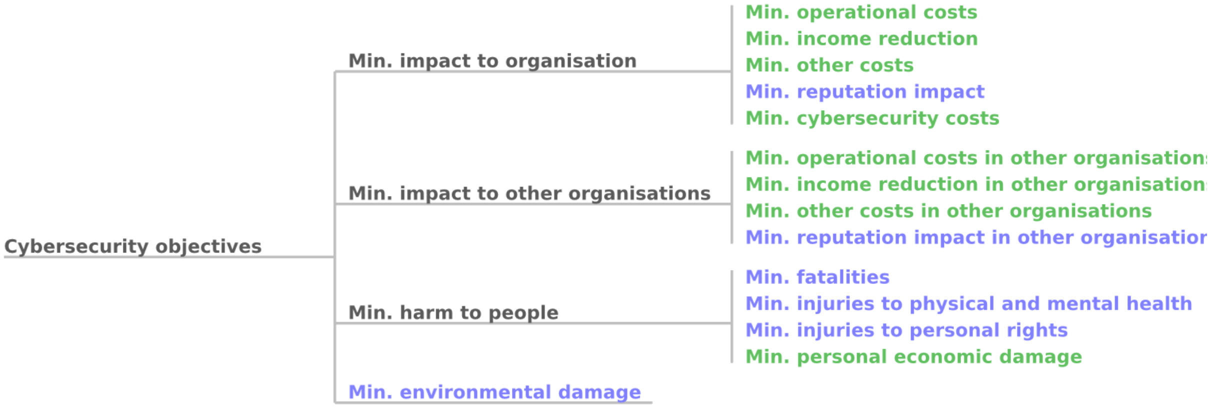

Cybersecurity objectives
Min. impact to organisation
Min. operational costs
Min. income reduction
Min. other costs
Min. reputation impact
Min. cybersecurity costs
Min. impact to other organisations
Min. operational costs in other organisation:
Min. income reduction in other organisation
Min. other costs in other organisations
Min. reputation impact in other organisatior
Min. harm to people
Min. fatalities
Min. injuries to physical and mental health
Min. injuries to personal rights
Min. personal economic damage
Min. environmental damage

Figure 7: Cybersecurity objectives. Green (measurable in monetary terms), blue (non-monetary).

Preference modelling. From the previous list of cybersecurity objectives, the incumbent stakeholders could choose the relevant ones. Then, we need a procedure to model stakeholder preferences and risk preferences. For this, we use the classic concepts of measurable multi-attribute value function (Dyer and Sarin, 1979) and relative risk aversion (Dyer and Sarin, 1982). First we aggregate the attributes with a muli-attribute value function of the type

$$u(c) = \sum_{i=1}^{q} k_i u_i(c_i),$$

where k_i is the **weight** for the attribute *I*; we then elicit the risk attitude assuming constant risk aversion or proneness and then combine both approaches adopting a functional form:

$$u(c) = 1 - \exp\left(-\rho \sum u_i(c_i)\right), \quad \rho > 0.$$

$$u(c) = \sum u_i(c_i).$$

$$u(c) = 1 + \exp\left(\rho \sum u_i(c_i)\right), \quad \rho > 0\,;$$

This is the approach that we follow in CYBECO to facilitate modelling the Defender's preferences, as it is not overly demanding cognitively and it is relatively general in its assumptions.

In Annex 2, the whole process is illustrated and we also include procedures to forecast the cybersecurity impacts. We also include the construction of a default utility function used in the Toolbox.

Attacker preference model

We briefly draw attention now over our perspective on the attacker's preference assessments concerning the consequences of the decision-making problem, that is, the random utilities. We shall usually have some information about his multiple interests, e.g. when dealing with terrorism problems, Keeney (2007) and Keeney and von Winterfeldt (2010) present extensive classifications of criteria amongst which to choose. The above list for the Defender may be also used for similar purposes.

Keeney (2007) then advocates that standard utility methods may be adopted by interviewing experts in the problem at hand, therefore developing utility functions modelling A's preferences. However, note that such preferences are not directly elicited from A, but rather through a surrogate. Thus, intrinsically, there is uncertainty about A's preferences. An alternative approach is to aggregate the objectives with a weighted measurable value function, as in Dyer and Sarin (1979). As an example, we could consider an additive value function for the Attacker. The uncertainty about the weights could be modelled using a Dirichlet distribution, so that we may estimate their value and then associate random variables with the corresponding means with one further judgement, e.g. fixing one of the parameter's variance. Finally, using the relative risk aversion concept (Dyer and Sarin, 1982), we could assume different risk attitudes when modelling the attacker's utility function, where we would typically assume risk proneness. Further uncertainty about the risk coefficient may be modelled e.g. through uniform distributions. In any case, to determine all the required distributions, we may ask experts to directly elaborate such distributions or request them to provide point estimates of the weights and coefficients and build the distributions from these.

The general approach is described in Annex 8 and an illustration is available in Annex 1. As in the Defender case, we include default Attacker utility functions in the Toolbox.

6 Computational enhancements

In this section, we describe computational enhancements to improve computations in our cybersecurity framework. We first sketch the general procedure proposed (available in a broader context in Annex 1); then, the enhancements in relation with the application of augmented probability simulation (available in detail in Annex 6); and finally the conversion of attack trees into belief networks to facilitate cybersecurity resource allocation (available in detail in Annex 7).

General computational description

The optimisation process may be cumbersome and we discuss here how we may implement it. First, recall that we need to compute the expected utility of each portfolio of security controls, recovery controls and insurance. For each portfolio, we may approximate by Monte Carlo its expected utility, where sampling from the required distributions is easily performed by forward sampling based on the influence diagram in Figure 6. We can also use importance sampling to avoid generating new samples for each portfolio that we wish to evaluate.

As for the optimisation part, when the number of portfolios is small, we just need to approximate the expected utility at each portfolio and find the optimal one. When the number of portfolios is large, or we have continuous portfolios, we may proceed in, at least, three ways:

- Evaluate the expected utility at some portfolios, approximate the expected utility through a regression metamodel and optimize the regression surface to obtain the approximate optimal portfolio.
- Use an optimization model which requires only functional evaluations, like the classic Nelder-Mead algorithm, and let the algorithm run until a local optimal portfolio is detected.
- Use an augmented simulation algorithm, sample from it through an MCMC approach until convergence is detected and, then, find the approximate sample mode of the marginal distribution on the portfolios.

This is fully described in Annex 1. The current version of the Toolbox incorporates the case in which the number of portfolios is small.

Augmented probability simulation

One of the computational enhancements alluded before is in the realm of algorithmic game theory, Nisan et al (2007), in that we aim at providing algorithms to approximate solutions to game theoretic problems, within the ARA approach. The key contribution is to avoid the two step ARA procedure and try to accelerate computation. As a by-product, we also provide APS approaches to approximate standard Nash equilibria solutions. Therefore, we explore how augmented probability simulation (APS) may be used to compute game theoretic solutions. APS is a powerful simulation based methodology used to approximate optimal solutions in decision analytic problems, see Bielza et al (1999). We start by defining an augmented distribution proportional to the product of the utility and the original distribution and, then, come out with a method to simulate from the augmented distribution. The mode of the marginal in the decision of the augmented distribution coincides with the optimal decision. Note that most of the emphasis in the ARA literature has been on foundational issues with little emphasis on computational challenges in complex problems as we may have to face in cybersecurity risk management. Therefore, we provide a complete outline of the role of APS for game theoretic computations.

APS is based on treating the decision variables as random and converting the optimization problem into a simulation one in the joint space of both decision variables and random variables. Simulating from the augmented distribution of decision and states simultaneously solves for the expectation of the objective function and optimization problem: the marginal mode over the decision variable provides the optimal decision. The strategy is very general in that it can accommodate arbitrary probability models and non-negative utility functions.

We provide APS approaches to approximating game theoretic solutions for the sequential defend-attack problem, under common knowledge. They use the idea of a nested APS framework similar to folding back a tree. If the game theoretic solution is not robust, we need to address the issue. One way forward is to perform an ARA approach. For this, we weaken the common knowledge assumption. Again, we introduce a nested approach.

This is fully described in Annex 6.

Integration of risk scenarios and risk mitigation strategies through a Bayesian network

Annex 7 uses an equivalent approach to the one presented in Section 3, except for the strategic assessment of attackers. The analysis builds on Bayesian networks, which provide a sound framework for probabilistic risk assessment by representing cyber threat scenarios as combinations of cascading events stemming from multiple attack trees. We, basically, build a model to quantify the risk scenarios and optimize the different mitigation strategies to identify the most cost-efficient one, considering budget and technical constraints or risk acceptability thresholds. We apply the model to a case study in electric power cybersecurity.

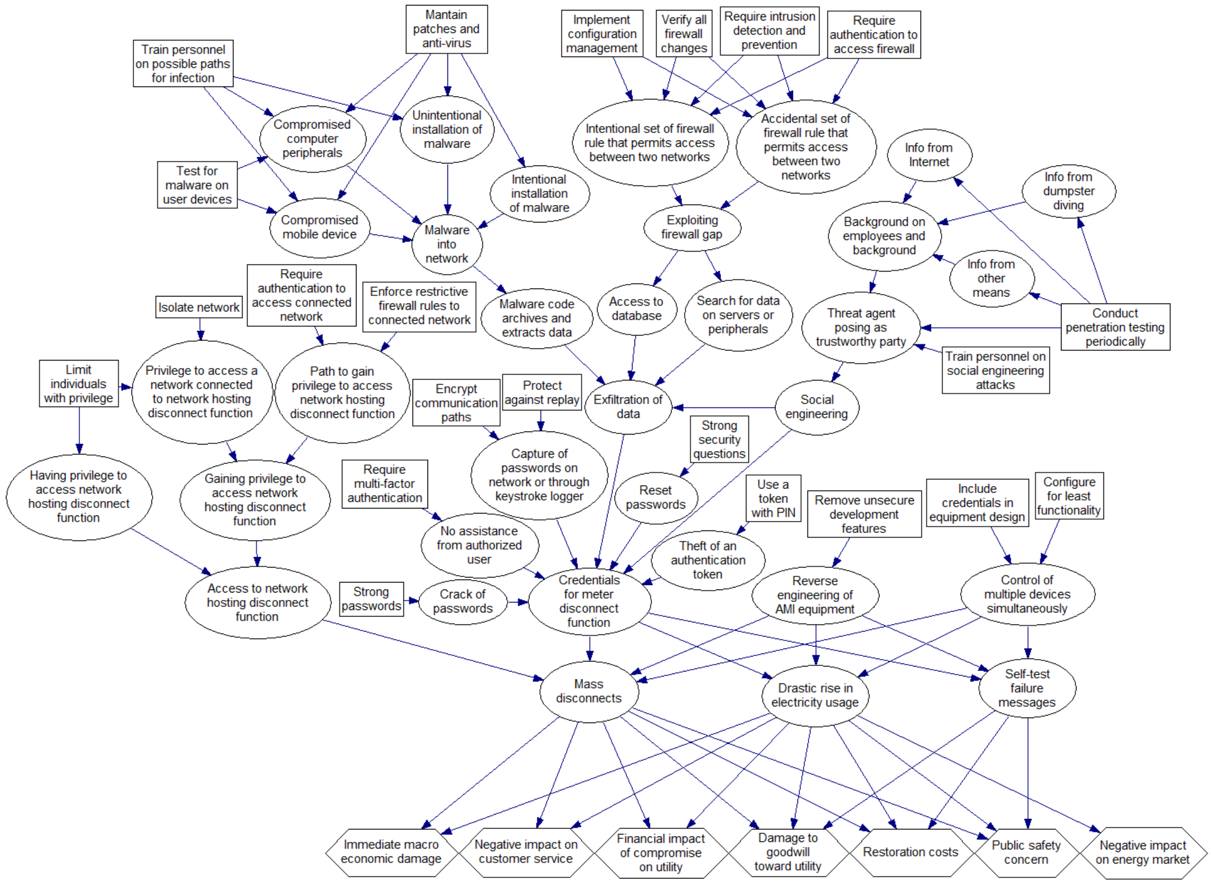

Mantain patches and anti-virus
Implement configuration management
Verify all firewall changes
Require intrusion detection and prevention
Require authentication to access firewall
Train personnel on possible paths for infection
Compromised computer peripherals
Unintentional installation of malware
Intentional set of firewall rule that permits access between two networks
Accidental set of firewall rule that permits access between two networks
Info from Internet
Info from dumpster diving
Test for malware on user devices
Intentional installation of malware
Compromised mobile device
Malware into network
Exploiting firewall gap
Background on employees and background
Info from other means
Require authentication to access connected network
Enforce restrictive firewall rules to connected network
Malware code archives and extracts data
Access to database
Search for data on servers or peripherals
Threat agent posing as trustworthy party
Conduct penetration testing periodically
Isolate network
Train personnel on social engineering attacks
Limit individuals with privilege
Privilege to access a network connected to network hosting disconnect function
Path to gain privilege to access network hosting disconnect function
Encrypt communication paths
Protect against replay
Exfiltration of data
Social engineering
Strong security questions
Having privilege to access network hosting disconnect function
Gaining privilege to access network hosting disconnect function
Require multi-factor authentication
Capture of passwords on network or through keystroke logger
Reset passwords
Use a token with PIN
Remove unsecure development features
Include credentials in equipment design
Configure for least functionality
No assistance from authorized user
Theft of an authentication token
Access to network hosting disconnect function
Strong passwords
Crack of passwords
Credentials for meter disconnect function
Reverse engineering of AMI equipment
Control of multiple devices simultaneously
Mass disconnects
Drastic rise in electricity usage
Self-test failure messages
Immediate macro economic damage
Negative impact on customer service
Financial impact of compromise on utility
Damage to goodwill toward utility
Restoration costs
Public safety concern
Negative impact on energy market

Figure 8: Influence diagram for the basic underlying model

This framework can be introduced as a novel practice for assessing cyber risks and supporting risk-based decisions on resource allocation to protect systems. Possible extensions need to be investigated, such as modelling strategically the threat agent(s) through ARA or the analysis of the cyber resilience, meaning the ability of the system to continuously deliver the intended outcome despite adverse cyber events.

The enhancement relies on the optimization algorithm. First, it selects the portfolios that fulfil the constraints (e.g., technical, acceptability threshold). After that, it selects the *non-dominated portfolios*, in the sense of the Pareto condition (a portfolio is Pareto-dominant against other portfolio if and only if the former reduces one or more risks without increasing any other risk). Finally, it identifies those portfolios that minimize the cost of deployment. The selection of the optimal solution is through the computation of additional analysis such as enumeration or evolutionary algorithms.

Annex 7 details tis developments with applications to cybersecurity in electric distribution systems. Some of the ideas in relation with handling constraints are incorporated into the Toolbox.

7 Transitioning to the CYBECO Toolbox

The purpose of the model introduced in Section 3 is to provide the best security controls and insurance portfolio, given a budget or other constraints, and for a certain planning period, say a year. As argued in D3.1, it entails more work than in traditional cybersecurity frameworks. To facilitate its implementation, we build a methodology to implement a decision support system (DSS), the CYBECO Toolbox, for cybersecurity risk management which we fully describe in D5.2. On the supply side, the potential users of such DSS would be insurance providers, insurance brokers, or cybersecurity consultants wanting to facilitate an organisation a decision about how to better allocate its cybersecurity resources. Such companies would constitute the demand side of the system. On the whole, by better allocating resources and incentivising the adoption of security controls and insurance, we aim at helping to contribute towards a more cybersecure environment.

Here we sketch how we have implemented the framework described in Section 3 with the aid of R routines to be adapted in the CYBECO Toolbox prototype.

Software algorithms in R

We have implemented as algorithms in R the following models:

- The one in relation with the case study in the paper 'An Adversarial Risk Analysis Framework for Cybersecurity' (Annex 1 of D3.1).

- The one related to case study in the paper 'A Decision Support System for Cyber Risk and Cyber Insurance Management' (Annex 1 of this deliverable).

They consist of the following elements:

1. Definition of R functions that model the different assessments over the organisation's non-strategic beliefs and preferences (defender problem) as well as the random beliefs and preferences for the attacker (attacker problem). We also include the different decisions as variables.

2. Definition of the inputs and outputs of these R functions so that they reflect the conditionality expressed in the involved influence diagrams. This way, we are able to calculate probabilities for different events and values modelled in the R functions.

3. Implementation of algorithms that calculate the random optimal decision of the attacker (solution of the attacker problem) and obtain the security control and insurance portfolio that maximises expected utility (solution of the defender problem). These require the calculations mentioned in item 2. Additionally, an important parameter in these algorithms is the number of simulation iterations.

4. Implementation of other algorithms to provide useful information for the risk analysis: conditional probabilities (i.e., probability of an event taking into account

the probability of its causing events) or conditional statistics (e.g., expected impacts taking into account the probability of its causing events).

All of this is coded in several R scripts that generate tables with the relevant risk analysis information of the case.

Integration into the Toolbox

The CYBECO Toolbox of WP5 provides a set of risk analysis templates based on these scripts (details on its architecture and functions in D5.2).

Ways of implementing the risk analysis cases

From the software architecture point of view, we have defined three ways of implementing the risk analysis cases. Specifically:

1. A risk analysis template that stores the analysis results in the Toolbox. In this case, we run the R scripts to generate tables that will be stored in the Toolbox, so that it is not necessary to run a simulation in R saving computational resources and time. The downside of this approach is that it is only useful for simple risk analysis or for consulting pre-calculated information as demo cases.

2. A risk analysis template that performs simple calculations in the Toolbox. In this case, there is no interaction with the R scripts. The downside is that this approach is only useful for simple risk analysis.

3. A risk analysis template that interacts with R. In this case, the Toolbox provides input parameters to R, and R runs the risk analysis script for a specific number of simulation iterations. Once the simulation finishes, R provides its output to the Toolbox. This allows for a more granular risk analysis but may require computational resources and time.

Integration with the Toolbox

Basically, the interaction between the Toolbox and the R algorithm is as follows:

1. The CYBECO Toolbox obtains the input from the user through the user interface (e.g., the selection of assets, threats or security controls and parameters related to them such as their cost or value).

2. The CYBECO Toolbox writes this input to a *frontend input* file in a format readable by the R algorithm.

3. Additionally, the administrators of the Toolbox can provide additional input to the R algorithm through a *backend input* file (e.g., the size of the simulation, parameters for the utility functions).

4. The Toolbox calls R to run the algorithm. The output of the algorithm is a table with the results of the analysis.

5. The Toolbox stores the results table, so that its users can consult the results through the Toolbox user interface.

Annex 9 provides a skeleton of the scripts and some examples of the code which are fully integrated in the Toolbox as described in Workpackage 5.

8 Other modelling efforts

We have outlined within Annex 1 other possible cyber insurance uses of our general model, with a view on the supply side of the Toolbox (cyber insurance providers).

Cyber insurance design decisions

The first one refers to cyber insurance design decisions and tries to answer questions such as:

- Given the current cybersecurity environment and the features of an organisation, including its risk aversion, what would be the maximum price that such organisation would be willing to pay for a certain cyber insurance product. We may include constraints to reflect market conditions (what eventual competitors are doing).
- We may extend the methodology to determine the minimum coverage acceptable given a price.
- Finally, we may combine design decisions concerning price and coverage, via Pareto optimisation.

Market segmentation

As the computations involved are relevant, we may use the same general (regression) model to perform a market segmentation to have faster answers for customers. The idea is to run the model for a handful of customers and then try a regression model which relates features of the company and their decisions. We could use such approach to provide an initial solution to the company and, if necessary, perform a more detailed study.

9 Conclusions

This deliverable has presented the improved CYBECO modelling framework for cybersecurity risk management. We have been able to move beyond current cybersecurity risk management frameworks in the following directions:

- Unlike most of them, which are essentially based on risk matrices, we focus on detailed and careful analysis of likelihoods and multiple impacts of cyber threats, going beyond oversimplified ordinal models which may lead to inferior decisions.

- Unlike most of them, we are capable of taking into account the intentionality of some of the cyber threats, using the framework of adversarial risk analysis, combined with other risk analysis models.

- We incorporate various references to cyber insurance as a part of a cyber risk management strategy.

- We include behavioural aspects of cyber risk managers (and attackers), including preferences and attitudes towards risk, through the utility functions included.

- We detail the identification of cybersecurity objectives and how to integrate them into preference and decision-making models.

- We outline how to cope with an eventual lack of data through structured expert judgement approaches.

- We develop software algorithms that can integrate a sophisticated risk analysis case with a user-friendly CYBECO Toolbox aimed at informing SME users about the cyber risks they are facing and the potential palliative actions they can implement.

We have focused on the use cases in WP 4, and presented graphically and analytically the model referring to the decision of a company about its security resource allocation, including an eventual cyber insurance product. Our approach, no doubt, entails more work than traditional cyber risk management approaches, however in many organisations the economic, environmental, political, stakes at play are so large at that it such additional effort would be worth being implemented.

To facilitate application, we have moved in three directions. The first one refers to providing a generic cybersecurity preference model, based on identifying a set of generic cybersecurity objectives from a defender perspective, from which a risk manager may choose, as well as generic utility function which covers the above objectives (and caters for risk attitudes); a similar approach is taken for attackers. The second one refers to developing the CYBECO Toolbox (Work package 5), for which we provide a high-level description of the models to be implemented, the inputs required and the outputs produced. The third one sketches computational strategies that might alleviate the proposed initial computational scheme.

The core of this document has aimed at describing in an accessible manner the above developments, which are detailed in the enclosed technical appendices:

- Annex 1: A decision support system for cyber risk and cyber insurance management
- Annex 2: Cybersecurity preference models. The defender case.
- Annex 3: Insider threat modelling: An adversarial risk analysis approach
- Annex 4: Assessing Supply Chain Cyber Risks
- Annex 5: Structured Expert Judgement Issues in a Supply Chain Cyber Risk Management System
- Annex 6: Augmented Probability Simulation Methods for Non-cooperative Games
- Annex 7. Risk-based Selection of Mitigation Strategies for Cybersecurity of Electric Power Systems
- Annex 8: Adversarial Risk Analysis for Structured Expert Judgement Modelling
- Annex 9: Skeleton and examples of the new R Routines
- Annex 10: Recommendations from focus groups, advisory board and reviewers
- Annex 11: Validation of ARA for belief formation

References

Allodi, L., Massacci, F. 2017. “Security Events and Vulnerability Data for Cybersecurity Risk Estimation.” *Risk Analysis*, Vol. 37, pp. 1606–1627.

Andress, J. and Winterfeld, S. 2013. *Cyber Warfare: Techniques, Tactics and Tools for Security Practitioners*. Elsevier.

Banks, D., Ríos, J. and Ríos Insua, D. 2015. *Adversarial Risk Analysis*. Francis and Taylor.

Balchanos, M.G. 2012. *A Probabilistic Technique for the Assessment of Complex Dynamic System Resilience*. Ph.D. Thesis, Georgia Institute of Technology (USA).

Bielza, C., Müller, P., and Rios Insua, D., 1999. “Decision Analysis by Augmented Probability Simulation”. *Management Science*, Vol. 45, No 7, pp 995-1007.

Brenner, J.F. 2013. “Eyes Wide Shut: The Growing Threat of Cyber Attacks on Industrial Control Systems”. *Bulletin of the Atomic Scientists*, Vol. 69, No. 5, pp. 15-20.

Cardenas, A., Amin, S., Sinopoli, B., Giani, A., Perrig, A. and Sastry, S. 2009. “Challenges for Securing Cyber Physical Systems”. *Workshop on Future Directions in Cyber-Physical Systems Security*.

Central Communication and Telecommunication Agency (UK). 2003. *Risk Analysis and Management Method*.

Clemen, R. T. and Reilly, T. 2013. *Making Hard Decisions with Decision Tools*. Cengage Learning.

Cloud Security Alliance. 2016. *Cloud Controls Matrix*.

Command Five PTY LTD (Australia). 2011. *Advanced Persistent Threats: A Decade in Review*.

Cooke, R. and Bedford., T. 2001. *Probabilistic Risk Analysis: Foundations and Methods*. Cambridge University Press.

Cox, L. A. 2008. “What’s Wrong with Risk Matrices?.” *Risk Analysis*, Vol. 28, No. 2, pp. 497–512.

Dantu, R., Kolan, P., Akl, R. and Loper, K. 2007. “Classification of Attributes and Behavior in Risk Management Using Bayesian Networks”. *IEEE Intelligence and Security Informatics 2007*, pp. 71-74.

Defense Science Board, DoD (USA). 2013. *Task Force Report: Resilient Military Systems and the Advanced Cyber Threat*.

DeNardis, L. 2015. “Five Destabilizing Trends in Internet Governance”. *I/S: A Journal of Law and Policy for the Information Society*, Vol. 12, No. 1, pp. 113-133.

Dyer, J. and Sarin, R. 1979. “Group Preference Aggregation Rules Based on Strength of Preference.” Management Science, Vol. 25, No. 9, pp. 822-832.

Dyer, J. and Sarin, R. 1982. “Relative Risk Aversion.” Management Science, Vol. 28, No. 8, pp. 875-886.

European Telecommunications Standards Institute. 2015. *ETSI GS ISI 002 v1.2.1 - Information Security Indicators; Event Model, A Security Event Classification Model and Taxonomy*.

European Union Agency for Network and Information Security. 2007. *IT Business Continuity Management – An Approach for Small Medium Sized Organizations*.

European Union Agency for Network and Information Security. 2007. *Business and* .

Herley, C. and Florêncio, D. 2010. "Nobody Sells Gold for the Price of Silver: Dishonesty, Uncertainty and the Underground Economy". *Economics of Information Security and Privacy*, pp. 33-53. Springer.

Hubbard, D.W. and Selersen, R. 2016. *How to Measure Anything in Cybersecurity Risk*. Wiley.

Greenberg, A. 2018. "The Untold Story of notpetya. The Most Devastating Cyberattack in History". *Wired*, August 2018.

Gobierno de España (Spain). 2018. *Real Decreto-Ley 12/2018, de 7 de septiembre, de Seguridad de las Redes y Sistemas de Información*.

Information Security Forum. 2016. *Information Risk Assessment Methodology 2*.

International Organization for Standardization. 2013. *ISO/IEC 27005:2013 – Information Security Risk Management*.

International Organization for Standardization. 2013. *ISO/IEC 27001:2013 – Information Security Management Systems – Requirements*.

International Organization for Standardization. 2014. *ISO 55000:2014 – Asset Management - Overview, Principles and Terminology*.

International Organization for Standardization. 2014. *ISO/IEC 19770-1:2017 – Part 1: IT asset management systems -- Requirements*.

International Organization for Standardization. 2015. *ISO 22317: Societal Security - Business Continuity Management Systems – Guidelines for Business Impact Analysis*.

International Organization for Standardization. 2018. *ISO 31000:2018 – Risk Management - Guidelines*.

Keeney, R. 2007. "Modeling Values for Anti-terrorism Analysis." Risk Analysis, Vol. 27,

No. 3, pp. 585–596.

Keeney, R. and Gregory, R. 2005. "Selecting attributes to measure the achievement of objectives." *Operations Research*, Vol. 53, No. 1, pp 1–11.

Keeney, R. and von Winterfeldt, D. 2011. "A Value Model for Evaluation Homeland Security Decisions." Risk Analysis, Vol. 31, No. 9, pp 1470–87.

Koller, D., and Milch, B. 2003. "Multi-agent influence diagrams for representing and solving games". *Games and Economic Behavior*, Vol. 45, No. 1, pp 181-221.

Lee, R. M., Assante, J. and Conway, T. 2014. ICS Defense Use Case Dec 301, 2014 - German Steel Mill Cyber Attack. SANS Institute (USA).

Li, Z., Liao, Q. and Striegel, A. (2009). "Botnet Economics: Uncertainty Matters". *Managing Information Risk and the Economics of Security*, pp. 245-267. Springer.

Lloyd's (UK). 2017. *Counting the Cost - Cyber Exposure Decoded.*

Low, P. 2017. "Insuring Against Cyber-Attacks". *Computer Fraud & Security*, Vol. 2017, No. 4, pp. 18-20.

Lund, M.S., Solhaug, B. and Stølen, K. 2010. *Model-driven Risk Analysis: the CORAS Approach*. Springer.

Marotta, A., Martinelli, F., Nanni, S., Orlando, A., and Yautsiukhin, A. 2017. "Cyber-Insurance Survey". *Computer Science Review*, Vol. 24, pp. 35-61.

McAfee (USA). 2014. *Net Losses: Estimating the Global Cost of Cybercrime.*

Merrick, J. and Parnell, G.S., "A Comparative Analysis of PRA and Intelligent Adversary Methods for Couterterrorism Risk Management." *Risk Analysis*, Vol. 31, No. 9, pp 1488-1510.

Ministerio de Hacienda y Administraciones Públicas (Spain). 2012. *Methodology of Analysis and Risk Management Information Systems, version 3.*

National Institute of Standards and Technology (USA). 2012. *NIST SP 800-30 Rev. – Guide for Conducting Risk Assessments.*

National Technical Authority for Information Assurance (UK). 2012. *HMG IA Standard Number 1.*

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V.V. (Eds.). 2007. *Algorithmic Game Theory*, Vol. 1. Cambridge University Press.

Organization for Economic Cooperation and Development. 2017. *Enhancing the Role of Insurance in Cyber Risk Management.*

Ortega, J., Rios Insua, D., and Cano, J. 2017. "Adversarial Risk Analysis for Bi-agent Influence Diagrams". *XXXVI Congreso Nacional de Estadística e Investigación Operativa.*

Rios Insua, D., et al. 2017. "Forecasting and Assessing Consequences of Aviation Safety Occurrences." Unpublished.

Rios, D., Couce-Vieira, A. and Musaraj, K. 2018. "Some Risk Analysis Problems in Cyber Insurance Economics". *Estudios de Economía Aplicada*, Vol. 36, No. 1, pp. 181--194.

Rios Insua, D., Couce-Vieira, A. Pieters, W., Labunets, K., Rubio, J.A., Rasines, D.G. 2019. "An Adversarial Risk Analysis Framework for Cybersecurity". *Risk Analysis* [Accepted]

Romanosky, S., Ablon, L., Kuehn, A., Jones, T. 2018. "Content analysis of cyber insurance policies:how do carriers write policies and price cyber risk?." Rand Co.

Sastry, S., Cardenas, S., and Amin, A.A. 2008. "Research Challenges for the Security of Control Systems". *Proceedings of the 3rd Conference on Hot Topics in Security*, pp. 6:1-6:6.

Shachter, R.D. 1986. "Evaluating Influence Diagrams". *Operations Research*, Vol. 34, No. 6, pp 871-882.

Tetlock, P.E., and Gardner, D. 2015. *Superforecasting: The Art and Science of Predicition.* Broadway Books.

The Common Criteria Recognition Agreement Members. 2012. *Common Criteria for Information Technology Security Evaluation, Version 3.1 Release 4.*

The Open Web Application Security Project. "OWASP Risk Rating Methodology" https://www.owasp.org/index.php/OWASP_Risk_Rating_Methodology [Ret. 04/18]

Thomas, P. 2013. "The Risk of Using Risk Matrices." Master's thesis, University of Stavanger, Norway.

Trend Micro. 2015. *Understanding targeted attacks. What is a targeted attack*.

UK Government's National Technical Authority for Information Assurance (CESG). 2012. *HMG IA Standard Number 1*.

World Economic Forum. 2017. *The Global Risks Report 2017*.

Yaqoob, I., Ahmed, E., Ur Rehman, M.H., Ahmed, A.I.A., Al-Garadi, M.A., Imran, M., and Guizani, M. 2017. "The Rise of Ransomware and Emerging Security Challenges in the Internet of Things". *Computer Networks*, [In Press].

Acronyms and Abbreviations

APS	Augmented Probability Simulation
ARA	Adversarial Risk Analysis
BAID	Bi-agent Influence Diagram
DDoS	Distributed Denial of Service
DSS	Decision Support System
GDP	Gross Domestic Product
ICT	Information and Communication Technologies
ID	Influence diagram
IT	Information Technologies
MAID	Multi-agent Influence Diagram
MC	Monte Carlo Simulation
MCMC	Markov Chain Monte Carlo Simulation
PII	Personally Identifiable Information
SCRM	Supply Chain Risk Management
SME	Small and Medium Enterprise
WP	Work Package

Annexes

We provide the following annexes:

- Annex 1: Paper: A decision support system for cyber risk and cyber insurance management
- Annex 2: Paper: Cybersecurity preference models. The defender case.
- Annex 3: Paper: Insider threat modeling: An adversarial risk analysis approach
- Annex 4: Paper: Assessing Supply Chain Cyber Risks
- Annex 5: Paper: Structured Expert Judgement Issues in a Supply Chain Cyber Risk Management System
- Annex 6: Paper: Augmented Probability Simulation Methods for Non-cooperative Games
- Annex 7. Paper: Risk-based Selection of Mitigation Strategies for Cybersecurity of Electric Power Systems
- Annex 8: Paper: Adversarial Risk Analysis for Structured Expert Judgement Modelling
- Annex 9: Skeleton and examples of the new R Routines
- Annex 10: Recommendations from focus groups, advisory board and reviewers
- Annex 11: Validation of ARA for belief formation

Annex 1: Paper: A decision support system for cyber risk and cyber insurance management

A Decision Support System for Cyber Risk and Cyber Insurance Management

Abstract

We describe a decision support system that facilitates cybersecurity risk resource allocation to an organisation, including the determination of a cyber insurance. We provide first a descriptive scheme for the cybersecurity risk management problem. We derive from it our basic underlying model. We then sketch our system and provide a case study solved with it. We complete the paper with a description of further uses of the system.

1 INTRODUCTION

Cybersecurity is increasingly perceived as a major global problem as reflected e.g. in the WEF [36] Global Risk Reports. It is becoming even more important as companies, administrations and individuals get more interconnected, thus facilitating the spread of cyber threats. Famous examples of recent cyber events include the Target data breach, [22], in which a cyber attack to that company through one of its suppliers caused the loss of 70 million credit card details, entailing major reputation damage; and, the NotPetya malware [16] which affected thousands of organisations worldwide with an estimated cost of more than €8 billion EUR.

Given the importance of this problem, numerous frameworks have been developed to support cybersecurity risk management. Some examples are ISO 27005 [19] or CORAS [21]. Similarly, several compliance and control assessment frameworks, like ISO 27001 [20] or CCM [4], provide guidance about the implementation of cybersecurity best practices. They have many virtues, particularly their extensive catalogues of threats, assets and controls, and their detailed guidelines for the implementation of countermeasures to protect digital assets. However, much remains to be done regarding cybersecurity risk analysis from a methodological point of view: a detailed study of the main approaches to cybersecurity risk management reveals that they often rely on risk matrices with shortcomings well documented in, e.g. Cox [7] and Thomas et al. [32]. Moreover, with few exceptions like IS1 [25], those methodologies do not explicitly take into account the intentionality of certain threats, in contrast with the relevance that organisations, see [18] or [33], and legislation, e.g. [17], start to give to targeted threats. As a consequence, ICT owners may obtain unsatisfactory results in relation with the prioritisation of cyber risks and the controls they should implement, even more if we take into account the increasing variety of threats, as well as the growing complexity of countermeasures for risk management. This includes the recent emergence of cyber insurance products [35] for risk transfer purposes.

In [28], we have presented an approach to support cybersecurity resource allocation. When compared with standard frameworks, our proposal provides a more comprehensive method leading to a more detailed modelling of risk problems, yet, no doubt, more demanding in terms of analysis. We believe though that the stakes at play are so high at so many organisations that this additional work should be worth the effort. To facilitate its implementation, we present a generic actionable model from which we build a methodology to implement a decision support system (DSS) for cybersecurity risk management at strategic level. The objective of such DSS would be to provide the best portfolio of security controls and insurance for a certain organisation, given a budget and other technical and legal constraints for the relevant planning period. On the supply side, the potential users of such DSS would be insurance companies, insurance brokers, and cybersecurity consultants aiming at facilitating an organisation a decision about how to better allocate its cybersecurity resources. Such organisations would constitute the demand side for the system. On the whole, by better allocating resources and incentivising the adoption of cybersecurity controls through insurance, we would aim at helping to contribute towards a more cybersecure environment.

The structure of the paper is as follows. First, we introduce the basic underlying concept and model for our DSS, which is an evolution of that in [28], with a more detailed, better structured and better typified description of the cyber situation, as well as a better adaptation to cyber standards, in particular as reflected by the Information Security Forum (ISF) [18] and the National Institute of Standards and Technologies (NIST) [24]. We then describe how the model is specified and implemented in a DSS architecture. Next we present a case study to which we apply the DSS. Finally, we provide a discussion of other uses of our system.

2 A GENERIC SPECIFICATION OF CYBERSECURITY RISK MANAGEMENT PROBLEMS

We start by providing an increasingly detailed view of the cybersecurity risk management problem. It will lead to the definition of our basic underlying model in Section 3 and will also help us in defining an input scheme to the DSS as well as specifying its architecture in Section 4.

In a nutshell, as reflected in Figure 1, the cybersecurity risk management problem refers to an organisation which faces potential cyber threats that may have impacts on it. We aim at choosing an appropriate cybersecurity portfolio to manage such risks as best as possible. The portfolio might include a cyber insurance contract to eventually transfer risks.

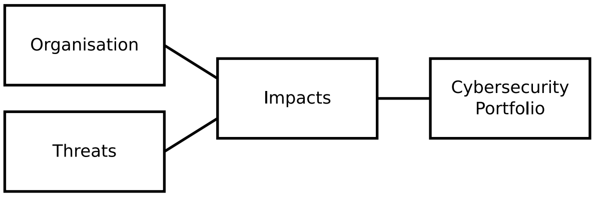


Figure 1: Input to the DSS - categories.

As reflected in Figure 2, these elements can be further specified:

- The *organisation* is basically described in terms of its *profile* and *assets*, together with what we shall call *other organisational features*.
- With respect to *threats*, we reflect the ISF [18] classification:
  - *Environmental threats*: Incidents in external systems outside the control of the organisation.
  - *Accidental threats*: Failures or human errors in systems within the control of the organisation.
  - *Non-targeted cyber threats*: Cyber attacks that target organisations opportunistically, e.g., through the most vulnerable target, or randomly.
  - *Targeted cyber threats*: These differentiate from non-targeted ones because another organisation, called attacker, devotes resources to specifically harm the defending organisation.
- Impacts will be separated according to:
  - *Insurable impacts*, which may be partly covered by an eventual insurance contract.
  - *Non-insurable ones*, which will not be typically covered by insurance contracts.
- Finally, the cybersecurity portfolio will include:
  - *Security controls*, put in place by the organisation covering protection and prevention activities as well as detection of and response to threats, partly reducing the likelihood of threats, partly mitigating their impact.
  - *Recovery controls*, typically implemented to respond and recover from the attacks, reducing their eventual impact.
  - *Insurance contracts* aimed at reducing the burden of the attacks, serving to transfer risks.

As in ISF classification, we shall further separate the security controls between *procedural* (referring to practices and procedures to enhance security), *technical* (concerning digital protection technologies) and *physical* (dealing with physical protection means). The above instruments will typically have to satisfy certain constraints in relation with available cybersecurity budgets, compliance requirements and others.

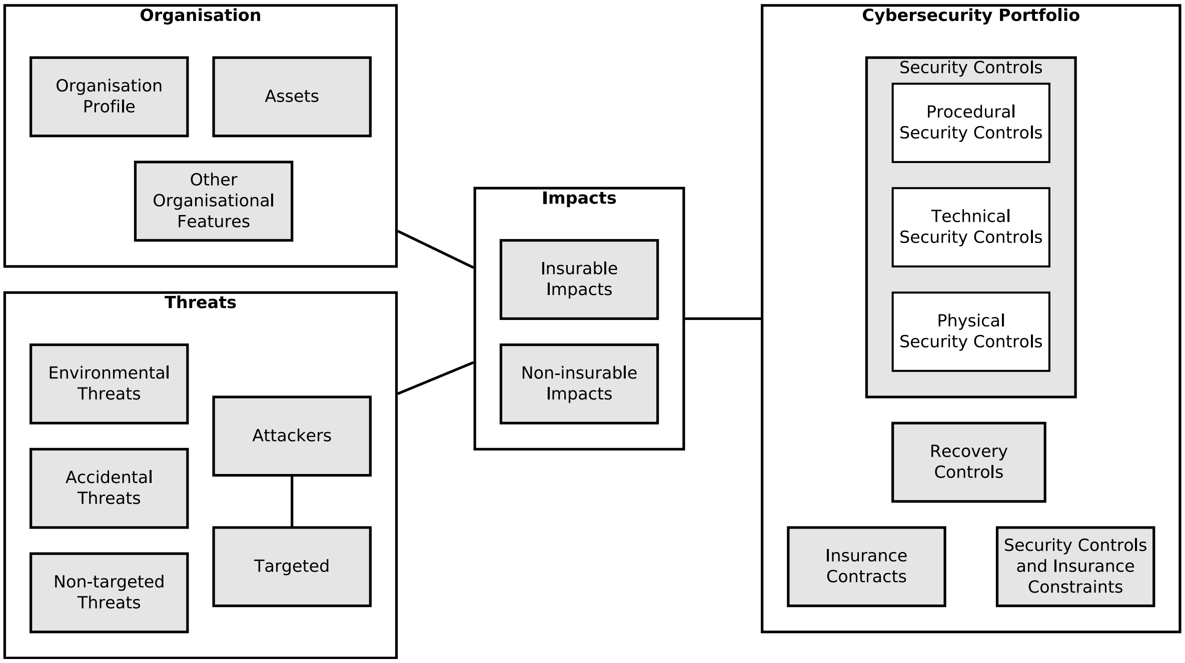


Figure 2: Input to the DSS - subcategories.

Once we have found the relevant categories and subcategories, we may consider the elements within the last ones. For this, we may take into account various catalogues from e.g. the methodologies included in Section 1, as exemplified in Figure 3, which includes just a couple of instances per category.

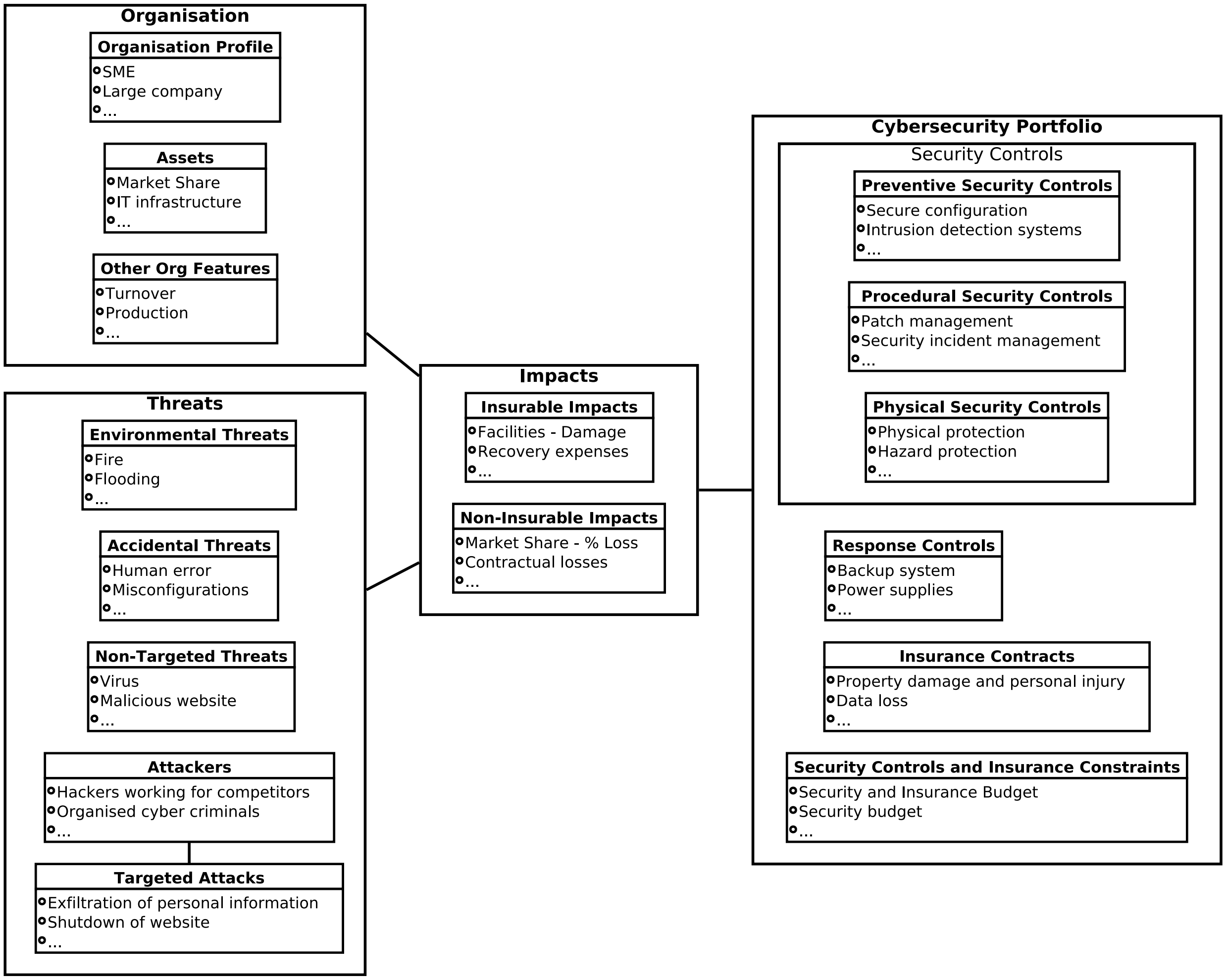


Figure 3: Input to the DSS - subcategories with examples.

A final important defining issue refers to how do we link the corresponding assets with the relevant impacts which will characterise cybersecurity risk management from the perspective of the incumbent organisation. Conceptually, we just need to link the identified relevant assets with the corresponding identified relevant impacts as reflected in Figure 4.

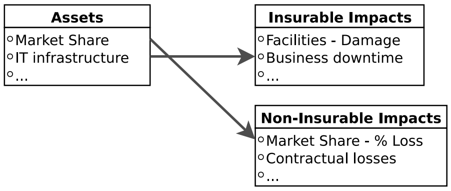


Figure 4: Linking assets and impacts

Clearly, the complexity of the cybersecurity risk management demands decision support tools to facilitate decision making.

3 BASIC UNDERLYING MODEL

Following the schematic description in Section 2, we formulate the basic qualitative model underlying our DSS and describe how it may be solved once we obtain the required quantitative inputs.

3.1 Model formulation

We start by describing our generic qualitative model for cybersecurity risk management. The aim is to support an organisation in its cybersecurity resource allocation process. Such organisation will be called Defender. For the moment, we also distinguish one Attacker which will try to perpetrate attacks to the Defender in pursue of certain goals.

We represent the problem as a bi-agent influence diagram (BAID) in Figure 5, with the terminology used in Banks et al. (2015): the diagram includes oval nodes which represent uncertainties relevant for the problem; hexagonal utility nodes which model preferences over consequences; rectangle nodes portraying decisions modelled through the set of relevant alternatives at such point; and, finally, double oval nodes representing deterministic factors relating the values of the antecessors of the nodes. The diagram also includes arrows pointing to decision nodes (meaning that such decisions are undertaken knowing the values of its antecessors) or chance and value nodes (meaning that the corresponding events or consequences are influenced by the antecessors). Light coloured nodes designate issues relevant to just the Defender problem; darker ones to the Attacker problem; stripped ones are relevant to both agents' decisions. We detail now the various nodes, which correspond with the elements presented in Section 2.

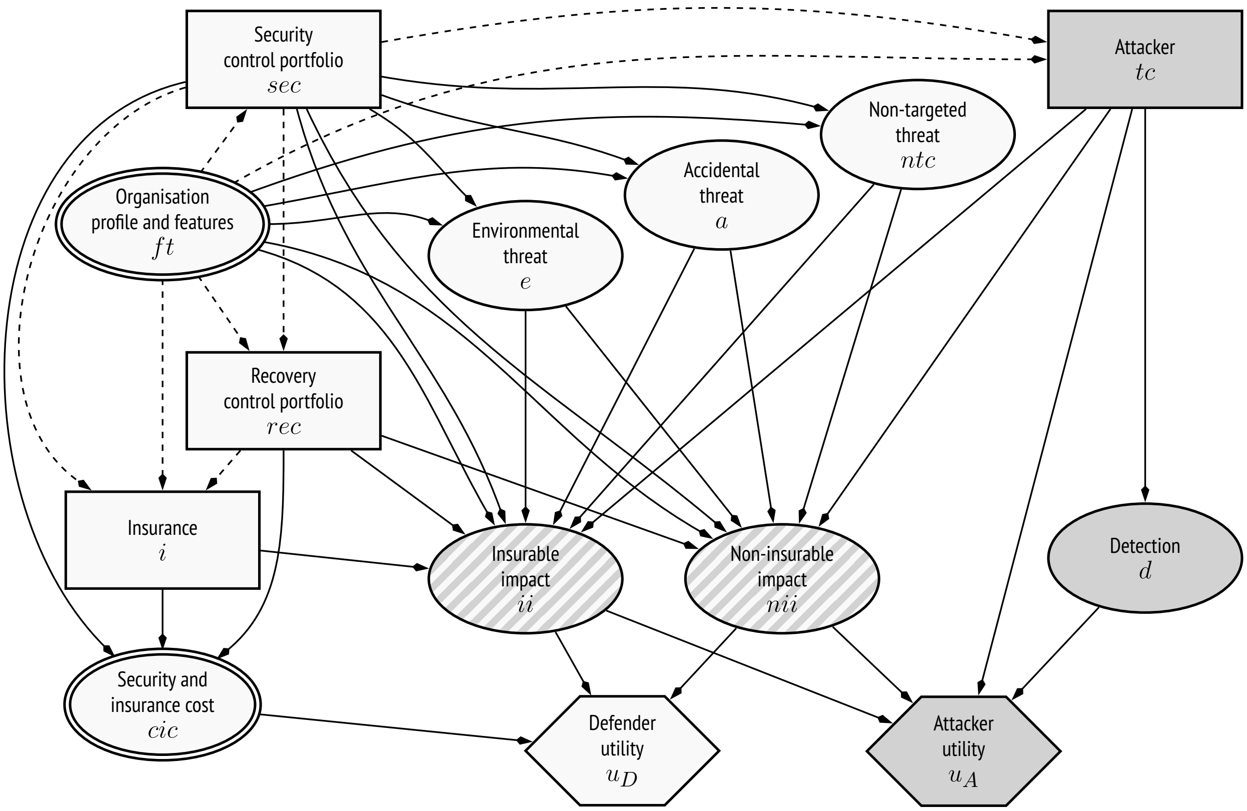


Figure 5: BAID for the basic underlying model.

We start with a description of the organisation profile and features, which will be deterministic; we designate them ft. We then identify the threats relevant to the organisation: we distinguish between *environmental*, designated through e; *accidental*, called a; and *non-targeted cyber threats*, designated ntc, all of them are modelled with uncertain nodes. Besides, we shall also consider *targeted cyber threats*, designated tc; we model them as decisions associated with a different agent, the Attacker. Having identified the threats and relevant assets, we may identify the impacts which we separate between *insurable*, designated ii, and *non-insurable* ones, referred to as nii.

Then, the organisation may identify the actions that may be undertaken to mitigate the likelihood and/or outcome of the threats. The three types of instruments, *security* (sec) or *recovery controls* (rec) and *insurance* (i), are modelled with decision nodes. They will have *(security and insurance) costs*, which will be typically deterministic. We designate them by cic. The above instruments may have to satisfy certain constraints. With all the relevant elements in place, we may build the preference model for the Defender through her utility u_D, identified with a value node.

We turn now to the remaining elements faced by the Attacker. Note first that it perpetrates its attacks with information concerning ft and sec. We then focus mainly on the *detection and identification of the* *attacker*, which we designate by d. Note that detection could depend on the controls implemented, but we may ignore such fact for most attackers because they typically undertake the attack remotely and their identification, in the sense of being able to prosecute them, depends more on their mistakes as attackers (e.g., the language they use in the code) than on the defender forensic capabilities. The exception are insiders, who are more exposed to being detected (in this case, we link the security control portfolio with the insider detection node). Finally, with all his relevant elements in place, we consider the preference model for the Attacker through the utility u_A of the Attacker through a value node.

The above may be seen as a basic template which may be simplified, for example, by eliminating the *rec* node if it is not deemed relevant, or by including several Attacker blocks, as we illustrate in our case in Section 4.

3.2 Model resolution

From the BAID describing the general cybersecurity problem, we derive two problems referred to as the Defender and the Attacker problems, which we represent as influence diagrams, respectively, in Figures 6 and 7.

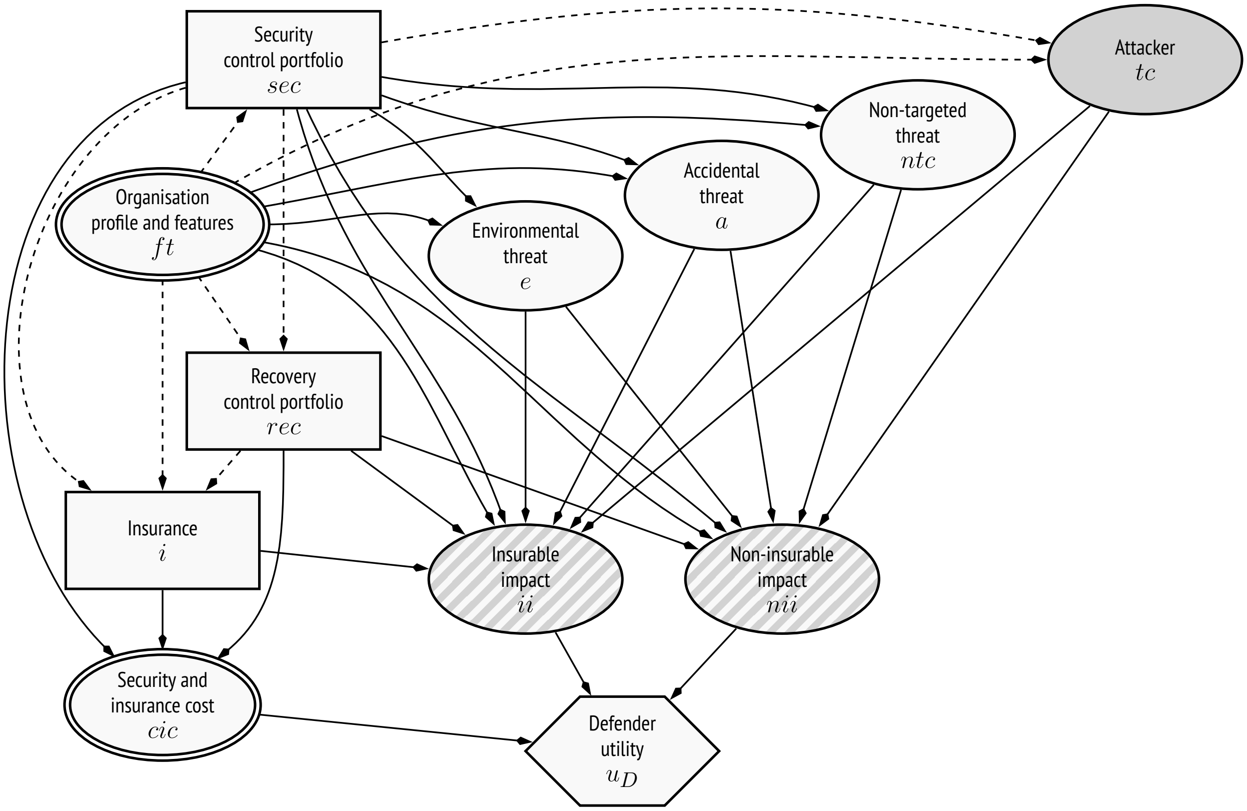

Security control portfolio sec
Attacker tc
Non-targeted threat ntc
Accidental threat a
Organisation profile and features ft
Environmental threat e
Recovery control portfolio rec
Insurance i
Insurable impact ii
Non-insurable impact nii
Security and insurance cost cic
Defender utility u_D

Figure 6: ID for the basic underlying model. Defender problem

We describe now how we use them to guide model elicitation and solution.

Model specification. The first stage when solving the Defender problem, Figure 6, is the provision of the quantitative models required at various nodes. We need to provide the following elements:

- At *chance nodes*, we define the relevant events and the corresponding conditional probability distributions. For example, at the environmental threat node we need to elicit the relevant events e that may affect the organisation and the distribution $p(e|sec, ft)$, which represents the probability that environmental threat event e happens, assuming that the preventive portfolio sec has been implemented by an organisation with features ft. The threats would be chosen from a catalogue and the distributions from another one of relevant parametrised distributions as illustrated in Section 4.

- At *deterministic nodes*, we define the functions relating the values of the node with its antecessors, if they exist. For example, at the security and insurance costs node we would have a function $cic = g(sec, rec, i)$, which represents the costs associated with the portfolios sec, rec and insurance i, typically aggregated additively.

- At *decision nodes*, we must define the alternatives available at the corresponding decision point. For example, at node *Security control portfolio* we must define the sec portfolios relevant to the organisation, characterised by its features ft. The features induce which controls from an available catalogue are relevant, from which we choose those consider incumbent. Based on them, we generate the portfolios satisfying the relevant constraints.

- At the utility node, we must define the relevant preference model which, in our case, would be a function $u_D(cic, ii, nii)$ adopting the general form in [6] outlined in Section 3.3.

The above assessments are standard and may be based on data and/or expert judgement, see e.g. [3], except for $p(tc|sec, ft)$ which models the beliefs concerning the targeted cyber threats given the security control portfolio implemented, as it entails strategic thinking. We provide an approach to facilitate its assessment through the Attacker problem.

The Attacker problem. The influence diagram in Figure 7 reflects the fact that the Attacker has means to detect sec and ft.

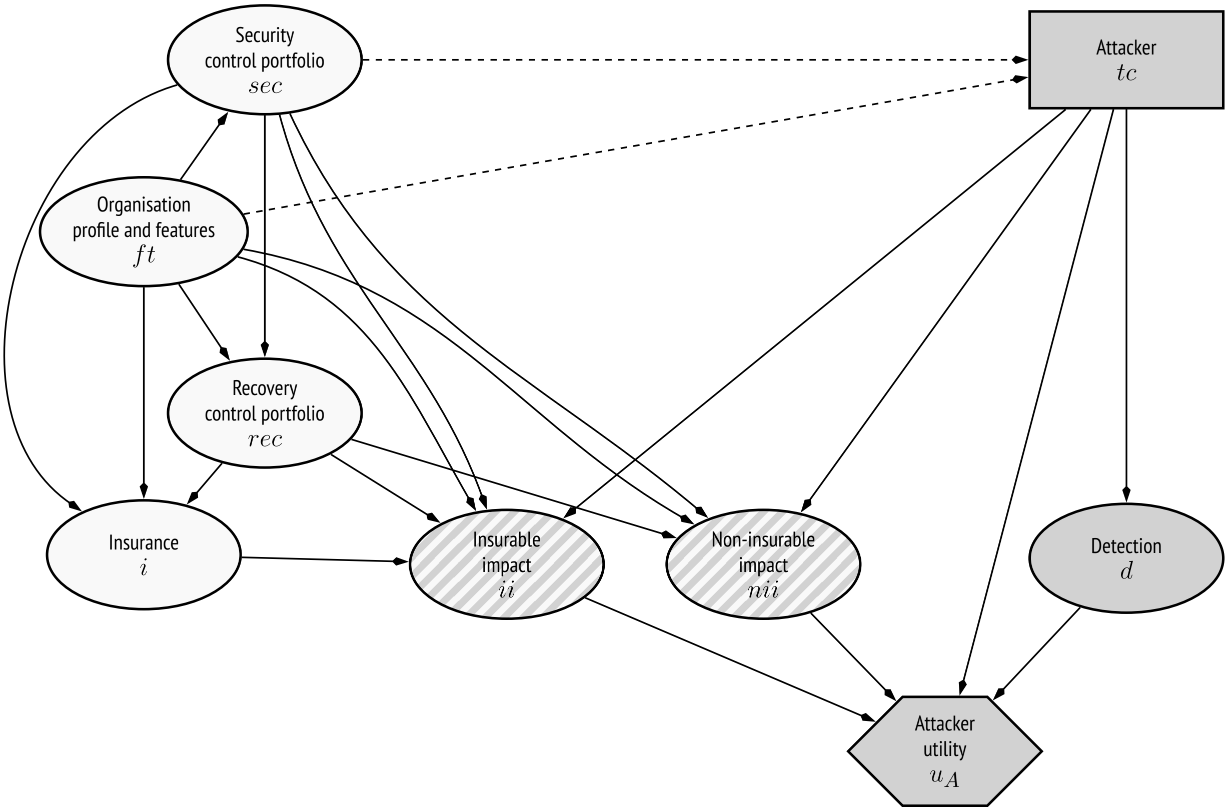


Figure 7: ID for the basic underlying model. Attacker problem

Since we typically have no access to the attacker, we shall have uncertainty about his beliefs and preferences which we model as follows.

- At *uncertainty nodes*, we use random probability models to assess our uncertainty about the attacker beliefs. For example, as we do not have the Attacker available to elicit $p_A(d|tc)$, we model our uncertainty about such distribution through a random distribution $P_A(d|tc)$.

- At the *value node*, we must define the relevant preference model. As we do not have the attacker available to assess the corresponding utility function, we model our uncertainty about it with a random utility function $U_A(ii, nii, d, tc)$.

Then, given a portfolio of security controls sec and features ft, we find the random expected utility associated with the attack tc through

$$\Psi_A(tc|sec, ft) = \int ... \int U_A(ii, nii, d, tc) P_A(nii|tc, sec, rec, ft) P_A(ii|tc, sec, rec, i, ft) P_A(i|sec, rec, ft)$$

$$P_A(d|tc) P_A(sec|ft) P_A(rec|ft, sec) P_A(i|sec, rec, ft) P_A(f) \; \mathrm{d}nii \; \mathrm{d}ii \; \mathrm{d}i \; \mathrm{d}d \; \mathrm{d}sec \; \mathrm{d}rec \; \mathrm{d}f,$$

and determine the random optimal attack by maximising it, which we may perform through computing the random optimal attack

$$TC^*|sec, ft = \arg\max_{tc} \Psi_A(sec, ft, tc). \tag{1}$$

Finally, we estimate the required probability distribution through

$$p_D(tc|sec, ft) = Pr(TC^* = tc|sec, ft).$$

We then use such distribution to feed in and solve the Defender problem.

Model solution. Based on the above assessments, we associate with each feasible combination of security controls and insurance products (sec, rec, ii) their expected utility and find the portfolio with maximum expected utility through the following steps:

1. Remove the deterministic node cic, computing the utilities

$$u_D\big(sec, rec, i, ii, nii\big) = u_D\big(g(sec, rec, i), ii, nii\big).$$

2. Compute the expected utility of the portfolio, removing the uncertainty nodes[1]

$$\psi(sec, rec, ii|ft) = \int ... \int u_D\big(g(sec, rec, i), ii, nii\big) p_D(nii|tc, ntc, a, e, sec, rec, ft) \times$$
$$p_D(ii|tc, ntc, a, e, sec, rec, i, ft) p_D(tc|sec, ft) p_D(ntc|sec, ft) p_D(a|sec, ft) p_D(e|sec, ft) \, \mathrm{d}nii \, \mathrm{d}ii \, \mathrm{d}cic \, \mathrm{d}tc \, \mathrm{d}ntc \, \mathrm{d}a \, \mathrm{d}e. \tag{2}$$

3. Remove the decision node i, registering the optimal cyber insurance, given the security portfolio:

$$i^*(sec, rec|ft) = \arg\max \psi_D(sec, rec, i|ft),$$

$$\psi_D(sec, rec|ft) = \psi_D(sec, rec, i^*(sec, rec|ft)|ft).$$

4. Remove the decision node rec, registering the optimal recovery controls, given the optimal security controls:

$$rec^*(pec|ft) = \arg\max \psi_D(sec, rec|ft),$$

$$\psi_D(sec|ft) = \psi_D(sec, rec^*(sec|ft)|ft).$$

[1] We could reduce the uncertainty nodes one by one. However, in order to save space we remove them in batch.

5. Remove the decision node sec, determining the optimal security controls, given the features of the organisation
$$sec^*(ft) = \arg\max \psi_D(sec|ft).$$

The Defender's optimal resource allocation would then be $(sec^*(ft)\ , rec^*(sec^*(ft)|ft)\ , i^*(rec^*(ft), rec^*(sec^*(ft)|ft))$.

3.3 Implementation

We now describe three important themes for the basic model required for the implementation in our DSS.

Preference models A key ingredient refers to the preference models used for the Defender and the Attacker as we describe next.

Defender preference model In [6], we introduced a generic utility function for a cyber defender whose form was strategically equivalent to

$$u_D(c) = a(1 - \exp(\rho_D c) + b,$$

where c were the Defender costs. By making, $u_D(0) = 1$, $u_D(c^*) = 0$, $u_D(c = c^*/2) = p$, with c^* the worst cost and p elicited from the decision maker, we have that for $x = \frac{1 \pm \sqrt{1-4p(1-p)}}{2(1-p)}$, we estimate the utility parameters with $\rho_D = \ln x/c$, $\quad a = \frac{1}{x^2-1}$, $b = 1$. In the particular case in which the impacts are economic (m) and personal, measured through the number r of records exfiltrated, we specify it through

$$u_D(m, r) = a\left(1 - \exp\left(\rho_D\left(m + c_r r\right)\right)\right) + b$$

where we assessed $c_r = 825$ EUR as the monetisation of an exfiltrated register. To fully adjust the utility function, we determine the worst reasonable cost $c_* = m_* + 825 r_*$, where m_* is the sum of the maximum cost of the impacts and the security budget and r_* is the maximum number of records that can be exfiltrated. We then need to assess the utility of $c_*/2$ and adjust the function [2].

Attacker preference model We provide now a generic model for the preference of an Attacker, based on similar attributes. We consider first as attributes the costs (now gains) inflicted to the Defender by the

[2] In the case in [6], for a certain organisation, m_* is estimated at €2.000.000 and r_* is estimated at 5000, the worst cost being €6.125.000. The best cost was $c^* = 0$, for $m^* = r^* = 0$. Furthermore, for $c_1 = \frac{1}{2}c_*$, we obtain $u(c_1) = 0.8$. The only valid root is $x = 4$, so that $a = 0.066$, $\rho = 4.5267 * 10^{-7}$ and $b = 1$, and the utility function is $u(m, r) = 0.066 * \left(1 - \exp\left(4.5267 * 10^{-7}\left(m + 825r\right)\right)\right) + 1$. which we shall use as default utility function in our DSS when required

Attacker defined through $m = i–d–at$, where i is the economic impact over the Defender, d are the costs if detected and at are the costs of implementing the attack. As before, we aggregate them through $c = m+e\times r$, where e is the reputation monetisation, which would depend on the Attacker. Assuming that this one is risk prone, we use the functional form

$$u_A(c) = a \times \exp(\rho_A \times c) + b.$$

For the attacker, we may estimate the worst c_* and best c^* values and associate with them utilities 0 and 1, respectively. We thus have the system

$$a \times \exp(\rho_A \times c^*) + b = 1,$$

$$a \times \exp(\rho_A \times c_*) + b = 0.$$

We, then assess, for $c = 0$ its utility u leading to

$$a + b = u.$$

The solution of the system comes from solving

$$b = u - a$$

$$\rho_A = \frac{1}{c_*} \log \frac{-b}{a}$$

$$a \left(\frac{u-a}{a} \right)^{\frac{c^*}{c_*}} + (u - a) = 1.$$

We may obtain an explicit expression if we make $\frac{c^*}{c_*} = -2$ in which case we have the relation

$$(3u - 1)a^2 + (2u - 3u^2)a + (u^3 - u) = 0.$$

From it we deduce a, then b and, finally, ρ_A. Note though that, as we do not have available the attacker to elicit u, we shall typically assess that $u \sim \mathcal{U}(u_1, u_2)$ and we would obtain the corresponding distributions over a, b and ρ_A, accordingly.

Multiple Attackers When building the cyber risk model, we shall typically consider several attackers. From a graphical point of view, this corresponds to introducing replicates of the grey part of the influence diagram in Figure 5 for each of the attackers. Then, assuming their conditional independence, given *sec* and

ft, we would solve the i-th Attacker problem as in (1) to obtain $p_D(tc_i|sec, ft)$, $i = 1, ..., k$ and use

$$p_D(tc_1, ..., tc_k|sec, ft) = \prod_{i=1}^{k} p_D(tc_i|sec, ft),$$

which would replace $p_D(tc|sec, ft)$ in the Defender analysis in Section 3.2. We illustrate this in the case study in Section 5.

Implementing the optimisation The optimisation process may be cumbersome and we discuss here how we implement it. First, recall that we need to compute the expected utility $\psi(sec, rec, ii|ft)$ based on (2), which we rewrite as

$$\psi(sec, rec, ii|ft) = \int u_D(x, \theta) p_D(\theta|x) d\theta = \psi(x), \tag{3}$$

where x represents the cybersecurity portfolio (sec, rec, ii), θ represents the involved random variables $(nii, ii, cic, tc, ntc, a, e)$ and we ignore dependence on ft as it is fixed for this discussion. For each portfolio x, we approximate by Monte Carlo the expected utility through

$$\widehat{\psi(x)} = \frac{1}{n} \sum_{i=1}^{n} u_D(x, \theta_i),$$

where $(\theta_i)_{i=1}^{n}$ is a sample of size n from $p_D(\theta|x)$. Note that sampling from such distribution is simple by forward sampling based on the corresponding influence diagram in Figure 6, according to the following algorithm:

```
Given $ft, sec, rec$ and $i$.

  1. Generate $e \sim p_D(e|sec, ft)$, $a \sim p_D(a|sec, ft)$, $ntc \sim p_D(ntc|sec, ft)$, $tc \sim p_D(tc|sec, ft)$.

  2. Generate $nii \sim p_D(nii|tc, ntc, a, e, sec, rec, ft)$, $ii \sim p_D(ii|tc, ntc, a, e, sec, rec, i, ft)$
```

Note also that changing from portfolio x to portfolio x' is facilitated through importance sampling [15] via

$$\int u_D(x', \theta) p_D(\theta|x') d\theta = \int u_D(x', \theta) \frac{p_D(\theta|x')}{p_D(\theta|x)} p_D(\theta|x)\, d\theta,$$

so that

$$\widehat{\psi(x')} = \frac{1}{n} \sum_{i=1}^{n} u_D(x', \theta_i) \frac{p_D(\theta_i|x')}{p_D(\theta_i|x)},$$

for the same sample $(\theta_i)_{i=1}^{n}$ above. Thus, we would just need the initial sample.

Finally, remember that we need to obtain the random optimal attacks to forecast the attacker's actions. We do this by computing

$$\Psi_A(tc|sec, ft) = \int U_A(tc, \beta) P_A(\beta|tc) d\beta,$$

$$TC^*|se, ft = \arg\max_{t_c} \Psi_A(tc|sec, ft),$$

where $\beta = (nii, ii, i, d, rec)$. We generate from such distribution by sampling from (U_A, P_A) and then solving the corresponding optimisation problem, which is structurally similar to the Defender optimisation problem, which we discuss now.

When the number of portfolios is sufficiently small, we just need to approximate the expected utility at each portfolio and find the optimal one. When the number of portfolios is large, or we have continuous portfolios, we may proceed in, at least, three ways:

- Evaluate the expected utility at some portfolios, approximate the expected utility through a regression (meta)model [23] $\widehat{\psi(x)}$ and optimise the regression surface to obtain the approximate optimal portfolio.
- Use an optimisation model which requires only functional evaluations, like the classic Nelder-Mead algorithm [26], and let the algorithm run until a local optimal portfolio is detected.
- Use an augmented simulation algorithm [14], based on the artificial distribution $h(x, \theta) \propto u_D(x, \theta) \times p_D(\theta|x)$ sample from it through an MCMC approach until convergence is detected and, then, find the approximate sample mode of the marginal distribution on the portfolios.

4 DSS ARCHITECTURE

We describe in this section the architecture of the DSS developed to implement the above framework for cybersecurity risk management. Our toolbox adopts the form of an online calculator to guide the user into analysing their current cybersecurity risk level and the optimal cybersecurity strategy for their specific needs, including a cyber insurance product. The calculator is viewed as a multi-step online visually-enriched form which asks the pertinent questions and finally offers the best option for the user. The tool has two modes: one for SME[3] users that are not experts in cybersecurity and another one for more expert users. They differ in the information that they ask to the user but both are implementations of the proposed model. For such purpose, the user selects and provides from various menus a parametric description of its problem.

The current implementation includes the options available in Table 1. It does not clearly exhaust all possible controls or insurance products currently available in the market, all possible impacts and so on. For

[3] Small and medium enterprise

Table 1: Table identifying the general components of the current CYBECO Toolbox model

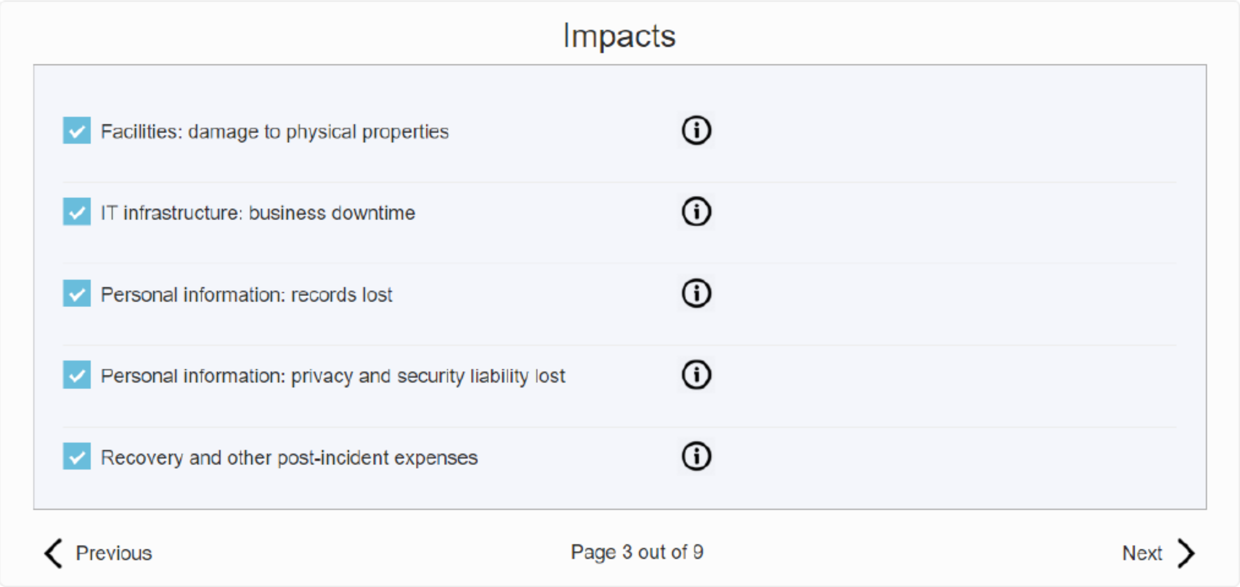
Category	**Component**	**Node**
Assets	Number of computers $ft_{computers}$ Number of servers $ft_{servers}$ Number of personal identifiable information (PII) records $ft_{records}$	ft
Other organisation features	Turnover of the organisation in Euros $ft_{turnover}$ Number of employees $ft_{employees}$	ft
Environmental threats	Fire e_{fire}	e
Accidental threats	Employee error $a_{emperror}$	a
Non-targeted cyber threats	Malware $ntc_{malware}$	ntc
Targeted cyber threats: Attackers	Hacktivists $(ha, hd, \dots)$ Cybercriminals $(ka, kd, \dots)$	
Targeted cyber threats: Targeted Attacks	Targeted data exfiltration, ha_{targex} and $ka_{targexf}$ Targeted data manipulation, $ha_{targman}$ and $ka_{targman}$ Targeted denial of services, $ha_{targdos}$ and $ka_{targdos}$	ha and ka
Attacker uncertainties	Detection	hd and kd
Impacts	IT infrastructure: damage to physical properties, $ii_{physical}$ and $nii_{physical}$ IT infrastructure: business downtime, $ii_{downtime}$ and $nii_{downtime}$ Personal information: records with personal information exposed, nii_{recexp} Personal information: privacy and security liability lost, $ii_{piiliab}$ and $nii_{piiliab}$ Recovery and other post-incident expenses, $ii_{postinc}$ and $nii_{postinc}$	ii or nii
Security controls	Boundary firewalls and internet gateways $sec_{boundary}$ Secure configuration $sec_{secconf}$ Access control sec_{access} Malware protection $sec_{malprot}$ Patch management and vulnerability management $sec_{patchman}$ Hazard protection $sec_{hazprot}$	sec
Insurance contracts	Property damage and personal injury ins_{damage} Data loss $ins_{dataloss}$	ins
Preferences	The defender utility node u_D The hacktivists utility node u_H The cybercriminals utility node u_K	

example, we only include fire within the environmental threats or employee error within accidental threats[4]; we cover only the controls identified in the UK Cyber Essentials [34]. However, the toolbox is open to being expanded by providing the appropriate information and augmenting the menus illustrated in Table 1. Thus, the user needs to specify the information concerning the corresponding nodes as reflected in Figure 8, by clicking or inserting the appropriate entries. Some of the items require the provision of certain parameters as exemplified in Figure 9. Specifically, we have four types of parameters that populate the analysis:

- The first type refers to the parameters provided by the end-user of the toolbox, the supported SME that needs a cybersecurity recommendation, through the user interface, since these parameters refer to data about their organisation (e.g., number of employees or annual turnover). Table 2 summarises these parameters with demonstrative values for the case study presented in Section 5. Note that the first seven parameters are typical of the user whether the remaining ones would be typically fed from a database of cybersecurity and cyber insurance products.
- The second one consists of those parameters provided by cybersecurity or risk analysis experts (e.g., utilities) or related to business but not displayed in the user interface. They can be defined by the

[4] To mitigate this, an 'Other threats' option is included in the corresponding menu

Figure 8: Screen to facilitate selection of relevant impacts.



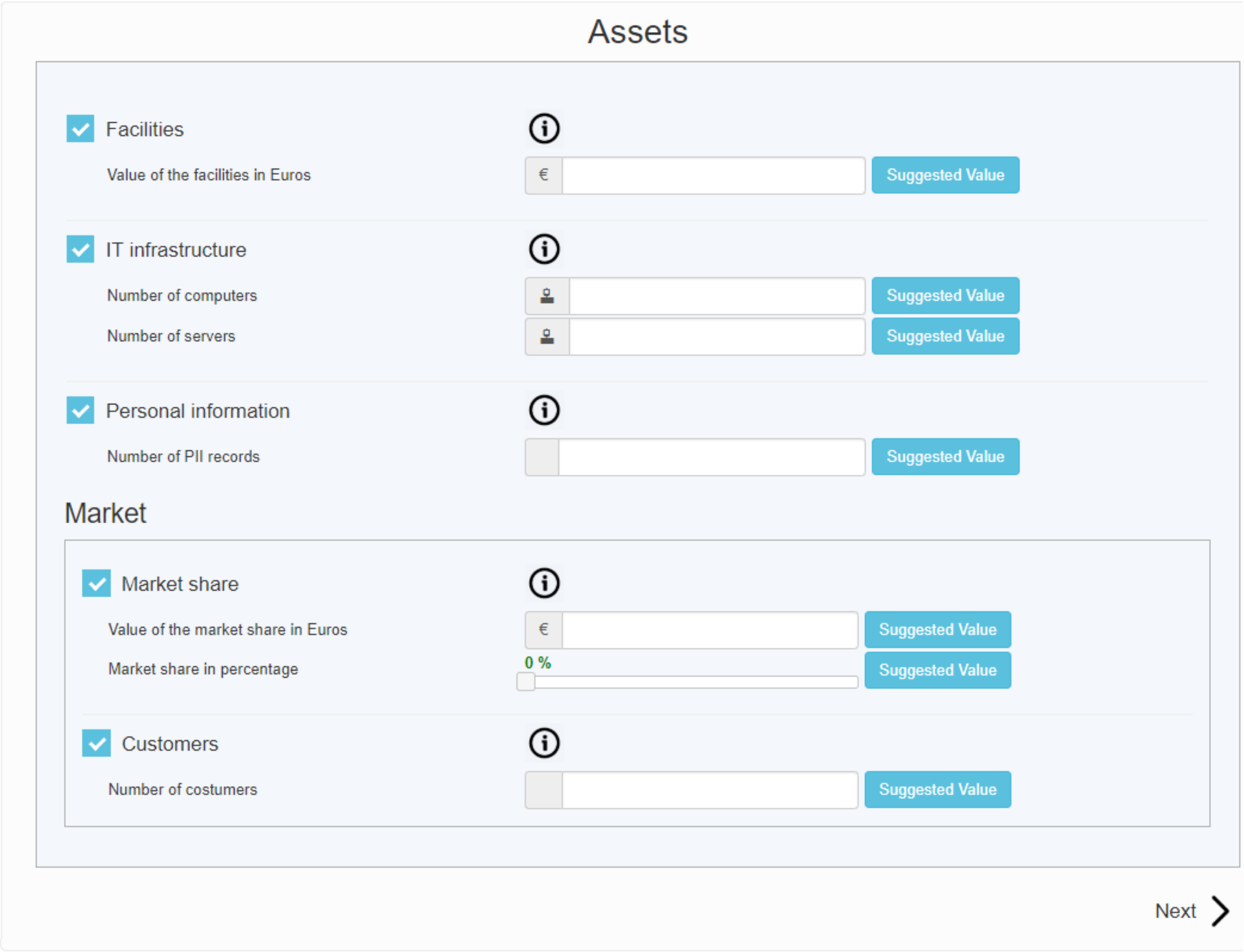
Impacts

- ☑ Facilities: damage to physical properties
- ☑ IT infrastructure: business downtime
- ☑ Personal information: records lost
- ☑ Personal information: privacy and security liability lost
- ☑ Recovery and other post-incident expenses

analysts in a configuration file of the Toolbox. Table 3 summarises these parameters with demonstrative values for our use case. The defender utility values are there presented may be seen as defaults ones, based on our earlier comments in Section 3.3; for sophisticated users we could elicit such values. Similarly, the utility values for the attacker may be seen as default values around which we describe a uniform distribution. Again, with a sophisticated user we could use the procedure described in Section 3.3.

- The third group refers to parameters derived from the previous ones (e.g., the annualised costs of security controls that derive from the OPEX and CAPEX). Table 4 summarises them.
- The fourth type refers to parameters or variables assessed for the models (e.g., probability of fire). They are detailed in the corresponding sections below and built in within the system, but open to being evolved as learning from data takes place.

Figure 9: Screen to facilitate selection of assets and provision of parameters.



Assets

☑ Facilities

Value of the facilities in Euros € Suggested Value

☑ IT infrastructure

Number of computers Suggested Value

Number of servers Suggested Value

☑ Personal information

Number of PII records Suggested Value

Market

☑ Market share

Value of the market share in Euros € Suggested Value

Market share in percentage 0 % Suggested Value

☑ Customers

Number of costumers Suggested Value

Next ❯

Table 2: Table summarising parameters defined by end user, with the demonstrative values in our use case. Capital expenses (CAPEX) refer to the cost of acquiring the product. In this use case, they are amortised in seven years. Operational expenses (OPEX) refer to ongoing or subscription costs during a year. All costs in euros.

Parameter	**Definition**
$ft_{facilities} = 3147560$	Monetary value of the facilities and their contents
$ft_{computers} = 20$	Number of computers
$ft_{servers} = 14$	Number of servers
$ft_{records} = 2000$	Number of personal identifiable information records
$ft_{turnover} = 3346706$	Turnover (year)
$ft_{employees} = 20$	Number of employees
$constraint_{budget} = 8000$	Security and insurance budget (year)
$capex_{boundary} = 1000$ $opex_{boundary} = 700$	Cost of boundary firewalls and internet gateways.
$capex_{secconf} = 800$ $opex_{secconf} = 700$	Cost of implementing a secure configuration.
$capex_{access} = 400$ $opex_{access} = 250$	Cost of implementing access control.
$capex_{malprot} = 0$ $opex_{malprot} = 600$	Cost of malware protection.
$capex_{patchman} = 0$ $opex_{patchman} = 800$	Cost of implementing patch and vulnerability management.
$capex_{hazprot} = 500$ $opex_{hazprot} = 400$	Cost of hazard protection.
$cic_{damage} = 300$	Insurance premium: Property damage and personal injury
$coverage_{damage} = 0.9$	Insurance coverage: Property damage and personal injury
$cic_{data} = 200$	Insurance premium: Data loss
$coverage_{data} = 0.9$	Insurance coverage: Data loss

Table 3: Table summarising parameters defined by Toolbox analyst with demonstrative values in our use case.

Parameter	**Definition**
$hourrate_{cybsec} = 100$	Hourly rate of cybersecurity team
$amortisation_{capex} = 5$	Number of years for CAPEX amortisation
$workload_{year} = 1500$	Annual hours worked by an employee
$\rho_D = -4.5267 \times 10^{-7}$	Defender risk aversion coefficient
$a_D = 0.066$	Defender utility parameter
$\rho_H = 2.901 \times 10^{-7}$	Hacktivists risk aversion coefficient
$a_H = 0.079$	Hacktivists utility parameter
$\rho_K = 3.402 \times 10^{-7}$	Cybercriminals risk aversion coefficient
$a_K = 0.07$	Cybercriminals utility parameter

Table 4: Table summarising parameters derived from those provided by the Toolbox.

Parameter	**Definition**
$productivity_{employee} = \frac{ftturnover}{ft_{employees}}$	Productivity (annual income per employee)
$productivity_{hour} = \frac{productivity_{employee}}{workload_{employee}}$	Productivity (income per hour worked by employee)
$value_{pii} = 825$	Personal value of personal identifiable information record
$liability_{piirecord} = 150$	Liability cost per record (average value)
$cic_* = opex_* + \frac{opex_*}{amortisation_{capex}}$	Security control annualised cost

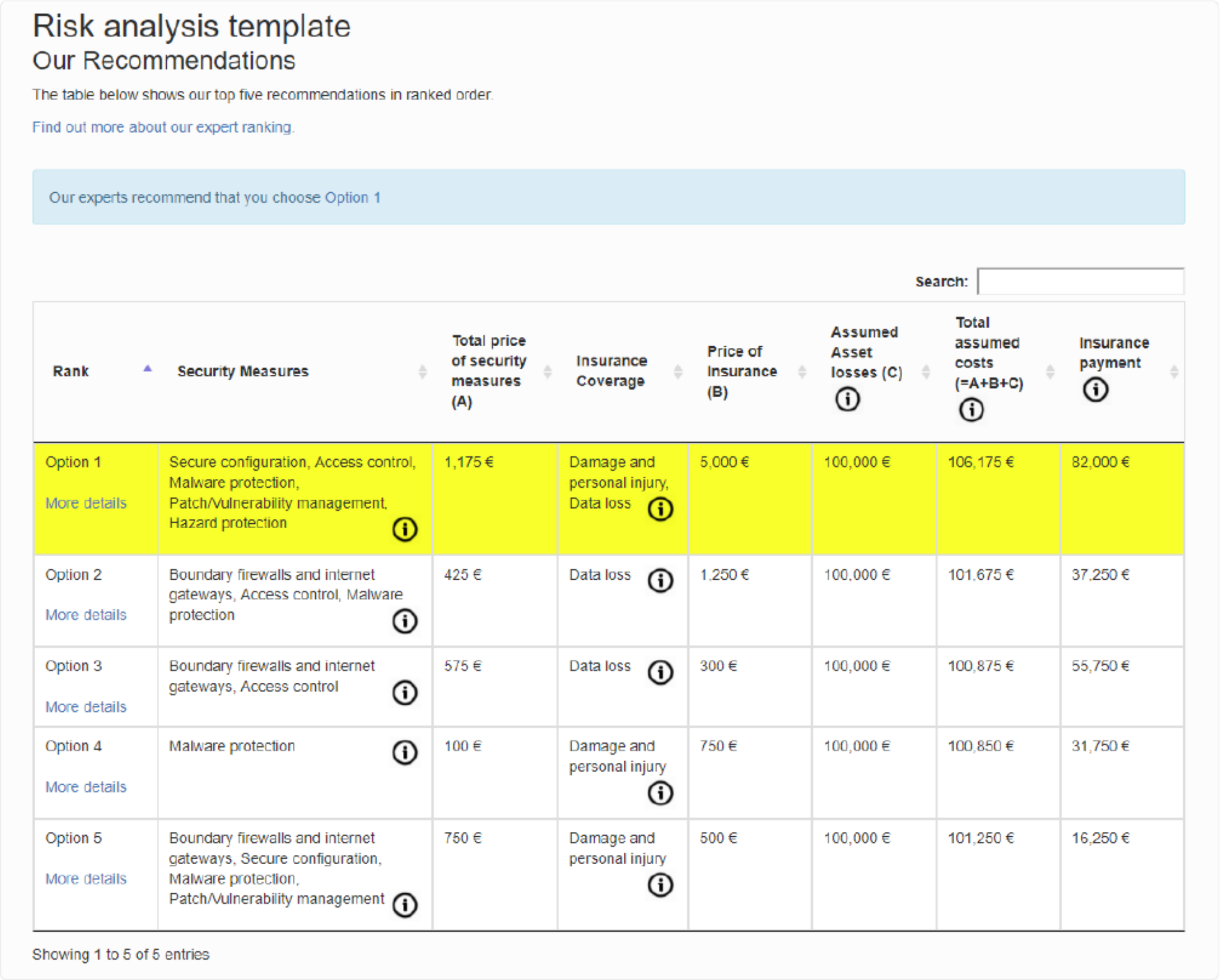
The system includes a help facility as well as a knowledge base with concepts in cybersecurity and cyber insurance.

In such a way, the organisation gets defined through its features, assets, non targeted threats, its potential attackers, targeted threats and attacker uncertainties, its relevant impacts, potential security controls and insurance contracts and preference models. The Toolbox distinguishes between technical, procedural and physical security controls in its display; however, this differentiation does not affect the structure of the underlying influence diagram in Figure 5. It also includes constraints concerning security and insurance budgets, only enabling portfolios satisfying such constraints.

Once the user concludes his selection, the DSS runs a simulation to forecast the likely actions of each of the considered attackers as reflected in (1). Such forecasts then feed in the Defender problem which is run to calculate the optimal choice of security controls and insurance as reflected in (3). When the optimisation is completed, the tool informs the user about such optimal portfolio and the likely attacks. Figure 10 provides an example of the recommendation in a specific case. Both the attackers' and the Defender's problems run on parametrised models at various nodes. For example, we have a parametric model (4) which relates the impact of cybersecurity controls on malware attack probabilities, as illustrated in Section 5.2.

In order to enhance the usability, visual appearance of outputs, and general user-friendliness of the DSS, three types of user-oriented validations have been undertaken to collect relevant feedback. First, we have designed and implemented a behavioural economic experiment with a sample of 2,000 potential users of the calculator (workers in SMEs in managerial or cybersecurity related positions) in Germany, Poland, Spain

Figure 10: Screen displaying the optimal portfolio.



Risk analysis template

Our Recommendations

The table below shows our top five recommendations in ranked order.

Find out more about our expert ranking.

Our experts recommend that you choose Option 1

Search:

Rank	Security Measures	Total price of security measures (A)	Insurance Coverage	Price of Insurance (B)	Assumed Asset losses (C)	Total assumed costs (=A+B+C)	Insurance payment
Option 1 More details	Secure configuration, Access control, Malware protection, Patch/Vulnerability management, Hazard protection	1,175 €	Damage and personal injury, Data loss	5,000 €	100,000 €	106,175 €	82,000 €
Option 2 More details	Boundary firewalls and internet gateways, Access control, Malware protection	425 €	Data loss	1,250 €	100,000 €	101,675 €	37,250 €
Option 3 More details	Boundary firewalls and internet gateways, Access control	575 €	Data loss	300 €	100,000 €	100,875 €	55,750 €
Option 4 More details	Malware protection	100 €	Damage and personal injury	750 €	100,000 €	100,850 €	31,750 €
Option 5 More details	Boundary firewalls and internet gateways, Secure configuration, Malware protection, Patch/Vulnerability management	750 €	Damage and personal injury	500 €	100,000 €	101,250 €	16,250 €

Showing 1 to 5 of 5 entries

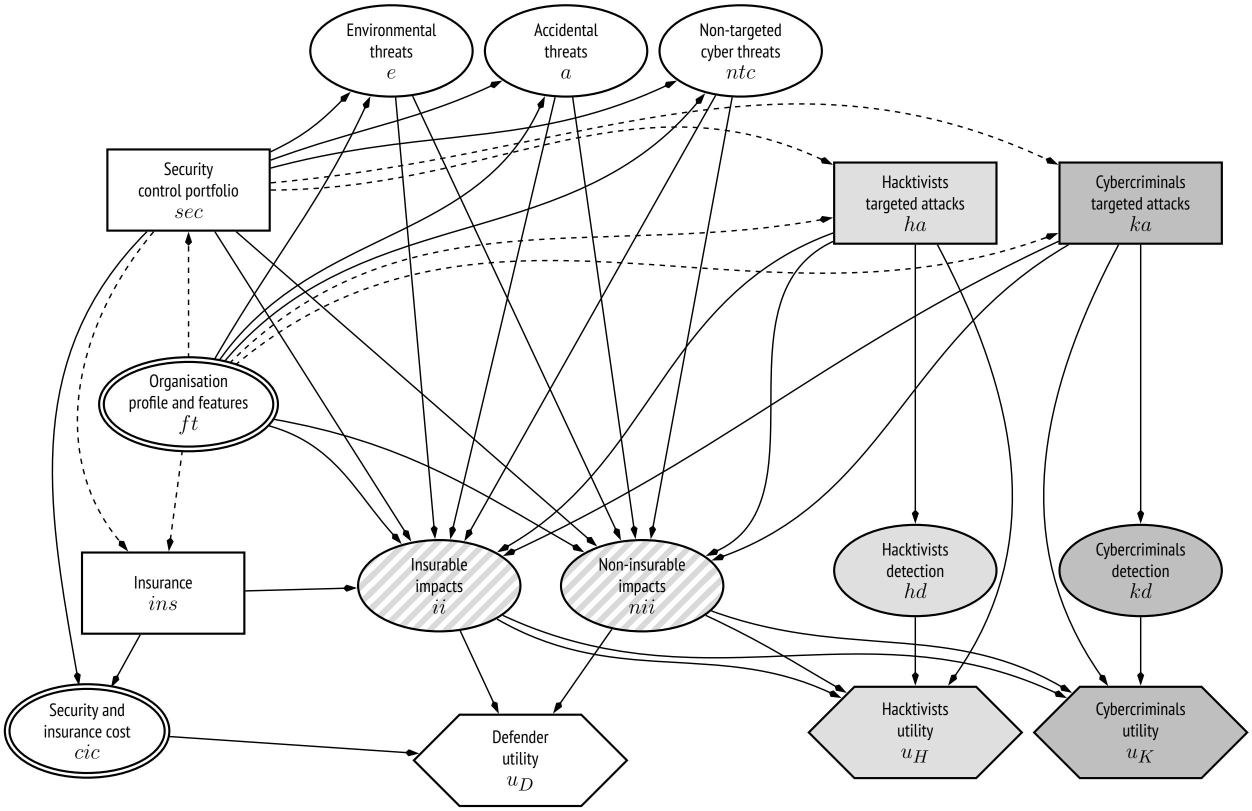
and UK. In a gamified controlled environment, the participants were asked to define the cyber-protection and cyber-insurance strategies of an SME using five different framings of the calculator output. Another evaluation target has been the user navigation paths, offered by the toolbox, which were evaluated by two focus groups with about 50 actual users, which helped to improve its visual aspect. Finally, a rich set of use cases has been developed and applied as usage patterns to crosscheck the correct implementation of the cyber risk analysis algorithms, one of which we next illustrate.

5 CASE STUDY

5.1 Introduction

We apply the framework described to a case study developed for the Toolbox. The cybersecurity risk management problem refers to an organisation, an SME which we call the Defender, which faces potential cyber threats. We must support them in choosing an appropriate cybersecurity portfolio, possibly including a cyber insurance. We consider a one-year planning horizon. We structure the problem through the multi-agent influence diagram (MAID) in Fig. 11. As variations over the basic model in Figure 5, we consider two targeted cyber threats from hacktivists ha and cybercriminals kf, represented with different colours; moreover, we do not consider reactive controls. In this case, we consider all components in Table 1.

Figure 11: Case study as a MAID. White nodes correspond solely to the defender problem. Light grey nodes correspond to the hacktivist problem. Darker grey nodes to the cybercriminal problem. Stripped nodes affect several agents.


Environmental threats e
Accidental threats a
Non-targeted cyber threats ntc
Security control portfolio sec
Hacktivists targeted attacks ha
Cybercriminals targeted attacks ka
Organisation profile and features ft
Insurance ins
Insurable impacts ii
Non-insurable impacts nii
Hacktivists detection hd
Cybercriminals detection kd
Security and insurance cost cic
Defender utility u_D
Hacktivists utility u_H
Cybercriminals utility u_K

5.2 Assessing the defender non-strategic beliefs and preferences

We first provide the quantitative assessment of the Defender beliefs and preferences not requiring strategic analysis. We follow the order of the influence diagram, so that parent nodes are defined before their child nodes.

Organisation profile and features. The organisation profile and features are described through

$$ft = (ft_{facilities}, ft_{computers}, ft_{servers}, ft_{records}, ft_{turnover}, ft_{employees}),$$

and specified in Table 2 for our case study.

Modelling security controls. The security controls portfolio is described through the variables

$$sec = (sec_{boundary}, sec_{secconf}, sec_{access}, sec_{malprot}, sec_{patchman}, sec_{hazpot}).$$

For each option, we have the binary choice of implementing it or not. We thus have 64 portfolios, which could be constrained by the security and insurance budget. The precedence relation $sec|ft$ is informative, indicating that the Defender knows his profile and features before she selects the security controls.

Modelling insurance contracts. The insurance portfolio is described through

$$ins = (ins_{damage}, ins_{dataloss}).$$

Each insurance product is defined by price and coverage, [29]. For all options, we have the binary choice of implementing it or not. We thus have 4 insurance portfolios, which could be constrained by the security and insurance budget. Note that we have included just one product per line of impact; should we have more than one of them, we would introduce constraints to limit the acquisition of just one product per impact. The precedence relation $ins|ft, sec$ is informative, indicating that the Defender knows his profile and features and their security control selection before the insurance choice.

Modelling security and insurance costs. The security and insurance costs are described through

$$cic = (cic_{boundary}, cic_{secconf}, cic_{access}, cic_{malprot}, cic_{patchman}, cic_{hazpot}, cic_{damage}, cic_{dataloss}).$$

The precedence relation $cic|sec, ins$ indicates that the cost is incurred only if the corresponding security control or insurance product is implemented. Based on the parameters and formulas in Tables 2, 3 and 4 and using its annualised values, we have that

$$cic = (900, 960, 330, 660, 800, 500, 300, 200).$$

Introducing constraints. The security and insurance budget is an user-defined parameter that constrains the security and insurance portfolio. In our example, such budget b is 8000, Table 2. Then, we exclude the combinations of portfolios that exceed the budget

$$\{(sec_i, ins_j) : cic|sec_i, ins_j \geq b\}.$$

Additionally, some insurance products may require the implementation of certain security controls. In our case, the property damage and personal injury insurance requires the hazard protection control, whereas the data loss insurance requires all controls to be implemented (i.e., boundary firewalls, secure configuration, access control, malware protection and patch and vulnerability management), since they are considered essential. Therefore, the following combinations of portfolios are excluded:

$$\{(sec_i, ins_j) : sec_{hazprot} = 0 \ \wedge \ ins_{damage} = 1\},$$

$$\{(sec_i, ins_j) : (sec_{boundary} = 0 \vee sec_{secconf} = 0 \vee sec_{access} = 0 \vee sec_{malprot} = 0 \vee sec_{patchman} = 0) \wedge ins_{data} = 1\}.$$

Modelling environmental threats. In our case, the environmental threats are described as [5]

$$e = e_{fire}.$$

The precedence relation $e|ft, sec$ indicates that the materialisation of a threat into an incident could depend on the security controls implemented and the profile and features of the organisation. Following our assessment in [28], we have that there is a 2.2% probability of an industrial fire during a year in a typical installation. Therefore, we model the number of annual fires through a Poisson distribution $\mathcal{P}(0.022)$.

Modelling accidental threats. The accidental threats are described as

$$a = a_{emperror}.$$

Based on our assessment[6], we have that there is a 5.92% probability of a human error during a year. Therefore, we model the number of employee errors per year through a Poisson distribution $\mathcal{P}(0.0595)$.

[5] Recall that typically there would be more threats, but we leave just one for simplicity of exposition.

[6] A high percentage of cybersecurity incidents involve human errors (some reports estimating it at around 90%). However these incidents also involve malware or hacking. The only indicative figure we have found [27] is that around 47% root causes of data breaches are malicious, 28% human errors and 25% system failures. The Eurostat 2010 survey on Digital Economy and Society statistics (`https://ec.europa.eu/eurostat/web/digital-economy-and-society`) indicates that around 1% of small enterprises (10-49 employees) experienced ICT related security incidents resulting in disclosure of confidential data in electronic form by employees (whether intentionally or not) and that 1% due to intrusion, phishing attacks and similar. Taking the previous percentages of root causes, we infer that around 0.37% of the disclosures (2% per year) come from human errors, i.e., 0.74% per year. The survey does not provide data about unintentional destruction or unavailability of data by employees, but assuming the same proportions as in data disclosure, we assess a 2.22% for unavailability and 2.96% for destruction or corruption. Adding these three probabilities we obtain the final percentage of 5.95%.

Modelling non-targeted cyber threats. The non-targeted cyber threats are described as

$$ntc = ntc_{malware}.$$

The precedence relation $ntc|ft, sec$ indicates that the materialisation of a cyber threat into an incident could depend on the security controls implemented and the profile and features of the organisation. Based on our assessment [7], we estimate that there is a 5.28% probability of malware during a year.Additionally, we have to take into account the effects of the cybersecurity controls[8], through a risk reduction coefficient, red, which we describe through

$$red = 0.34+0.66(1-0.13sec_{patchman})(1-0.5sec_{boundary})(1-0.5sec_{secconf})(1-0.5sec_{access})(1-0.9sec_{malprot}). \tag{4}$$

Finally, we model the number of malware attacks through a Poisson distribution $\mathcal{P}(1, 0.0528 \times red)$.

Modelling impacts on the assets. The non-insurable impacts on the assets are described as

$$nii \sim (nii_{recexp}, nii_{downtime}, nii_{postinc}),$$

whereas the insurable ones are

$$ii \sim (ii_{physical}, ii_{piiliab}),$$

as expressed in Table 1. The precedence relation $ii|ft, ins, e, a, ntc, ha, ka$ indicates that the impacts could be affected by the organisation features and profile, their insurance choice and the threats (including the adversarial ones that we assess in Sec. 5.3).

We model the damage to physical properties as

$$ii_{physical} \sim (e_{fire}, sec_{hazprot}).$$

Following our assessment in [28], we assume that the fire duration $duration_{fire}$ depends on whether the

[7] Malware represents 34.1% of attacks (37.7% , if we exclude attacks classified as unknown). From the Eurostat Digital economy and society data we have that 14% of small enterprises experienced security incidents excluding disclosure from employees, and assess a probability of malware infection of 5.28% per year. Data obtained from the Hackmageddon (`https://www.hackmageddon.com/category/security/cyber-attacks-statistics/`) portal specialised in information security statistics

[8] An assessment of the cybersecurity essentials against a set of commodity-level attacks [31] indicates that patch management was effective 87% of the time, whereas anti-malware only 10%. No percentage was provided for the other controls, so we assume an effectiveness of 50%. Additionally, commodity-level attacks are the majority but not all of the attacks, specially against small business. We assume in this case that at least two thirds are commodity-level and one third is bespoken or based on sophisticated tools

hazard protection measures are implemented or not. Specifically, if $sec_{hazprot} = 1$, $duration_{fire}$ is described through a Triangular distribution $\mathcal{T}ri(0.8, 63, 10)$, whereas if $sec_{hazprot} = 0$, $duration_{fire}$ is a Gamma distribution $\Gamma(0.85, 0.01099)$. Based on our assessments[9], the fire impact on facilities is modelled through

$$ii_{physical} = 0.8 ft_{facilities} \min\left(1, \frac{duration_{fire}}{60}\right)$$

When it comes to exfiltrated records, most models apply to certain topologies of networks and computers. Our model just asks for the number of computers and servers and is applied to an SME (and we assume that it is not extremely difficult to hijack computers and servers once one of them has been hacked from the exterior). Based on that, for each type of attack we model the exfiltration impact as follows:

- For the targeted exfiltration (hacktivists) $nii_{recexp|ha_{targexf}} \sim \mathcal{U}(0, ft_{records})$, if $ha_{targexf} = 1$ (similarly for the cybercriminal targeted exfiltration).
- For malware[10], $nii_{recexp|ntc_{malware}} = 0.4\, x_{recexp}\, ntc_{malware}, \quad x_{recexp} \sim \mathcal{U}(0, ft_{records})$, if $ntc_{malware} > 0$.
- For the error[11], $nii_{recexp|a_{emperrror}} = 0.1243\, y_{recexp}\, a_{emperror}, \quad y_{recexp} \sim \mathcal{U}(0, ft_{records})$, if $a_{emperror} > 0$.

Each of these threats could expose a different number of records. Although this involves that some record could be exposed in various attacks, we assume that the final number of records exposed comes from the maximum of the previous incidents, i.e.

$$nii_{recexp} = \arg\max\Big(nii_{recexp}|ha_{targexf}, nii_{recexp}|ka_{targexf}, nii_{recexp}|ntc_{malware}, nii_{recexp}|a_{emperror}\Big).$$

A derived, and potentially insurable, impact is the liability due to the records exposed. This relates to the liability per record exposed[12], i.e.,

$$ii_{liability} = liability_{piirecord} \times nii_{recexp}$$

[9] Typically, 20 % to 30% of the real estate value corresponds to the land value. Should the building be destroyed, this value still remains. Additionally, we assume that the offices of an SME (several hundreds of square meters) would degrade in less than an hour, based on our assessment [28] in which we elicited from experts that a fire would degrade a typical industrial warehouse in less than two hours.

[10] Malware attacks affecting confidentiality represents around 40%, based on the Eurostat Digital economy and society data.

[11] Based on the data we describe in the footnotes when modelling accidental threats, these threats represent 12.43% of cybersecurity incidents caused by accidental human errors.

[12] For this value we take the demonstrative figure of €150 for $liability_{piirecord}$, based on an study that estimated the cost of data breach per record in \$ 143 in Italy and \$ 195 in Germany [27]

For business downtime, we first calculate the downtime for each of the threats that can cause them as

$$nii_{downtime|threat} = x_{downtime|threat} \times occurrences_{threat} \times ft_{employees} \times productivity_{employee},$$

where *threat* could be any of the threats, except the targeted exfiltration as these do not usually cause downtime. The downtime caused by a threat is modelled as $x_{downtime|threat} \sim \mathcal{U}(0.5y_{downtime|threat}, 2y_{downtime|threat})$, where $y_{downtime|threat}$ is the average downtime in hours per threat occurrence[13]. We finally aggregate all the downtimes caused by different threats as

$$nii_{downtime} = nii_{downtime|fire}+nii_{downtime|emperror}+nii_{downtime|malware}+nii_{downtime|targexf}+nii_{downtime|targman}.$$

For the post incident cost, we follow a similar approach to the business downtime: we first calculate the cost for each of the threats based on the cybersecurity experts rate as

$$nii_{postinc|threat} = x_{postinc|threat} \times hourrate_{cybsec},$$

where *threat* could be any of the threats above. The post incident cost caused by a threat is modelled through $x_{postinc|threat} \sim \mathcal{U}(0.5y_{postinc|threat}, 2y_{postinc|threat})$, where $y_{postinc|threat}$ is the average hour to solve a cyber incident in hours per threat occurrence[14]. We finally aggregate all downtimes caused by the different threats as

$$nii_{postinc} = nii_{postinc|emperror} + nii_{postinc|malware} + nii_{postinc|targexf} + nii_{postinc|targman} + nii_{postinc|targdos}$$

Defender utility. We use the utility function defined in [6], and revised in 3.3. In our case,

$$m = cic + ii_{physical} + ii_{downtime} + ii_{piiliab} + ii_{postinc} + nii_{physical} + nii_{downtime} + nii_{piiliab} + nii_{postinc},$$

$$r = nii_{recexp}.$$

We use the parameter settings in Table 3, therefore using our default utility function.

[13] Following expert consultation, we have that the average downtime are 28 hours for the fire, 2.5 for the employee error or the malware, and 10 for the targeted denial of services or targeted manipulation.

[14] Following reports [1], we have that the average hours of postincident work are 55.2 for the targeted exfiltration, 61 for the targeted manipulation, 20 for the targeted denial of service, 16.8 for the accidental error and 14.6 for the malware.

5.3 Assessing the Hacktivists' random beliefs and preferences

We assess now the random judgements available from the first Attacker group.

Targeted attacks. This decision node models whether the hacktivists undertake any of their attacks against the Defender, specifically

$$ha = (ha_{targexf}, ha_{targman}, ha_{targdos}).$$

For all the options, they have the binary choice of undertaking them or not. Thus, they have 8 attack portfolios. The precedence relation $ha|sec$ is informative, indicating that the hacktivists observe the security controls from the Defender. In general, we have $p(ha|sec) = p(ha_{result}|sec)[ha_{decision}|sec]$, where $ha_{decision}|sec$ designates the hacktivist decisions and $p(ha_{result}|sec) = red_{attacker}$ is the probability that the hacktivists are successful in their attack. This is conceptually equivalent to the risk reduction coefficient s defined when we modelled the non-targeted cyber threats. However, in this case, it represents the belief of the attacker about his success in case he observes security controls. If the hacktivist does not observe security controls then $red_{attacker} = 1$; otherwise, we add uncertainty modelling $red_{attacker}$ as an uniform distribution[15] $\mathcal{U}(0.016250.5)$.

Impacts. We base our estimate on that of the Defender nodes (Sect. 5.2), considering only the impacts caused by the hacktivists and adding some uncertainty through random probability distributions.

Hacktivists detection. This represents the chance of the hacktivists being detected, given the type of attack executed. The precedence relation $hd|ha$ indicates that the detection depends on whether the hacktivists launch an attack

$$hd \sim (ha_{targexf}, ha_{targman}, ha_{targdos}).$$

Following our assessment through expert judgement [16] in [28], we model the costs of attacker detection hd through $hd_{probability}$ hd_{cost}, where $hd_{probability} \sim \mathcal{B}(1, 0.002)$ and $hd_{costs} \sim \mathcal{U}(300000, 450000)$.

[15] Based on what we described when assessing the effectiveness of the security controls against malware, the simplest situation for the attacker is when there is only one security control with a 50% effectiveness, whereas the hardest situation would be when all the security measures are in place, providing a combined effectiveness of 98%

[16] We have that detection and identification of the attacker has a probability of 0.2%, and thus $p(hd|ha) = 0.002(ha_{targexf} + ha_{targman} + ha_{targdos})$, indicating that the attacker has more chances if he decides to undertake several attacks. Taking our estimations in [28], we have that the hacktivists may face legal costs, penalties and indemnities, around the expected cost of detection for the hacktivists are around €380.000. To add some uncertainty, we model them as a uniform distribution $\mathcal{U}(300000, 450000)$.

Hacktivists utility. We tke the utility function defined in Section 3.3 where m is the monetary impact caused to the defender, minus the detection costs and the costs of undertaking the attacks (€800 for the exfiltration, €1250 for the manipulation and €600 for the denial of service). We also have that r is the impact on personal rights, monetised with $e = 825$, ρ_A is the risk proneness coefficient and a and b are parameters used to scale the utility in [0,1].

We use the parameter settings defined in Table 3 so that the utility function is

$$u_D(m_{defender}, r_{defender}) = 0.066 * \Big(1 + \exp\Big(4.5267 * 10^{-7}\Big(m_{defender} + 825 r_{defender})\Big)\Big) + 1$$

with uncertainty around a, b and ρ_A.

Simulating the Hacktivists' problem. As described in Sect. 3.3, we simulate from the hacktivists problem to forecast his actions and estimate the probability distribution $p(ha|sec)$ of random optimal attacks by the hacktivist, represented in Table 5.

Table 5: Conditional probability table generated by simulating the hacktivists problem. It displays the attack decision probabilities, conditional on the protection implemented by the defender.

Does the defender implement any technical security control?	**Probability of hacktivists deciding to launch targeted exfiltriation**	**Probability of hacktivists deciding to launch targeted data manipulation**	**Probability of hacktivists deciding to launch targeted denial of service**
Yes	50.3%	73.8%	72.5%
No	52.4%	89.1%	95.7%

5.4 Assessing the Cybercriminals' random beliefs and preferences.

We proceed in a similar fashion as in the hacktivists problem. All the assessments and figures are equivalent (they can execute the same attacks and both have the same skill levels) except some costs that increase due to their organisational nature[17]. The result of the simulation is the probability distribution $p(ka|sec)$ of random optimal attacks by the cybercriminals, represented in Table 6.

5.5 Solving the defender problem

After defining the non-strategic beliefs of the Defender and simulating the attackers' problems, we proceed to find the optimal portfolio of security controls and insurance for the Defender, as described in Sect. 3.3,

[17] Since they are a cybercriminal organisation they may face additional costs if detected. Taking our estimations in [28], we have that they may face legal costs and indemnities such as the hackers, but also reputational costs and suspension costs since most of the time cybercriminal organisations have a legal façade as a legit business. In this case the cost of detections are around €2.430.000 but to add some uncertainty we model them as a uniform distribution $\mathcal{U}(2000000, 3000000)$.Additionally the cost of undertaking the attacks are a higher (at least double) as they operate as a business.

Table 6: Conditional probability table generated by simulating the cybercriminals problem. It displays the attack decision probabilities, conditional on the protection implemented by the defender.

Does the defender implement any technical security control?	Probability of cybercriminal deciding to launch targeted exfiltriation	Probability of cybercriminal deciding to launch targeted data manipulation	Probability of cybercrimianl deciding to launch targeted denial of service
Yes	51.6%	71.4%	70.5%
No	52.9%	87.9%	94.2%

obtaining the the expected utility of each of the portfolios from which we can obtain the best one and establish a ranking among them (Table 7).

Table 7: Five best portfolios of security controls and insurance based on maximising the Defender's expected utility.

Boundary firewalls and Internet gateways	Implement secure configuration	Implement access control	Malware protection	Implement patch and vulnerability management	Hazard protection	Insurance: property damage and personal injury	Insurance: data loss	Ranking based on maximising expected utility
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1
Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	2
Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	3
Yes	Yes	Yes	No	Yes	Yes	Yes	No	4
Yes	Yes	No	No	Yes	Yes	Yes	No	5

5.6 Sensitivity analysis

As discussed in [28] we could perform sensitivity analysis to check the robustness of the proposed cybersecurity portfolio by observing the impact of perturbations on the output of the analysis.

6 ADDITIONAL USED

We describe additional computational problems that may be based on our models and are very useful for cyber insurance providers.

6.1 Cyber insurance product design

A major decision for insurance companies is to determine the price and coverage that we put to a cyber insurance product. [29] provide a description of such decision, which is frequently driven by the market,in the sense that companies do not want to deviate too much from what competitors do.

We start with the pricing decision. Recalling (3), we rewrite the optimisation problem for the company as

$$\max_{x\in R} \int u_D(x, c(ii), \theta) p_D(\theta|x) d\theta,$$

where $R = [a, b]$ is the relevant price range, as suggested by the prices of the competitors, and $x = (sec, rec, ii)$ and we make now explicit the dependence of the decision through the utility including the insurance price. For a given insurance product that we are interested in selling to the specific organisation considered, we would consider the maximum price $c(ii)$ for which the company would include such insurance product within its optimal portfolio. We could undertake this computation by searching in a grid, starting from the right extreme b, checking whether for such price the optimal decision includes the insurance product and, if so, move to the next cheaper price in the grid, until it is not included.

For a given price, we could perform a similar approach to design the coverage of the insurance product, looking for the minimum coverage given the price that the company would be willing to accept to include the product in the portfolio. The approach would be similar, but now we make explicit the dependence $p_D(\theta|x, p(ii))$ of the impacts on the coverage.

Finally, we could explore Pareto efficient insurance products by combining the above procedures simultaneously dealing with price and coverage.

6.2 Market segmentation

Another important problem refers to market segmentation. By this we mean determining clusters of organisations, defined by their features ft and risk aversion coefficient ρ, which would choose similar cybersecurity portfolios. This would facilitate marketing operations, automate computations and partly relieve from the need to perform very expensive simulation-optimisation problems.

This may be undertaken as follows. For a set of organisations characterised by $\{ft_i, \rho_i\}_{i=1}^m$, we would compute with our system their respective optimal decisions x_i. Based on such data, we learn the parameters $\hat{w}$ of a model $x_i \approx \psi(ft, \rho, w)$ and use $\psi(ft, \rho, \hat{w})$ to forecast the decision to be made by companies. We could implement the above approach for the whole portfolio x or for parts of it, in particular, the cyber insurance component in them.

7 DISCUSSION

We have provided a conceptualisation of the cybersecurity risk management problem leading to a model and a DSS that supports its implementation. Key elements are the consideration of adversarial threats, the inclusion of cyber insurance within the security portfolio and the use of parametrised models for beliefs and preferences. With this system, mainly we may support an organisation seeking advice on which portfolio of security controls and insurance to implement. The current version of the system just illustrates a few of the threats, impacts and attacker types. However, it is open for future extensions. We mentioned above the various validations undertaken. In particular, our tests showed that the potential users of the CYBECO toolbox use it as an information source to make such a decision in a better informed manner.

Acknowledgements Work supported by the EU's Horizon 2020 project 740920 CYBECO (Supporting Cyber insurance from a Behavioural Choice Perspective). The work of DRI is supported by the Spanish Ministry of Economy and Innovation program MTM 2017-86875-C3-1-R and the AXA-ICMAT Chair on Adversarial Risk Analysis.

References

[1] Accenture. 2017. 2017 Cost of Cyber Crime Study.

[2] Banks, D., Ríos, J. and Ríos Insua, D. 2015. *Adversarial Risk Analysis.* Francis and Taylor.

[3] Clemen, R., Reilly, T. 2013. *Making Hard Decisions.* Cengage Learning.

[4] Cloud Security Alliance. 2016. *Cloud Controls Matrix.*

[5] The Common Criteria Recognition Agreement Members. 2009. *Common Criteria for Information Technology Security Evaluation, Version 3.1 Release 4.*

[6] Couce, A., Rios Insua, D. 2019. Cybersecurity Preference Models: The Defender Case. Technical Report [Annex of CYBECO Deliverable 3.2].

[7] Cox, L. A. 2008. "What's Wrong with Risk Matrices?". *Risk Analysis*, Vol. 28, No. 2, pp. 497–512.

[8] The CYBECO Consortium. 2018 *Deliverable 3.1 – Modelling Framework for Cyber Risk Management.*

[9] The CYBECO Consortium. 2019 Deliverable 3.2 –Improved Modelling Framework for Cyber Risk Management.

[10] The CYBECO Consortium. 2018 *Deliverable 4.1 –Cyber-insurance Use Cases and Scenarios.*

[11] The CYBECO Consortium. 2018 *Deliverable 4.2 –Use Case Evaluation of the Methodology and Framework.*

[12] The CYBECO Consortium. 2018 *Deliverable 5.4 –CYBECO Prototype 2.0.*

[13] The CYBECO Consortium. 2018 *Deliverable 6.3 –Report with Findings of Experiments and Policy Implications.*

[14] Ekin, T., Naveiro, R., Torres, A., Rios Insua, D. 2019. Augmented Simulation for Game Theoretic Problems. Technical Report [Annex of CYBECO Deliverable 3.2].

[15] Gelman, A., Meng, X.L. 1998. Simulating Normalizing Constants: from Importance Sampling to Bridge Sampling to Path Sampling, *Statistical Science*, pp. 163-185.

[16] Greenberg, A. 2018. "The Untold Story of notpetya. The Most Devastating Cyberattack in History." In *Wired*, August.

[17] Gobierno de España (Spain). 2018. *Real Decreto-Ley 12/2018, de 7 de septiembre, de Seguridad de las Redes y Sistemas de Información.*

[18] Information Security Forum. 2016. *Information Risk Assessment Methodology 2.*

[19] International Organization for Standardization. 2013. *ISO/IEC 27005. Information Security Risk Management.*

[20] International Organization for Standardization. 2013. *ISO/IEC 27001 – Information Security Management Systems - Requirements.*

[21] Lund, M.S., Solhaug, B. and Stølen, K. 2010. *Model-driven Risk Analysis: the CORAS Approach* Springer.

[22] Manworren, N., Letwat, J., and Daily, O. 2016. "Why You Should Care About the Target Data Breach." In *Business Horizons*, Vol. 59, No. 33, pp. 257–266.

[23] Mehdad, E., Kleijnen, J. 2018. *Stochastic Intrinsic Kriging for Simulation Metamodeling*, Vol. 34, pp. 322-337.

[24] National Institute of Standards and Technology (USA). 2018. *Cybersecurity Framework.*

[25] National Technical Authority for Information Assurance (UK). 2012. *HMG IA Standard Number 1.*

[26] Nelder, J. and Mead, R. 1965 "A Simplex Method for Function Minimization." *Comp. Journal*, Vol. 7, pp. 308-318.

[27] Ponemon Institute (USA). 2017. *2017 Cost of Data Breach Study. Global Overview.*

[28] Rios Insua, D., Couce-Vieira, A. Pieters, W., Labunets, K., Rubio, J.A., Rasines, D.G. 2019. "An Adversarial Risk Analysis Framework for Cybersecurity". *Risk Analysis* [accepted for publication].

[29] Romanosky, S., Ablon, L., Kuehn, A., Jones, T. 2018. "Content analysis of cyber insurance policies: how do carriers write policies and price cyber risk?". *Rand Co.*.

[30] Shachter, R.D. 1986. "Evaluating influence diagrams". *Operations Research* Vol. 34, No. 6, pp 871–882.

[31] Such, J.M., Vidler, J., Seabrook, T., Rashid, A. 2015. *Cyber Security Controls Effectiveness: A Qualitative Assessment of Cyber Essentials.* Technical Report SCC-2015-02, Security Lancaster, Lancaster University.

[32] Thomas, P., Bratvold, R. B. & Bickel, J. E. (2014). The Risk of Using Risk Matrices. In Society of Petroleum Engineers (Orgs.) Annual Technical Conference and Exhibition, Sept 30–Oct 02 2013, New Orleans.

[33] Trend Micro. 2015. *Understanding targeted attacks. What is a targeted attack.*

[34] United Kingdom Government. 2014. *Cyber Essentials Scheme: Requirements for Basic Technical Protection from Cyber Attacks*

[35] Woods, D. and Simpson, A. 2017. "Policy Measures and Cyber Insurance: a Framework, Journal of Cyber Policy, Vol. 2, No. 2, pp 209–226.

[36] World Economic Forum. 2018. *The Global Risk Report.*

Annex 2: Paper: Cybersecurity preference models. The defender case.

Cybersecurity preference models. The defender case

Abstract

We have introduced a model for cybersecurity risk management which facilitates an organisation to decide which countermeasures to adopt, possibly including a cyberinsurance. To facilitate its implementation and the development of the corresponding decision support system, we introduce a generic model for the preferences and risk attitudes of the organisation undertaking the analysis regarding the cybersecurity objectives.

1 Introduction

At present, all kinds of organisations are critically impacted by cyber threats, from private corporations to governmental facilities, going through critical infrastructures (Andress and Winterfeld, 2013). Risk analysis is a fundamental tool to help managing such problems (Cooke and Bedford, 2001). With it, organisations can assess the risks affecting their assets and what security controls should be implemented to reduce the likelihood of such threats or their impacts, in case they take place.

We have proposed a rigorous framework for risk management in cybersecurity (Rios Insua et al., 2019) which overcomes several of the defects in current standards by modelling cyber risks in more detail and including adversarial threats and insurance. The framework emphasises adversarial aspects for better prediction of threats, mitigates lack of data through structured expert judgement techniques and includes cyber insurance within the risk management portfolio. Although we included a template case study, its implementation may entail extensive work within large organisations. Towards the aim of facilitating a decision support system to aid in implementing our cybersecurity risk analysis framework, we propose here a general cybersecurity preference model for the organisation, the ICT owners undertaking a cybersecurity risk analysis, which we call Defender. Specifically, our aim is to provide:

- A generic tree of potential cybersecurity objectives for ICT owners. We describe potential attributes corresponding to each objective. Its purpose is to support the identification of all potential impacts of cybersecurity incidents in terms of relevant stakeholders' assets.

- A description of relevant models to forecast the outcomes in the involved attributes, when a threat takes place.

- A generic multi-attribute utility function to translate the previous objectives in quantified assessments of stakeholders' preferences and risk attitudes.

Our approach is inspired by earlier work in counter-terrorism, homeland security, aviation safety risk management and cybersecurity financial risk management. Specifically, Keeney (2007) identifies and structures preferences in antiterrorism analysis from the perspective of a government and we aim at a similar purpose in the cybersecurity domain from the perspective of a general organisation; Keeney and von Winterfeldt (2011) provided a value model to assess homeland security decisions and we pursue a similar aim to assess cybersecurity decisions, both in private and public administrations; Rios Insua et al. (2019) provide a value model for aviation safety at a state agency, including models to forecast and assess the impacts in aviation safety events, and our purpose is for cybersecurity events in a general organisation; Eling and Wirfs (2019) provide models for some of the cybersecurity financial costs and we complement them with other financial and nonfinancial impacts.

The rest of the document is structured as follows. We first provide a generic objective tree for cybersecurity risk management; ideally this would be shown to cybersecurity risk managers who would pick the relevant objectives for their problem at hand or, alternatively, use it to complete their own cybersecurity objectives tree. For each of the objectives, we provide potential attributes that measure or estimate objective achievement. We also describe several forecasting models for cybersecurity objectives, with a focus on cases in which there is little available incident data. Once the objectives and their attributes are specified, we can build a utility function to build up the model. We provide a generic model for this purpose, in which its parameters are obtained through a series of questions that would need to be addressed by the cybersecurity risk manager. We include an illustrative example based on the case study in our template.

2 Cybersecurity risk management objectives

Cybersecurity occurrences may entail very negative consequences in terms of costs, loss of reputation or even, in some cases, casualties. We track them through performance measures that we want to optimise, which we designate objectives. Through risk management, we aim at implementing treatments, possibly including a cyber insurance, to perform optimally with respect to such objectives, which will depend on the incumbent organisation. They will typically vary from state organisation to private ones and, among these, will differ for, say, a standard small enterprise, an Information and Communication Technology (ICT) based small enterprise, a medium enterprise or a large company. They may also vary in different countries and domains (e.g., air traffic management, healthcare, manufacturing). With each objective we associate, at least, one attribute with which to assess it.

We present here a generic list of objectives from which an organisation may choose when undertaking their cyber risk management process. The context of our problem is an organisation that aims at introducing a cyber risk management strategy, including possibly a cyberinsurance, to improve cybersecurity. Our framing for this problem may be seen in detail in Rios Insua et al. (2019).

2.1 General concepts

As Brownlow and Watson (1987) points out, structuring objectives in trees can help a risk manager (RM) overcome the cognitive overload brought by the volume of information which needs to be integrated into the solution of large, complex issues as in cybersecurity risk management. An analyst can work with RMs to build such objective hierarchy or tree in several ways, as detailed in Keeney (1992) and Clemen and Reilly (2013), including a brainstorming process or a questionnaire to identify the relevant objectives. There are several requirements that these must meet if they are to be useful for decision support (Keeney and Gregory, 2005): *comprehensive*, covering the whole range of relevant consequences for the incumbent organisation; *measurable*, either objectively or subjectively; *non-overlapping*, two objectives should not measure similar impacts; *relevant*, in the sense of being capable of distinguishing between alternatives; *unambiguous,* having a clear relationship between impacts and their description using the corresponding objective; *understandable,* the objective should be presented so that a reader reasonably familiar with risk or business can easily comprehend it.

The lowest tree nodes provide a series of dimensions, say q of them, which may be used to describe the consequences of alternatives, cybersecurity policies in our case, and uncertain scenarios. Each of these objective scales may be quantified in an *attribute*, allowing each consequence to be represented as a vector of attribute levels $c = (c_1, c_2, \dots, c_q)$. We distinguish three types of scales.

- A *natural attribute* gives a direct measure of the objective involved and is universally understood. An example of this, typical with cyber incidents to SMEs, are costs in relation with ICT support services that repair, reinstall or recover desktop computers and measured in EUROs.

- *Constructed attributes* are created for a specific decision context and are not universally understood. They are based on an artificially built ordinal scale, say 1 to 10. For example, in the case of image loss, level 1 could be associated with a case of minimal impact, e.g., a cyber attack with no loss of image even internally at the organisation. Level 10 could be associated with a maximum impact accident and a full compromise of the information assets of the organisation, like the exposure of thousands of private or personal information about customers, with appearance in global media. Henceforth, we would associate each of the levels with a qualitative description of severity with respect to image.

- *Proxy attributes* are used because of its perceived relationship with the objective, when no natural attributes are available and constructed scales are deemed too ambiguous. Variations in a proxy attribute are perceived to correlate with the issue of concern. For example, in online businesses the proxy attribute *website downtime* usually correlates with lost online sales.

2.2 Initial objectives

A popular approach to describing cybersecurity objectives is in terms of the information security attributes of confidentiality, integrity and availability (Mowbray, 2013). However, such objectives may be difficult to interpret from a business perspective: they are useful for expressing security from an information security perspective, in which ICT systems are described in terms of sets of pieces of information that are stored, processed or transmitted. We can think of this as the technical perspective through which we can express cyber risk. Yet the business perspective focuses more on assets and activities relevant for the organisation and its stakeholders; this is even more relevant if we reflect the general principles introduced above: objectives should cover the consequences over relevant organisational assets and activities expressed in variables directly used, or understandable, in the language of the incumbent organisation.

Some cybersecurity frameworks provide catalogues of concepts analogue to our objectives, mostly those addressing business impact analysis in cybersecurity, including ETSI GS ISI 002 v1.2.1 (2015), ISO 22317 (2015)[1], OWASP business impacts (2017), the OECD cyber losses types (2017), the ENISA Information Package for SMEs (2007), the ENISA report on ICT business continuity management for SMEs (2010), SABSA (2009) or MAGERIT (2012). We also include the list of impacts identified in Hubbard and Seiersen (2016) and the CYBECO deliverable on defining cyber insurance scenarios (Musaraj et al., 2018). They depict a few general categories of impacts (legal and regulatory, productivity, financial reputation and loss of customers) with some examples or subcategories. However, they do not meet well the requirements earlier established. Most of them provide a list of recurrent or important business impacts rather than a comprehensive list encompassing less typical impacts (e.g., cyber attack physical impacts). Similarly, they provide types of objectives or impacts that somehow overlap: most of them affect monetary objectives and, thus, some categorisation among them is recommended. For instance, some costs affect specific assets (e.g., activity interruption), whereas others are more general (e.g., competitive advantage, reputation).

Of course, creating a comprehensive and non-overlapping set of objectives may have disadvantages, namely, the addition of more concepts. One example in business terms is that income generation is a clear and main objective for companies to make money through sales. However, companies have alternative means to earn money, which may be even more relevant in other types of organisations such as NGOs, including grants, investments or licenses. A second example refers to the emerging and potential impacts of cyber risks involving physical and psychological aspects. Thus, third-party impacts such as health and environment should be also taken into account.

As a consequence, our approach is to list objectives and impacts in cybersecurity and sort them in a hierarchy of objectives in a more comprehensive, measurable, non-overlapping, relevant, unambiguous and understandable manner. As mentioned, comprehensiveness and non-overlapping involve, mostly, careful addition of novel concepts. Relevance and understandability are more related with translating cybersecurity impacts from the confidentiality, integrity and availability realms to another one focused on assets, activities and stakeholders.

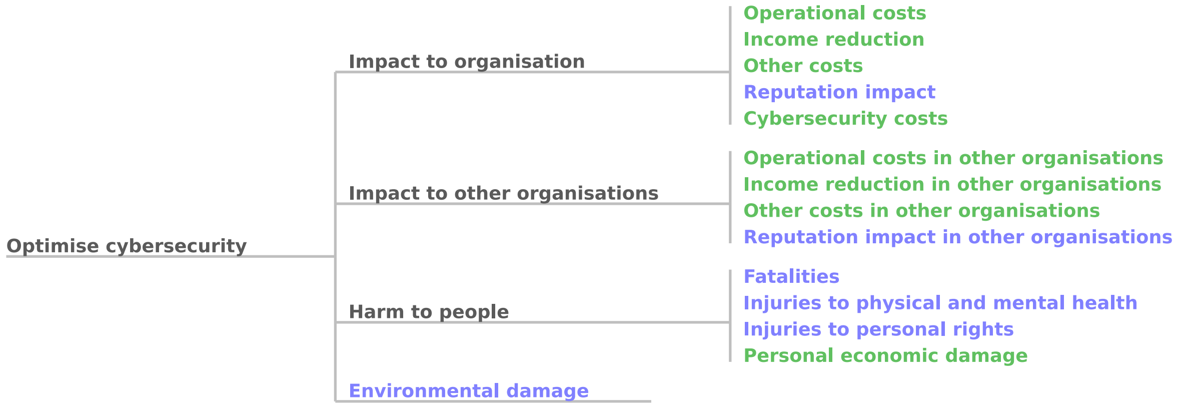
Besides the existing lists of cybersecurity impacts, the main influences on ours come from asset management and law. The first discipline, ISO 55000 (2014) on asset management in general or ISO 19770 for ICT assets (2015), helps us in conceptualising the different status that an asset could attain, so that engineers could characterise how an asset affects a system or the organisation in terms of reliability and predictability. The

[1] Standards in the ISO 22300 family are the continuation of BS 25999 (2007), one of the most popular standards in business continuity management.

second influence comes from law, in particular, the distinction between damages on property (economic or pecuniary) and persons (general or non-pecuniary). This facilitates the distinction between objectives that can be measured or evaluated in monetary terms (directly or through estimation) and others that are non-monetary and, thus, need special consideration when it comes to their evaluation. It also helps with the distinction between the objectives' owners (e.g., health and environmental damages are suffered by third parties besides the monetary, legal or reputational consequences that such damages could cause to the organisation).

We thus have developed a generic tree of cybersecurity objectives for a generic organisation, summarised in Fig. 1 and aimed at reaching the properties mentioned in Section 2.1. When it comes to comprehensiveness, we have evaluated existing categories of impacts to, at least, have categories that cover them. Solving the overlapping problem would mean creating more abstract concepts. We think that this question should be actually addressed when performing the actual risk assessment. Should a risk involve impacts on several categories, it would be necessary to check that impacts included in one category are not included in different ones. We have also tried to bring more general terms for the objectives rather than more domain-specific (e.g., organisation instead of business). This may add a little more ambiguity and less understandability compared to domain-specific IT or business categories, but it provides a more comprehensive approach.

Figure 1: Cybersecurity objectives. Green, assessed in monetary terms; blue, not directly measurable in monetary terms (e.g., health, environmental); grey, with both types of sub-objectives.


Optimise cybersecurity
Impact to organisation
Operational costs
Income reduction
Other costs
Reputation impact
Cybersecurity costs
Impact to other organisations
Operational costs in other organisations
Income reduction in other organisations
Other costs in other organisations
Reputation impact in other organisations
Harm to people
Fatalities
Injuries to physical and mental health
Injuries to personal rights
Personal economic damage
Environmental damage

The rest of this section describes in some detail the rationale behind such objectives. Note that some of them refer to impacts that may last several years and, for instance, those measured in monetary terms should be dealt with net present values (NPV) (French and Ríos Insua, 2000). We also provide an appendix that maps some of the previously mentioned catalogues to these cybersecurity objectives. As expressed in Figure 1, all objectives refer to minimisation, for example when mentioning *impact to the organisation* we understand *minimising impact to organisation*. Finally, unless explicitly mentioned, the objectives will be expressed in monetary terms.

2.2.1 Impact to the organisation

This objective consists of the following sub-objectives:

- **Operational costs**. We refer to those costs related with the assets and activities involved in the organisation's operations, i.e., the area responsible for producing goods or delivering services, the cost of degradation, malfunction, abuse, unavailability, elimination, recovery and unrecoverability of their assets and activities. We focus on assets such as software, ICT devices, documents, and equipment; and activities such as serving food, delivering merchandise, writing a report, or supporting administrative acts with citizens. All of these impacts can be represented with a monetary attribute. We include:
  - Degradation if the asset or activity performs its function in a less productive or more costly manner, e.g., a text processor running slower than normal as an asset degradation, or slower document production as an activity degradation.
  - Malfunction if the asset or activity has disturbances or a hazardous behaviour, e.g., a text processor producing errors when writing several pages.
  - Abuse if the asset or activity is maliciously manipulated, e.g., a malicious macro exfiltrating the document edited in the text processor.
  - Unavailability of the asset or activity, e.g., the employees cannot run the text processor.
  - Recovery as the actions and resources to restore an asset or activity to a normal situation. Note though that some assets might be unrecoverable (e.g. a piece of art) and this might have an operational impact (e.g., uninstallation of a text processor with several macros tailored to the business that cannot be reprogrammed because a programmer left the company).
- **Income reduction**: Impacts that reduce the income obtained by the organisation. In synthesis, minimizing loss of sales, contracts, market share, funding or licenses. In a business context, they typically involve marketing and commercial aspects related to sales. However, we also have to take into account that some income does not necessarily come from sales, e.g. in public and non-profit organisations. All of them can be assessed in monetary terms. We include:
  - Income reduction over sales flow, involving sales but also leads, quotes, post-sale and customer service.
  - Loss of market share, which can be expressed through the reduction over the sales flow. However, it might also be considered as an asset with an estimated economic value that can drop if market share is reduced.
  - In some cases, when the contracts are few but big, loss of contracts might be a more practical indicator than sales and market share.
  - Loss of funding not directly related with sales flow, e.g., through investments, grants or public funding.
  - Loss of licenses. It has a compliance origin but their loss could reduce income.

- **Other costs**: These refer to other impacts that affect an organisation. It includes some strategic, compliance and financial assets or potential costs. Although their identification or estimation might be difficult, all of them may translate into income (e.g., technological advantages) or costs (e.g., less advertisement for a well-known brand). All of them can be represented through monetary attributes. We include:
  - Loss of competitive advantage caused by leaked, spied, or publicly disclosed sensitive information, including intellectual property or commercial secrets. Although it could be correlated with income reduction or reputation impact, it is also considered an intangible but defined asset that can be estimated Raggio and Leone (2019) or sold.
  - Depreciation, abuse, unavailability or elimination of financial assets. Examples are changes in stock value, financial blackmail, extortion or ransom, theft of financial assets, including money or financial instruments.
  - Costs from non-compliance with contracts, regulations, standards or any other enforceable policy. Examples are fines and regulatory penalties, contractual and agreement penalties and litigation costs.
- **Reputation impact**: We refer to impacts over reputation that affect the trustworthiness of the organisation as an institution, rather than those more directly measurable in monetary terms that impact brand value, reduce income or operations or the activities towards recovering the reputation. In principle, these impacts cannot be represented with monetary attributes.
- **Cybersecurity costs**: It is practical to separate the costs related with managing cybersecurity, since this is the activity we aim to support in our decision-making model (Rios Insua et al., 2019). It covers the costs of preventive and reactive controls as well as the eventual cyber insurance. All of them can be represented in monetary terms.

2.2.2 Impact to other organisations

A cybersecurity incident in our organisation might affect other organisations and, thus, the organisation objectives also involve minimising damage to them. It replicates the objectives for our organisation except for minimisation of cybersecurity costs, since we are not supporting their cybersecurity decision-making. Therefore, it consists of the following sub-objectives: **Operational costs to other organisations**; **Income reduction to other organisations**; **Other costs to other organisations**; **Reputation impact in other organisations** (non-monetary).

2.2.3 Harm to people

A cybersecurity incident might also affect people such as employees, customers, or local communities. Therefore, the organisation objectives could also involve minimising harm to people. Some of the sub-objectives proposed in Figure 1 entail impacts which have been very rare, so far, in cybersecurity. For example, cyber attacks with physical impact are unusual but the emergence of industrial systems and smart infrastructures brings these risks to the fore, recall, e.g, Stuxnet. We include: **Fatalities** (non-monetary);

Physical and/or mental health injuries (non-monetary); **Injuries to personal rights** (non-monetary), e.g., dignity or privacy; **Personal economic damage**.

2.2.4 Environmental damage

Similar to damage inflicted to people, the environment might be affected by cyber attacks against systems with physical operations. Here we model the impact over the natural environment as such (e.g., the costs of cleaning contamination are an impact to organisations or people).

2.3 Attributes for quantifying the non-monetary objectives

We have identified several objectives that were not measurable in monetary terms. We describe here how we may proceed for each of them:

1. We could start with the identification of the main scenarios that a cybersecurity incident could cause. These what-if scenarios should be comprehensive in terms of covering all feasible types of impacts, related to the objective that the relevant stakeholders, assets and activities of the organisation may suffer if attacked.

2. Once these scenarios are identified, they should be quantified for their use in the model, following the approach depicted in Sect. 2.1 based on natural, constructed or proxy attributes.

2.3.1 Impact on reputation

Hubbard and Selersen (2016) discuss how to assess reputational damage in cybersecurity. The authors did not find strong evidence linking data breaches and stock prices of an attacked company, but observed that a relevant cyberattack cost is related to control the damage to limit reputation effects. As mentioned, this objective may impact brand value, reduce income or service or recovery costs from a public image perspective. However, reputation also encompasses aspects related with trustworthiness, legitimacy and image.

In the organisational theory literature, several authors apply an overall measure of reputation (Fombrun, 2012) whereas others use an attribute-specific measure (Jensen et al., 2012; Greenwood et al., 2005), since organisations may have multiple types of reputations. Carpenter and Krause (2011) provide four categories: *moral reputation* referring to how the organisation treats stakeholders; *procedural*, related to the extent the organisation follows legal and social norms; *performative*, referring to the capability of the organisation for performing their job; and, finally, *technical* related to the capability of the organisation for dealing with complex environments different from their business as usual status. We can use the same four categories with changes in names to facilitate understanding in the context of our model: moral, compliance, performative and adaptability reputations.

Common ways of measuring or building attributes for concepts like reputation are interviews with representatives of stakeholder groups or surveying a representative sample of such groups (van Riel and Fombrun, 2007). Indeed, measuring reputation is meaningful when it is done for specific groups of stakeholders (Jensen et al., 2012; Greenwood et al., 2005), relative to a competitor or a similar organisation (Fombrun, 2012) or past reputation performance (Jensen and Roy, 2008).

If we proceed with the constructed-attribute approach, we should first identify the scenarios taking into account the previously mentioned components (e.g., what type of reputation? with respect to which stakeholders?). Once these scenarios are identified, they should be ordered from the most to the least preferred. Table 1 provides a simple example of different reputation situations for a particular organisation ranked from best to worst.

Table 1: Example of reputational impact scenarios constructed scale.

Rank	**Impact on reputation**
1	No impact
2	Loss of moral or compliance reputation in up to 10% of employees, customers or the general public.
3	Loss of performative or adaptability reputation in more than 25% of customers or general public.
4	Loss of moral or compliance reputation in up to 50% of employees, customers or the general public.
5	Loss of moral and compliance reputation in more than 50% of employees, customers or the general public.

Alternatively, if we proceed with a proxy attribute, we could focus on the salience of cybersecurity incidents in news, media and social networks or the cost of handling the reputation impact of the incident.

2.3.2 Harm to people: Fatalities and injuries to physical and mental health

Cybersecurity incidents may pose a physical risk and, thus, triggering incidents that may affect people's health. Indeed, they are a major concern in medical devices (Fu and Blum, 2013), industrial control systems (Macaulay and Singer, 2011) or self-driving vehicles (Taeihagh and Lim, 2018). Additionally, mental health might be a relevant issue too, for instance, in relation to cyber bulling (Vandebosch and van Cleemput, 2008).

Our first sub-objective, minimising fatalities, could be assessed with a natural attribute such as the number of fatalities, as long as we do not distinguish between different types of victims. For the others, for example, the *WHO*[2] *International Classification of Diseases* (World Health Organization, 2018) provides a list with all types of injuries, diseases and disorders and, together with the object or substance producing them, the place of occurrence and the activity when injured. These classifications provide thousands of events or injuries. However, in a real case, our assessment will be more straightforward. Usually, the physical risks of cybersecurity would be a new causing or facilitating event of an already existing safety risk that, most of the time, has been documented by the organisation through industrial or occupational assessments.

Risk analysis typically distinguish between major and minor injuries. We could use them as the two natural attributes. They are also suitable for a constructed-attribute approach. There are several methods that may help us to create an ordinal scale (Hasler

[2] World Health Organisation

et al., 2012), such as the *Injury Severity Score* to assess the severity of injuries or the *Global Assessment of Functioning* or the *WHO Disability Assessment Schedule* (Ustün et al., 2010) for physical or mental functioning. Table 2 provides an example of different levels of mental and physical impacts, based on some of the previous scoring systems but excluding those scores related to fatalities.

Table 2: Example of physical and mental impact scenarios constructed scale.

Rank	**Injuries to physical and mental health**
1	No injury, emergency or functional impairment.
2	Minor emergency that does not require medical intervention (NACA I); or minor injury ($4 > \text{ISS} > 0$); or absent or minimal psychological or physical symptoms, no more than everyday problems or concerns (GAF 81-90).
3	Slight to moderate non life-threatening emergency that requires medical intervention (NACA II and III); or moderate or serious injury ($16 > \text{ISS} >= 4$); or mild and moderate psychological or physical symptoms, causing slight to moderate impairment in social or occupational functioning (GAF 51-80).
4	Serious emergency that may be life-threatening and which requires medical care (NACA IV-VI); or severe to maximal (currently untreatable) injury ($\text{ISS} >= 16$); or serious psychological or physical symptoms or persistent danger causing serious to persistent inability in several areas of functioning including family, mood, relations, thinking or even danger of hurting self or others (GAF 1-50).

Alternatively, we could use the number of people entering into hospital in relation with the cybersecurity event as a proxy attribute.

2.3.3 Harm to people: Injuries to personal rights

Cyber attacks may harm our dignity or privacy, accidentally or intentionally. Furthermore, large scale activities of governments or companies on the Internet have become a major issue on this topic, such as the US NSA surveillance (Margulies, 2013), the Great firewall of China (Lee and Liu, 2012) or the scandal of Cambridge Analytica (Kurtz et al., 2018). In this context, governments and international institutions are pushing for a more secure and governable cyberspace. Namely, the UN Human Rights Council has stated that “the same rights that people have offline must also be protected online” (United Nations Human Rights Council, 2015). See also the recent GDPR (REF) in Europe.

These rights could be identified from national jurisprudence, but the collection of by UN provides an international and overreaching framework. For our purposes it might be useful the classification system in the *Universal Human Rights Index Database*(United Nations Human Rights Council, 2016) which covers the human rights recognised by UN under categories such as civil and political rights; economic, social and cultural rights; or rights to specific persons or groups.

The constructed-attribute approach may be the best one to operationalise this subobjective. However, the nature of these rights, hardly commensurable, and their relatively big number makes this task demanding. One approach could be creating a hierarchy inspired on (Maslow, 1943) pyramid of needs. Most criticisms of this hierarchy focus on its last two categories (esteem and self-actualization); for instance, differences between individuals and societies on what constitutes esteem and self-actualization or even whether they consider the latter more basic than the former. Based on that, Table 3 provides an example of different impact levels over personal rights, using our modification of Maslow's pyramid.

Table 3: Example of personal rights impact scenarios constructed scale.

Rank	Injuries to personal rights
1	No personal right violation
2	Violation of personal rights that may affect esteem and self-actualisation needs.
3	Violation of personal rights that may affect social belonging needs.
4	Violation of personal rights that may affect safety needs.
5	Violation of personal rights that may affect physiological needs, including safety needs that also affect physiological needs.

Alternatively, we could use as a proxy attribute the number of legal actions against the organisation due to personal rights violations or the number of personal identifiable information records exposed.

2.3.4 Environmental damage

As in subsection 2.3.2, cybersecurity incidents may trigger incidents with environmental impact. Indeed, there is a high number of potential environmental risks. We have two relevant types of classifications: focused on the environmental impact of normal operations and on the environmental impact of incidents. For instance, the European eco-management and audit scheme (EMAS) (European Commission, 2017) or the British environmental key performance indicators (Department for Environment, Food and Rural Affairs (UK), 2006) provide suggestions to assess the environmental impact of normal activities such as land use, energy efficiency or emissions to air. In our case, these might be useful to identify impact scenarios in which the environmental performance of the organisation is disrupted by a cyber incident. Additionally, frameworks like the Irish (Environmental Protection Agency (Ireland), 2010) and British Common Incident Classification Scheme (CICS) (Environment Agency (UK), 2006) provide frameworks to identify environmental incidents such as the preservation of natural sites or habitats or contamination of water. They also provide severity scores that might be helpful for building a constructed scale for this objective. Note though that they include impacts that we classify in other sections such as human health or agricultural losses.

Based on the British frameworks (Department for Environment, Food and Rural Affairs (UK), 2006), (Environment Agency (UK), 2006), we can build a constructed attribute for the environmental impacts. Table 4 provides a simple example of different environmental impacts based on these two frameworks.

Alternatively, the quantitative nature of environmental performance indicators might serve us to use them as proxy attributes. For example, we could employ the variation in percentage of the most affected environmental indicator.

Table 4: Example of environmental damage constructed scale.

Rank	Environmental damage
1	No environmental impact.
2	Disturbance in the environmental performance indicators of the organisation.
3	Limited environmental damage, corresponding to CICS category 3 incidents.
4	Significant environmental damage, corresponding to CICS category 2 incidents.
5	Major environmental damage, corresponding to CICS category 1 incidents.

2.4 Summary

Table 5 summarises the cyber security risk management objectives and attributes that we include in our study.

Table 5: Summary of objectives and attributes.

Objective	Natural attribute	Constructed attribute	Proxy attribute
Min. operational costs Min. income reduction Min. other costs Min. operational costs in other orgs. Min. income reduction in other orgs. Min. other costs in other orgs. Min. personal economic damage	Monetary units		
Min. reputation impact Min. reputation impact in other orgs.		Yes	Media salience Public relations cost
Min. fatalities	Number of fatalities		
Min. injuries to physical and mental health	No.injured people	Yes	No.people in hospital
Min. injuries to personal rights		Yes	Num. of legal actions against the organisation Num. of personal identifiable information records exposed
Min. environmental damage		Yes	Percentage of variation in environmental indicator

3 Impact forecasting models

Besides assessing the consequences, we need models to forecast them. Some of the above, especially those referring to monetary consequences, may be dealt with relatively standard actuarial and financial models, as described in Eling and Wirfs (2019). However, we face a problem of lack of data, as companies seem reluctant to reveal details when they are attacked. Thus, we shall typically need to rely on structured expert judgement techniques (Dias et al., 2018). In particular, we could use several experts possibly combined through a weighted additive combination with weights depending on the experts' performance, e.g. based on Cooke's classical model (Cooke, 1991). We illustrate forecasting the non-monetary attributes of interest.

3.1 Service unavailability

We focus first on service unavailability given an attack. This is a relevant component to forecast income reduction and operational costs as well as, given its relevance in supply chain risks, in the same objectives in relation with third parties.

Under reasonable assumptions and given its flexibility in modelling various shapes, we model the duration of the downtime i_s through a gamma distribution $f(i_s|\mu_1, \mu_2) \sim Ga(\mu_1, \mu_2)$. The objective would be to obtain good estimates of μ_1 and μ_2. Given the lack of data, we aim at obtaining them through expert judgement. For this, we may ask the experts for the first p_{25} and third p_{75} quartiles of the i_s distribution and infer the parameters by solving

$$\min_{\mu_1, \mu_2} \left\{ (p_{25} - \text{cdf}(.25,\, \mu_1,\, \mu_2))^2 + (p_{75} - \text{cdf}(.75,\, \mu_1,\, \mu_2))^2 \right\}, \tag{1}$$

where cdf$(\cdot)$ designates the cumulative distribution function of the gamma distribution with parameters μ_1 and μ_2. After obtaining the corresponding parameters μ_1^* and μ_2^*, if necessary we could approximate the downtime e.g. through the expected value of the distribution $Ga(\mu_1^*, \mu_2^*)$, which is $\frac{\mu_1^*}{\mu_2^*}$. We would perform consistency checks based on other quantiles. As service unavailability data become available, we could incorporate such information with the data updating the distribution in a Bayesian fashion.

For aggregation purposes, we would typically multiply the downtime duration by the estimated expected cost of each downtime unit.

3.2 Reputation impact

As discussed in Section 2.3.1, there is no natural attribute that allows us to assess this impact. Our focus will be therefore in its consequence, the loss in market share in the customer induced by the attack over the organisation. We designate it by l_s, the proportion of customers abandoning to a competitor. The following considerations can be made: if the organisation is dominant in such service, the loss in market share would be negligible, and we shall assume that $l_s = 0$; if, on the contrary, there are alternative service suppliers, the market loss could be non-negligible, essentially depending on the reputation loss: the bigger it is, the bigger l_s will be. Under reasonable assumptions, we assume that l_s follows a beta distribution with parameters α and β. We would perform one assessment for each level of l_s, say those in Table 1. We would proceed in a similar fashion to Section 3.1 by asking two quartiles to experts and, subsequently, approximating the parameters, based on a least squares cdf approximation, after appropriate consistency checks. Finally, the expected predictive proportion of customers lost could be approximated through $\widehat{l_s} = \frac{\alpha}{\alpha+\beta}$. Again, we could introduce schemes to learn about α and β trough Bayesian updating when data is actually available. All in all, the loss could be quantified as $\widehat{l_s} \times k \times n \times c$, where k is the current market share, n is the market size and c is the income produced per customer in the relevant risk management period, which would be available from in house accounting experts.

A similar approach could be adopted for the reputational impact to third parties.

3.3 Fatalities

The approach adopted considers that a cyberattack triggers an incident which causes the failure of a cyber-physical system possibly leading to fatalities. As a consequence, in a given application we just need to forecast the number of fatalities associated with an accident in the corresponding system, besides the probability that the event triggers the accident. As described by Palali and de Jong (2015), there are three general types of safety models: sequence of discrete events (e.g., fault trees, failure modes and effect analysis or bow ties), epidemiological with latent and active failures (e.g., Human Factors Analysis and Classification System), and systemic for understanding the structure and behaviour of a system (e.g., System Theoretical Accident Model and Processes or the System Theoretical Process Analysis). Other relevant formal models include multivariate unobserved component models (Carnero and Pedregal, 2010).

We could adapt the fatality forecasting model for aviation safety accidents in Rios Insua et al. (2019). Thus, the context is that a cyber attack induces an accident in an airplane, causing some fatalities. We predict the number n_F of fatalities in an accident triggered by a cyber attack with a model $n_F = p_F \cdot q \cdot M$, where p_F designates the proportion of fatalities; q, the installation occupancy degree; and, finally, M, its maximum occupancy. The parameters would depend on the type of installation.

For the proportion p_F, we may use a mixture model $p_F \sim \tau_1 I_0 + \tau_2 \mathcal{B}e(a,b) + \tau_3 I_1$, where τ_1 designates the proportion of accidents with no fatalities; τ_2, the proportion of accidents with both fatalities and survivors; and, finally, τ_3, the proportion of accidents with no survivors, with $\tau_1 + \tau_2 + \tau_3 = 1$, $\tau_i \geq 0, i = 1,2,3$. I_0 is the degenerate distribution at 0 (no occupant dies); $\mathcal{B}e(a,b)$ models the distribution of the proportion of fatalities in accidents when there are fatalities and survivors; and, finally, I_1 is the degenerate distribution at 1 (all occupants die). A priori, $(\tau_1, \tau_2, \tau_3) \sim \mathcal{D}ir(a_1, a_2, a_3)$, $p_F \sim \mathcal{B}e(a,b)$. For the occupancy proportion q, the prior distribution is $q \sim \mathcal{B}e(c,d)$. We would assess all these parameters with expert judgement as in the earlier cases.

In presence of data, we make inferences about the weights τ_i with a Dirichlet-multinomial model; about p_F, when $0 < p_F < 1$, with a Beta-binomial model; and, about the occupancy proportion with a Beta-binomial model.

As an alternative, we could use an expert judgement based approach using a Poisson distribution by asking two quantiles much as we did in Section 3.1.

Finally, we could use the concept of statistical value of life (Viscusi and Aldy, 2003) to monetise the fatalities.

3.4 Injuries

Some cyber attacks might produce injuries. As in Section 2.3.2, we distinguish between minor and major injuries. We consider three proportions p_{h_i}, $i = 1,2,3$ for the three types of survivors ($i = 1$, minor injured; $i = 2$, major injured; $i = 3$, uninjured), following a model

$$p_H = (p_{h_1}, p_{h_2}, p_{h_3}) \sim \alpha \cdot I(0,0,1) + (1-\alpha) \cdot \mathcal{D}ir(h_1, h_2, h_3),$$

where α designates the proportion of occurrences in which none is injured and $I(0,0,1)$ is the degenerate distribution in which there are no injuries. Initially, we may assume $\alpha \sim \mathcal{B}e\,(a,b)$ prior.

As data becomes available, we would make inference about the weight α with a Beta-binomial model; about the proportions of injured occupants, with a Dirichlet-multinomial model. The initial estimation of q and p_F would be made as in Section 3.3.

Then, the number $n_H = (n_{h_1}, n_{h_2}, n_{h_3})$ of injuries for an occurrence is predicted with a model

$$n_H = p_H \cdot q \cdot (1 - p_F) \cdot M,$$

where p_H designates the proportions of injuries and p_F, q and M as in Section 3.3.

We could use the concept of statistical value of an injured person to monetise the injuries, EUROCONTROL (2013).

3.5 Personal rights

There are several models in the literature related to forecasting incidents that can impact personal rights, such as the chance of cybersecurity breaches (Liu et al., 2015; Sarabi et al., 2016) or some proposal regarding litigation forecasting (Brown et al., 2004).

We could focus on forecasting the number n_p of personal identifiable information records exposed in a cyberattack. For that we could use a model

$$n_p = q_p N_p,$$

where q_p is the proportion of exposed records and N_p is the maximum number of records. A priori we could use a beta model for q_p and introduce a beta-binomial model to learn about it as data accumulates. We could segment the models for q_p, e.g., depending on the economic sector considered or other organisational features.

Similar to the minimization formula in (1), we would infer the distribution parameter through expert judgement by asking .

3.6 Impact over the natural environment

Forecasting environmental impacts depend on the specifics of the sector and territory under assessment (e.g., the European Food Safety Authority models European Food Safety Authority (2017)). A general method that help forecasting environmental impacts is life cycle assessments (LCA) (Hellweg and Milá i Canals, 2014). As in the case with fatalities, we could view the cyber attack as triggering event which would start an environmental accident and then use an environmental risk assessment model.

4 Utility model

Section 2 identified a comprehensive list of cybersecurity objectives. From it, the incumbent organisation could choose the objectives relevant in its problem. Then, we need a procedure to model preferences over such impacts, as we do now through a utility function. A brief review on utility functions may be seen in Ortega et al. (2018). We use the classic concepts of measurable multi-attribute value function (Dyer and Sarin, 1979) and relative risk aversion (Dyer and Sarin, 1982).

The approach that we adopt, as it is not overly demanding cognitively, is relatively general in its assumptions and is easy to assess in practice is as follows. Under sufficiently general conditions (Ortega et al., 2018), the utility must have the following structure:

1. $u(c) = 1 - \exp(-\rho \sum v_i(c_i)), \quad \rho > 0.$
2. $u(c) = \sum v_i(c_i).$
3. $u(c) = 1 + \exp(\rho \sum v_i(c_i)), \quad \rho > 0.$

where ρ is the risk aversion coefficient and the v_i's are measurable value functions.

We discuss now how to assess the parameter ρ, facilitating scaling the utility to $[0, 1]$. In our case, the relevant attributes may be viewed as costs, which are decreasing. We make $c = -d$, to make the attribute increasing. The minimum cost is 0 and suppose the maximum cost is c^*. The utility function has to be strategically equivalent to

$$u(d) = 1 - e^{-\rho d} = 1 - e^{\rho c}.$$

This means that its form should be

$$u(c) = a(1 - e^{\rho c}) + b.$$

We make

$$u(0) = 1,\ \ a(1 - e^{\rho 0}) + b = 1 \Rightarrow b = 1$$
$$u(c^*) = 0,\ \ a(1 - e^{\rho c^*}) + 1 = 0 \Rightarrow a(1 - e^{\rho c^*}) = -1.$$

We need one more judgement for a certain cost, which we fix at $c = c^*/2$. We use, for example, the probability equivalent method (Farquhar, 1984). In order to acheve so, we ask the cyber risk manager to provide the probability p such that she finds equally interesting the lotteries

$$\begin{pmatrix} 1 \\ c \end{pmatrix} \sim \begin{pmatrix} 1-p & p \\ c^* & 0 \end{pmatrix}. \tag{2}$$

Then,

$$u(c) = (1 - p)u(c^*) + pu(0) = p,$$

and we have the system

$$\begin{cases} a(1 - e^{\rho c}) = p - 1, \\ a(1 - e^{\rho c^*}) = -1, \end{cases}$$

from which

$$\frac{1 - e^{\rho c}}{1 - e^{\rho c^*}} = \frac{p - 1}{-1}.$$

This leads to

$$e^{\rho c} + (p - 1)e^{\rho c^*} - p = 0.$$

Taking $x = e^{\rho c}$, we have

$$(p - 1)x^2 + x - p = 0, \tag{3}$$

whose solution is $x = \frac{1 \pm \sqrt{1 - 4p(1-p)}}{2(1-p)}$. We then make

$$\rho = \ln x/c,$$

and

$$a = \frac{1}{x^2 - 1}.$$

4.1 Utility model in the CYBECO system

In the CYBECO project, we have identified several scenarios relevant for cybersecurity (Musaraj et al., 2018) and integrated within the CYBECO Toolbox (The CYBECO Consortium, 2018). We have identified and synthesised a number of impacts of cybersecurity incidents, mapped in Table 6 against the cybersecurity objectives in Section 2. All impacts are linked with monetary costs, except the loss of personal records, which is linked with injuries to personal rights. Assuming a risk averse organisation, then if we apply the utility function defined in the previous section, we use

Table 6: CYBECO impacts mapped onto our objectives.

Impacts in CYBECO Toolbox	Cybersecurity objective
Facilities: damages to physical properties	Min. operational costs
IT infrastructure: business downtime	Min. operational costs
Market share: percentage lost	Min. income reduction
Personal information: records lost	Min. injuries to personal rights
Personal information: privacy and security liability lost	Min. other costs
Customers: loss of customers due to brand reputation and damage	Min. income reduction
Production: interruption of provided services or produts	Min. income reduction
Contractual and regulatory losses	Min. other costs
Recovery and other post-incident expenses	Min. operational costs Min. cybersecurity costs

$$u(m,r) = 1 - \exp\Big(-\rho\Big(v_m(m) + v_r(r)\Big)\Big)$$

where m is the monetary impact, r is the impact on personal rights and $v_m(m)$ and $v_r(r)$ are their corresponding value functions. To operationalise this function, we could use the quantitative attributes that measure such subobjectives, so that the utility function can be described as

$$u(m,r) = 1 - \exp\Big(\rho\Big(m + c_r r\Big)\Big)$$

The first one, m, is measured through a natural attribute (monetary units) that we shall express in €. Note that this also includes the security costs of security controls and insurance, since they are related to the objective *Min. cybersecurity costs*.

The second one, r, is measured with a proxy attribute (records exposed), associated with the parameter c_r. To elicit this parameter, we should provide an economic value to privacy. The legal costs of injuries to personal rights are part of the monetary costs. However, there is no solid estimations for the *value of privacy* (Acquisti et al., 2013). Estimations based on British (Godel et al., 2017) and American (Hann et al., 2007) customers reveal that consumers' value of their personal information is up to £7.25 and $44.62 respectively. Assuming that they are risk neutral and they assign a probability of less than one percent to a data exposure then taking the more conservative British figure (equivalent to €8.25), we shall use that at least they value their personal information at €825. Risk aversion would reduce this figure slightly, whereas the American figures or a lower perception of the likelihood would increase it (e.g., more than €4.000 with the American figures or €1.650 if we assume a probability of breach of less than 0,5 percent). Therefore, we use €825 as a conservative estimate of the economic value of privacy per record.

Then, the utility function that we shall be using is strategically equivalent to

$$u(m, r) = 1 - \exp\Big(\rho\Big(m + 825r\Big)\Big)$$

To adjust it, we determine the worst reasonable cost $c_* = m_* + 825r_*$, where m_* is the sum of the maximum cost of the impacts and the security budget and r_* is the maximum number of records that can be exfiltrated. Suppose that for a certain organisation, m_* is estimated at €2.000.000 and r_* is estimated at 5000, so that the worst cost is €6.125.000. We also determine the best cost, which is $c^* = 0$, for $m^* = r^* = 0$. Further suppose that, for $c_1 = \frac{1}{2}c_*$, we obtain $u(c_1) = 0.8$, through the probability equivalent method. We then obtain that the only valid root in 3 is $x = 4$, so that $a = 1/15 = 0.066$, $\rho = 4.5267 * 10^{-7}$ and $b = 1$, and the utility function is

$$u(m, r) = 0.066 * \Big(1 - \exp\Big(4.5267 * 10^{-7}\Big(m + 825r\Big)\Big)\Big) + 1.$$

5 Discussion

In earlier work, we have presented an adversarial risk analysis (ARA) framework for cybersecurity risk management that provides a formal method supporting all relevant steps when undertaking a comprehensive cybersecurity risk analysis. To facilitate its implementation in case studies, as well as to develop a decision support system facilitating its implementation, we have introduced a generic objective tree for cybersecurity risk management from the Defender erspective, with the corresponding objectives and attributes, some ideas on the pertinent forecasting models and a generic preference model, illustrated with specific examples. From it, a cybersecurity risk manager could choose the relevant objectives to proceed in a risk analysis, formulate his preference model by responding a few simple questions and have an orientation on the forecasting models to be implemented, facilitating his analysis.

In future work, we shall undertake a similar approach for cybersecurity attackers, using the concept of random utility functions.

Acknowledgements

Work supported by the EU's Horizon 2020 project 740920 CYBECO (Supporting Cyberinsurance from a Behavioural Choice Perspective). The work of DRI is supported by the Spanish Ministry of Economy and Innovation program MTM2014-56949-C3-1-R, MTM2017-86875-C3-1-R and RTC-2017-6593-7, and the AXA-ICMAT Chair on Adversarial Risk Analysis.

References

[1] A. Acquisti, K. J. Leslie, and G. Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013.

[2] J. Andress and S. Winterfeld. *Cyber Warfare: Techniques, Tactics and Tools for Security Practitioners.* Elsevier, 2013.

[3] British Standards Institution. BS 25999-2:2007 Specification for Business Continuity Management. 2007.

[4] S. Brown, S. A Hillegeist, and K. Lo. Management forecasts and litigation risk. *SSRN Electronic Journal*, 2004.

[5] S. Brownlow and S. Watson. Structuring multi-attribute value hierarchies. *Journal of the Operational Research Society*, 38(4):309–317, 1987.

[6] M. C Carnero and D. J. Pedregal. Modelling and forecasting occupational accidents of different severity levels in spain. *Reliability Engineering and System Safety*, 95 (11):1134–1141, 2010.

[7] D. P. Carpenter and G. A. Krause. Reputation and public administration. *Public Administration Review*, 72(1):26–32, 2011.

[8] R. T. Clemen and T. Reilly. *Making Hard Decisions with Decision Tools.* Cengage Learning, 2013.

[9] R. M Cooke. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press, 1991.

[10] R. M. Cooke and T. Bedford. *Probabilistic Risk Analysis: Foundations and Methods.* Cambridge University Press, 2001.

[11] Department for Environment, Food and Rural Affairs (UK). Environmental Key Performance Indicators Reporting Guidelines for UK Business. Technical report, 2006.

[12] L. C. Dias, A. Morton, and J. Quigley. Elicitation: State of the art and science. In L. C. Dias, A. Morton, and J. (eds) Quigley, editors, *Elicitation: The Science and Art of Structuring Judgement*, pages 1–14. Springer International Publishing, 2018.

[13] J. Dyer and R. Sarin. Group preference aggregation rules based on strength of preference. *Management Science*, 25(9):822–832, 1979.

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

[15] M. Eling and J. Wirfs. What are the actual costs of cyber risk events. *European Journal Oper. Research*, 272:1109–1119, 2019.

[16] Environment Agency (UK). Incidents and their Classification: the Common Incident Classification Scheme (CICS), Version 12. Technical report, 2006.

[17] Environmental Protection Agency (Ireland). Guidance to Licensees/COA Holders on the Notification, Management and Communication of Environmental Incidents. Technical report, 2010.

[18] European Commission. Commission Decision (EU) 2017/2285 of 6 December 2017 Amending the user's guide setting out the steps needed to participate in EMAS, under Regulation (EC) No 1221/2009 of the European Parliament and of the Council on the voluntary participation by organisations in a Community eco-management and audit scheme (EMAS). Technical report, 2017.

[19] European Food Safety Authority. EFSA Guidance Document for predicting environmental concentrations of active substances of plant protection products and transformation products of these active substances in soil. Technical report, 2017.

[20] European Network and Information Security Agency. Information Package for SMEs. Technical report, 2007.

[21] European Network and Information Security Agency. IT Business Continuity Management – An Approach for Small Medium Sized Organisationss. Technical report, 2010.

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

[23] C. J. Fombrun. The building blocks of corporate reputation: Definitions, antecedents, consequences. In *The Oxford Handbook of Corporate Reputation*, pages 94–113. Oxford University Press, 2012.

[24] S. French and D. Ríos Insua. *Statistical Decision Theory.* Wiley, 2000.

[25] K. Fu and J. Blum. Controlling for cybersecurity risks of medical device software. *Communications of the ACM*, 56(10):35–37, 2013.

[26] M. Godel, W. Landzaat, and J. Suter. Research and Analysis to Quantify the Benefits Arising from Personal Data Rights under the GDPR – Report to the Department for Culture, Media & Sport. Technical report, London Economics, may 2017.

[27] R. Greenwood, S. X. Li, R. Parkish, and D. L. Deephouse. Reputation, diversification, and organizational explanations of performance in professional service firms. *Organization Science*, 16(6):661–673, 2005.

[28] I.-h. Hann, H. Kai-Lung, T. L. Sang-Yong, and P. L. P. Ivan. Overcoming information privacy concerns: An information processing theory approach. *Journal of Management Information Systems*, 24:13–42, 2007.

[29] R. M. Hasler, C. Kehl, A. K. Exadaktylos, R. Albrecht, S. Dubler, R. Greif, and N. Urwyler. Accuracy of Prehospital Diagnosis and Triage of a Swiss Helicopter Emergency Medical Service. *Journal of Trauma and Acute Care Surgery*, 73(3): 709–715, 2012.

[30] S. Hellweg and L. Milá i Canals. Emerging approaches, challenges and opportunities in life cycle assessment. *Science*, 344(6188):1109–1113, 2014.

[31] D. W. Hubbard and R. Selersen. *How to Measure Anything in Cybersecurity Risk.* Wiley, 2016.

[32] Industry Specification Group. ETSI GS ISI 002 V1.1.1 Information Security Indicators (ISI); Event Model, A Security Event Classification Model and Taxonomy Annex B1.8 – With what kind of impact. Technical report, European Telecommunications Standards Institute, 2015.

[33] International Organisation for Standardization. ISO 55000:2014 – Asset Management – Overview, Principles and Terminology. 2014.

[34] International Organisation for Standardization. ISO 19770-5:2015 – IT Asset Management – Overview and Vocabulary – Part 5. 2015.

[35] International Organisation for Standardization. Societal Security – Business Continuity Management Systems – Guidelines for Business Impact Analysis. Technical report, 2015.

[36] M. Jensen and A. Roy. Staging exchange partner choices: When do status and reputation matter? *Academy of Management Journal*, 51(3):495–516, 2008.

[37] M. Jensen, H. Kim, and B. K. Kim. Meeting expectations: A role-theoretic perspective on reputation. In M. L. Barnett and T. G. Pollock, editors, *The Oxford Handbook of Corporate Reputation*, pages 140–159. Oxford University Press, 2012.

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

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

[40] R. L. Keeney. On the foundations of prescriptive decision analysis. In W. Edwards, editor, *Utility theories: Measurements and Applications*, pages 57–72. Springer, Dordrecht, 1992.

[41] R. L. Keeney and R. S. Gregory. Selecting attributes to measure the achievement of objectives. *Operations Research*, 53(1):1–11, 2005.

[42] C. Kurtz, M. Semmann, and W. Schulz. Towards a framework for information privacy in complex service ecosystems. In *39th Int. Conf. on Information Systems*, 2018.

[43] J. A. Lee and C. U. Liu. Forbidden city enclosed by the great firewall: The law and power of internet filtering in china. *Minnesota Journal of Law, Science & Technology*, 13(1):125–151, 2012.

[44] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *24th USENIX Security Symposium*, pages 1009–1024, 2015.

[45] T. Macaulay and B. L. Singer. *Cybersecurity for Industrial Control Systems: SCADA, DCS, PLC, HMI, and SIS.* Auerbach Publications, 2011.

[46] P. Margulies. The NSA in Global Perspective: Surveillance, Human Rights, and International Counterterrorism. *Fordham Law Review*, 82(5):2137–2167, 2013.

[47] A. H. Maslow. A theory of human motivation. *Psychological Review*, 50(4):370–96, 1943.

[48] Ministerio de Hacienda y Administraciones Públicas (Spain). Metodología de análisis y gestión de riesgos de los sistemas de información, version 3. Technical report, 2012.

[49] T. J. Mowbray. *Cybersecurity: Managing Systems, Conducting Testing, and Investigating Intrusions.* Wiley, 2013.

[50] K. Musaraj, M. Cousin, V. Melvin, C. Melvin, A. Couce, D. Rios Insua, J. Vila, W. Pieters, and K. Labunets. CYBECO Deliverable D4.1 Cyber insurance use cases and scenarios. Technical report, The Cybeco Consortium, 2018.

[51] OECD. *Enhancing the Role of Insurance in Cyber Risk Management.* OECD Publishing, Paris, 2017.

[52] J. Ortega, V. Radovic, and D. Rios Insua. Utility elicitation. In L. C. Dias, A. Morton, and J. (eds) Quigley, editors, *Elicitation: The Science and Art of Structuring Judgement*, pages 241–264. Springer International Publishing, 2018.

[53] B. Palali and J. de Jong. Evaluating contemporany safety views and models in incident/accident investigation reports. Master's thesis, Amsterdam University of Applied Sciences, 2015.

[54] R.D. Raggio and R.P. Leone. Drivers of brand value, estimation of brand value in practice and use of brand valuation: Introduciton to the special issue. *Journal of Brand Management*, 17(1):1–5, 2019.

[55] D. Rios Insua, C. Alfaro, J. Gomez, P. Hernandez-Coronado, and F. Bernal. Forecasting and assessing consequences of aviation safety occurrences. *Safety Science*, 111:243–252, 2019.

[56] D. Rios Insua, A. Couce-Vieira, J. A. Rubio, W. Pieters, K. Labunets, and D.G. Rasines. An adversarial risk analysis framework for cybersecurity. 2019.

[57] A. Sarabi, P. Naghizadeh, Y. Liu, and M. Liu. Risky business: Fine-grained data breach prediction using business profiles. *Journal of Cybersecurity*, 2(1):15–28, 2016.

[58] A. Taeihagh and H. S. M. Lim. Governing autonomous vehicles: Emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews*, 39(1):103–128, 2018.

[59] The CYBECO Consortium. CYBECO Deliverable D5.2 CYBECO content and data collection Manual. Technical report, 2018.

[60] The Open Web Application Security Project. The OWASP risk rating methodology. Technical report, 2017.

[61] The SABSA Institute. The SABSA White Paper. Technical report, 2009.

[62] United Nations Human Rights Council. Resolution on the promotion, protection and enjoyment of human rights on the internet. 2015.

[63] United Nations Human Rights Council. Universal human rights index database. 2016. URL http://uhri.ohchr.org/search/guide.

[64] T.B. Ustün, N. Kostanjsek, S. Chatterji, and J. Rehm. *Measuring Health and Disability: Manual for WHO Disability Assessment Schedule, WHODAS 2.0*. World Health Organization, Geneva, 2010.

[65] C. B. M van Riel and C. J. Fombrun. *Essentials of Corporate Communication*. Routledge, 2007.

[66] H. Vandebosch and K. van Cleemput. Defining cyberbullying: A qualitative research into the perceptions of youngsters. *CyberPsychology & Behavior*, 11(4): 499–503, 2008.

[67] World Health Organization. International statistical classification of diseases and related health problems, 11th revision. Technical report, 2018.

Appendix

Mapping of CYBECO D4.2 definitions to our cybersecurity objectives tree

Table 7: CYBECO D4.1 mapping to cybersecurity objectives tree

CYBECO D4.1(2018)	**Cybersecurity Objectives Tree**
Impact/consequence - loss of data and software	Operational costs
Impact/consequence - loss or damage to physical properties	Operational costs
Impact/consequence - product recall	Operational costs (in case of substitution) or Income reduction (in case of retrieval)
Impact/consequence - fraud	Other costs
Impact/consequence - theft of money, securities	Other costs
Impact/consequence - extortion	Other costs
Impact/consequence - privacy liability	Other costs
Impact/consequence - identity theft	Operational costs
Impact/consequence - failure to render service	Operational costs
Impact/consequence - security liability	Other cost
Impact/consequence - property damage	Other cost and Impact to other organisations
Impact/consequence - personal injury	Other costs and Injuries to physical and mental health
Impact/consequence - media liability	Operational costs
Impact/consequence - product liability	Operational costs
Impact/consequence - failure to supply	Operational costs
Impact/consequence - management liability	Other cost
Impact/consequence - breach of duty	Other cost
Impact/consequence - non-compliance with regulation	Other cost
Impact/consequence - brand and reputational damage	Other costs (if brand is treated as asset) and/or Reputatiol impact
Impact/consequence - non-compliance with regulation	Other costs
Impact/consequence - business interruption	Operational costs

Table 8: CYBECO D4.2 mapping to cybersecurity objectives tree (cont.)

CYBECO D4.2(2018)	**Cybersecurity Objectives Tree**
Potential loss - revenue loss (all except compensatory payments)	Income reduction
Potential loss - compensatory payments to costuemrs and suppliers	Other costs
Brand and reputation damage	Other costs (if brand is treated as asset) and/or Reputation impactt
Financial penalties	Other costs
Loss of competitivity and productivity	Operational costs
Collateral expenses	Operational costs

Mapping of MAGERIT valuation criteria to our cybersecurity objectives tree

Table 9: MAGERIT mapping to cybersecurity objectives tree

MAGERIT(2012)	**Cybersecurity Objectives Tree**
Personal information	Injuries to personal rights (impacts to persons), Other costs (impacts to organisation due to non-compliance regarding personal information) and Operational costs (information asset degradation).
Legal obligations	Other costs
Security	Cybersecurity costs
Commercial or economic interests	Income reduction or other costs (if strategic)
Service interruption	Operational costs
Public order	For most organisations is Impact to other organisations. For those organisations responsible for public order it might be necessary to create a new cybersecurity objective of of non-monetary nature for evaluating the potential states of public order: *Max. public order.*
Operations	Operational costs
Administration and management	Operational costs
Loss of confidence (reputation)	Reputation impact
Prosecution of crimes and law enforcement	For most organisations is min. impact to other organisations. For those organisations responsible for these tasks it is related with Operational costs
Service recovery time	Operational costs
Classified information	As a characteristic of information assets, Operational costs

Mapping of SABSA high-level general business attributes to our cybersecurity objectives tree

Table 10: SABSA mapping to cybersecurity objectives tree

SABSA(2009)	**Cybersecurity Objectives Tree**
Financial - Accounted	Other costs
Financial - AML compliant	Other costs
Financial - Auditable	Other costs
Financial - Benefit-evaluated	Income reduction.
Financial - Cash-flow forecasted	Income reduction
Financial - Credit controlled	Other costs
Financial - Credit risk managed	Other costs
Financial - Investment returnable	Other costs
Financial - Liquidity risk managed	Other costs
Financial - Market risk managed	Other costs (understood as financial market risks)
Financial - Profitable	Income reduction
Financial - Reporting compliant	Other costs
Physical (all attributes)	Operational costs. Note that some characteristics are related to security/risk characteristics of the assets (access controlled, damage protected, defended, secure, theft protected).
Human (all sub-attributes)	Characteristics related to human capital, which could be classified as an asset. Therefore, the related objective is Operational Costs
Process (all sub-attributes)	Other costs
Strategic - Administered	Other costs
Strategic - Branded	Other costs
Strategic - Communicated	Other costs
Strategic - Competitive	Other costs
Strategic - Compliant	Other costs
Strategic - Financed	Other costs
Strategic - Goal oriented	Other costs
Strategic - Governed	Other costs
Strategic - Logistically managed	Operational costs
Strategic - Market penetrated	Income reduction
Strategic - Market positioned	Income reduction
Strategic - Reputable	Reputation impact.
Strategic - Supply chain managed	Operational costs
System (all attributes)	Operational costs. Note that some characteristics are related to security/risk characteristics of the assets (access controlled, incident managed, risk managed).

Annex 3: Paper: Insider threat modeling: An adversarial risk analysis approach

Insider threat modeling: An adversarial risk analysis approach

Abstract

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

1 Introduction

Insider threats are encountered in many risk analysis areas including international security, geo-politics, business, and cyber security. They are not only widely perceived to be significant ([Schulze, 2018], [Ware, 2017]), but also often considered to be more damaging and more likely than outsider attacks ([Schulze, 2018], [CERT, 2012]). Moreover, it is feared that the impact of the insider threat problem actually known is only the tip of an iceberg as many organizations are choosing not to report such incidents unless required to do so by law ([Wood et al., 2016]): as described in [Hunker and Probst, 2009], it is a field in which little data is available, specially in the cyber security domain. Protection from insider threats is challenging as the perpetrators might have access to sensitive resources and privileged system accounts. Finally, solutions to insider threat problems are considered to be complex ([Lee and Rotoloni, 2015]): technical solutions do not suffice since insider threats are fundamentally a people issue, as thoroughly discussed in [Sarkar, 2010] and [Greitzer et al., 2012].

In its simplest form, it is natural to view the insider threat problem as a two player game. We may call the first player *the organization* (which could refer to a single business or military unit or a similar entity, but also to a whole country or a coalition of entities or countries) and the second one, *the employee* (which could refer to one or more employees, contractors, or persons who have significant access to the organization and have been trusted with such access). A typical scenario would be as follows: since insider threats are a well-known phenomena, it will frequently be the case that several measures would have already been implemented by the organization (at least, in case of sufficiently mature organizations) to prevent or deter an insider attack. As an example, [Silowash et al., 2012] provide a catalog of best practices against insider threats in cybersecurity. The employee will typically be aware of the measures in place and plans an attack accordingly. Once the attack has been carried out and detected, the organization will undertake actions to end the attack and mitigate any damage caused, possibly based on the resources deployed at the first stage. This type of interactions have been named sequential Defend-Attack-Defend games, see e.g. [Brown et al., 2006].

It is therefore natural that game-theoretic models of the insider threats phenomenon have been explored. For example, [Liu et al., 2008a] model the problem as a two-player, zero-sum dynamic game. At each discrete time point, both players make decisions resulting in a change of state and opposite (given the zero-sum property) rewards to them. The authors then look for *Nash equilibria* (NE). This model is oversimplified in several respects. For example, there could be multiple attackers, the attacker pay-offs might not be immediate to obtain and the game might not be zero-sum. Also, in most cases, the defender would have already employed measures to prevent an insider attack and, therefore, the problem should be modeled as a sequential Defend-Attack-Defend game instead and not as a simultaneous one.

A more realistic approach is described in [Kantzavelou and Katsikas, 2010] who consider an insider threat problem in cybersecurity, trying to model the continuous interactions between an intruder and an intrusion detection system (IDS). They assume bounded rationality on them, use quantal response equilibria instead of the standard NE and assign pay-offs through utilities to assess the outcomes. However, their model focuses on a particular application and is not immediately generalizable. Moreover, the game does not consider multiple players and carries on even after detecting an attack as the detection causes the attack to be stopped, but does not eliminate the attacker from the game. [Tang et al., 2011] also model insider threats to IT systems considering bounded rationality and combine game theory with an information fusion algorithm to improve upon traditional IDS based methods by being able to consider various types of information. [Feng et al., 2015] and [Hu et al., 2015] propose three player games to model the use of Advanced Persistent Threats (APT) by a malicious insider. They employ a two layer game and show the existence of NE.

While game theory has been the typical choice to model interactions between two or more strategic adversaries, limitations of such theory, e.g. [Gintis, 2009], [Camerer, 2003], or [Raiffa et al., 2002], have long been pointed out, focusing on common knowledge assumption and the conservative nature of its solutions. Limitations of conventional risk analysis in security have been pointed out as well; [Cox, 2009] and [Brown and Cox, 2011] warn that it is inappropriate to model, say, terrorist actions in the same way as hurricanes. Therefore, in this paper, we shall propose adversarial risk analysis (ARA), [Insua et al., 2009], approaches to insider threats. ARA does not assume common knowledge and solves the problem from the point of view of just one of the players - typically, the defender, taking into account their knowledge and uncertainties regarding the choices available to them, to their adversaries, the possible outcomes, their payoffs/utilities and their opponents payoffs/utilities. Since its introduction, it has been used to model a variety of problems such as network routing for insurgency ([Wang and Banks, 2011]), international piracy ([Sevillano et al., 2012]), or autonomous social agents ([Esteban and Insua, 2014].

ARA takes into account the expected utilities for the defender as well as the random expected utilities for the opponents, incorporating uncertainty regarding the strategic reasoning of the opponents. However, an ARA solution to insider threats has not yet been developed.

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

2 A Defend-Attack-Defend model for the insider threat problem

We start with a Defend-Attack-Defend model to deal with the insider threat problem, which considers a defender D (the organization, she) and an agent A (the employee, he). Our model is based upon the graphical framework described in [Banks et al., 2015]. Figure 1 presents the problem using a bi-agent influence diagram (BAID) where decisions are represented by square nodes, uncertainties using circular nodes and utilities with hexagonal nodes. Nodes corresponding to D are not shaded; those corresponding to A are diagonally shaded; and, finally, the shared chance node S is shaded using horizontal dashed lines. Dashed arrows indicate that the involved decisions are made with the corresponding agent knowing the values of the preceding nodes, whereas solid arrows indicate probabilistic or value dependence of the corresponding node with respect to the predecessors.

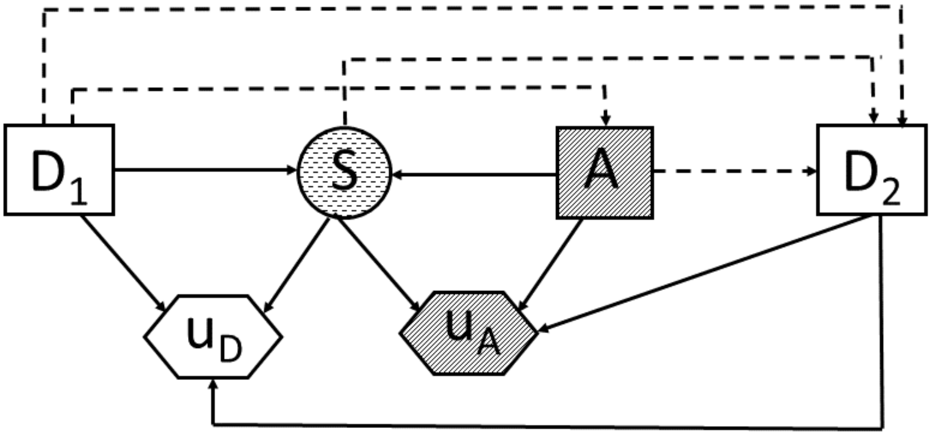


Figure 1: BAID for the Defend-Attack-Defend insider threat game

The action and outcome sets are as follows. Initially, the organization must choose one of the available preventive measures d_1 in the set $\mathcal{D}_1$. Having observed the preventive measure taken, the employee will adopt one of the actions a in $\mathcal{A}$; this set could consist of either 'no attack' or 'attack' or different types/intensities of attacks or other attack options. The set $\mathcal{S}$ consists of the possible outcomes s that can occur as a result of the preventive measure d_1 and the attack a adopted. Once the attack has been detected, the organization will choose to carry out one of the possible actions d_2 in the set $\mathcal{D}_2$ to end the attack, limit any damage and possibly pre-empt future attacks leading to the final outcomes of both agents, respectively, evaluated through their utility functions u_D and u_A. Note that all three sets $\mathcal{D}_1$, $\mathcal{A}$ and $\mathcal{D}_2$ could contain a *do nothing* action.

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

1. The distribution $p_D(a|d_1)$ modeling her beliefs about the attack a chosen at node A by the employee given the chosen defense d_1.

2. The distribution $p_D(s|d_1, a)$ modeling her beliefs about the outcome s of the attack, given a and d_1.

3. Her utility function $u_D(d_1, s, d_2)$ which evaluates the consequences associated with their first (d_1) and second (d_2) defensive actions as well as the outcome s of the attack.

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

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

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

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

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

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

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

The above analysis requires the defender to elicit $p_D(a|d_1)$. This can either be done using risk analysis based approaches such as [Ezell et al., 2010] or by modeling the strategic analysis process of the attacker. To parallel the attacker's analysis, the defender should assess the attacker's utility function $u_A(a, s, d_2)$ and probability distributions $p_A(s|a, d_1)$ and $p_A(d_2|d_1, a, s)$. However, since the corresponding judgments will not be available to the defender, we could model her uncertainty about them through a random utility function $U_A(a, s, d_2)$ and random probability distributions $P_A(s|a, d_1)$ and $P_A(d_2|d_1, a, s)$. Once these random quantities are elicited, the defender solves the attacker's decision problem using backward induction. This is done by following a process similar to how they solved their own decision problem but taking into account the randomness in judgments. First, the defender finds the random expected utility for each $d_2 \in \mathcal{D}_2$

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

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

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

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

$$A^*(d_1) = \arg\max_{a \in \mathcal{A}} \boldsymbol{\Psi}_{\boldsymbol{A}}(d_1, a). \tag{6}$$

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

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

Note that, in the above analysis, we have assumed that all the involved quantities are continuous. Should some of the quantities be discrete, the corresponding integrals would be replaced by sums. Further, in Section 4, we illustrate how $P_D(a|d_1)$ can be approximated using Monte-Carlo methods.

3 An ARA model for the insider threat problem with segmented employees

The sequence of interactions between an organization and an employee could be more complex for various reasons. Firstly, it has been described ([Moore et al., 2015], [Liu et al., 2008b], [Martinez-Moyano et al., 2008]) that the measures in D_1 can have unintended negative consequences. If the employee feels that the measures introduced by the organization to mitigate insider threats are intrusive or micromanaging or even aggressive, that could lead him to react in unintended ways. This could include not reporting suspicious activities or misusing the reporting processes either accidentally or intentionally. At worst, it could even motivate an employee to go rogue. Secondly, although we have treated the group employee as a single entity, in reality, this group could typically include a large number of people and therefore, the organization may be faced with multiple actors taking multiple actions. Note that, usually, a majority of employees will not take any action that would harm the organization. In fact, some of them would actively help prevent an insider attack. For example, one of the possible insider actions in A could be to correctly follow the processes or measures set out by the organization possibly resulting in the successful prevention of the imminent attack altogether. Finally, the actions by employees could be dependent (sequential) or independent (simultaneous).

We shall focus on considering the issue of modeling different types of employees. [Liu et al., 2008b] provide a segmentation with inadvertent and malicious insiders. We shall classify the employees as A_1 (*the good*), A_2 (*the bad*) and A_3 (*the ugly*), with S_1, S_2 and S_3 being the corresponding outcome sets. Each group of employees generates a relevant game as shown in Figure 2. Specifically, we consider that:

- A_1 are the employees who correctly and promptly perform their duties including following any procedures to prevent insider attacks. They have a positive impact on the productivity and work culture of the organization and will correctly report any suspicious activity, thus helping the organization to protect itself. Therefore, their actions will be positive to it.

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

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

3.1 Sequential action

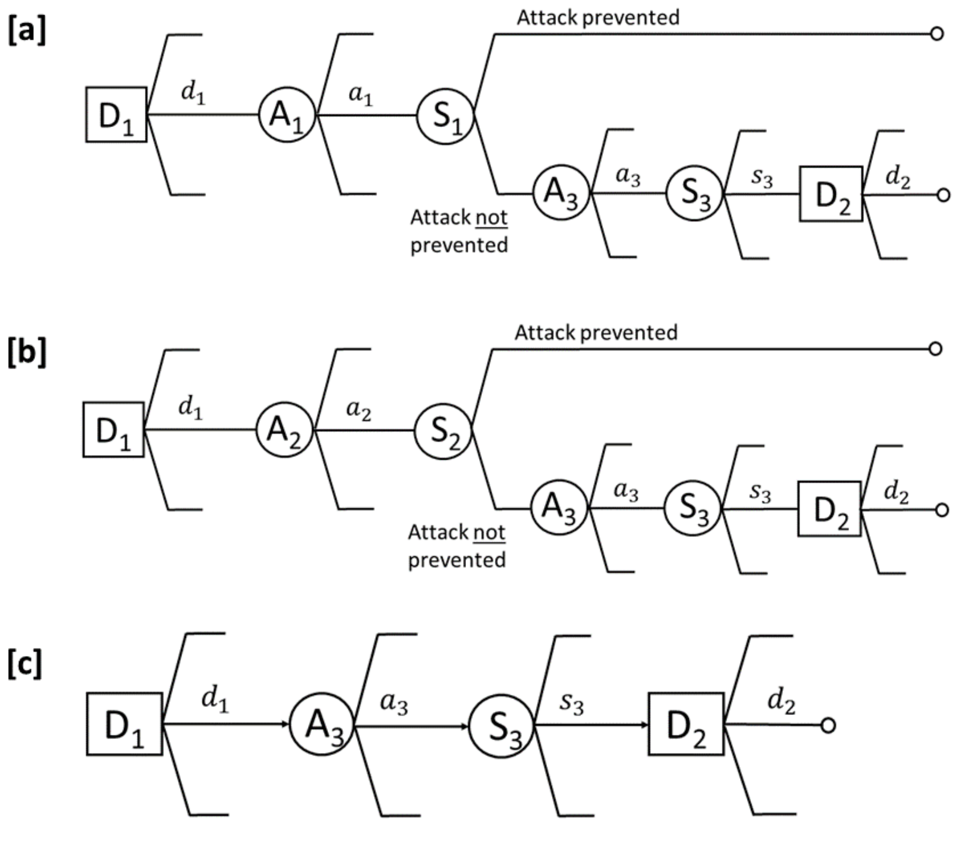


Figure 2: Decision trees for the three games in the insider threat problem with segmented employees.

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

The game represented in Figure 2[b] considers the role played by the 'bad' employee. Their action set $\mathcal{A}_2$ includes intentional or unintentional misuse of defensive procedures, possibly leading that suspicious activities are either not reported or reported through external/ unauthorized channels which, in turn, could cause significant harm to the organization. At worst, such a culture could even motivate an employee to launch an insider attack. The outcome set $\mathcal{S}_2$ consists of the same events as in $\mathcal{S}_1$. In case that the attack was detected/prevented, we assume that no further action was required from any of the players and hence the game ends. However, in the event that the attack was not detected/prevented, the 'ugly' employee will proceed with their chosen action $\mathcal{A}_3$ which could consist of an attack of a certain level resulting in the outcome set $\mathcal{S}_3$. Upon detection, the organization will take whatever actions $\mathcal{D}_2$ necessary to end the attack and contain any damage.

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

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

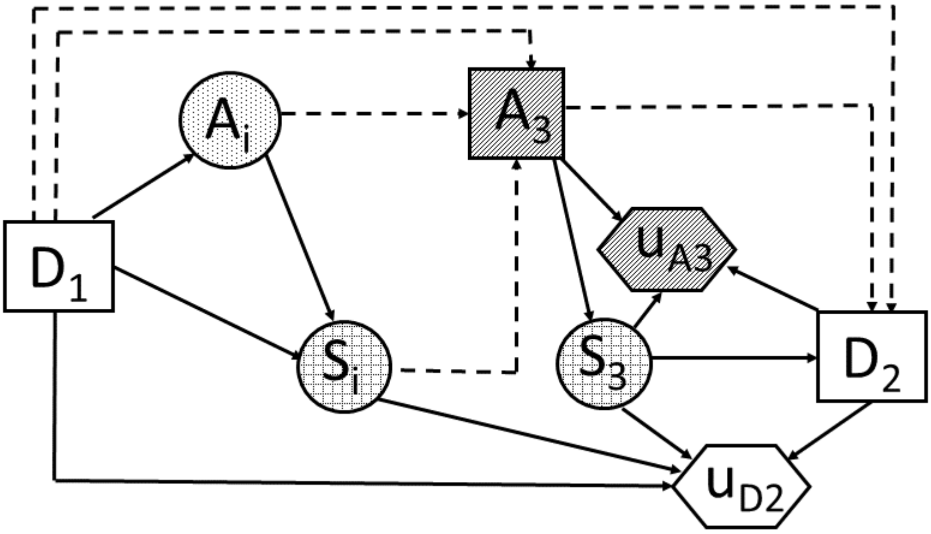


Figure 3: MAID for decision trees [a] and [b] in the segmented employees insider threat game

defender must first quantify the following.

1. Her predictive distribution $p_D(a_i|d_1)$ about the action that will be chosen at node A_i given the defense d_1.

2. Her predictive distribution $p_D(s_i|d_1, a_i)$ about the outcome of such action, given a_i and d_1.

3. Her predictive distribution $p_D(a_3|d_1, a_i, s_i)$ about the attack that will be chosen at note A_3 given the outcome s_i and actions a_i and d_1.

4. Her predictive distribution $p_D(s_3|d_1, a_i, s_i, a_3)$ about the outcome of the attack, given outcome s_i, actions a_3, a_i and d_1.

5. The utility function $u_D(d_1, a_i, s_i, a_3, s_3, d_2)$ given their first and second defensive actions, the outcomes of the attack s_3 and s_i and the actions a_3 and a_i.

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

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

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

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

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

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

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

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

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

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

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

The above analysis requires the defender to elicit $p_D(a_3|d_1, a_i, s_i)$ and $p_D(a_i|d_1)$. Of these, $p_D(a_i|d_1)$ refers to the actions by the *good* or *bad* employees, neither of whom intend to strategically harm the organization *per se*. Therefore, action A_i can be considered to be non-strategic. For this reason, A_i is represented as a random node in the MAID in Figure 3. Further, A_i being non-strategic, $p_D(a_i|d_1)$ can be elicited using historical data/research on employee behavior, where available. Eliciting $p_D(a_3|d_1, a_i, s_i)$ is, however, less straightforward. The defender could model the attacker's strategic analysis process by assuming that the attacker will perform an analysis similar to the defender to find their optimal action a_3^*. The attack A_3 will only go ahead if the outcome s_i has not resulted in it being prevented. Provided that attack A_3 can take place, while A_i and S_i may have an effect on the probability of an attack, we assume that the choice of an attack depends only on the defender action d_1, that is, $p_D(a_3|d_1, a_i, s_i) = p_D(a_3|d_1)$. To elicit it, the defender must assess $U_A(a_3, s_3, d_2)$, $P_A(s_3|a_3, d_1)$ and $P_A(d_2|d_1, a_3, s_3)$. These random utilities and distributions could be elicited in several ways, outlined in [Ríos Insua et al., 2019]. Once elicited, the defender solves the attacker's decision problem using backward induction - similar to how they solved their own decision problem. First, the defender finds the random expected utilities for each action $d_2 \in \mathcal{D}_2$

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

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

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

and, finally, compute the random optimal attack

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

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

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

Note that, in the above analysis, we have assumed that all the involved quantities are continuous. Should some of the quantities be discrete, the corresponding integrals would be replaced by sums. Further, in Section 4, we illustrate how $P_D(a|d_1)$ can be approximated using Monte-Carlo methods.

3.2 Simultaneous actions

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

Figure 4 is the MAID for a game in which both the *good* and the *bad* employees act simultaneously and after having observed D_1. Similar to Section 3.1, these actions are considered as non-strategic and hence are represented as a joint random node A_1A_2 in the MAID. These actions result in the random outcome S_{12}. The *ugly* employee observes these actions and the outcome before launching their attack. For this game, the ARA solution proceeds in an identical manner to the solution described in Section 3.1. The decision maker first seeks to find action $d_2^*(d_1, a_1, a_2, s_{12}, a_3, s_3)$ which will maximize their utility

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

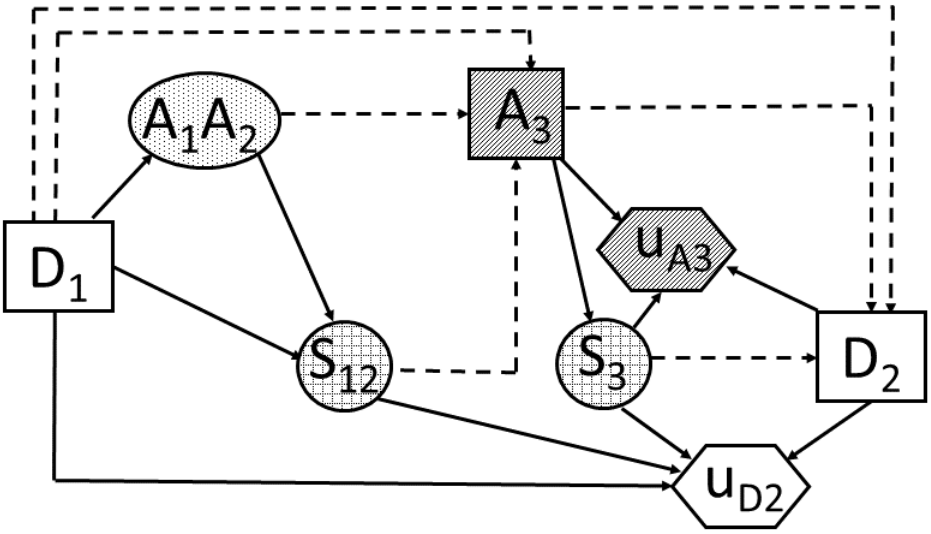


Figure 4: MAID for the game where A_1 and A_2 act simultaneously followed by A_3.

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

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

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

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

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

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

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

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

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

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

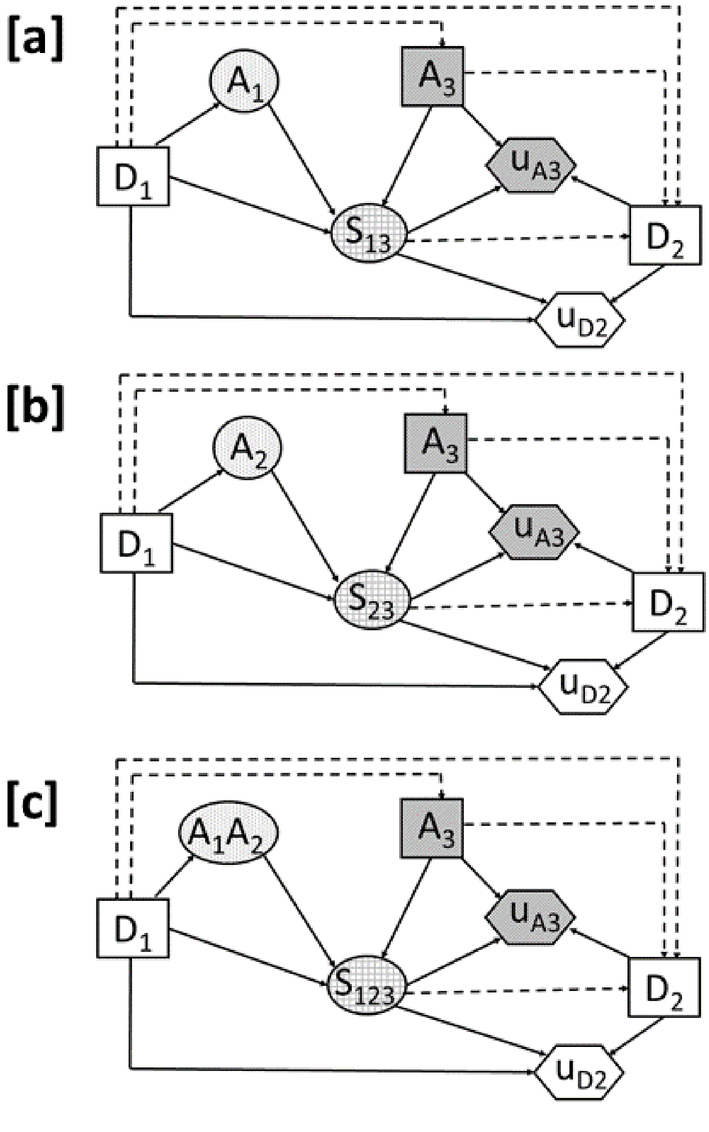


Figure 5: MAIDs for games where A_3 acts simultaneously with either A_1 or A_2 or both act simultaneously followed by A_3.

$p_D(a_3|d_1, a_1, a_2, s_{12})$ can be elicited by following Equations 13 to 16 after replacing a_i with (a_1, a_2) and s_i with s_{12}.

The MAIDs in Figure 5 correspond to extensions of the Defend-Attack-Defend game described in Section 2. In these games either two (a, b) or all three (c) types of employees act simultaneously after observing D_1 and the ARA solution proceeds in a similar manner to the solution described in Section 2. The main difference is that now the Defender must quantify the joint probabilities for the actions of two or all three types of employees concerned. For example, for the MAID in Figure 5 [a], the defender must quantify the distributions $p_D(a_1, a_3|d_1)$ and the distribution $p_D(s_{13}|d_1, a_1, a_3)$. If the actions of the *good* and *ugly* employees can be considered to be independent given d_1 then, $p_D(a_1, a_3|d_1) = p_D(a_1|d_1) \times p_D(a_3|d_1)$, where, $p_D(a_1|d_1)$ can be elicited using historical data and research on employee behavior (again, actions by the *good* and *bad* employees are considered as non-strategic and therefore represented by a chance node in the MAIDs) and $p_D(a_3|d_1)$ can be elicited by modeling the strategic analysis process of the *ugly* employee as detailed in Section 2.

3.3 Model uncertainty

We have seen how one can find an optimal action, that is, the ARA solution for the organization given a specific game/model. The expected utility $\psi_D(d_1)$ that we find in each of the models is, in fact, $\psi_D(d_1|M)$, where M refers to the game under consideration. In reality though, the exact scenario will be unknown. It will not be known if the *ugly* employee is able to act without being affected by the actions of the *good* and/or the *bad* employees and if so, whether such interaction is sequential or simultaneous. A Bayesian approach allows the organization to incorporate model uncertainty into the analysis and identify the expected utility taking model uncertainty into account, [Draper, 1995].

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

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

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

4 Example

We consider an insider threat scenario motivated by [Martinez-Moyano et al., 2008] in which the malicious insider attempts to harm the incumbent organization without getting caught. The organization focuses on information/data collection and needs to protect itself against both insider and outsider attacks. It already has its sites and IT systems protected so that only authorized personnel are able to access them. However, anticipating attacks, the organization is considering implementing an additional security layer to defend itself. The defensive actions (D_1) under consideration are

1. anomaly detection/data provenance tools;

2. information security measures and employee training; and

3. carrying out random audits.

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

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

4.1 Using the defend-attack-defend model

We first analyze the problem using the model in Section 2. We start by assessing the defender's utility function $u_D(d_1, a, s, d_2)$. We assume here that the defender's utilities depend not only on the outcome s (and d_1 and d_2) but also on the attack a. Indeed, we assume that u_D aggregates the monetary costs $c(d_1)$ and $c(d_2)$ associated with actions d_1 and d_2 respectively and the monetized perceived utilities associated with every (a, s) combination through

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

The costs and perceived utilities are listed in Tables 1 and 2. They are scaled from -100 to 100.

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

Table 1: Costs associated with defensive actions d_1 and d_2.

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

Table 2: Monetized perceived utility for every combination (a, s).

The global utility u_D can then be computed for each combination; for example, $u_D(d_1 = \text{Rand. aud.}, a = \text{Med. scale}, s = \text{Not success.}, d_2 = \text{No upgr.}) = -50+60+0 = 10$.

We next elicit the probabilities $p_D(s|d_1, a)$. Suppose that they are as listed in Table 3, with probabilities of failed attacks obtained through $p_D(\text{not successful } |d_1, a) = 1 - p_D(\text{successful } |d_1, a)$.

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

Table 3: Probabilities $p_D(S = \text{successful } |d_1, a)$ elicited for every (d_1, a).
.

In order to implement the ARA solution, the defender must first identify the action $d_2^*(d_1, s)$ maximizing their utility. In this case, d_2^* turns out to be 'No upgrade', being the cheapest option and will therefore maximize u_D. Then, they must compute the expected utility $\psi_D(d_1, a)$ using (2). The expected utility $\psi_D(d_1, a)$ can now be computed, for example, $\psi_D(\text{Random audits, Medium scale}) = -100 \times 0.4 + 10 \times 0.6 = -34$. The $\psi_D(d_1, a)$ values are given in Table 4.

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

Table 4: $\psi_D(d_1, a)$ for every combination of a and s.
.

Then, we need to compute the expected utility for each $d_1 \in D_1$ using (3) and the predictive distribution $p_D(a|d_1)$ about what the malicious insider may do. Assume first that the defender has elicited $p_D(a|d_1)$ using her own beliefs as in Table 5. The defender's expected utility $\psi_D(d_1)$ for each action d_1 is computed using Equation (3). For example, $\psi_D(\text{Random audits}) = -47.5 \times 0.5 - 34 \times 0.4 - 24 \times 0.1 = -39.75$. Similarly, the expected utility for *Anomaly detection and Data provenance* is -69.005, whereas for *Information security and training* it is -29. This implies that the optimal option for the organization is to invest in information security and staff training and, if the attack was to happen, then, irrespective of whether it was successful or not, the optimal follow-up would be not to upgrade their defenses.

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

Table 5: $p_D(a|d_1)$ elicited by defender using their beliefs
.

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

Table 6 lists the distributions elicited for $U_A(a, s, d_2)$ by the defender, with utilities between -100 and $+100$. We assume the defender thinks that the attacker believes that the defender has a short-sighted view and their utilities are a direct function of the costs involved in establishing the upgrades D_2. Whereas they will find it less valuable to upgrade their defensive mechanisms if the attack had, in fact, failed, no upgrade will on average be the least attractive option given that an attack was detected (whether successful or not).

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

Table 6: Distributions $U_A(a, s, d_2)$ elicited by defender.

Table 7 lists the distributions elicited for $P_A(d_2|d_1, a, s)$ by the defender, consistent with the utilities $U_A(a, s, d_2)$. For example, since an upgrade is considered to be less valuable in the event of a failed attack, the defender is unlikely to upgrade in the wake of a failed attack, reflected in the $Dir(1, 9, 90)$ distribution elicited for it. On the other hand, the $Dir(5, 4.9, 0.1)$ indicates that an upgrade is considered almost certainly likely in the wake of a successful large attack (irrespective of D_1).

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

Table 7: Distributions $P_A(d_2|d_1, a, s)$ elicited by defender for every combination (a, s, d_1). First row, D_1 = Anom. det. & Data prov; second, D_1 = Inf. sec.; third, D_1 = Random audit. For each Dirichlet distribution, $Dir(\alpha_1, \alpha_2, \alpha_3)$, α_1 relates with probability of major upgrade, α_2 to minor and α_3 to no upgrade.

Finally, Table 8 lists the distributions elicited for $P_A(\text{successful }|d_1, a)$ by the defender. For example, she believes that the attacker thinks that an attack is much more likely to succeed if D_1 is *Information security and training* compared to the other options. Also, they believe that the attacker thinks that a small attack is much more likely to succeed than a medium or large attack.

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

Table 8: $P_A(S = \text{successful }|d_1, a)$ elicited by defender for every combination (a, s).

.

For each combination of d_1, a and s, we simulate $N = 1000$ samples from $U_A(a, s, d_2)$ and $P_A(d_2|d_1, a, s)$ to obtain samples from the attacker's expected utility $\mathbf{\Psi_A}(d_1, a, s)$ using (4) and, then, a sample of the attacker's expected utility $\mathbf{\Psi_A}(d_1, a)$

using (5). Then, for each of the simulations, we find the optimal defense d_1 maximizing $\mathbf{\Psi_A}(d_1, a)$ and, finally, estimates $p_D(a|d_1)$ by counting how many times (out of N) would the attacker choose a particular attack given d_1. These are presented in Table 9. We can now use these estimates of $p_D(a|d_1)$ to compute the defender's

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

Table 9: $P_D(a|d_1)$ elicited by defender modeling the strategic analysis of the attacker

.

expected utility $\psi_D(d_1)$ using 3. The expected utility for the *Anomaly detection and Data provenance* comes out to be -36.115, for *Information security and training* be -39.051, and, finally, for the *Random audit* be -34.566. This implies that in this case, the optimal option for the organization is to invest in conducting random audits and, if the attack was to happen irrespective of whether it was successful or not, the optimal follow-up action would be not to upgrade their existing defenses.

Observe, therefore, that the $p_D(a|d_1)$ elicited by modeling the attacker's strategic analysis (Table 9) turns out to be quite different from that elicited using their own belief and knowledge (Table 5), leading to different optimal decisions.

4.2 Using the segmented employees model

We now analyze this problem using the model discussed in Section 3 by assuming three types of employees. Again, we start by assessing the defender's utility function $u_D(d_1, a_i, s_i, a_3, s_3, d_2)$. Just like we did with the model in Section 4.1, we assume that u_D adopts the form

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

where $c(d_1)$ and $c(d_2)$ are as defined in Table 1 and $u(a_3, s_3)$ is the same as $u(a, s)$ defined in Table 2. To define $u(a_i, s_i)$, we first define the values that A_i and S_i can take for $i = 1, 2$. As described in Section 3, the outcome sets are $\mathcal{S}_i = \{$attack prevented, not prevented$\}$. In reality, the set $\mathcal{A}_1$ could consist of various actions that a good employee can take, for example, $\mathcal{A}_1 = \{$diligently perform all tasks, follow appropriate processes, be vigilant, etc.$\}$. Similarly, $\mathcal{A}_2 = \{$ misuse of policies, incorrectly following processes, actions affecting culture of organization, actions affecting productivity, etc.$\}$. The exact actions undertaken will affect the likelihood of an attack being prevented or not. Also, the utility $u(a_i, s_i)$ could depend on every combination of the actions a_i and outcomes s_i. However, for the sake of simplicity, we assume that individual actions do not affect the outcome or the utilities, but only the nature of the actions (desirable or not) does. Therefore, we do not distinguish between different desirable actions and consider them to be represented by a single a_1 and similarly, represent all non desirable actions using a single a_2. The $u(a_i, s_i)$ values are thus elicited as in Table 10, which indicates a preference to desirable actions irrespective of the outcome. We are now able to calculate $u_D(d_1, a_i, s_i, a_3, s_3, d_2)$ using Tables 1, 2 and 10.

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

Table 10: $u(a_i, s_i)$ for every combination of a_i and s_i.

.

In order to implement the ARA solution to this problem, the defender must use backward induction and first identify action $d_2^*(d_1, a_i, s_i, a_3, s_3)$ which will maximize their utility. In this case, again, d_2^* turns out to be 'No upgrade', as it is the cheapest option and will therefore maximize u_D. Next, we need to elicit the probabilities $p_D(s_3|d_1, a_i, s_i, a_3)$. Note that these probabilities are only defined when $s_i =$ *not prevented*. We further assume that if the attack could not be prevented; then, the probabilities of its success are irrespective of the actions a_i encountered. Under that assumption, $p_D(s_3|d_1, a_i, s_i = \text{not prevented}, a_3) = p_D(s_3|d_1, \text{not prevented}, a_3) = p_D(s_3|d_1, a_3)$. Therefore, these are considered to coincide with $p_D(s|d_1, a)$ in Table 3. We are now able to compute $\psi_D(d_1, a_i, \text{not prevented}, a_3)$ using (9). These are listed in Table 11. We next seek to compute the expected utility $\psi_D(d_1, a_i, \text{not prevented})$,

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

Table 11: $\psi_D(d_1, a_i, \text{not prevented}, a_3)$ values for A_1 and A_2.

.

which requires us to elicit $p_D(a_3|d_1, a_i, \text{not prevented})$. This can be elicited either using the defender's knowledge, experience or guess or by modeling the malicious insider's strategic analysis process using Equations (13) to (16). Consider the first case; assume that the nature of the attack is independent of the type of employee A_1 or A_2 encountered earlier. Under this assumption, $p_D(a_3|d_1, a_i, \text{not prevented})$ is considered to coincide with $p_D(a|d_1)$ in Table 5. $\psi_D(d_1, a_i, \text{not prevented})$, thus calculated, is listed in Table 12. The defender now seeks to compute the expected

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

Table 12: $\psi_D(d_1, a_i, \text{not prevented})$ elicited by defender using their beliefs and knowledge and $u_d(d_1, a_i, \text{prevented})$ for each combination of D_1 and A_i.

.

utility $\psi_D(d_1, a_i)$ by integrating out $p_D(s_i|d_1, a_i)$. Note that $\psi_D(d_1, a_i, \text{prevented}) = u_D(d_1, a_i, \text{prevented})$, since the game does not proceed any further if the attack was indeed prevented. $u_D(d_1, a_i, \text{prevented})$ are also listed in Table 12. Suppose that the defender considers that the probability of the attack being prevented only depends on the type of employee encountered and is independent of d_1. Suppose that the chances of preventing the attack was considered to be 50% if the attacker encountered the *good* employees and just 10% is the attacker encountered by the *bad* employees, that is, $p_D(S_1 = \text{prevented}|d_1, a_1) = 0.5$ and $p_D(S_2 = \text{prevented}|d_1, a_2) = 0.1$. $\psi_D(d_1, a_i)$ thus computed, is listed in Table 13. Finally, the defender needs to integrate out $p_D(a_i|d_1)$ to compute the expected utility $\psi_D(d_1)$ of his defensive actions D_1 so as to identify the optimal action d_1^* that will maximize this expected utility. Suppose the defender believes that the *good* and the *bad* employees are randomly and evenly spread throughout their entire workforce and, therefore, $p_D(a_i|d_1)$ is independent of d_1. Suppose the defender guesses that 80% of the employees are good ones and the rest are bad. Then, $\psi_D(\text{Anom.det. \& Data prov.}) = -71.2229$, $\psi_D(\text{Info. sec. \& train.}) = -31.22$ and $\psi_D(\text{Random audits}) = -33.255$. Thus, based on the elicited utilities and probabilities, the optimal defensive action is to *invest in Information security* and *training of the staff*.

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

Table 13: $\psi_D(d_1, a_i)$ computed by the defender.

.

We now consider the case where $p_D(a_3|d_1, a_i, \text{not prevented})$ is elicited by modeling the attacker's strategic thinking process. As discussed earlier, given that the attack has not been prevented, the choice of the attack will be independent of the type of employee (A_1 or A_2) encountered. We assume that the choice of an attack depends only on the defender actions d_1 and d_2. That is, $p_D(a_3|d_1, a_i, s_i) = p_D(a_3|d_1)$; to elicit it, the defender must assess $U_A(a_3, s_3, d_2)$, $P_A(s_3|a_3, d_1)$ and $P_A(d_2|d_1, a_3, s_3)$. It is also reasonable to assume that attacker's preferences and uncertainties are also independent of the type of employee encountered. Therefore, we consider $U_A(a_3, s_3, d_2)$ to coincide with $U_A(a, s, d_2)$ in Table 6, $P_A(s_3|a_3, d_1)$ to be exactly same as $P_A(s|a, d_1)$ elicited in Table 8 and, finally, $P_A(d_2|d_1, a_3, s_3)$ with $P_A(d_2|d_1, a, s)$ elicited in Table 7.

We follow Equations (13) to (16) to compute $P_D(A_3 = a_3|d_1, a_i, \text{not prevented})$, which, as expected turns out to be $p_D(A = a|d_1)$ elicited in Table 9. We now compute the random expected utility $\mathbf{\Psi_D}(d_1, a_i, s_i)$, and Equation (10) and proceed to compute the random expected utility $\mathbf{\Psi_D}(d_1)$ using Equations (11) and (12). In this case, we have $\psi_D(\text{Anom. det. \& Data prov.}) = -52.147$, $\psi_D(\text{Info. sec. \& train.}) = -37.049$ and $\psi_D(\text{Random audits}) = -30.248$. Therefore, based on the elicited utilities and probabilities, the optimal defensive action is to invest in performing random audits.

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

4.3 Model uncertainty

Suppose now that the defender is not certain if the malicious insider will be able to act on their own or whether his actions will be affected by other employees. He decides to consider two models: M_1, the model in Section 4.1 and M_2, that in Section 4.2. He elicits a prior probability $p_D(M_1) = 0.3$, which implies that $p_D(M_2) = 0.7$. He would then perform the ARA analysis and arrive at his expected utilities $\psi_D(\text{Anom.det. \& Data prov.}, \text{Info.sec. \& train.}, \text{Random audits}|M_1) = (-36.115, -39.051, -34.556)$, when $p_D(a|d_1)$ is elicited by modeling the attacker's strategic thinking, for model M_1 as illustrated in Section 4.1. Similarly, he arrives at his expected utilities $\psi_D(\text{Anom. det. \& Data prov.}, \text{Info. sec. \& train.}, \text{Rand. aud.}|M_1) = (-52.147, -37.049, -30.248)$, for model M_2 as illustrated in Section 4.2. Then using (22), the expected utilities $\psi_D(d_1)$ taking into account the the model uncertainty are $\psi_D(\text{Anom.det. \& Data prov.}, \text{Info.sec. \& train.}, \text{Random audits}|M_1) = (-47.337, -37.65, -31.541)$. Thus, investing in random audits is the optimal strategy for the defender taking into account their model uncertainty.

5 Discussion and Further Work

Insider threats constitute a major security problem worldwide. We have provided two ARA based models to deal with them illustrated with a data security application. The first one is a Defend-Attack-Defend model; the second one includes a segmentation of employees in three groups and considers several interaction possibilities over time.

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

Insider threats come in many different forms. For example, the three player versions could consist of two attackers and one defender or the other way around or even an attacker, a defender and a victim (which would be a third party). For example, a three player case consisting of a malicious insider, the APT and the organization consists of two attackers and a defender. But the malicious insider could be also someone who uses their privileges to exploit, abuse or harm a third party, which could be clients, customers, patients, etc. A recent well-known example of this type is that of the USA gymnastic team osteopathic physician Dr. Nassar who was convicted for sexual abuse of young athletes under the pretext of treating them for their injuries. Therefore, an important extension would be to develop ARA solutions to such complex three player games. This could provide a much more realistic alternative to the game theoretic models proposed in [Feng et al., 2015] and [Hu et al., 2015].

Players are not always entirely rational and hence incorporating bounded rationality may make the model more realistic. ARA is naturally equipped to incorporate attackers with different reasonings, such as non-strategic thinking, Nash equilibrium, level-k thinking and the mirror equilibrium ([Banks et al., 2015]). However, a general ARA solution using bounded rationality has not yet been developed.

ARA relies on the elicitation of the adversary's utilities and probabilities. Robustness analysis of ARA to these elicitations is necessary, but has yet to be developed. [Ríos Insua et al., 2016] highlight the need and illustrate how a robustness analysis can be performed in principle for ARA. It is important to be able to investigate the sensitivity of the ARA outcome - the optimal strategy - to any errors or mis-specifications in the utilities and the probabilities elicited for the analysis.

Acknowledgments

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

References

[Banks et al., 2015] Banks, D., Rios, J., and Insua, D. R. (2015). *Adversarial Risk Analysis*. CRC Press, first edition.

[Brown et al., 2006] Brown, G., Carlyle, M., Salmeron, J., and Wood, R. (2006). Defending critical infrastructure. *Interfaces*, 36:530–544.

[Brown and Cox, 2011] Brown, G. G. and Cox, Jr., L. A. (2011). Making terrorism risk analysis less harmful and more useful: Another try. *Risk Analysis*, 31(2):193–195.

[Camerer, 2003] Camerer, C. (2003). *Behavioural Game Theory*. Princeton University Press.

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

[Cox, 2009] Cox, Jr., L. A. (2009). Game theory and risk analysis. *Risk Analysis*, 29(8):1062–1068.

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

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

[Ezell et al., 2010] Ezell, B., Bennett, S., Winterfeldt, D., Sokolowski, J., and Collins, A. (2010). Probabilistic risk analysis and terrorism risk. *Risk Analysis*, 30(4).

[Feng et al., 2015] Feng, X., Zheng, Z., Hu, P., Cansever, D., and Mohapatra, P. (2015). Stealthy attacks meets insider threats: A three-player game model. In *MILCOM 2015 - 2015 IEEE Military Communications Conference*, pages 25–30.

[Gintis, 2009] Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of Behavioural Sciences.* Princeton University Press.

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

[Hu et al., 2015] Hu, P., Li, H., Fu, H., Cansever, D., and Mohapatra, P. (2015). Dynamic defense strategy against advanced persistent threat with insiders. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 747–755.

[Hunker and Probst, 2009] Hunker, J. and Probst, C. (2009). Insiders and insider threats: An overview of definitions and mitigation techniques. *Journal Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 2(1):4–27.

[Insua et al., 2009] Insua, I. R., Rios, J., and Banks, D. (2009). Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854.

[Kantzavelou and Katsikas, 2010] Kantzavelou, I. and Katsikas, S. (2010). A game-based intrusion detection mechanism to confront internal attackers. *Computers & Security*, 29(8):859 – 874.

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

[Liu et al., 2008a] Liu, D., Wang, X., and Camp, J. (2008a). Game-theoretic modeling and analysis of insider threats. *International Journal of Critical Infrastructure Protection*, I:75 – 80.

[Liu et al., 2008b] Liu, D., Wang, X., and Camp, J. (2008b). Mitigating inadvertent insider threats with incentives. *BUSCAR!!*, ??:??

[Martinez-Moyano et al., 2008] Martinez-Moyano, I., Rich, E., Conrad, S., Andersen, D., and Stewart, T. (2008). A behavioral theory of insider-threat risks: a system dynamic approach. *ACM Transactions on Modeling and Computer Simulation*, 18(2).

[Moore et al., 2015] Moore, A., Novak, W., Collins, M., Trzeciak, R., and Theis, M. (2015). *Effective Insider Threat Programs: Understanding and Avoiding Potential Pitfalls.* White Paper.

[Raiffa et al., 2002] Raiffa, H., Richardson, J., and Metcalfe, D. (2002). *Negotiation Analysis.* Havard University Press.

[Ríos Insua et al., 2019] Ríos Insua, D., Banks, D., Ríos, J., and Ortega, J. (2019). *Adversarial Risk Analysis as an Expert Judgement Methodology*, pages –. Springer International Publishing.

[Ríos Insua et al., 2016] Ríos Insua, D., Ruggeri, F., Alfaro, C., and Gomez, J. (2016). *Robustness for Adversarial Risk Analysis*, pages 39–58. Springer International Publishing.

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

[Schulze, 2018] Schulze, H. (2018). *Insider Threat, 2018 report.* ca Technologies.

[Sevillano et al., 2012] Sevillano, J. C., Insua, D. R., and Rios, J. (2012). Adversarial Risk Analysis: The Somali Pirates Case. *Decision Analysis*, 9(2):86–95.

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

[Tang et al., 2011] Tang, K., Zhao, M., and Zhou, M. (2011). Cyber insider threats situation awareness using game theory and information fusion-based user behavior predicting algorithm. *Journal of Information & Computational Science*, 8(3):529 – 545.

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

[Ware, 2017] Ware, B. (2017). *Insider Attacks, 2017 insider threat study.* Haystax.

[Wood et al., 2016] Wood, P., Nahorney, B., Chandrasekar, K., Wallace, S., and Haley, K. (2016). *Internet Security Threat Report*, volume 21. Symantec.

Annex 4: Paper: Assessing Supply Chain Cyber Risks

Assessing Supply Chain Cyber Risks

Abstract

Risk assessment is a major challenge for supply chain managers, potentially affecting business factors such as service costs, supplier competition and customer expectations. The increasing interconnection between organizations has put into focus methods for supply chain cyber risk management. We introduce a framework for such activity which takes into account various techniques of attacking an organisation and its suppliers, as well as the impacts of such attacks. Since data is lacking in many respects, we use structured expert judgment methods to facilitate its implementation. We couple a family of forecasting models to enrich risk monitoring. The framework may be used to set up risk alarms, negotiate service level agreements, rank suppliers and identify insurance needs, among other management possibilities.

Keywords: Risk Analysis, Supply Chain Risks, Cybersecurity, Expert Judgment

1 INTRODUCTION

Earthquakes, economic crises, strikes, terrorist attacks and other events may disrupt supply chain operations with significant impact over the performance of organizations. As examples, it is reported that Ericsson lost 400 million EUR after their supplier's semiconductor plant caught on fire in 2000 (Latour, 2001) and that Apple lost many customer orders during a supply shortage of DRAM chips after a 1999 earthquake in Taiwan (Lovejoy, 2016).

Supply chain risk management (SCRM) has come into place to implement strategies to manage risks in a supply chain with the goal of reducing vulnerabilities and avoid service and product disruptions. As in other risk analysis application areas, SCRM usually involves four processes: identification, assessment, controlling and monitoring of risks (Bedford & Cooke, 2001). Tang and Tomlin (2008) define the field as *the management of supply chain risks through coordination or collaboration among supply chain partners to ensure profitability and continuity* and consider four basic areas to mitigate their impact: supply, demand, product and information management. Ritchie and Brindley (2007) developed a framework that categorizes risk drivers and integrates risks dimensions in supply chains. Sharland et al. (2003), Jüttner (2005) and Zsidisin and Ritchie (2008) identify key issues in SCRM through surveys, presenting best practices. Hallikas et al. (2004) proposed a risk management process in network environments giving a more holistic view; they provide a risk matrix approach, although this type of tool has well known shortcomings, Cox (2008). Thekdi and Santos (2016) introduce interdependency modeling through an input-output model accross multiple sectors for assessing the social and economic factors associated with SCRM. Kern et al. (2012) illustrate how supply chains would benefit from the capacity of predicting service unavailability early enough, so that interruptions may be mitigated. DiMase et al. (2016) provide traceability considerations using high risk parts prioritization, certificates of conformance, adoption of standards and considering resilience to recover its original functional state after a disruption. Foerstl et al. (2010), Bandaly et al. (2012), Kern et al. (2012) and Ghadge et al. (2013) develop frameworks that cover identification, assessment, response management and performance outcomes. Curkovic et al. (2015) identify how companies may manage supply chain risks through FMEA. Aqlan and Lam (2016) and Fattahi et al. (2017) used optimization and simulation to deal with deterministic and stochastic features in SCRM. Zahiri et al. (2017) and Song and Zhuang (2017) are recent examples of areas where SCRM has a major impact in real applications.

Due to the proliferation of cyber attacks and the increasing interconnectedness of organisations, a major feature of recent interest refers to new cyber threats affecting supply chain operations in what we shall call Supply Chain Cyber Risk Management (SCCRM). As an example, Target suffered in 2013 a major cyber breach through their air conditioning supplier losing up to 70 million credit and debit cards of buyers, with a massive loss of reputation, McGrath (2014). Another relevant attack was Wannacry which took over, among many others, Telefonica and the UK NHS producing the unavailability of numerous services, which entailed costs estimated to have reached \$4 billion, Berr (2016).

We present here a general framework for SCCRM. Section 2 presents a general description, covering models to forecast attacks and their impacts, integrating such information to provide relevant risk indicators. Due to lack of data, we need to rely on expert judgment to assess the involved parameters. Section 3 outlines its implementation. Section 4 covers a numerical example. We end up with some discussion in Section 5.

2 A FRAMEWORK FOR SCCRM BASED ON EXPERT JUDGMENT

Consider a company c interconnected with its suppliers s, pertaining to a set $\mathcal{S}$. Both the company and the suppliers are subject to various types of attacks $a \in \mathcal{A}$, which will be the set of incumbent attacks. Examples include attacks through botnets or based on stolen login information. Attacks to the supplier could also be transferred to the company. As an example, imagine a case in which one of the company's suppliers is infected through malware. The attacker could then scan the supplier's network and send infected emails to the company which would be more likely to get infected as the received software originates from a legitimate source.

We have access to a threat intelligence system (TIS) (Tittel, 2017) which collects data $\mathbf{n}^a_{c,s}$, $a \in \mathcal{A}$, about the cyber attack vectors gathered from the company and its suppliers. The vectors could include, e. g., information such as IPs of botnet infected devices or number of malware infections found. The TIS may also provide data about the security environment, including, for instance, the number of negative mentions in hacktivist blogs, and the security posture, covering for example the patch cadence or the number of vulnerabilities. All of the above information will be available for both the company and its suppliers.

Based on such data, and other available information, we aim at assessing:

- the probabilities that the company's suppliers are attacked;
- the probability that the company is attacked, either directly or through its suppliers;
- the impacts that such attacks might induce over the company,

and aggregate such information to facilitate cyber risk assessment to the company in relation with its suppliers so as to support cyber risk management decisions.

We describe a framework to obtain, combine and apply in practice the required model ingredients. Given the reluctance of companies to provide data concerning sufficiently harmful attacks for reputational reasons, we shall cope with the eventual lack of data to fit the proposed models by extracting information from cybersecurity specialists through structured expert judgment techniques (Cooke, 1991; O'Hagan et al., 2006). Here, the term "sufficiently harmful" indicates that the attack is relevant and has caused significant damage to the company or its suppliers[1].

2.1 Probability of a Sufficiently Harmful Attack

We start by describing how to estimate the probability that the supplier, or the company, is successfully attacked, given the information scanned through the TIS. For simplicity, we do not include for the moment the security environment and posture in the model, which we cover in Section 2.2. We would undertake the proposed approach for each attack type $a \in \mathcal{A}$.

We assume that the probabilities for the attack vectors may be modelled with logistic regression models (Hosmer et al., 2013) with the number of parameters depending on the number of levels corresponding to the attack vector. For example, suppose that one of the vectors which includes h severity levels; we would include $h + 1$ parameters, as in

$$Pr(y = 1 \mid \boldsymbol{\beta}, \mathbf{n}) = \frac{\exp(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{n})}{1 + \exp(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{n})}, \tag{1}$$

[1] For example, this could correspond to attacks that need to be declared to the supervisory authority, as in the recent General Data Protection Regulation, EU (2018).

where $y = 1$ indicates that the attack was successful; n_i is the number of i-th level infections of the incumbent attack vector, $i = 1, 2, ..., h$; $n = (n_1, \ldots, n_h)$; and, finally, $\beta = (\beta_1, \ldots, \beta_h)$.

The parameters will be indirectly estimated through expert judgment, for which we provide appropriate questions and consistency checks. We illustrate the approach with the above specific case with h infection categories. We use elicitation techniques, Clemen and Reilly (1999), to assess, for example,

$$\widehat{p}_0 = Pr(y = 1 \mid \beta_0, \boldsymbol{\beta}, \mathbf{n} = [0, \ldots, 0]) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}.$$

A typical question that could be posed to experts is:

> *Assuming that the TIS has detected no evidence in the network concerning such infection (i.e., $n_i = 0$, $i = 1, \ldots, h$), what would be the probability p_0 of actually suffering a sufficiently harmful attack due to such type of infection?*

We then make

$$\log\left(\frac{\widehat{p}_0}{1 - \widehat{p}_0}\right) = \widehat{\beta}_0.$$

Note that we may introduce interactive schemes to elicit the pertinent judgments.

Then, we further assess through expert judgment

$$\widehat{p}_1 = Pr(y = 1 \mid \beta_0, \boldsymbol{\beta}, \mathbf{n} = [1, 0, \ldots, 0]) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)},$$

obtaining

$$\widehat{\beta}_1 = \log\left(\frac{\widehat{p}_1}{1 - \widehat{p}_1}\right) - \widehat{\beta}_0.$$

We would extract $\widehat{p}_i$ from the experts in a similar manner to obtain $\widehat{\beta}_i$, $i = 2, \ldots, h$. Finally, we would check for consistency based on assessments such as, e. g.,

$$\widehat{p} = Pr(y = 1 \mid \beta_0, \boldsymbol{\beta}, \mathbf{n} = [2, 0, \ldots, 0]),$$

checking whether

$$\log\left(\frac{\widehat{p}}{1 - \widehat{p}}\right) \simeq \widehat{\beta}_0 + 2\widehat{\beta}_1.$$

If not, we would need to reassess some of the judgments and modify the parameters accordingly.

Besides the attack probabilities, we also need to assess the probability of a type a attack being transferred from a supplier to the company, denoted by q^a. We define an attack to a supplier as transferred successfully if it is immediately followed by a second attack to the company, taking advantage from either the information gathered in the first attack or the compromised infrastructure. The probabilities of transferring an attack are different for each of the types; thus, we elicit them directly from the experts. An example of a typical question for assessing such probabilities, in relation e. g. with malware, would be

> *Suppose that there is an attack to the supplier based on malware, what would be the probability of the customer suffering another one, taking advantage of the supplier's attack?*

As before, we introduce consistency checks and interactive procedures to evaluate such assessment. We assume that the probabilities are the same for every supplier, mainly to reduce the cognitive load over the experts.

2.2 Taking into Account the Security Environment and Posture

We describe now how we incorporate information about the security environment and posture of the supplier and the company within the attack probabilities. Essentially, we introduce indices for the corresponding variables through multicriteria value functions (González-Ortega et al., 2018) and then apply the approach in Section 2.1 to extract the required coefficients.

We define first an index e which assesses the security environment of, say, the supplier based on the k environment variables captured by the TIS. Let e_i be the *i-th* variable, $i = 1, \dots, k$, rescaled to $[0, 1]$. With no loss of generality, assume that the bigger e_i is, the worse is the security environment. We use a multicriteria value function, $e = \sum_{i=1}^{k} \lambda_i e_i$ with $\sum_{i=1}^{k} \lambda_i = 1$, $\lambda_i \geq 0$, $i = 1, \dots, k$. We determine the λ_i weights by asking experts to compare pairs of security environment contexts leading to a system of equations

$$\begin{aligned} \delta_1^1 \lambda_1 &= \delta_1^2 \lambda_2, \\ &\vdots \\ \delta_{k-1}^1 \lambda_{k-1} &= \delta_{k-1}^2 \lambda_k. \end{aligned} \tag{2}$$

For example, given a reference value δ_1^1 for the first environment variable, we obtain the first equation by asking the expert about the value δ_1^2 of the second environment variable such that the following two security environments are perceived as equally unsafe

$$(\delta_1^1, 0, \dots, 0) \sim (0, \delta_1^2, \dots, 0).$$

As an example, given that $\delta_1^1 = 0.7$, the expert answer could be $\delta_1^2 = 0.3$, leading to the equation $0.7\lambda_1 = 0.3\lambda_2$. We can introduce interactive schemes to obtain the δ_i's and perform consistency checks.

Then, from system (2), we obtain the $k-1$ equations

$$\lambda_i = \frac{\delta^1_{i-1}}{\delta^2_{i-1}}\lambda_{i-1} = r_{i-1}\lambda_{i-1}, \quad i = 2, \dots, k,$$

with

$$r_i = \delta^1_i/\delta^2_i, \quad i = 1, \dots, k-1.$$

Taking into account that $\sum_{i=1}^k \lambda_i = 1$, we solve for λ_1 to get

$$\lambda_1 \Bigl(1 + \sum_{i=1}^{k-1} \prod_{j=1}^{i} r_j\Bigr) = 1,$$

so that

$$\lambda_1 = \frac{1}{1 + \sum_{i=1}^{k-1} \prod_{j=1}^{i} r_j},$$

and

$$\lambda_i = \prod_{j=1}^{i-1} r_j \lambda_1, \quad i = 2, \dots, k.$$

We proceed in a similar manner to aggregate the security posture, defining an index l and assessing such posture through a multicriteria value function

$$l = \sum_i v_i l_i,$$

with $\sum_i v_i = 1$ and $v_i \geq 0 \; \forall i$, where l_i is the *i-th* security posture variable.

Once we have built the environment and posture value functions, we assess the corresponding β parameters. We adopt common parameters for all attack types. For example, for the specific case of a harmful attack with h levels of infection, as in equation (1), we use

$$Pr(y = 1 \mid \boldsymbol{\beta}, \mathbf{n}, l, e) = \frac{\exp(\beta_0 + \boldsymbol{\beta} \cdot [\mathbf{n}, l, e])}{1 + \exp(\beta_0 + \boldsymbol{\beta} \cdot [\mathbf{n}, l, e])},$$

where $[,]$ designates the concatenation of the corresponding vectors. We can adopt a reference value for l, say v_1, which we associate with $l = [1, 0, \ldots, 0]$ and ask about the corresponding probability $\widehat{p}_{h+1}$, when, e. g., $n_1 = n_2 = 1$ and $n_i = 0$, $i = 3, \ldots, h$. Then, with $e = 0$, it would be

$$\widehat{p}_{h+1} = \frac{\exp(\beta_0 + \beta_1 1 + \beta_2 1 + \beta_3 0 + \cdots + \beta_h 0 + \beta_{h+1} v_1 + \beta_{h+2} 0)}{1 + \exp(\beta_0 + \beta_1 1 + \beta_2 1 + ... + \beta_{h+1} v_1 + \beta_{h+2} 0)}.$$

Since we have already elicited $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ we easily obtain $\widehat{\beta}_{h+1}$, through

$$\widehat{\beta}_{h+1} = \frac{1}{v_1}\left(\log\left(\frac{\widehat{p}_{h+1}}{1 - \widehat{p}_{h+1}} \right) - \widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2 \right).$$

We would proceed similarly to obtain β_{h+2}, introducing the corresponding consistency checks.

As mentioned, we implement this approach for all attack types $a \in \mathcal{A}$, both for the company and its suppliers.

2.3 Impacts over Supplier and Company

We describe now the models used to predict the impacts of attacks over the company. The relevant impacts might depend on the organization. In our supply chain application area we have included: the suppliers service unavailabilities, as they would induce a cost in the company due to lacking such services; the company's service unavailability, which would also induce a cost, typically, higher than the earlier ones; and, finally, the loss of company reputation which might induce a loss of customers. We use the same distributions for all attack types. Hubbard and Seiersen (2016) discuss how to measure other relevant impacts in cybersecurity.

2.3.1 *Supplier and Company Unavailability*

We focus first on supplier and company service unavailability, given a sufficiently harmful attack. Their durations will be designated by i_s and i_c, respectively. We model the corresponding downtimes through gamma distributions

$$f(i_s \mid k_s, \theta_s) \sim \text{Ga}(k_s, \theta_s); \quad f(i_c \mid k_c, \theta_c) \sim \text{Ga}(k_c, \theta_c).$$

Given the lack of data, we aim at obtaining estimates of k and θ through expert judgment. For this, we may ask the experts for the first p_{25} and third p_{75} quartiles of the i_s and i_c distributions and infer the parameters by solving

$$\min_{k,\theta} \left\{ (p_{25} - \text{cdf}(.25,\, k,\, \theta))^2 + (p_{75} - \text{cdf}(.75,\, k,\, \theta))^2 \right\},$$

where $\text{cdf}(\cdot,\, k,\, \theta)$ designates the cumulative distribution function of the gamma distribution with parameters k and θ. This leads to the corresponding optimal parameters k^* and θ^*. We perform consistency checks based on other quantiles and use interactive procedures to obtain the involved quantiles. When required, we may approximate the downtimes through the expected value of the distributions, respectively,

$$\bar{i}_s = \mathbb{E}[f(i_s \mid k_s^*, \theta_s^*)] = k_s^* \theta_s^*; \;\; \bar{i}_c = \mathbb{E}[f(i_c \mid k_c^*, \theta_c^*)] = k_c^* \theta_c^*.$$

We undertake this approach for all the suppliers and the company.

2.3.2 *Reputation*

We consider now the impact of an attack over the company's reputation. We assume that reputational impacts occur only if there is a direct attack to the company. There is no natural attribute that allows us to assess reputation loss, see the discussion in Hubbard and Seiersen (2016). Our focus is therefore in its business consequence which we consider to be the loss in market share induced by a harmful attack over the organization.

Let us designate by d the proportion of customers abandoning to a competitor due to the incumbent loss of reputation, which we model as a beta distribution with parameters a and b, that is $d \sim \text{Beta}(a, b)$. We proceed in a similar fashion to Section 2.3.1, by asking two quartiles to experts and, subsequently, approximating the parameters a^*, b^*, based on a least squares cdf approximation, after appropriate consistency checks. When required, the expected proportion of customers lost could be approximated through the expected value

$$\bar{d} = \frac{a^*}{a^* + b^*}.$$

2.3.3 *Aggregating Impacts*

We describe now how we aggregate all relevant impacts to compute the expected impact over the company. We focus on approximations based on means to allow for fast computations, appropriate for continuous monitoring. A more detailed analysis would be based on Monte Carlo approximations based on samples from the distributions of i_c, i_s and d. In any case, we would require τ, the proportion of market share for the company; η, its (monetary) market size; and κ_s and κ_c, the average cost per hour of supplier and company service unavailability, respectively. We may assess them from data and/or experts.

We compute the average downtime cost of the supplier s and company c after a sufficiently harmful attack through

$$C_{i_s} = \kappa_s \times \bar{i}_s,$$
$$C_{i_c} = \kappa_c \times \bar{i}_c.$$

On the other hand, the average reputational cost after a successful attack is approximated through the cost associated with clients abandoning the company, which would be

$$C_d = \bar{d} \times \tau \times \eta.$$

Recall now that there are three types of attacks:

1. Direct attacks to the company, with expected cost $C = C_d + C_{i_c}$.

2. Attacks to the supplier that disrupt its service but are not transferred to the company, with expected cost $C = C_{i_s}$.

3. Attacks to the supplier that disrupt its service and are transferred to the company. The expected cost in this case would be $C = C_d + C_{i_c} + C_{i_s}$.

To cater for the company's risk attitude, we may use a constant risk averse utility model, González-Ortega et al. (2018),

$$u(C) = (1 - e^{-\rho C})/\rho,$$

where ρ is extracted from experts within the company.

2.4 Model Outputs

We describe now the model outputs that may be relevant for risk management, including the probabilities of various attacks, the expected impacts and the expected utilities derived from the above results. Having available sufficient computational time and resources, we would replace the proposed indicators by their Monte Carlo approximations.

2.4.1 *Attack Probabilities*

For the incumbent company c and its suppliers, we rewrite the probability of a type $a \in \mathcal{A}$ security event resulting in a sufficiently harmful attack in a more compact form through

$$p_i^a = \frac{\exp(\beta_0^a + \boldsymbol{\beta}^a \cdot \mathbf{n}_i^a)}{1 + \exp(\beta_0^a + \boldsymbol{\beta}^a \cdot \mathbf{n}_i^a)}, \tag{3}$$

where $\mathbf{n}_i^a$ represents the attack vector count, including the environment and posture indicators, and $i \in \{c,\, s\}$. Assuming that the attack types are independent, the probability of a direct attack to the company will be approximated through

$$\mathrm{AP}_c = \sum_{k=1}^{K_{\mathcal{A}}} \sum_{\mathcal{I} \in \mathcal{C}_{\mathcal{A},k}} \left(\prod_{a \in \mathcal{I}} p_c^a \prod_{a \in \mathcal{A} \setminus \mathcal{I}} (1 - p_c^a) \right),$$

where $\mathcal{C}_{\mathcal{A},k}$ is the set of all possible combinations of k elements taken from $\mathcal{A}$ and $K_{\mathcal{A}} \leq |\mathcal{A}|$ is the maximum number of simultaneous attacks taken into account.

For each supplier $s \in \mathcal{S}$, we approximate the induced attack probability from s to c as

$$\mathrm{IAP}_c^s = \sum_{k=1}^{K_{\mathcal{A}}} \sum_{\mathcal{I} \in \mathcal{C}_{\mathcal{A},k}} \left[\underbrace{\left(\prod_{a \in \mathcal{I}} p_s^a \prod_{a \in \mathcal{A} \setminus \mathcal{I}} (1 - p_s^a) \right)}_{\text{Probability of direct attack}} \underbrace{\left(1 - \left(1 - \prod_{a \in \mathcal{I}} (1 - q^a) \right) \right)}_{\text{Probability of transferring attack}} \right],$$

recalling that q^a is the probability of an attack of type a being transferred from s to c.

Finally, the global attack probability to the company can be approximated through

$$\mathrm{GAP}_c \simeq \mathrm{AP}_c + (1 - \mathrm{AP}_c) \sum_{k=1}^{K_{\mathcal{S}}} \sum_{\mathcal{I} \in \mathcal{C}_{\mathcal{S},k}} \left(\prod_{s \in \mathcal{I}} \mathrm{IAP}_c^s \prod_{s \in \mathcal{S} \setminus \mathcal{I}} (1 - \mathrm{IAP}_c^s) \right),$$

where $\mathcal{S}$ is the set of all possible suppliers of company c and $K_{\mathcal{S}} \leq |\mathcal{S}|$ is the maximum number of suppliers that can reasonably transfer an attack over the same time period. Note that in the above approximation, we assume that when an attack is direct to the company it supersedes the indirect attacks as its consequences will be much more important.

2.4.2 *Risk Measures: Expected Impacts of Attacks*

Recall that we are assuming that if an attack is successfully transferred from a supplier, there are unavailability and reputational costs, whereas if the attack is not transferred, we only consider supplier unavailability costs. We now approximate the expected impacts. These include the impact due to direct attacks to the company, expressed as

$$\mathrm{R}_c = \sum_{k=1}^{K_{\mathcal{A}}} \sum_{\mathcal{I} \in \mathcal{C}_{\mathcal{A},k}} \left(\prod_{a \in \mathcal{I}} p_c^a (C_d + C_{i_c}) \prod_{a \in \mathcal{A} \setminus \mathcal{I}} (1 - p_c^a) \right),$$

and the impact induced by attacks to the supplier s, which is

$$\mathrm{IR}_c^s = \sum_{k=1}^{K_{\mathcal{A}}} \sum_{\mathcal{I}\in\mathcal{C}_{\mathcal{A},k}} \Bigg\{ \Bigg(\prod_{a\in\mathcal{I}} p_s^a \prod_{a\in\mathcal{A}\setminus\mathcal{I}} (1-p_s^a) \Bigg) \Bigg[\overbrace{\Bigg(1 - \prod_{a\in\mathcal{I}}(1-q^a)\Bigg)(C_{i_s})}^{\text{Attack not transferred}} + \underbrace{\Bigg(1 - \Bigg(1 - \prod_{a\in\mathcal{I}}(1-q^a)\Bigg)\Bigg)(C_d + C_{i_c} + C_{i_s})}_{\text{Attack transferred}} \Bigg] \Bigg\}.$$

Finally, the total impact would be

$$\mathrm{TR}_c = R_c + \sum_{s\in\mathcal{S}} \mathrm{IR}_c^s.$$

In a similar fashion, we may consider approximations to expected utilities replacing the previous impacts by the corresponding utilities.

As mentioned, we could refine the analysis by using Monte Carlo approximations to compute expected utilities, should we have sufficient computational time.

2.5 Forecasting Risk Indicators

The previous approach is used periodically based on collecting data through the TIS and aggregating the results to assess supply chain cyber risks. As a relevant complement, observe that the proposed approach focuses on studying several risk indicators X_j (attack probabilities, expected impacts, expected utilities) to monitor SCCR at the company, in reference to time j.

The ensuing analysis focuses on just one of the indicators, but applies to all of them. D_j represents the data available until time j. $X_{j+1}|D_j$ will represent a forecasting model for the risk index at time $j+1$ and summarises all information available at time period j concerning such index.

We employ Dynamic Linear Models (DLMs) to support forecasting tasks in risk monitoring. We briefly sketch the basic DLM results we use. For further details see West and Harrison (2013) and Petris et al. (n.d.). We adopt the general, normal DLM with univariate observations X_j, characterized by the quadruple $\{F_j, G_j, V_j, W_j\}$, where, for each j, F_j is a known vector of dimension $m \times 1$, G_j is a known $m \times m$ matrix, V_j is a known variance, and W_j is a known $m \times m$ variance matrix. The model is written as

$$\begin{aligned} \theta_0 | D_0 &\sim N(m_0, C_0), \\ \theta_j | \theta_{j-1} &\sim N(G_j \theta_{j-1}, W_j), \\ X_j | \theta_j &\sim N(F_j' \theta_j, V_j). \end{aligned}$$

Because of the relative stability of the type of series considered, we use a trend (second order polynomial) model which is a specification of the above DLM with constant F_j and G_j through

$$F = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

West and Harrison (2013) summarize the basic features of DLMs for forecasting purposes that we use. They are based on the one-step ahead predictive distributions which, for each j, have normal distribution

$$X_j | D_{j-1} \sim N(f_j, Q_j),$$

for mean f_j and variance Q_j recursively defined. k-steps ahead forecasts are also based on normal models and will be used below.

2.6 Uses

We sketch now various uses of the above outputs that facilitate SCCRM:

- *Issue alarms.* Based on the above mentioned forecasting models, we may use point forecasts at time j for the risk indicator given by $\mathbb{E}[X_j|D_{j-1}] = f_j$ and interval forecasts determined by $[l_j = f_j - z_{1-\alpha/2}Q_j^{1/2},\ u_j = f_j + z_{1-\alpha/2}Q_j^{1/2}]$, where u_j and l_j respectively represent the upper and lower bounds of the interval; $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution; and, finally, α is the desired probability level of the predictive interval. We then issue an alarm about an unexpected increase, or decrease, in the incumbent risk indicator when the corresponding next observation x_j does not fall in the predictive interval $[l_j,\ u_j]$. Should this happen repeatedly over time, the alarm could be modulated.

  Another type of alarms may be raised when the predictive interval captures a sufficiently high risk level y. For this, we perform predictions k steps ahead based on standard DLM forecasting results to try to forecast sufficiently in advance critical risk issues, i.e. we try to detect whether $Pr(X_{j+k} \geq y)$ will be sufficiently high for a certain k.

- *Rank suppliers.* We consider for this the induced risks, induced expected impacts or induced attack probabilities of various suppliers over the company. For example, we could say that supplier s_1 is preferred to supplier s_2 if its induced expected impact is smaller, that is, if $\mathrm{IR}_c^{s_1} \leq \mathrm{IR}_c^{s_2}$. This supports identifying and dealing with riskier suppliers for communication purposes, increasing transparency and forecasting critical situations.

- *SLA negotiations between the company and its suppliers.* Using the same indices as before, this could be based, for example, on the company requiring the supplier to preserve its risk indicator X_j to remain below a certain maximum level m_j, as well as monitoring and forecasting whether such maximum acceptable level is attained. Repeated violations of the agreed maximum risk level could lead to a contract breaching and a penalty, possibly encompassing a supplier to better manage cybersecurity. For example, the company could require the supplier to preserve $\mathrm{IR}_c^s \leq m$, over time.

- *Insurance.* The above measures allow us to properly apportion the cyber risks to which a company is subject to, including those related with third parties, thus facilitating the negotiation of insurance contracts, both with the suppliers and the insurer. For example, the company could apply for an insurance premium reduction if TR_c is preserved below a certain agreed level throughout a year.

3 IMPLEMENTATION

We describe how we implement the proposed approach and couple it with an available TIS. Essentially, we first calibrate several experts concerning their cyber security knowledge; we then obtain their judgments and combine them to obtain the corresponding parameters; and, finally, during operation, we assess attack probabilities, risks, issue alarms and forecast risks for the next period. This will require some data preprocessing.

3.1 Security Expert Calibration

We obtain judgments from m experts which we calibrate using Cooke (1991) classical model. We first ask the experts several general questions related with cyber security. As an example, one of the questions we used is:

> *Which was the number of new ransonware types over the last year?*

The experts' answers describe the intervals that, they believe, cover with high probability the value, as well as their median. Based on such assessments, we obtain the expert scores $(\omega_i)_{i=1}^m$, $\omega_i \geq 0$, $\sum_{i=1}^m \omega_i = 1$.

3.2 Security Expert Assessment. Generic Questions

Once we have calibrated the experts, we extract from them through a questionnaire the relevant attack probabilities based on hypothetical scenarios in which sufficently harmful attacks may occur. For example, one of the questions referring to a scenario pertaining to a botnet based type of attack we use is:

> *According to you, what would be the probability of actually suffering a sufficiently harmful attack through botnet infected devices if the TIS did not detect any of them in the company's network? And if the TIS detected* 0.25% *of such infected devices in their network?*

The i-th expert provides a probability p_i for the corresponding question, which we aggregate through

$$p = \sum_{i=1}^{m} \omega_i p_i.$$

Based on such type of assessments we obtain the β parameters, covering with such information the stages described in Sections 2.1 and 2.2.

3.3 Company Expert Assessment. Specific Questions

We then extract the specific parameters which refer to impacts to the company (direct or over the supplier) through a questionnaire. These parameters include, among others, the current market share, the market size, or the expected number of customers lost after a successful attack. Some of them will be based on available data; for others, we could ask experts as described above. For instance, a question we use is:

> *According to you, what would be the average cost of one hour of service unavailability of such supplier?*

In such a way, we cover the aspects described in Section 2.3.

3.4 Data Preprocessing

The data received through the TIS are preprocessed via exponential smoothing (Brown, 2004). This allows us to control the growth rate of various security indicators and partly mitigate fluctuations associated with random variations. Given $\{x_j\}$, the raw data sequence, n_j will represent the security indicator at time j, and h the smoothing factor through

$$n_0 = x_0, \quad n_j = h \times x_j + (1-h)n_{j-1}.$$

Observe that

$$n_j = h \times x_j + (1-h)hx_{j-1} + \cdots + (1-h)^{k-2}hx_{j-(k-1)} + (1-h)^k hn_{j-k}.$$

Since $(1-h)^{k^*}$ will be small enough for k^* sufficiently large, we have that

$$n_j \simeq \sum_{i=0}^{k^*} h(1-h)^i x_{j-i},$$

which effectively entails preserving the last k^* scans $(x_j, x_{j-1}, \ldots, x_{j-k^*})$ and consolidating them in n_j.

3.5 Operation

With the above information, we are capable of putting the proposed framework under operation. We summarize first all the required information, after having calibrated the experts:

1. The coefficients $\boldsymbol{\lambda}$, $\mathbf{v}$, and $\boldsymbol{\beta}$ are obtained (indirectly) from the experts as described in Sections 2.1, 2.2 and 3.2.

2. The probability q^a of each attack type transferring from the supplier to the customer is obtained directly from the experts as described in Section 2.1.

3. The information needed to compute both the downtime cost of the company and supplier, C_c and C_s, must be provided by the company. This involves assessing every supplier s to obtain estimates of its downtime distribution i_s, as well as i_c.

4. The information needed to compute the reputational cost C_d is obtained from experts within the company.

Once we have obtained those parameters, we may start operations by essentially scanning through the TIS, performing the risk computations, updating the forecasting model and issuing alarms when required. The suppliers and company data obtained periodically by scanning the network through the TIS are preprocessed as in Section 3.4. In summary, at every time step j we perform the following computations:

1. Scan network for attack, posture and environment vectors $\mathbf{n}$, e and l, for the company and its suppliers.
2. Assess the attack probabilities AP_c, IAP_c^s and GAP_c.
3. Assess the risks R_c, IR_c^s and TR_c.
4. Issue alarms, if required.
5. Display risk outputs.
6. Compute risk forecasts for next period.

4 A NUMERICAL EXAMPLE

We describe now a numerical example using simulated data. Assume that we are able to scan information regarding four attack vectors, $|\mathcal{A}| = 4$. For the first security event, one expert provided attack probabilities $\widehat{p}_0 = 0.05$, when the scan did not detect any infected devices, and $\widehat{p}_1 = 0.25$, when the scan detected 1% of infected devices in the company's network. We then make

$$\widehat{\beta}_0 = \log\left(\frac{\widehat{p}_0}{1-\widehat{p}_0}\right) = -2.94,$$

and

$$\widehat{\beta}_1 = \log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right) - \widehat{\beta}_0 = 1.85.$$

	p_c	p_{s_1}	p_{s_2}	q
atk0	0.057	0.468	0.383	0.103
atk1	0.187	0.164	0.350	0.107
atk2	0.131	0.166	0.143	0.056
atk3	0.236	0.481	0.200	0.084

Table 1. Probabilities of successful attack for each attack type

If we detect now that 0.075% of the devices are actually infected at $t = 1$, we estimate the attack probability of that specific security event as

$$p_c = \frac{\exp(-2.94 + 1.85 \times 0.075)}{1 + \exp(-2.94 + 1.85 \times 0.075)} = 0.057.$$

We show in Table 1 columns 1-3 the probabilities of each of the security events for the company and two of its suppliers s_1 and s_2. These probabilities are obtained through equation (3), as demonstrated above for the company and the first security event. We also show in the fourth column the probabilities of an attack being transferred from a supplier to the company.

Given the above, we can compute the direct attack probability to the company over the next period, which is AP $= 0.491$; the attack probabilities induced by the suppliers $\text{IAP}_1 = 0.111$, $\text{IAP}_2 = 0.098$; and, finally, the global attack probability GAP $= 0.592$. Note that having two suppliers increases the chances for the company of receiving attacks. If both suppliers were offering the same service, say, both were Internet providers, we could take into account their IAPs to decide whether to contract the services from one or the other. In the example, we might be more inclined to work with the second supplier as receiving an attack through it seems less likely.

To transform the previous probabilities into expected costs, we assume the following information: the average cost per hour of downtime of our company is 20k EUR; the market size is 2922 billion EUR; and, finally, the current market share of the company is 18%. In addition, the company estimates that one hour of service unavailability of both suppliers costs 25k EUR. The first and third quartiles for the distribution of the downtime are 2, 6 hours for the company and 1, 4 hours for the suppliers.

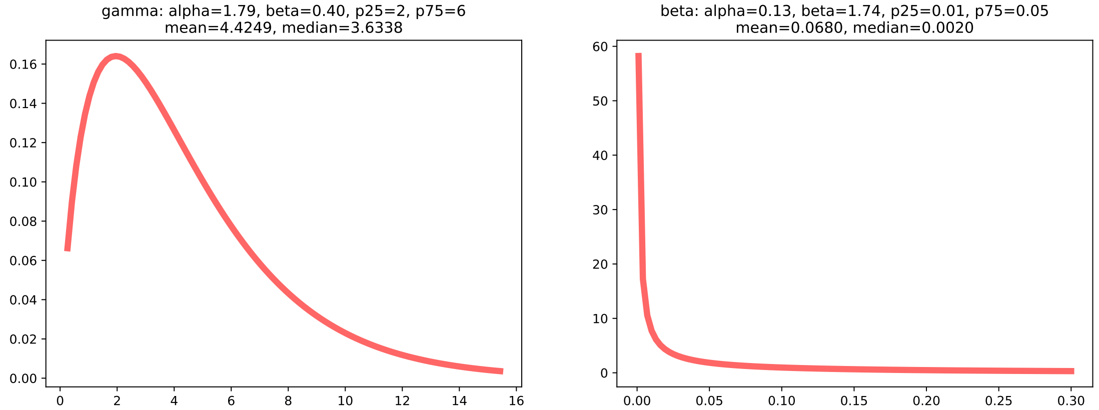


Figure 1. Distributions of company downtime duration (left) and proportion of customers lost (right)

With the previous information, using the procedure in Section 2.3, we infer the distributions of the downtime duration and the proportion of customers lost, Figure 1. The downtimes are modeled as a Gamma(1.79, 0.40) distribution for the company and a Gamma (1.21, 0.42) for the suppliers. Finally, the corresponding quartiles for the proportion of customers lost distribution after a successful attack are assessed as 0.01 and 0.05, corresponding to a Beta(0.13, 1.74) distribution.

We aggregate the previous costs and compute the expected direct cost for the company R = 517.16k EUR; the expected cost induced by the suppliers $\text{IR}_1 = 116.73\text{k}$, $\text{IR}_2 = 103.38\text{k}$ EUR; and the total expected cost TR = 737.27k EUR. Note again that the first supplier seems worse as its entailed expected loss is bigger.

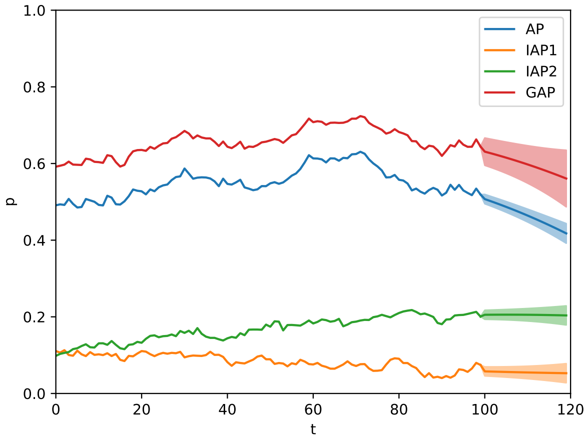


Figure 2. Evolution of AP, IAP and GAP

Finally, the procedure described above would be run periodically, obtaining new values for the probabilities and costs. An example for $T = 100$ time steps is shown in Figure 2, where we plot the evolution of the attack probability, the induced attack probabilities and the global attack probability. Here we can see how, at $t = 0$, we may prefer supplier 2 over supplier 1 since it seems to induce less risk to the company. However the probability induced by supplier 2 gets worse over time, reverting the situation. We also fit DLMs for the four probability indices, as described in Section 2.5. This allows us to forecast the different attack probabilities k-steps ahead. Figure 2 shows the expected value of the predictive distribution $X_{100+k}|D_{100}$ for $k = 1, ..., 20$, and the corresponding 95% predictive intervals.

5 DISCUSSION

As shown by the recent presence of several products in the market, SCCRM is a very relevant and current managerial problem given the increasing importance of cyber attacks and interconnectedness of organizations in modern economy. This has motivated us to provide a framework to support SCCRM. Given the reluctance of companies to release attack data, we have described how we may extract knowledge from security experts to obtain the parameters required to assess risk scores, based on information coming from a TIS in relation with attack vectors as well as the security posture and environment from a company and its suppliers. Besides, we have incorporated forecasting models that allow us to monitor risk and issue alarms. We may also use the provided information to rank suppliers, negotiate SLAs or use them for insurance purposes.

There are several ways to further advance this work. For example, should data about attacks be revealed, we could introduce schemes to learn about the involved parameters using our assessments as priors, through Markov Chain Monte Carlo procedures, see e. g. French and Rios Insua (2000). We could also incorporate additional impacts such as loss of productivity, loss of revenue or the increase of working hours.The model might also consider collaboration among suppliers to mitigate consequences when a service is disrupted or monitoring information sharing between suppliers to avoid unfair competition. Finally, we have included only direct suppliers to the company but we could also consider suppliers of suppliers, and beyond.

REFERENCES

Aqlan, F., & Lam, S. S. (2016). Supply chain optimization under risk and uncertainty: A case study for high-end server manufacturing. *Computers & Industrial Engineering*, *93*, 78–87.

Bandaly, D., Satir, A., Kahyaoglu, Y., & Shanker, L. (2012). Supply chain risk management-I: Conceptualization, framework and planning process. *Risk Management*, *14*(4), 249–271.

Bedford, T., & Cooke, R. M. (2001). *Mathematical Tools for Probabilistic Risk Analysis.* Cambridge University Press.

Berr, J. (2016). *Wannacry ransomware attack losses could reach $4 billion.* `https://www.cbsnews.com/news/wannacry-ransomware-attacks-wannacry-virus-losses/`.

Brown, R. G. (2004). *Smoothing, Forecasting and Prediction of Discrete Time Series.* Courier Corporation.

Clemen, R. T., & Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, *45*(2), 208–224.

Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford University Press.

Cox, A. L. (2008). What's wrong with risk matrices? *Risk Analysis*, *28*, 497–512.

Curkovic, S., Scannell, T., & Wagner, B. (2015). *Managing supply chain risk: Integrating with risk management.* CRC Press.

DiMase, D., Collier, Z. A., Carlson, J., Gray, R. B., & Linkov, I. (2016). Traceability and risk analysis strategies for addressing counterfeit electronics in supply chains for complex systems. *Risk Analysis*, *36*(10), 1834–1843.

EU. (2018). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

Fattahi, M., Govindan, K., & Keyvanshokooh, E. (2017). Responsive and resilient supply chain network design under operational and disruption risks with delivery lead-time sensitive customers. *Transportation Research Part E: Logistics and Transportation Review*, *101*, 176–200.

Foerstl, K., Reuter, C., Hartmann, E., & Blome, C. (2010). Managing supplier sustainability risks in a dynamically changing environment—sustainable supplier management in the chemical industry. *Journal of Purchasing and Supply Management*, *16*(2), 118–130.

Ghadge, A., Dani, S., Chester, M., & Kalawsky, R. (2013). A systems approach for modelling supply chain risks. *Supply Chain Management: An International Journal*, *18*(5), 523–538.

González-Ortega, J., Radovic, V., & Insua, D. R. (2018). Utility Elicitation. In Dias L., Morton A., Quigley J. (Ed.), *Elicitation: The Science and Art of Structuring Judgement* (pp. 241–264). Springer.

Hallikas, J., Karvonen, I., Pulkkinen, U., Virolainen, V.-M., & Tuominen, M. (2004). Risk management processes in supplier networks. *International Journal of Production Economics*, *90*, 47–58.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

Hubbard, D. W., & Seiersen, R. (2016). *How to Measure Anything in Cybersecurity Risk.* John Wiley & Sons.

Jüttner, U. (2005). Supply chain risk management: Understanding the business requirements from a practitioner perspective. *The International Journal of Logistics Management*, *16*, 120–141.

Kern, D., Moser, R., Hartmann, E., & Moder, M. (2012). Supply risk management: model development and empirical analysis. *International Journal of Physical Distribution & Logistics Management*, *42*(1), 60–82.

Latour, A. (2001). *A Fire in Albuquerque Sparks Crisis For European Cell-Phone Giants.* https://www.wsj.com/articles/SB980720939804883010.

Lovejoy, B. (2016). *Apple chipmaker TSMC says earthquake damage will hit production, full impact unclear as yet.* https://9to5mac.com/2016/02/15/tsmc-earthquake/.

McGrath, M. (2014). *Target Data Breach Spilled Info On As Many As 70 Million Customers.* https://www.forbes.com/sites/maggiemcgrath/2014/01/10/target-data-breach-spilled-info-on-as-many-as-70-million-customers/#2d90e3b2e795.

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

Petris, G., Petrone, S., & Campagnoli, P. (n.d.). *Dynamic Linear Models with R.* Springer.

Ritchie, B., & Brindley, C. (2007). Supply chain risk management and performance: A guiding framework for future development. *International Journal of Operations & Production Management*, *27*(3), 303–322.

Sharland, A., Eltantawy, R. A., & Giunipero, L. C. (2003). The impact of cycle time on supplier selection and subsequent performance outcomes. *Journal of Supply Chain Management*, *39*(2), 4–12.

Song, C., & Zhuang, J. (2017). Modeling a government-manufacturer-farmer game for food supply chain risk management. *Food Control*, *78*, 443–455.

Tang, C., & Tomlin, B. (2008). The power of flexibility for mitigating supply chain risks. *International Journal of Production Economics*, *116*(1), 12–27.

Thekdi, S. A., & Santos, J. R. (2016). Supply chain vulnerability analysis using scenario-based input-output modeling: Application to port operations. *Risk Analysis*, *36*(5), 1025–1039.

Tittel, E. (2017). *Comparing the top threat intelligence services.* `https://searchsecurity.techtarget.com/feature/Comparing-the-top-threat-intelligence-services`.

West, M., & Harrison, J. (2013). *Bayesian Forecasting and Dynamic Models.* Springer New York.

Zahiri, B., Zhuang, J., & Mohammadi, M. (2017). Toward an integrated sustainable-resilient supply chain: A pharmaceutical case study. *Transportation Research Part E: Logistics and Transportation Review*, *103*, 109–142.

Zsidisin, G. A., & Ritchie, B. (2008). *Supply Chain Risk: A Handbook of Assessment, Management, and Performance.* Springer Science & Business Media.

Annex 5: Paper: Structured Expert Judgement Issues in a Supply Chain Cyber Risk Management System

Structured Expert Judgement Issues in a Supply Chain Cyber Risk Management System

Abstract

The escalation of cyber threats is a major problem for supply chain managers with potentially enormous impacts affecting service availability and reputation, among other performance indicators. We sketch a framework and system to support supply chain cyber risk management. As data regarding impacts of cyber attacks are scarce and difficult to obtain, we describe how we acquire the required operational parameters through structured expert judgement techniques. We then describe how the whole framework is set up and implemented.

Keywords: Supply Chain Risk Management, Cybersecurity, Structured Expert Judgement.

1 Introduction

Organisations worldwide are suffering cyber attacks with important consequences. This is increasingly perceived as a major global problem as reflected e.g. in the World Economic Forum (2018) Global Risk Report, and becoming even more important as companies, administrations and individuals get more and more interconnected, facilitating the spread of cyberthreats. As an example, the recent WannaCry attack affected around 45,000 systems in large organisations worldwide, including the UK NHS, Renault and Telefónica, causing major service interruptions. Its ransonware caused estimated financial losses of nearly $4 billion, Evans et al. (2017). Another relevant example is the Target data breach, McGrath (2014), in which a cyber attack to that company through one of its suppliers leading to the loss of 70 million credit card details, entailing major reputational damage.

As a consequence, organisations face significant risks due to the need of using interconnected suppliers for their services. To alleviate such problem, the discipline of Supply Chain Cyber Risk Management (SCCRM) aims at implementing strategies to oversee cyber risks with the objective of mitigating service interruptions and decreasing their eventual impact, Redondo et al. (2018). To further complicate matters organisations are reluctant to release information concerning attacks for reputational reasons, as this could affect relations with stakeholders and entail a loss of customers (Pelteret and Ophoff, 2016). In order to supplement such lack of data, we may appeal to structured expert judgement elicitation techniques, Cooke (1991), O'Hagan et al. (2006), Clemen and Reilly (2013) and WHO (2019), exploiting the knowledge available from cyber security experts to support cyber risk management.

This paper briefly sketches a framework for SCCRM judgement in Section 2. Data regarding occurrences and impacts of cyber attacks are scarce and difficult to obtain, as companies are reluctant to reveal them for reputational reasons. Therefore, we need to rely on various expert judgement techniques to assess the various parameters[1] in the required impact and preference models in Sections 3, 4 and 5. Section 6 illustrates operational aspects of our framework and system. We end up with a discussion in Section 7.

2 A framework for SCCRM

We aim at supporting a company c interconnected with k suppliers in its supply chain cyber risk management activities. We briefly sketch the framework that we use for such purpose, with full technical details available in Redondo et al. (2018). Our focus will be on the expert judgement techniques and processes used to extract beliefs and preferences from experts to make the framework operational.

The company faces three cyber attack scenarios: *direct attacks*; *attacks to its suppliers not transferred to the company*, but affecting it through the unavailability of the corresponding product or service; and, finally, *attacks targeting the suppliers that are eventually transferred to the company*. Some of these attacks could be successful in the sense of producing noticeable harm in the company. We assume we have access to a Threat Intelligence Service (TIS) (Tittel, 2017) which compiles data, both for the company and its suppliers, about: *attack vectors*, such as the number of malware infected devices or the number of phishing attempts suffered; *the security environment*, e.g, the number of negative tweet mentions about the company and its suppliers; and, finally, *the security posture*, like the corresponding number of open ports or installed firewalls.

Based on the TIS data, and other available information, we aim at assessing the following basic ingredients for our SCCRM framework: the probabilities that the company

[1] For confidentiality reasons, data have been conveniently masked when presented

and its suppliers are attacked; the probability that an attack to a supplier gets transferred to the company; the impacts over the company associated with eventual attacks, direct or indirect, during the relevant security planning period; and how does the company evaluate various impacts. We then integrate such assessments to evaluate the supply chain cyber risks that the company could face and support risk management decisions both at strategic and operational level.

To begin with, we start by estimating the probability that the company and its suppliers are attacked through various attack vectors. First, we aggregate the information about the security environment and posture of the company and the suppliers through two different indicators which are a linear combination of several variables.Then, attack vectors are considered conditionally independent given the posture and the environment. We then assume that all attack probabilities may be modelled through logistic regressions with explanatory variables referring to the indicators corresponding to that attack vector and the security environment and posture. As companies are reluctant to provide their attack data, we indirectly estimate the corresponding logistic regression weights based on expert judgement as described below. Besides, we need to be able to assess the probabilities that eventual attacks to suppliers get transferred to the company, which we obtain directly also from experts. With all this information, we may assess the relevant attack probabilities directly to the company or indirectly through its suppliers, duly apportioning their sources.

We next need to estimate the impacts that an attack might have over the company, taking into account the three types of attacks mentioned above. The relevant impacts may vary across organisations. Some examples are the costs associated with the rupture of a service provided by a supplier, as in the Wannacry case with Telefonica; the costs associated with the unavailability of the company's service or product, as in the Wannacry case with the UK NHS; or the loss of reputation associated with a major attack, which might induce a loss of customers or stock value, as in the Target case. We typically use continuous distributions assessed based on quantiles obtained from experts. We then aggregate various impacts through a multi-attribute utility function, if we need to cater for risk attitudes, González-Ortega et al. (2018).

Based on the above probability and preference models, we assess the expected impacts and risks, duly apportioning them to various sources (suppliers, transferred attacks from suppliers, or direct attacks) and use such assessments to rank suppliers, negotiate service level agreements, or allocate cyber security risk management resources, including cyber insurance products, among other possibilities.

We present now how the expert judgement elicitation tasks described above are actually implemented and how we integrate all the information for risk management purposes.

3 Expert Calibration

We start by calibrating eight experts available based on their cyber security knowledge. After a training session, we passed them a questionnaire which served for weighting purposes.

3.1 The calibration process

We used reports, such as Kaspersky (2016) or Imperva (2016), to elaborate the questionnaire about cyber security attacks impacting SMEs and large companies. The questionnaire was built using the Google Forms tool and was ran initially with two colleagues to check for comprehensibility. It included ten questions concerning attack likelihoods and impacts. Two examples are:

> *Which was the number of new ransomware types over the last year?*

> *Which was the average cost in dollars of a ransomware incident over the last year?*

Before interviewing the actual experts, we suggested that they watched the YouTube *It's a Risky Life* videos 2, 3 and 4 to refresh the basic issues and concepts required for the session. When beginning the session, we also provided a review of the concepts, objectives and process to be followed. Some of the experts were interviewed physically, the remaining ones through the communication tool Skype. We introduced the process as follows

> *We present here a few general questions in relation with cyber security attacks, their likelihood and impacts. Answer them with what represents for you the quantities described.*
>
> *At each question, we shall ask you about an interval which covers with high probability (0.90) the actual value based on the 5% and the 95% quantiles and what is, according to you, the median value. For example, the interval could be [30-40] and the median value 35, so the answer would be [30-40], 35.*

Several motivating and warming up examples were included to further facilitate understanding, together with additional explanations about cognitive and motivational biases. In such a way, we tried to make sure that the experts understood the questions and response format correctly. They were also encouraged to ask for further clarification whenever they felt like. We also provided graphical support (fortune wheels) to facilitate the assessments. In the end, we verified whether the experts had answered all questions according to the given instructions and checked that the results had been submitted correctly, allowing them to modify responses upon reflection. One of the experts (Ex2) actually misunderstood the concepts underlying some of the questions, so we decided to suppress his responses from the study. We also eliminated questions $Q\{6, 7\}$ as the answers were astray, possibly because of inadequate wording on our behalf. The average duration of the sessions was approximately 1 hour, covering both the calibration and elicitation questions.

3.2 Expert response exploratory analysis

We start with some exploratory data analysis about the expert responses. We display their point and interval responses in Table 1, as well as the actual observations. We double checked whether some of the questions had been misunderstood (consider e.g. the responses Ex8-Q1, Ex6 for $Q\{9,10\}$ but the participants confirmed their results. Incidentally, this pointed out towards somewhat unknown areas about which even security experts are not sufficiently aware of.

Experts	Q1	Q2	Q3	Q4	Q5	Q8	Q9	Q10
Ex1	1,15,50	2,50,100	20,50,100	10,45,90	50,60,100	500,750,1000	50,60,75	30,40,40
Ex3	5,10,10	40,45,50	50,60,60	10,10,20	40,50,50	1000,1000,3000	75,80,85	10,10,15
Ex4	2,2,3	20,20,40	2,3,5	25,30,35	30,40,50	5000,15000,400000	70,75,80	70,75,80
Ex5	5,7,8	6,10,50	18,30,35	40,40,50	90,95,99	20000,1500000,2000000	20,25,30	1,2,6
Ex6	3,4,5	10,30,30	30,40,50	60,80,100	40,50,60	100,10000,15000	1,2,5	1,2,5
Ex7	1,3,5	1,40,100	1,90,100	1,10,100	1,80,100	1,12000,100000	1,80,100	1,60,100
Ex8	500,750,1000	70,90,100	10,15,25	50,65,80	70,80,90	50000,70000,150000	10,15,20	20,40,45
Obs	62	22	42	32	77	700	64	36

Table 1: Responses of experts and observations for the 7 retained experts and 8 retained questions

The boxplots in Figure 1, in which we have normalised the answers of question Q8 to [0,100], shows that, globally, the expert responses tend to cover the observations (blue line), except for Q1 (which was mostly underestimated) and Q8 (which was overestimated).

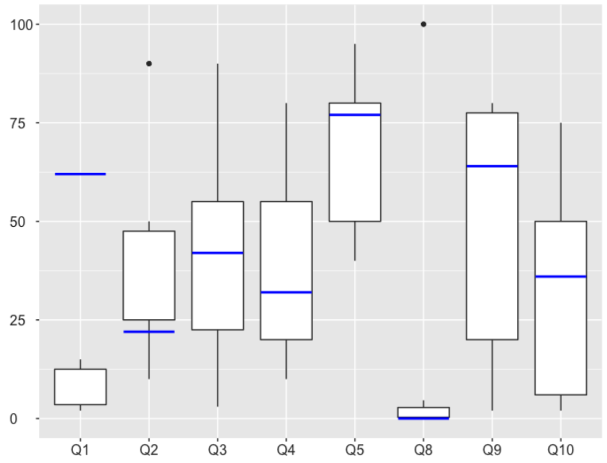


Figure 1: Expert answer box plots. Actual observations in blue. Extreme outliers removed

We next display the scatter plots of the experts' responses and their correlation matrix, Figure 2, in which we have also included the observations (as the responses of a ninth expert). We have removed the very extreme observation of Ex8 in Q1 from this representation.

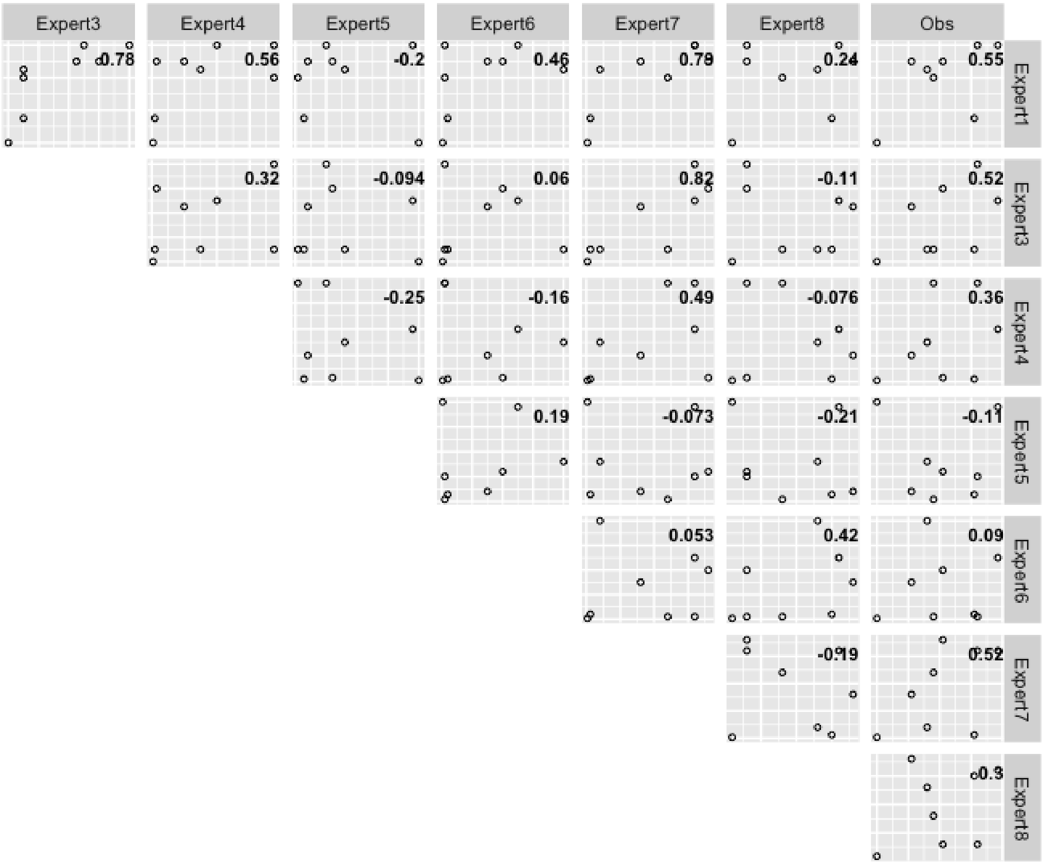


Figure 2: Scatterplot of expert answers and correlation matrix

We do not observe very high correlations. For example, if we use 0.5 as cutting value for noticeable correlations, only Ex1 with Ex$\{3, 4, 7\}$ and Ex3 with Ex7 show relevant correlation between themselves, and Ex$\{1, 3\}$ and Ex7 with the actual observations.

Table 2 summarises the performance of the experts on the 8 seed questions, presenting how many observed responses appeared in each of the intervals, compared with the expected responses.

Expert	Below 5th	5th to 50th	50th to 95th	Above 95th
Ex1	0	5	2	1
Ex3	4	0	0	4
Ex4	3	0	2	3
Ex5	3	0	1	4
Ex6	1	2	1	4
Ex7	0	6	1	1
Ex8	4	2	0	2
Exp	0.405	4.05	4.05	0.405

Table 2: Performance of experts on the 8 seed questions

3.3 Calibration

We used Excalibur (Lighttwist, 2018) to score the experts as described in Table 3, based on Cooke's classical method (Cooke, 1991), which provides also the calibration and information scores of the experts retained. We did not use DM optimisation and adopted a significance level of 0.001.

Expert	Calib.Sc.	Weight	Info.Sc.
Ex1	0.429	0.820	1.834
Ex3	0.000	0.000	3.440
Ex4	0.000	0.000	2.592
Ex5	0.000	0.000	2.143
Ex6	0.002	0.004	2.636
Ex7	0.145	0.176	1.168
Ex8	0.000	0.000	1.441

Table 3: Calibration, weights and information scores of experts

We performed a robustness analyses and found $Q\{5, 9\}$ to be the most influential over the results. Also, $Ex\{1, 7\}$ showed the lowest discrepancy.

4 Attack probabilities' assessment

We describe now how to extract the cyber security knowledge from the experts to enable us building our SCCRM model. For this, we created a second questionnaire with Google Forms. To start with, the questionnaire included a short introduction outlining the procedure to answer the questions:

> *The following questions will aid us in extracting your expertise on cyber security so as to build a model that allows us to forecast sufficiently important attacks to a company. Please feel confident. There are no right or wrong answers. We shall be posing questions that take advantage from your cyber security expertise.*

The questions were divided into two groups: first, attack probability questions and, then, questions related with the environment and the posture.

4.1 Attack probabilities

With this first group of questions, we aimed at obtaining for each expert i the probability p_i of various events. We then aggregate the probabilities through $p = \sum \omega_i p_i$ (effectively, $i \in \{1, 6, 7\}$), where the weights ω_i are the result of the calibration process in Section 3.3. Based on such probabilities, at this stage we extracted the judgements required to obtain the logistic regression parameters mentioned in Section 2. Each question included a description of a relevant scenario with the answer interpreted as a probability.

We illustrate the procedure for the specific attack due to malware infections. Our TIS is able to detect three types of malware. Thus, the model has three coefficients besides β_0 and the logistic regression model we use is

$$Pr(y = 1 | \beta_0, .., \beta_3, n_1, ..., n_3) = \frac{\exp(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{n})}{1 + \exp(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{n})} \tag{1}$$

where $\mathbf{n} = (n_1, n_2, n_3)$ is the vector containing the counts of the three types of malware, the β's are the logistic parameters and $y = 1$ indicates that the attack through malware was successful (sufficiently harmful). First, we ask the experts for the attack probability p_0 in a scenario in which no such infections were detected by the TIS, $\mathbf{n} = (0, 0, 0)$. The actual question posed to the experts was:

> *Assume that the TIS has detected no evidence of malware infections in your network, what would be the probability of actually suffering an attack based on malware?*

We then aggregate the responses of the (three) experts to obtain p_0. Since we are assuming that the attack probability follows equation (1) and no infections are found, we have $\beta_0 = \sigma(p_0)$, with $\sigma(p) = \log\left(\frac{p}{1-p}\right)$. Table 4 (first row) includes the responses of the three experts. We then obtain the assessment $p_0 = 0.011$ and, consequently, $\beta_0 = -4.521$.

j	Ex1	Ex6	Ex7	p_j	m	β_j
0	0.010	0.050	0.010	0.011	0	−4.521
1	0.020	0.100	0.020	0.020	5	3.746
2	0.150	0.100	0.020	0.127	5	4.135
3	0.350	0.200	0.050	0.296	1	3.655

Table 4: Responses of experts and estimated parameters.

We compute the remaining required coefficients in a similar manner, asking the experts to provide an estimate of the attack probability p_j if the TIS detects now a certain number m of infections belonging to the j-th level of the attack vector and none of the rest, aggregating the responses and solving for the corresponding β_j. We need to ask at least one question per coefficient to each expert. A typical question would be:

> *Assume that the TIS has detected 5 malware infected devices of level 1 in your network, what would be the probability of actually suffering an attack based on malware?*

The previous question proposes an scenario in which $n_k = m$ if $k = j$ and $n_k = 0$ if $k \neq j$, with $m = 5$ and $j = 1$. After aggregating the expert responses in p_j, we make

$$\beta_j = \frac{\sigma(p_j) - \beta_0}{m}.$$

Table 4 includes the responses of the experts, their aggregation and the corresponding parameters.

Note that the coefficients regarding different infection levels are independent between them and depend only on β_0. We also perform additional questions for each level j with different values of m to check whether the experts are consistent in their answers, as similar β_j values should be obtained.

We perform the above for each attack vector detectable bt the TIS..

4.2 Environment and posture

We describe now how to incorporate information about the security environment. Examples of the incumbent variables include the number of negative mentions about the company in major social media and the number of mentions in security blogs. To do so, we first define an index e which assesses the security environment through a multi-criteria value function, González-Ortega et al. (2018). Let e_i be the i-th environment variable, $i = 1, \dots, k$, captured by the TIS, rescaled to $[0, 1]$; we assume that the bigger e_i is, the worse the security environment is. We aggregate all the variables $e = \sum_{i=1}^{k} \lambda_i e_i$ with $\lambda_i \geq 0$, $i = 1, \dots, k$ and $\sum_{i=1}^{k} \lambda_i = 1$. A similar procedure is applied to combine the posture variables into a posture indicator, $l = \sum_i v_i l_i$, with $\sum_i v_i = 1$ and $v_i \geq 0$, $\forall i$.

We determine the weights λ_i by asking experts to compare pairs of security environment contexts, identifying the corresponding system of equations and solving it. For example, for the variables mentioned above we could pose the question:

> *How would you weight the relative importance of the number of mentions in security blogs and negative mentions in social media regarding the likelihood of receiving a successful harmful attack? Both numbers should add up to 100; the higher the weight, the more impact you give to such variable (in the sense of deeming more likely an attack).*

In this case, the expert's answer should be a pair of numbers (λ_1, λ_2), which add up to 100. If both are 50, the expert considers both variables equally relevant when assessing the security environment of the company, leading to the equation $\lambda_1 \times e_1 = \lambda_2 \times e_2$. We introduce iterative schemes to come out with the pairs.

Ex	T_1	T_2	T_3	w_1	w_2	w_3	w_4
1	(70, 30)	(30, 70)	(50, 50)	0.292	0.125	0.292	0.292
3	(60, 40)	(30, 70)	(60, 40)	0.235	0.157	0.365	0.243
4	(10, 90)	(10, 90)	(50, 50)	0.006	0.052	0.471	0.471
5	(15, 85)	(40, 60)	(85, 15)	0.060	0.340	0.510	0.090
6	(60, 40)	(30, 70)	(70, 30)	0.257	0.171	0.400	0.171
7	(60, 40)	(40, 60)	(50, 50)	0.273	0.182	0.273	0.273
8	(70, 30)	(30, 70)	(50, 50)	0.292	0.125	0.292	0.292
				0.202	0.165	0.372	0.262

Table 5: Environment responses of experts and weights

Table 5 includes the responses of our experts for the above questions with four environment variables and the seven retained experts. The minimum number of questions to be posed to each of them is $k-1$, as the k-th equation relies on the restriction that all weights should add up 100. We select overlapping pairs of environment variables for the questions, comparing variables 1 and 2; 2 and 3; and, so on, until the $(k-1)$-th and k-th variables are compared. To mitigate biases, the order in which the questions are posed is randomised. Moreover, additional questions using other combinations of variables are added to check for consistency. We then find the value function corresponding to each expert and aggregate them with equal weights[2].

Finally, we incorporate the environment and posture indices into the model using the

[2] Note that this refers to value judgements, not belief judgements as in Section 3

procedure described in Section 4.1. First, we construct an scenario where no infections are found by the TIS and ask the experts how much would the attack probability increase assuming a certain value for one of the security environment variables, as in the following example:

> *Assuming that the TIS has detected no evidence in the network concerning malware infections, how much would the attack probability increase if we detect a value for the first environment variable equal to 10 and none for the others?*

Since the rest of the environment variables are zero, we can compute the index value as $e = w_1 \times m$, where $m = 10$, and expand the attack vector. Continuing with the example of malware infections, we have $\mathbf{n} = (0, 0, 0, e, 0)$, where the last two elements correspond now to the environment and posture indices. Then, since we are asking for the increase in probability and assuming the expert's answer is Δ, we have $p_e = p_0 + \Delta$ and again

$$\beta_e = \frac{\sigma(p_e) - \beta_0}{m}.$$

We apply the same procedure to incorporate the security posture. As before, consistency questions are posed (using different environment variables and values for m).

4.3 Probability of attack transfer

Finally, we also ask the experts questions regarding the probability of an attack being transferred from a supplier to the company. There are as many questions as attack types, as we use the same probabilities for all suppliers. For each of the types, we aggregate the experts' probabilities. As an example, the probabilities for the transfer of malware attacks were 0.3, 0.1 and 0.2, respectively, for experts 1, 6 and 7. The final aggregated probability, using the weights in Section 3.3, is 0.282.

5 Attack impact assessment

The final group of questions refers to information concerning the impacts of a successful attack. They are different from the previous questions as they are company specific: for example, even if the consequence of a successful attack may be the same for two companies, like losing 1% of their customers, their economic impact in Euros will typically differ; the unavailability period will depend on the company's recovery capacity; and, finally, distinct companies assess impacts differently. Thus, these questions are answered by in-company experts instead of by general cyber security experts. In the same manner, such type of information from the suppliers may not be readily available as its experts could have no incentives to answer the required questions or even be unavailable for the necessary elicitation exercise. In this case, the in-company experts may try to estimate what would be the answer to the questions for the corresponding suppliers.

As before, we introduce a training session with the local experts as well as an eventual aggregation procedure, should there be several of them available.

5.1 Relevant impacts

Relevant cyber attack impacts might change across organisations. In our supply chain area, we have focused on downtimes, for both the company and its suppliers, and the induced reputational damage.

We model the downtimes in hours with Gamma distributions (for the company and suppliers). To obtain estimates of the parameters for these distributions, we ask the experts for at least two of its quantiles, for instance, the first and the third quartiles. An example question would be:

> *What is the duration of the downtime in hours due to malware at your or-*

ganisation such that you would expect 25% of the downtimes to be below this value?

Once we obtain the quantiles, we use a least squares approach to estimate the parameters of the distributions, as in Morris et al. (2014). We may ask additional quantiles to perform consistency checks. As an example, in one case, an expert provided as first and third quartiles, respectively, 2 and 6. The best fitting gamma distribution was Gamma(1.79, 0.40). After obtaining the distributions, we may compute centrality measures such as the mean or the median, if required, or use them for simulation purposes.

We performed similarly in relation with reputational damage, estimating the proportion of lost customers due to a certain type of attack with a Beta distribution.

5.2 Aggregating impacts

We aggregate the three types of impacts taking into account the costs associated with unavailable services and the percentage of lost customers due to reputational damage. The general expression would be $c = \kappa_s \times i_s + \kappa_c \times i_c + \kappa_d \times n_d \times d$, where κ_s, κ_c represent, respectively, the average cost per hour of supplier and company interruption; i_s and i_c, respectively, symbolise the supplier and company service unavailability duration in hours; κ_d, the cost lost per customer; n_d, the number of customers; and, d, the proportion of lost customers. For the required additional information, the corresponding questions are straightforward and directly posed to in-company experts.

5.3 Utility elicitation

Should we wish to cater for the company's risk attitude we would introduce an utility function. A simple but very useful form of utility function arises when the relative risk aversion is set to a constant, in which case we have $u(c) = 1 - \exp(-\rho\, c)$ with $\rho > 0$. To assess the risk tolerance parameter ρ, Keeney and Raiffa (1993), we ask the DM to determine the largest stake c_{max} for which she would accept the 50-50 gamble

$$\begin{cases} 2c_{max} & \text{with probability } \frac{1}{2}, \\ -c_{max} & \text{with probability } \frac{1}{2}. \end{cases}$$

This leads to the approximate expression $\rho \approx \frac{1}{2c_{max}}$ (González-Ortega et al., 2018). Consistency checks would lead us to elicit additional values and iterative attempts to assess such value.

6 Operational uses

We now have the necessary components to develop our SCCRM framework. We begin first by sketching some of its potential uses. We then describe how to implement it, and finally provide a numerical example.

6.1 Some uses

The above information may be summarised in several measures and indices that may be used for risk monitoring and management purposes. These include attack probabilities through different attack vectors, both through the various suppliers or the company, resulting in a successful attack; the direct attack to the company probability; the induced attack probabilities; and, the total attack probability. Recall that if an attack is successfully transferred from a supplier, there are unavailability and reputational costs. Thus, we include also the expected impact due to direct attacks, the expected impact induced when a supplier is attacked and the total expected impact generated. Finally, we would also employ the corresponding expected utilities.

As an example, we provide the expressions for two of such indices whose use we illustrate in Figure 3. First, the attack probability to the company c through a specific attack vector a is

$$p_c^a = \frac{\exp(\beta_0^a + \boldsymbol{\beta}^a \cdot \mathbf{n}_c^a)}{1 + \exp(\beta_0^a + \boldsymbol{\beta}^a \cdot \mathbf{n}_c^a)},$$

where $\mathbf{n}_c^a$ represents the a-th attack vector count for the company c, including the environment and posture indicators. Based on them, the direct Attack probability to the company is

$$\mathrm{AP}_c = \sum_{k=1}^{|\mathcal{A}|} \sum_{\mathcal{I} \in \mathcal{C}_{\mathcal{A},k}} \left(\prod_{a \in \mathcal{I}} p_c^a \prod_{a \in \mathcal{A} \setminus \mathcal{I}} (1 - p_c^a) \right),$$

where $\mathcal{C}_{\mathcal{A},k}$ is the set of all possible combinations of k elements taken from $\mathcal{A}$. We describe now how we use in our framework the risk indicators:

- *Risk management.* We set up warning and critical level alarms for the indices to advise when specially dangerous situations have been detected. When such levels are reached, as we have apportioned them to various sources (vector attacks and suppliers), we may point to the most critical ones to try to act over them.

- *Risk forecasting.* Each of the indices mentioned above may be viewed as an observation of a time series. We may, therefore, introduce forecasting models (specifically, we use dynamic linear models. Petris et al. (2009)) for such series to forecast whether we shall reach the critical levels (through long-term forecasts) or which levels should we normally expect in the near future (through short-term forecasts). These forecasts can also detect sudden changes in the behaviour of the series and, consequently, potential security issues.

- *Supplier negotiations.* We can use the indices produced to rank suppliers according to the risk they induce. We can also use them to negotiate minimum induced security requirements to demand actions to suppliers or negotiate service level agreements, say requiring to maintain a risk induced level below a certain value to preserve business continuity.

- *Insurance.* We may use the risk series generated to demonstrate low risk levels at the incumbent company and, consequently, negotiate lower insurance premiums. Alternatively, from the point of view of an insurance company, we could introduce insurance products with variable premium depending on the risk indices integrated over time. For example, an incentive could be introduced if, say, the average and maximum risk indices, fall below a certain level over the contracted period of time.

6.2 Operations

The above framework has been implemented to support a dynamic approach to SCCRM in conjunction with an available TIS. For a given company, the TIS periodically gathers data, and the system computes the risk indices, provides various forecasts, issues warnings and performs update operations as follows:

1. Obtain new attack vectors evaluating the security posture and environment of the company and its suppliers.

2. Compute attack probabilities for suppliers and company, for various attacks and globally.

3. Estimate the expected impacts and utility for the company.

4. Launch alarms depending on limits defined.

5. Display risks associated with attack vectors and suppliers.

6. Predict risks for the next k-periods ahead.

7. (Update the probability models).

8. Proceed to the next period.

All of the tasks have been described above, except for the seventh one which refers to updating the parameter distributions in the logistic regression and impact models through MCMC methods as standard in Bayesian inference, see French and Insua (2000). This would be possible as long as the company releases relevant data about attacks.

Figure 3 provides a trace of the model which runs periodically acquiring new probabilities and costs. Specifically, we show the evolution for $T = 100$ time steps of the direct attack probability (AP), the induced probabilities from two suppliers (IAP1, IAP2) and the global attack probability GAP. Here, we can observe that from $T = 0$ to $T = 40$ suppliers 1 and 2 induced similar risks. However, from $T = 40$ we may prefer supplier 1 since it seems to induce a lower risk to the company. We may fit DLMs (West and Harrison (2006); Petris et al. (2009)) to forecast the attack probabilities k-steps ahead. Figure 3 presents the predictive distribution for $k = 1, \ldots, 20$, from period 100, with the corresponding predictive intervals.

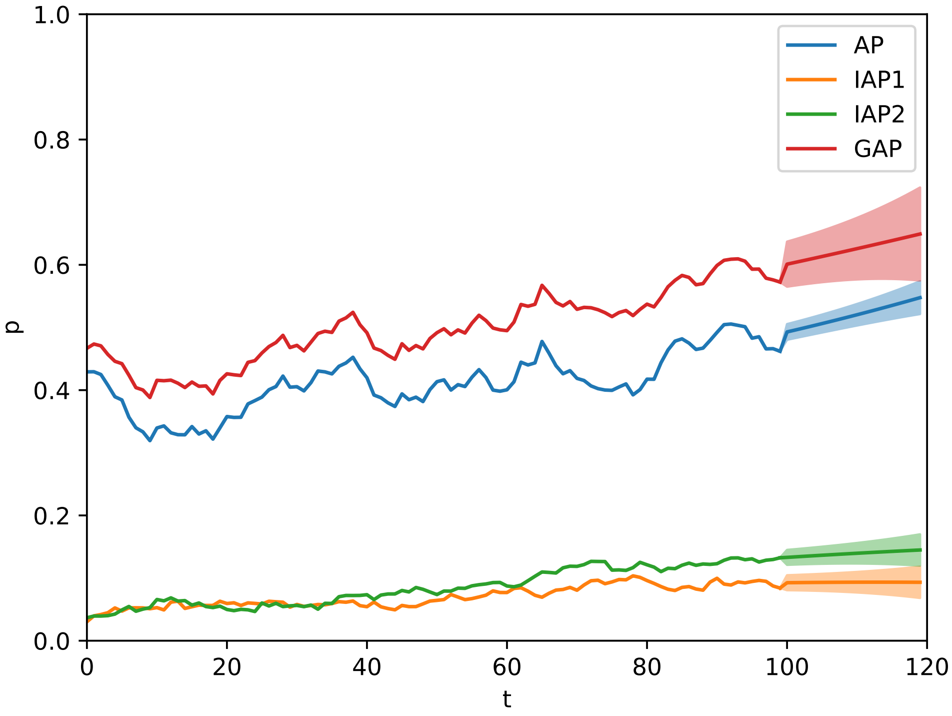


Figure 3: Trace of risk indices over time

7 Discussion

The proliferation of cyber attacks and the increasing interconnectedness of organisations is framing the new field of SCCRM with several commercial solutions available. For reputational reasons, organisations are reluctant to release data concerning attacks. Therefore, we have sketched an approach to SCCRM which uses SEJ techniques to assess the parameters required to make the approach implementable. We have focused on how suppliers may affect organisations, but the ideas extend to the impact of suppliers of suppliers, and so on.

We have presented here the SEJ aspects of the framework as well as its operational implementation. We have covered issues concerning calibration of experts; eliciting attack probabilities indirectly through logistic regression models; aggregating environment and posture variables through multi-attribute value functions; directly eliciting transfer attack probabilities; eliciting impact distributions through quantiles; and, finally, eliciting utilities to cater for risk attitudes. We have also described how such information is integrated for various risk management purposes. Mathematical details may be seen in Redondo et al. (2018).

The whole framework has been implemented through Python routines based on a specific TIS and is running successfully supporting several companies in their SCCRM duties. The experience gained will allow us to further refine the framework; improve and/or expand the attack vectors as well as the assessment of the environment and posture.

Acknowledgements

The authors acknowledge the support by the Spanish Ministry of Economy and Innovation program MTM2014-56949-C3-1-R, MTM2017-86875-C3-1-R and RTC-2017-6593-7, the ESF-COST Action IS1304 on Expert Judgement and the AXA-ICMAT Chair on Adversarial Risk Analysis. This work has also been partially supported by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Program for Centers of Excellence in R&D (SEV-2015-0554), and has received funding from the European Union's H2020 Program for Research, Technological Development and Demonstration, under grant agreement no. 740920 (CYBECO). AR work acknowledges the Spanish Ministry of Economy for his PTA position.

References

Clemen, R. T. and Reilly, T. (2013). *Making Hard Decisions*. Duxbury Press.

Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press.

French, S. and Insua, D. R. (2000). *Statistical Decision Theory*. Wiley.

González-Ortega, J., Radovic, V., and Insua, D. R. (2018). Utility Elicitation. In Dias L., Morton A., Quigley J., editor, *Elicitation: The Science and Art of Structuring Judgement*, pages 241–264. Springer.

Imperva (2016). DDoS Threat Landscape Report 2015 - 2016. `https://lp.incapsula.com/rs/804-TEY-921/images/2015-16%20DDoS%20Threat%20Landscape%20Report.pdf`.

Kaspersky (2016). Story of the year: The ransomware revolution. `https://securelist.com/kaspersky-security-bulletin-2016-story-of-the-year/76757/`.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Lighttwist (2018). Excalibur. `http://www.lighttwist.net/wp/excalibur`.

McGrath, M. (2014). Target Data Breach Spilled Info On As Many As 70 Million Customers. `https://www.forbes.com/sites/maggiemcgrath/2014/01/10/target-data-breach-spilled-info-on-as-many-as-70-million-customers/#2d90e3b2e795`.

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4.

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

Pelteret, M. and Ophoff, J. (2016). A review of information privacy and its importance to consumers and organizations. *Informing Science*, 19:277–301.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Springer.

Redondo, A., Torres-Barran, A., Rios Insua, D., and Domingo, J. (2018). Assessing supply chain cyber risk management. *Tech. Report*, pages 1–20.

Tittel, E. (2017). Comparing the top threat intelligence services. `https://searchsecurity.techtarget.com/feature/Comparing-the-top-threat-intelligence-services`.

West, M. and Harrison, J. (2006). *Bayesian Forecasting and Dynamic Models*. Springer.

WHO (2019). WHICH. In French, S., Nane, T., Hanea, A., Bedford, T., editor, *Structured Expert Judgement in Risk and Decision Analysis*. Springer.

World Economic Forum (2018). The Global Risks Report 2018. `http://www3.weforum.org/docs/WEF_GRR18_Report.pdf`.

Annex 6: Paper: Augmented Probability Simulation Methods for Non-cooperative Games

Augmented Probability Simulation Methods for Non-cooperative Games

Abstract

We provide an augmented probability simulation framework to solve non-cooperative games, focusing on sequential problems. We include approaches to approximate subgame perfect equilibria under common knowledge conditions, assess the robustness of such solutions and, finally, approximate adversarial risk analysis solutions when lacking common knowledge. Cyber security examples serve for illustration.

Keywords: Games, Decision analysis, Adversarial risk analysis, Augmented probability simulation, Markov chain Monte Carlo, Cyber security.

1 Introduction

Non-cooperative game theory refers to conflict situations in which two or more agents make decisions whose payoffs depend on the implemented actions of all of them, and, possibly, on some random outcomes. Agents aim at maximising their payoffs. Under common knowledge assumptions about the agents' preferences and beliefs, the analysis is pervaded by Nash equilibria, and related refinements, which constitute a prediction of the decisions to be made by the agents. Ozdaglar and Menache (2011) provide a review, whereas Heap and Varoufakis (2004) include an in-depth critical assessment. Adversarial risk analysis (ARA), Banks et al. (2015), provides an alternative decision analytic approach aimed at one-sided prescriptive support to one of the intervening agents based on a subjective expected utility model treating the adversaries' decisions as uncertainties. Their (random) optimal actions are predicted taking into account the uncertainty about the adversaries' probabilities and utilities in an expected utility model of their behaviour. In contrast with game theoretic approaches, the standard common knowledge hypothesis is not assumed.

Our realm in this paper is within algorithmic game theory, Nisan et al. (2007), and algorithmic decision theory, **?**, in that we aim at providing efficient algorithms to approximate solutions for game theoretic problems, both in the standard and the ARA approaches. For cases in which an analytical solution is not available or computationally expensive, simulation based approaches can be utilized. Among those, Monte Carlo (MC) methods are straightforward to use and widely implemented. However, they can be inefficient in certain conditions such as in presence of a high number of decision alternatives for the agents. For instance, counter-terrorism and cyber security may involve thousands of possible decisions, and there could be large uncertainties associated with the goals and resources of the terrorists. This can result in computational challenges especially in cases where model uncertainty dominates, Rios Insua et al. (2009). Sampling procedures that focus on high-probability events have the potential to handle such computational challenges. In particular, we explore how augmented probability simulation (APS) may be used to compute game theoretic solutions. APS is a powerful simulation based methodology used to approximate optimal solutions in decision analytic problems, Bielza et al. (1999), as reviewed in Appendix 1.

We focus on sequential games with two agents: one of the agents makes a first decision which is observed by the other one, who then makes his decision. We start by computing subgame perfect equilibria with APS. We then apply robustness concepts to assess such solution. If it is not robust, we use the ARA approach to find an alternative solution, with the aid of APS. Again, we criticize it through sensitivity analysis concepts. We illustrate the ideas with examples concerning cyber security and end up with some discussion.

2 Equilibria in sequential games through APS

2.1 The basic approach

We consider sequential games with two agents. As an example, consider a case in which a company deploys certain cyber security controls and then, having observed them, a hacker decides whether he launches a DDoS attack against such company. These games have received various names in the literature including sequential Defend-Attack (Brown et al., 2006) or Stackelberg (Gibbons, 1992).

To fix ideas, consider a case in which a Defender (she) chooses her defense $d \in \mathcal{D}$ and, then, an Attacker (he) chooses his attack $a \in \mathcal{A}$, after having observed d Except when noted, we assume that both $\mathcal{D}$ and $\mathcal{A}$ are finite.. The corresponding bi-agent influence diagram (Ortega et al., 2018) is shown in Figure 1. The arc between nodes D and A reflects that the Defender choice is observed by the Attacker. The consequences for both participants depend on the success θ of the attack. Each decision maker has a different assessment on the probability of the result of the attack, which depends on the defense and attack adopted, designated $p_D(\theta|d,a)$ and $p_A(\theta|d,a)$. The utility function of the Defender $u_D(d,\theta)$ depends on her chosen defense and the result of the attack. Similarly, the Attacker's utility function has the form $u_A(a,\theta)$.

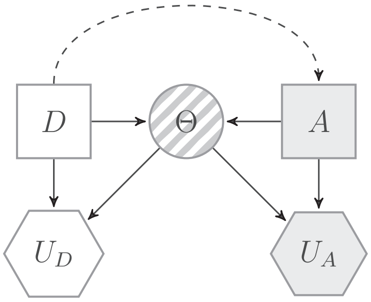


Figure 1: The two player sequential decision game.

The standard game theoretic solution does not require the Attacker to know the Defender's probabilities and utilities, since he observes the Defender's actions. However, the Defender must know the Attacker's utilities and probabilities (u_A, p_A), the common knowledge condition in this case. We, then, proceed as follows. First, we compute both agents' expected utilities at node Θ in Figure 1

$$\psi_A(a,d) = \int u_A(a,\theta)p_A(\theta|d,a)\,\mathrm{d}\theta, \tag{1}$$

$$\psi_D(d,a) = \int u_D(d,\theta)p_D(\theta|d,a)\,\mathrm{d}\theta. \tag{2}$$

Then, we compute the Attacker's best response to the Defender's action d

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

Knowing this, the Defender's optimal action is

$$d^*_{\mathrm{GT}} = \arg\max_{d \in \mathcal{D}} \, \psi_D(d, a^*(d)).$$

The solution $(d^*_{\mathrm{GT}}, a^*(d^*_{\mathrm{GT}}))$ is a Nash equilibrium and, indeed, a sub-game perfect equilibrium, Heap and Varoufakis (2004). Note that we are using backwards induction for both agents, switching from the Attacker to the Defender problem as required.

A generic MC approach to solving the problem is shown in Algorithm 1. From a computational perspective, it requires generating $|\mathcal{D}| \times (|\mathcal{A}| \times Q + P)$ samples, in addition to the cost of the final optimization and the $|\mathcal{D}|$ inner loop optimizations, where $|\cdot|$ designates the cardinal of the corresponding set. When the sets $\mathcal{A}$ and/or $\mathcal{D}$ of alternatives are continuous, we may introduce further discretization and/or sampling steps to appropriately

```
input: $P$, $Q$
for $d \in \mathcal{D}$ do
    for $a \in \mathcal{A}$ do
        Generate $Q$ samples $\theta_1, \dots, \theta_Q \sim p_A(\theta | a, d)$
        Approximate $\widehat{\psi}_A(a, d) = \frac{1}{Q} \sum u_A(a, \theta_i)$
    Find $a^*(d) = \arg\max_a \widehat{\psi}_A(a, d)$
    Generate $P$ samples $\theta_1, \dots, \theta_P \sim p_D(\theta | a^*(d), d)$
    Approximate $\widehat{\psi}_D(d) = \frac{1}{P} \sum u_D(d, \theta_i)$
Find $d^* = \arg\max_d \widehat{\psi}_D(d)$
```

Algorithm 1: MC approach to solve a sequential Defend-Attack problem

explore the alternatives available, or we could use a regression metamodel, as explained in our final discussion.

In some cases, when dealing with decision dependent uncertainties, as is the case of sequential Defend-Attack games, MC approaches may turn out to be extremely heavy computationally. They require sampling from $p_D(\theta|d,a)$ and $p_A(\theta|d,a)$ for the defender's and the attacker's problem respectively, and this entails loops over decision spaces $\mathcal{D}$ and $\mathcal{A}$. When $\mathcal{D}$ or $\mathcal{A}$ are high dimensional, considering the whole decision space as in MC, this is unfeasible. We propose APS algorithms to address such challenges.

2.2 An APS approach

Recalling that u_A and p_A are assumed to be known by the Defender, the required common knowledge assumption in this case, we provide an APS approach to approximate game theoretic solutions for sequential defend-attack games. Indeed, it is a nested APS framework similar to folding back a tree. For a given d, we introduce an artificial distribution, assuming that $u_A(a,\theta)$ is non-negative, in relation with Equation (1)

$$\pi_A(a,\theta|d) \propto u_A(a,\theta) p_A(\theta|d,a). \tag{3}$$

Its marginal on attacks satisfies $\pi_A(a|d) \propto \psi_A(a,d)$. Consequently, the optimal attack given the defense d is such that

$$a^*(d) = \mathrm{mode}(\pi_A(a|d)).$$

Moving backwards in the tree, we introduce the artificial distribution

$$\pi_D(d, \theta | a^*(d)) \propto u_D(d, \theta) p_D(\theta | d, a^*(d)). \tag{4}$$

Its marginal $\pi_D(d|a^*(d))$ is proportional to $\psi_D(a^*(d), d)$, so that

$$d^*_{GT} = \text{mode}(\pi_D(d|a^*(d))). \tag{5}$$

Consequently, we may introduce a APS scheme to sample from $\pi_D(d|\theta, a^*(d))$, such that, in a preprocessing phase, computes $a^*(d)$ for each d using another APS based on $\pi_A(a, \theta|d)$. In principle, we sample from both artificial distributions through a Gibbs sampler, Casella and George (1992): in the attacker's APS, for each d, we sample iteratively from $\pi_A(a|\theta, d)$ and $\pi_A(\theta|d, a)$, whereas, in the defender's APS, we sample iteratively from $\pi_D(d|\theta, a^*(d))$ and $\pi_D(\theta|d, a^*(d))$. The resulting procedure is summarized in Algorithm 2. This can be preferred to MC when the Attacker have high dimensional decision spaces, as we avoid iterating through the attacker's alternatives.

The computational complexity of Algorithm 2 does not depend on the dimension of the attacker's decision space. This could be crucial when $|\mathcal{A}|$ is very big or continuous. Recall that MC requires discretization of such space whereas APS does not. In particular, Algorithm 2 requires $2 \cdot (|\mathcal{D}| \cdot M + N)$ samples plus the cost of convergence checks and $|\mathcal{D}| + 1$ mode approximations. Thus, overall, Algorithm 2 would be more efficient than the MC approach whenever the number of attacker's alternatives is big (or continuous) and the number of defender's alternatives is small.

If the draws from the conditional distributions in Algorithm 2 are not readily available, we could use Metropolis-Hastings algorithm. The resulting algorithms for both the Defender and Attacker are provided in Appendix 2. Turning to Metropolis-Hastings sam-

```
input: $N$, $M$
initialize: $a^{(0)}$, $\theta^{(0)}$
for $d \in \mathcal{D}$ do
    for $j = 1$ to $M$ do
        Draw $\theta_A^{(j)}$ from $\pi_A(\theta | d, a^{(j-1)})$
        Draw $a^{(j)}$ from $\pi_A(a | \theta_A^{(j)}, d)$
    Compute mode of $M$ draws $\{a^{(j)}\}$ and record it as $a^*(d)$
initialize: $d^{(0)}$, $\theta^{(0)}$
for $i = 1$ to $N$ do
    Draw $\theta_D^{(i)}$ from $\pi_D(\theta | d^{(i-1)}, a^*(d^{(i-1)}))$
    Draw $d^{(i)}$ from $\pi_D(d | \theta_D^{(i)}, a^*(d))$
Compute the mode of $N$ draws $\{d^{(i)}\}$ and record it as $\widehat{d}^*_{GT}$
```

Algorithm 2: Nested APS

pling, also allows us to get rid of the loop over the defender's decision space. Let us call d and θ the current samples in the Metropolis scheme of the defender's APS. Within each iteration, we need to sample a candidate $\tilde{d}$ for the defender's alternative from the proposal distribution. Once it is sampled, we invoke an inner APS to compute $a^*(\tilde{d})$. Finally, we accept the sample with probability $\pi_D(\tilde{d}, \tilde{\theta} | a^*(\tilde{d})) / \pi_D(d, \theta | a^*(d))$, being $\tilde{\theta}$ the candidate for uncertainty parameter. The computational complexity of this new procedure does not depend on the dimensions of the attacker's and the defender's decision spaces. This will be the optimal choice when facing a problem where the cardinality of these spaces is very big or when they are continuous.

2.3 Sensitivity of the game theoretic solution

In the above setting, we could contend that since we are supporting the Defender, we know (u_D, p_D) reasonably well. However, information about (u_A, p_A) may be not that precise, since it essentially requires the Attacker to reveal such judgments. This is questionable in application areas in which information is concealed and hidden to adversaries, including cybersecurity.

We may perform a sensitivity analysis by considering that the Attacker's utilities and probabilities are modeled through classes $u \in \mathcal{U}_A$, $p \in \mathcal{P}_A$, which summarize the information available from the Attacker obtained through informants, leakage or earlier interactions. For each pair (u, p), we compute the Nash defense $d^*_{u,p}$, using the techniques from sections 2.1 and 2.2. After that, we need to assess whether the game theoretic solution remains reasonably stable for the allowed perturbations of u and p. One possibility could be focusing on the regret $r_{u,p}(d^*_{\mathrm{GT}})$ given by the difference in expected utility between the Nash defense d^*_{GT} and the Nash defense $d^*_{u,p}$ for (u, p). A small value of $\sup_{(u,p)\in\mathcal{U}_A\times\mathcal{P}_A} r_{u,p}(d^*_{\mathrm{GT}})$ would denote robustness with respect to the choice of utility and probability of the Attacker and, therefore, any pair (u, p) could be chosen with no significant changes in the attainable expected utilities. Otherwise, we have an issue which questions the relevance of the proposed Nash defense d^*_{GT}. At a deeper level, it also questions the appropriateness of the (u_A, p_A) assessment, actually serving to criticize the game theoretic assumption of common knowledge. Operationally, a *threshold* on the maximum regret might be fixed such that if exceeded, such assumption must be questioned. The whole procedure is illustrated in Algorithm 3, in which B is determined based on the available computational budget. We stress that, as this approach is implemented in an exploratory sense, we might not need very big sample sizes (in Algorithm 2) thus enabling us to allocate more resources in exploring a larger sample of (u, p)'s.

```
input: $d*_{GT}$, $\mathcal{U}_A$, $\mathcal{P}_A$, threshold, B
for $i = 1$ to $R$ do
    Randomly sample $u$ and $p$ from $\mathcal{U}_A$ and $\mathcal{P}_A$, respectively
    Compute $d^*_{u,p}$ using Algorithm 2 (or its variants)
    Compute $r_{u,p}(d^*_{\mathrm{GT}})$
    if $r_{u,p}(d^*_{GT}) >$ threshold then
        Robustness requirements not satisfied
        Stop
Robustness requirements satisfied.
```

Algorithm 3: Robustness assessment of the game theoretic solution

3 ARA solution

3.1 The basic approach

We thus need to address the case when the game theoretic solution is not robust. One way forward is to perform an ARA approach, Banks et al. (2015). For this, we weaken the common knowledge assumption: the Defender does not know (p_A, u_A). The problem she faces is depicted in Figure 2. To solve it, besides $p_D(\theta|d, a)$ and $u_D(d, \theta)$ available from our earlier discussion in Section 2, the Defender requires $p_D(a|d)$. This is her assessment

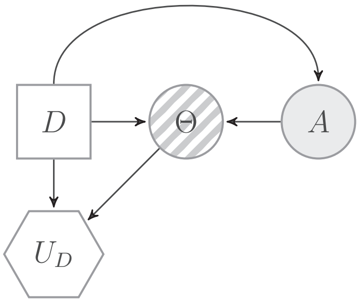


Figure 2: The decision problem as seen by Defender.

of the probability that the Attacker will choose attack a after having observed that she has chosen the defense d. Then, we can proceed as follows. First, the expected utility of d would be

$$\psi_D(d) = \int \psi_D(a,d) p_D(a|d)\,\mathrm{d}a = \int \left[\int u_D(d,\theta) p_D(\theta|d,a)\,\mathrm{d}\theta \right] p_D(a|d)\,\mathrm{d}a.$$

Finally, her optimal decision would be $d^*_{\mathrm{ARA}} = \arg\max_{d \in \mathcal{D}} \psi_D(d)$. This solution d^*_{ARA} does not need to correspond to a Nash equilibrium as both solutions are based on different assumptions and information, see for instance the example in Section 4.

Eliciting $p_D(a|d)$ is facilitated if the Defender analyzes the problem from the Attacker's perspective, Figure 3. The Defender will use all the information and judgment available to her about the Attacker's utilities and probabilities. Instead of using point estimates for p_A and u_A to find the Attacker's optimal decision $a^*(d)$ for a given d, as in Section 2, the Defender's uncertainty about the Attacker's decision would derive from her uncertainty about (p_A, u_A), through a distribution $F = (U_A, P_A)$ on the space of utilities and probabilities, which we designate random probabilities and utilities. This induces a distribution over the Attacker's expected utility in (1), where the random expected utility for A would

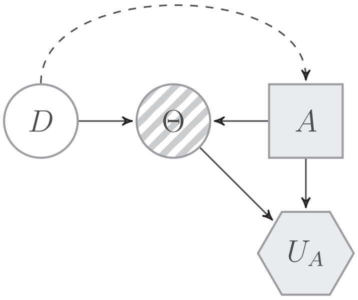


Figure 3: Defender's analysis of Attacker's problem.

be $\Psi_A(a,d) = \int U_A(a,\theta) P_A(\theta|a,d)\,\mathrm{d}\theta$. Then, the Defender would find

$$p_D(a|d) = \mathbb{P}_F \left[a = \arg\max_{x \in \mathcal{A}} \Psi_A(x,d) \right], \tag{6}$$

in the discrete case and, similarly, in the continuous one. In general, we use MC simulation to approximate $p_D(a|d)$ by drawing J samples $\{(P_A^i, U_A^i)\}_{i=1}^J$ from F and setting

$$\hat{p}_D(a|d) \approx \frac{\#\{a = \arg\max_{x \in \mathcal{A}}\ \Psi_A^i(x,d)\}}{J}, \tag{7}$$

where $\Psi_A^i(a,d) = \int U_A^i(a,\theta) P_A^i(\theta|a,d)\,\mathrm{d}\theta$.

Algorithmically, it consists of a sequential use of MC simulation to estimate $p_D(A^* = a|d)$ and, then, the use of expected utility maximization, as shown in Algorithm 4. Note that to solve the optimization problems in d and a, we shall also typically use an MC approach. Algorithm 4 requires generating $|\mathcal{D}| \cdot (|\mathcal{A}| \cdot Q \cdot J + P)$ samples, where Q and P are the number of samples required to approximate $\int u_A^i(a,\theta) p_A^i(\theta|a,a)\,\mathrm{d}\theta$ and $\max_d \int\int u_D(d,\theta) p_D(\theta|a,d) \hat{p}_D(A^* = a|d)\,\mathrm{d}\theta\,\mathrm{d}a$, respectively.

```
input: $J$
for $d \in \mathcal{D}$ do
    for $i = 1$ to $J$ do
        Sample $u_A^i(a,\theta) \sim U_A(a,\theta)$
        Sample $p_A^i(\theta|a,d) \sim P_A(\theta|d,a)$
        Compute $a_i^*(d)$ as $\arg\max_a \int u_A^i(a,\theta) p_A^i(\theta|a,a)\,\mathrm{d}\theta$
    $\hat{p}_D(A^* = a|d) = \frac{1}{J}\sum_{i=1}^J I[a_i^*(d) = a]$
Solve $\max_d \int\int u_D(d,\theta) p_D(\theta|a,d) \hat{p}_D(A^* = a|d)\,\mathrm{d}\theta\,\mathrm{d}a$
```

Algorithm 4: MC approach to solve the ARA problem

We now turn to the APS approach for the ARA problem. We first provide several observations concerning the relevant augmented probability models and, then, outline the algorithms. To start with, we replicate the argument in relation with Equation (3), with random utilities and probabilities. For a given d, we introduce the artificial random distribution

$$\Pi_A(a, \theta|d) \propto U_A(a, \theta) P_A(\theta|a, d),$$

whose marginal $\Pi_A(a|d)$ is proportional to the random expected utility $\Psi_A(a, d)$. Slightly abusing notation, the random optimal attack coincides with the mode of the marginal of the random augmented distribution

$$A^*(d) = \text{mode}\,(\Pi_A(a|d)).$$

Based on it, we build $p_D(a|d)$ as in Equation (6). Then, moving backwards, we introduce the artificial distribution

$$\pi_D(d, a, \theta) \propto u_D(d, \theta)\; p_D(\theta|a, d)\; p_D(a|d),$$

whose marginal in d satisfies $\pi_D(d) \propto \psi_D(d)$, so that

$$d^* = \text{mode}\,(\pi_D(d)).$$

Based on this argument, we propose a nested APS algorithm, Algorithm 5, which emulates the tree fold back approach, estimating directly $\pi_D(a|d)$ and then optimizing.

From a computational perspective, Algorithm 5 requires generating $|\mathcal{D}| \cdot (2 \cdot M \cdot J) + 3 \cdot N$ samples from multivariate distributions in addition to the cost of the convergence checks and mode computation.

```
input: $N, M, J$
for $d \in \mathcal{D}$ do
    for $j = 1$ to $J$ do
        Sample $U_A^j$, $P_A^j$ and define $\Pi_A^j$
        Initialize $\theta^0$
        for $i = 1$ to $M$ do
            Sample $a^{(i)}$ from $\Pi_A^j(a|\theta^{(i-1)}, d)$
            Sample $\theta^{(i)}$ from $\Pi_A^j(\theta|a^{(i)}, d)$
        Estimate $a_j^*$ as mode of $\{a^{(i)}\}$
    Estimate $p_D(a|d)$ from $\{a_j^*\}$
Initialize $(d^{(0)}, \theta^{(0)})$
for $i = 1$ to $N$ do
    Draw $d^{(i)}$ from $\pi_D(d|a^{(i-1)}, \theta_D^{(i-1)})$
    Draw $\theta_D^{(i)}$ from $\pi_D(\theta|a^{(i-1)}, d^{(i)})$
    Draw $a^{(i)}$ from $\pi_D(a|d^{(i)}, \theta_D^{(i)})$
Estimate $d^*$ as mode of $\{d^{(i)}\}$
.
```

Algorithm 5: Nested APS approach to solve the ARA problem

3.2 Sensitivity analysis of the ARA solution

The above approach leads to a decision analysis problem with the peculiarity that it includes a sampling procedure to forecast the adversary's actions.

A sensitivity analysis can be conducted with respect to the inputs of the Defender's decision analysis, which are $(u_D(d, \theta), p_D(\theta|a, d), p_D(a|d))$. Our focus is on the sensitivity to the last component $p_D(a|d)$, the most contentious one as it is obtained through adversarial calculations based on the random utility $U_A(a, \theta)$, and the random probability distribution $P_A(\theta|a, d)$. For that, we could define classes $\mathscr{U}_A$, $\mathscr{P}_A$ of random utilities and probabilities. For each pair U, P in such class, we define $p_D^{UP}(a|d)$ through the ARA approach which, in turn, leads to d^{*UP}_{ARA}. Then, we may consider the impact of the imprecision about U and P over the attained expected utility $\psi(d^{*UP}_{ARA})$, We would say that sensitivity holds if the maximum expected utility changes considerably when the input parameters change, in which case we also need to check whether the ARA solution changes as well.

If the ARA results are sensitive, we may opt for gathering additional information to reduce the classes $\mathscr{U}_A$ and $\mathscr{P}_A$. Once all possible sources of information have been exploited to increase robustness about d^*_{ARA} without success, an extra criterion would need to be introduced to make a decision and report a value about the quantity of interest. In any case, such decision should be reported with a warning of lack of robustness. We could consider, e.g., the decision d^*_R minimizing the maximum regret, i.e.

$$\min_d \max_{U \in \mathscr{U}_A, P \in \mathscr{P}_A} \left[\int \psi_D(a, d^{*UP}_{ARA}) p^{UP}_D(a|d^{*UP}_{ARA}) da - \int \psi_D(a,d) p^{UP}_D(a|d) da \right].$$

All of the above may be embedded in an appropriate simulation scheme, much as we did in Section 2.3 with the game theoretic approach.

4 An illustrative example

We illustrate the proposed framework through a simple sequential defend-attack cyber security problem. An organization needs to decide its security protocol, either through a safe, but costly route, or through cheaper, but more dangerous protocols with different protection levels, rendering business performance increasingly at risk.

4.1 Problem structure

Assume that the Defender has the following ten protocols: $d = 0:$ do nothing, i.e. no defensive action is taken; $d = 1:$ use a level 1 protection protocol; $d = 2, \ldots, 8:$ use increasing levels of protection; $d = 9:$ use the very safe but cumbersome protocol. In turn, the attacker has two alternatives: attack (denoted $a = 1$); no attack ($a = 0$.) Successful (unsuccessful) attacks are, respectively, denoted by $\theta = 1$ ($\theta = 0$). Clearly, when there is no attack ($a = 0$), we necessarily have $\theta = 0$.

	θ	
d	0	1
0	0.05	7.05
1	0.10	7.10
2	0.15	7.15
3	0.20	7.20
4	0.25	7.25
5	0.30	7.30
6	0.35	7.35
7	0.40	7.40
8	0.45	7.45
9	0.50	7.50

(a)

	a	
d	0	1
0	0.0	0.50
1	0.0	0.40
2	0.0	0.35
3	0.0	0.30
4	0.0	0.25
5	0.0	0.20
6	0.0	0.15
7	0.0	0.10
8	0.0	0.05
9	0.0	0.01

(b)

	θ	
a	0	1
0	0.00	0.00
1	-0.53	1.97

(c)

d	α	β
0	50.0	50.0
1	40.0	60.0
2	35.0	65.0
3	30.0	70.0
4	25.0	75.0
5	20.0	80.0
6	15.0	85.0
7	10.0	90.0
8	5.0	95.0
9	1.0	99.0

(d)

Table 1: (a) Defender's costs; (b) Successful attack probabilities; (c) Attacker's costs; and (d) Beta distribution parameters

4.2 Defender's non strategic judgments

We assess the Defender's non strategic judgments. Table 1a presents the costs c_D associated with each defender decision d and outcome θ of the success of the attack. The business expected valuation of the defender is assumed to be $7M$ euros. Each increase in security level is assumed to cost $0.05M$ euros. The Defender's probability vector of a successful attack for each of the attacking actions and defender decisions is shown in Table 1b, with complementary probabilities for unsuccessful attacks (for example, $p_D(\theta = 0|a = 1, d = 2) = 1 - 0.35$). The Defender is constant risk averse with respect to monetary costs. Thus, her utility function is strategically equivalent to $u_D(c_D) = -\exp(c * c_D)$ with $c > 0$. Suppose that c is equal to 0.4.

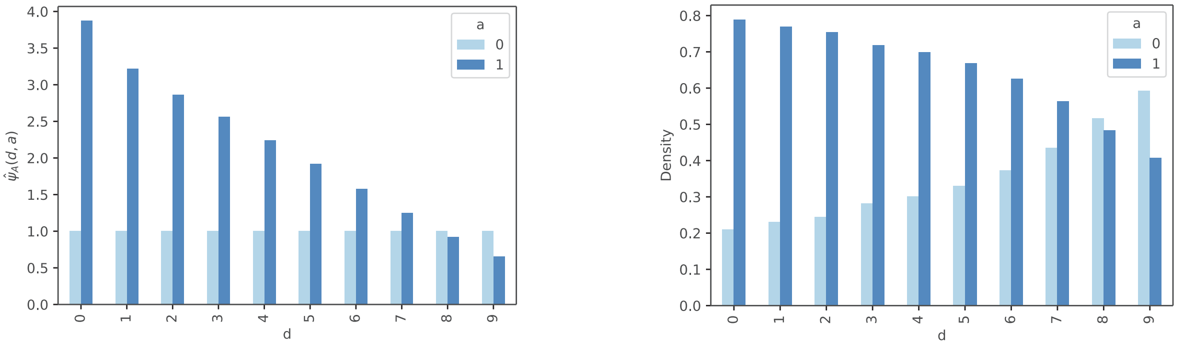
4.3 Attacker judgments

Consider now the Attacker problem. The average cost of an attack operation is estimated at $0.03M$ Euros and the eventual average benefit (market share obtained, potential ransom, etc.) is assumed to be $2M$ Euros. We assume that the cost of the attack is $0.5M$ Euros if the attack is repelled. Table 1c presents the attacker profit level c_A associated with each attack decision a and outcome θ of the attack. Qualitatively, the Defender thinks the Attacker is constant risk prone over money. Therefore, she considers that his utility function is strategically equivalent to $u_A(c_A) = \exp(e \cdot c_A)$ with e modeling the Attacker risk attitude.

4.4 Game theoretic approach

Assuming common knowledge, we set $p_A(\theta = 1|a, d) = p_D(\theta = 1|a, d)$. Particularly, the Attacker probabilities $p_A(\theta = 1|a = 1, d)$ are reflected in Table 1b. With respect to u_A, the value of e is set at 1.

To compute the solution we use the MC and nested APS approaches described in Algorithms 1 and 2, respectively. In the nested APS, we use a Metropolis-Hastings scheme



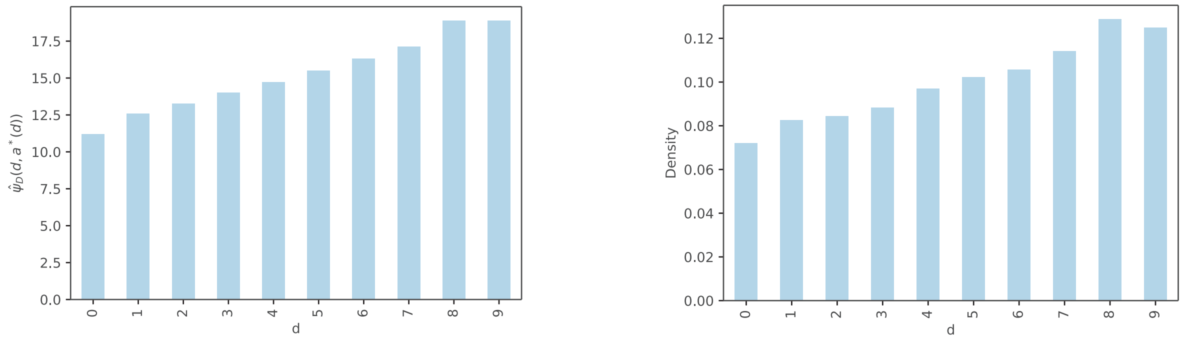
(a) MC solutions

(b) Nested APS solutions

Figure 4: Solutions of the attacker problem for each defense

to sample from the marginal conditionals. Firstly, for each defense strategy d, the optimal attack $a^*(d)$ has been computed. Figure 4a represents the expected MC estimation of the attacker expected utility for each d and a. The optimal attack for each d is that having maximum expected utility. For example, for $d = 5$ the attacker's optimal decision is to perform the attack, whereas for $d = 8$ he should decide not to attack. Note that when there is no attack ($a = 0$), the cost for the attacker is always null, and consequently the utility is constant and equal to 1. Figure 4b represents, for each d, the results of sampling from the augmented distribution $\pi_A(a, \theta|d)$ marginalized on a, in terms of frequency of occurrence. The mode of such distribution coincides with the optimal attack. For example, for $d = 9$ the mode is $a = 0$ and, consequently, the optimal solution is not to attack. As can be seen, for defenses running from 0 (no defensive action), until 7 (level 7 protocol), the attacker should attack. But if we adopt stronger defenses, the attack should be avoided.

Armed with $a^*(d)$, we compute the optimal defense, using again APS and MC. Figure 5a presents the MC estimation of $\psi_D(d, a^*(d))$ for each d; and 5b, the frequency of samples from the marginal augmented distribution $\pi_D(d|a^*(d))$. As expected, both methods agree



(a) MC solution

(b) APS solution

Figure 5: Solutions of the defender problem

that the optimal decision is acquiring level 8 protection, which is the cheapest decision that avoids being attacked, under the common knowledge hypothesis.

4.4.1 Robustness of the game theoretic solution

We next compute the optimal defense for $K = 1000$ small perturbations of $u_A(c_A)$ and $p_A(\theta|d,a)$. To build such perturbations, we sample $e' \sim \mathcal{U}(0,2)$, and use $u'_A(c_A) = \exp{(e \cdot c_A)}$; if $p_A(\theta|d,a)$, obviously $p_A(\theta = 1|d, a = 0) = 0$ for all d; to perturb $p_A(\theta|d, a = 1)$ we sample from a beta distribution with parameters α and β specified in Table 1d for each possible d.

Figure 6 reflects the frequency with which each d was optimal. $d^* = 8$ was the solution of the original problem and emerges around 36% of the time as optimal. However, the solution is not very stable since inducing small perturbations in the utilities and success probabilities leads to other solutions appearing 27% of the time ($d = 7$) and 21% of the time ($d = 9$). Nevertheless, variations in expected utility are not that big. The value of the maximum regret obtained in this case was 0.23, which represents a 1.21% of the total

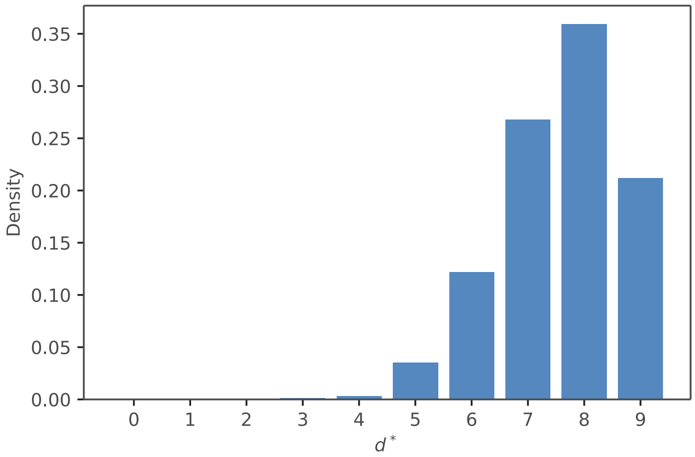


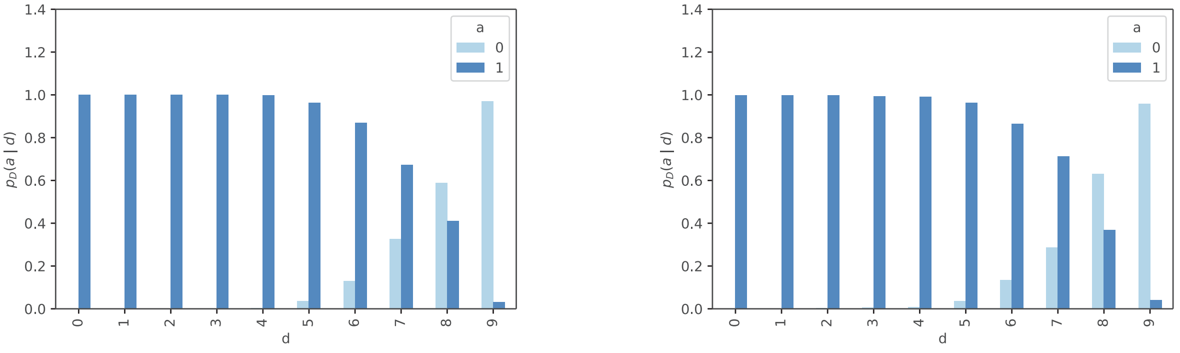
Figure 6: Sensitivity analysis of the game theoretic solution

expected utility. Thus, we can conclude that the game theoretic solutions is robust to these small perturbations.

4.5 ARA approach

We next relax the common knowledge assumptions describing the Defender's beliefs over the Attacker's judgments through P_A and U_A. Assuming that the Attacker knows the Defender's decision, the probability of success for the Attacker will be modeled as $P_A(\theta = 1|a = 1, d) \sim Beta(\alpha, \beta)$ with parameters α and β, dependent on the defense, specified in Table 1d; their expected values are set equal to $p_D(\theta = 1|a, d)$ from the game theoretic setting under common knowledge. In addition, we assume uncertainty over e with $e \sim \mathcal{U}(0, 2)$, which induces the uncertainty over $U_A(c_A)$.

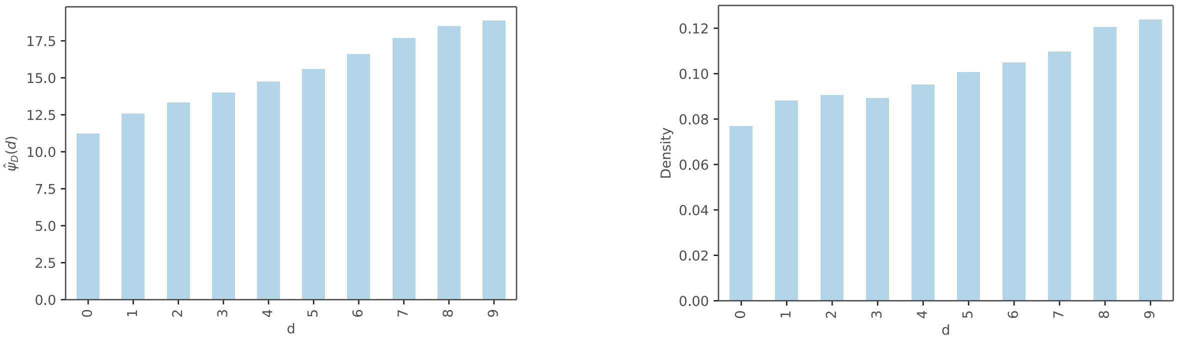
Their forecast $p_D(a|d)$ over attacks a, given the defenses d is presented in Figure 7. Figure 7a presents the estimates using MC, Algorithm 4. Figure 7b those with the nested APS approach in Algorithm 5. Both solutions coincide up to numerical errors. With the



(a) MC estimation of $p_D(a|d)$

(b) APS estimation of $p_D(a|d)$

Figure 7: Estimation of $p_D(a|d)$ through ARA



(a) MC solution

(b) APS solution

Figure 8: ARA solutions for the defender

forecast over the attacks, we compute the ARA optimal solution for the defender. Figure 8a shows the MC estimation of the defender's expected utility. Figure 8b shows the frequency of samples from the marginal augmented distribution $\pi_D(d)$. The mode of this distribution coincides with the optimal defense, $d^*_{ARA} = 9$, in agreement with the MC solution.

We emphasize that the ARA solution does not correspond with a Nash equilibrium. Note that in this case, the ARA solution appears to be more conservative, as it suggests a safer, although more expensive, defense. Of course, as in any decision analysis, we could perform sensitivity analysis with, e.g. alternate c values, say between 0.1 and 1.

5 Discussion

We have considered the problem of supporting a decision maker who faces adversaries such that the consequences attained are random and depend on the actions of all participating agents. The prevalent paradigm is Game Theory. We can also view the problem as a decision analytic one through ARA. We have presented a full approach to the problem, switching from the game theoretic to the ARA concept when the common knowledge assumption is questionable. Essentially, the procedure could be summarized as follows: under common knowledge assumptions, the game theoretic solution can be computed and subject to an appropriate sensitivity analysis. If stable, such solution may be used with confidence and no further analysis is required. Otherwise, the common knowledge assumption is questioned and we use ARA as an alternative decision analytic approach. If the ARA solution is found to be stable as a result of the sensitivity analysis, it may be used with confidence and the analysis stops. Otherwise, more data must be gathered and relevant classes must be refined, eventually declaring the robustness of the ARA solution; if not sufficient, a minimum regret (or other robust) analysis can be undertaken.

We have shown how MC and APS can be used for solving game theoretic and ARA models while discussing their computational complexity, see Table 2 for a summary. As can be seen, the number of MC samples depends on the cardinality of the Attacker's decision space, while this dependence is not present in APS. Thus, in problems in which the adversary's decision space is very big or even continuous, APS would be more efficient than performing MC.

	MC	**APS**
GT	$\lvert\mathcal{D}\rvert\Big(\lvert\mathcal{A}\rvert \cdot Q + P\Big)$	$2\Big(\lvert\mathcal{D}\rvert \cdot M + N\Big)$
ARA	$\lvert\mathcal{D}\rvert\Big(\lvert\mathcal{A}\rvert \cdot Q \cdot J + P\Big)$	$\lvert\mathcal{D}\rvert\Big(2 \cdot M \cdot J + 3 \cdot N\Big)$

Table 2: Number of samples required by MC and APS algorithms for game theoretic and ARA

In the game theoretic framework, Algorithm 7 can be used to remove dependence of the complexity on the cardinality of the Defender's decision space as well. This algorithm would thus be more efficient than MC in problems in which the defender's decision space is very big or continuous. In addition, in the case of continuous decision spaces, APS could provide solutions with arbitrary precision, while MC is limited to the precision of the discretization of the corresponding decision space. A promising approach in problems with continuous decision spaces could consist on using MC to limit the area of the decision space where the optimum is located, and then switch to an APS approach to search within that area in more detail.

Moreover, MC errors associated with approximating the expected utility can overwhelm the calculation of the optimal x^*. Moreover, samples from $p(\theta|x)$ will typically need to be recomputed for each x. In contrast, APS performs the expectation and the optimization simultaneously. It draws samples of x from decision regions with high objective function values, whereas draws of θ are tilted away from the conditional density $p(\theta|x)$ towards the artificial distribution: the approach concentrates on "smart" values of θ where the importance function is the objective function that tightens around the optimal x^*, when convergent. Overall, sampling in a utility-tilted way helps to draw the random parameter θ more frequently from where it leads to higher utility. This reduces the MC error as less optimization effort is wasted in parts of the parameter space with low objective function values, resulting also in typically reducing sample sizes.

Apart from computational issues, note also that when the surface of the expected utility function is flat, MC simulation may need many draws or may result in poor estimates. Something similar can be argued for non-symmetric distributions. For instance, in a cybersecurity one may have a low probability for a scenario which has catastrophic consequences. MC simulation may not be able to take that into account. APS can deal with those cases since it is based on sampling from the optimizing portions of the decision space. Even in cases with very flat expected utility surfaces, APS can be improved by replacing the expected utility surface by a power transformation that uses a more peaked surface without changing the solution of the problem (Müller, 2005). This property is similar to simulated annealing (Kirkpatrick et al., 1983) that powers up the function to be maximized to find its optimum. See Müller et al. (2004), Jacquier et al. (2007) and Aktekin and Ekin (2016) for such implementations.

Finally, we have focused on the sequential two stage game problem. The ideas may be extended to other types of games, like the simultaneous Defend-Attack one (Rios and Rios Insua, 2012). In addition, it may be worth exploring cases in which there are several adversaries.

Appendix

Appendix 1: Augmented probability simulation

We provide an outline of APS. It was initially proposed in Bielza et al. (1999) and further extended in Müller et al. (2004) to solve decision analysis problems with the aid of Markov chain Monte Carlo (MCMC) procedures. Ekin et al. (2014) utilized it to solve two stage stochastic programs with recourse, whereas Ekin et al. (2017) and Ekin (2018) used it for extensions within the simulation based stochastic programming context.

APS treats the decision variables as random and converts the decision analysis problem into a simulation one in the joint space of both decision and random variables, creating an auxiliary distribution proportional to the product of the utility function and the original distribution. Simulating from this auxiliary distribution solves simultaneously for the expectation of the objective function and the optimization problem: the marginal mode over the decision variable provides the optimal decision. The strategy can accommodate arbitrary non-negative utility functions and probability models.

Suppose that we aim at finding the decision $x^* \in \mathcal{X}$ maximizing the expected utility

$$\psi(x) = \int u(x,\theta) p(\theta|x) \, \mathrm{d}\theta, \tag{8}$$

where x is the decision to be made and $\mathcal{X}$ is the feasible set; $u(x,\theta)$ is the utility function; θ is the random state; and $p(\theta|x)$ is the incumbent probability distribution which depends on the decision made. Suppose that $u(x,\theta)$ is non-negative. We define an auxiliary distribution $\pi(x,\theta)$, which we call augmented probability, such that

$$\pi(x,\theta) \propto u(x,\theta) p(\theta|x). \tag{9}$$

Observe that its marginal distribution over the decisions is $\pi(x) \propto \int u(x,\theta)p(\theta|x)\,\mathrm{d}\theta$, thus being proportional to the expected utility $\psi(x)$. Then, the optimal x^* coincides with the mode of the marginal distribution $\pi(x)$.

Note that MC simulation approaches Equation (8) by first estimating the expected utility through an MC average, $\widehat{\psi(x)}$, computed using N independent MC samples from $p(\theta|x)$,

$$\widehat{\psi(x)} = \frac{1}{N}\sum_{i=1}^{N} u(x, \theta^{(i)}).$$

We then optimize $\widehat{\psi(x)}$ over x, Shao (1989).

We usually simulate from the augmented model $\pi(x,\theta)$ using MCMC methods, Gamerman and Lopes (2006). Particularly, we utilize Gibbs sampling which iteratively samples from the conditional distributions $\pi(x|\theta)$ and $\pi(\theta|x)$, resulting in samples from the joint distribution in the limit under appropriate conditions (Casella and George, 1992). Thus, this requires simulation from the conditional distribution $\pi(\theta|x)$, a 'tilted' version of $p(\theta|x)$ as $\pi(\theta|x) \propto u(x,\theta)p(\theta)$. In addition, we need to be able to simulate from $\pi(x|\theta)$. Non-standard conditional distributions may require the use of Metropolis-Hastings steps to draw samples (Chib and Greenberg, 1995).

The application of APS requires a non-negative utility function to get a proper probability density function. Adding a large enough number to the utility will do that without changing the nature of the aforementioned distributions. When the Markov chain has converged, the mode of the marginal samples of x approximates the optimal solution. Practical convergence may be assessed, e.g., with the aid of the Brooks-Gelman-Rubin (BGR) statistics, Brooks and Roberts (1998), among other possibilities.

Appendix 2

Recalling Equations (4) and (5), if we sample pairs (d, θ) from the augmented distribution $\pi_D(d, \theta | a^*(d))$ and marginalize them to the space of alternatives, we could approximate the game theoretic solution by the mode of such samples. Algorithms 6 and **??** propose MCMC Metropolis schemes to sample from $\pi_D(d, \theta | a^*(d))$ avoiding loops over decision spaces; Algorithm 6 serves as subroutine within Algorithm **??**.

```
input: $d$, $M$, $K$, $g_A$ symmetric distribution
initialize: $a^{(0)}$, $\theta^{(0)} \sim p_A(\theta | d, a^{(0)})$
for $i = 1$ to $M$ do
    Propose a new attack $\tilde{a} \sim g_A(\tilde{a} | a^{(i-1)})$.
    Draw $\tilde{\theta} \sim p_A(\theta | d, \tilde{a})$.
    Evaluate the acceptance probability
        $$\alpha = \min \left\{ 1, \frac{u_A(\tilde{a}, \tilde{\theta})}{u_A(a^{(i-1)}, \theta^{(i-1)})} \right\}.$$
    With probability $\alpha$ set $a^{(i)} = \tilde{a}$ and $\theta^{(i)} = \tilde{\theta}$.
Discard the first $K$ samples and compute mode of the rest of draws $\{a^{(i)}\}$. Record it
  as $a^*(d)$.
```

Algorithm 6: APS for solving the attacker's problem.

```
input: $N$, $J$,$g_D$ symmetric distribution
initialize: $d^{(0)}$
Compute $a^*(d^{(0)})$ using Algorithm 6 and store it.
Draw $\theta^{(0)} \sim p_D(\theta|d^{(0)}, a^*(d^{(0)}))$.
for $i = 1$ to $N$ do
    Propose a new defense $\tilde{d} \sim g_D(\tilde{d}|d^{(i-1)})$.
    Read $a^*(\tilde{d})$ if available, otherwise compute it using Algorithm 6 and store it.
    Draw $\tilde{\theta} \sim p_D(\theta|\tilde{d}, a^*(\tilde{d})$.
    Evaluate the acceptance probability
        $$\alpha = \min \left\{ 1, \frac{u_D(\tilde{d}, \tilde{\theta})}{u_A(d^{(i-1)}, \theta^{(i-1)})} \right\}$$
    With probability $\alpha$ set $d^{(i)} = \tilde{d}$, $a^*(d^{(i)}) = a^*(\tilde{d})$ and $\theta^{(i)} = \tilde{\theta}$.
Discard the first $J$ samples and compute mode of rest of draws $\{d^{(i)}\}$. Record it as
$\widehat{d}^*_{GT}$.
```

Algorithm 7: APS for solving the defender's problem.

References

Aktekin, T. and Ekin, T. (2016). Stochastic call center staffing with uncertain arrival, service and abandonment rates: A bayesian perspective. *Naval Research Logistics (NRL)*, 63(6):460–478.

Banks, D. L., Aliaga, J. M. R., and Insua, D. R. (2015). *Adversarial Risk Analysis*. CRC Press.

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

Brooks, S. P. and Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8(4):319–335.

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

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Ekin, T. (2018). Integrated maintenance and production planning with endogenous uncertain yield. *Reliability Engineering & System Safety*, 179:52–61.

Ekin, T., Polson, N. G., and Soyer, R. (2014). Augmented Markov chain Monte Carlo simulation for two-stage stochastic programs with recourse. *Decision Analysis*, 11(4):250–264.

Ekin, T., Polson, N. G., and Soyer, R. (2017). Augmented nested sampling for stochastic programs with recourse and endogenous uncertainty. *Naval Research Logistics (NRL)*, 64(8):613–627.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC.

Gibbons, R. (1992). *A Primer in Game Theory*. Harvester Wheatsheaf.

Heap, S. H. and Varoufakis, Y. (2004). *Game Theory*. Routledge.

Insua, D. R., Vieira, A. C., Rubio, J. A., Pieters, W., Labunets, K., and Rasines, D. G. (2019). An adversarial risk analysis framework for cybersecurity. *arXiv preprint arXiv:1903.07727*.

Jacquier, E., Johannes, M., and Polson, N. (2007). MCMC maximum likelihood for latent state models. *Journal of Econometrics*, 137(2):615–640.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Müller, P. (2005). Simulation based optimal design. *Handbook of Statistics*, 25:509–518.

Müller, P., Sansó, B., and De Iorio, M. (2004). Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, 99(467):788–798.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. (2007). *Algorithmic Game Theory*, volume 1. Cambridge University Press Cambridge.

Ortega, J., Rios, D., and Cano, J. (2018). Bi-agent influence diagrams from an adversarial risk analysis perspective. *Tech. Report*.

Ozdaglar, A. and Menache, I. (2011). *Network Games: Theory, models, and dynamics.* Morgan & Claypool Publishers.

Pincus, M. (1970). A Monte Carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research*, 18(6):1225–1228.

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

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

Shao, J. (1989). Monte Carlo approximations in Bayesian decision theory. *Journal of the American Statistical Association*, 84(407):727–732.

Annex 7: Paper: Risk-based Selection of Mitigation Strategies for Cybersecurity of Electric Power Systems

Risk-based Selection of Mitigation Strategies for Cybersecurity of Electric Power Systems

Abstract—Electric power systems extensively rely on cyber physical systems to control physical components through cyber-based commands. Thus, the vulnerability to cyber threats requires an efficient allocation of resources to mitigate the risk of attacks. Common practices guide the selection of mitigation actions by prioritizing the cyber threat scenarios through a qualitative assessment. These practices can result in sub-optimal allocations of resources to protect the system. To overcome these drawbacks, we quantify the risk of cyber threats to the system through a comprehensive analysis of the system vulnerabilities. This analysis relies on Bayesian networks, which provide a solid framework for probabilistic risk assessment by representing cyber threat scenarios as combinations of cascading events. In addition, we develop an optimization model to determine the non-dominated mitigation strategies to protect the system from cyber threats. Specifically, the minimization of the risk of cyber threats supports the selection of mitigation actions, considering budget and technical constraints. The optimization model provides additional insight into risk management at different budget levels.

Index Terms—Cyber physical systems, Cybersecurity, Electric power grids, Risk management, Multi-objective optimization.

1 Introduction

CYBER physical systems are physical systems in which operations are integrated, monitored and controlled through multi-core processors [1]. Such systems are increasingly employed in a wide range of industries, including electric power industry. Despite the substantial benefits to our society, the rapid proliferation of cyber physical systems also provides potential attackers with new opportunities to disrupt critical infrastructures [2].

Costly impacts can result from such attacks, for instance a cyber attack in 2015 caused the power outage of 225000 customers in Ukraine that lasted up to six hours. In that occasion, the operators at the three operations centers were unable to regain remote control of more than 50 substations affected by the incident. After the loss of over 130MW of load, the operators restored power by sending technicians to the substations and manually controlling the power system [3]. Besides critical infrastructures, cyber threats may affect all kind of institutions with potentially severe and costly impacts worldwide. For instance, the Petya and WannaCry cyber-attacks hit thousands of companies across the globe in 2017 [4]. Other relevant cases include the Stuxnet attack in 2010 to target an uranium enrichment centrifuge in Iran [5] and the attack on a German steel mill in 2014 to take over the plant control systems [6]. In recent years, cyber attacks have increased dramatically in terms of quantity, diversity and sophistication with significant economic losses [7].

These episodes prove the need for an effective deployment of security measures to mitigate the risk of cyber threats. Poolsappasit et al. [8] develop a mitigation strategy based on the likelihood of cyber attacks. A genetic algorithm supports the selection of a subset of mitigation actions by minimizing the cost of deployment and the expected damage to the system. Shameli-Sendi et al. [9] propose a dynamic framework for selecting optimal countermeasures to mitigate attacks. The selection is based on minimizing the cost of deployment and the impact on users and services. However, these optimization models do not consider the multiple impacts deriving from the cyber threat scenarios. Instead, mitigation strategies are selected on the cost and performance of individual actions. The resulting resource allocation could be sub-optimal for the cyber physical systems due to the lack of modeling the multiple impacts of cyber attacks [10]. Thus, the efficient allocation of resources to secure cyber physical systems involves challenges that we address in this paper.

Specifically, this paper fits into the first two functions of the National Institute of Standards and Technology (NIST) cybersecurity framework [11] in *detecting* system vulnerabilities and *protecting* the system from cybersecurity incidents. The NIST cybersecurity framework sets broadly accepted guidelines to improve the security of cyber physical systems. In this framework, we propose a methodology for the risk assessment of cyber threats based on a comprehensive analysis of the system vulnerabilities. This methodology relies on Bayesian networks that model a probabilistic representation of combinations of events, possibly leading to severe outcomes. This model responds to the need for intuitive and computationally efficient methods for risk analysis, combining expert judgment and statistical analyses for the quantitative assessment of risks [12].

The proposed methodology leads to select the optimal portfolios of mitigation actions, based on the minimization of the risk of multiple impacts of cyber attacks. In particular, this paper focuses on mitigation strategies for protecting electric power systems, yet the framework has broader applications on cyber physical systems. Recently, the Electric

Power Research Institute (EPRI) analyzed the cybersecurity failure scenarios and impacts for the electric sector [13]. The report provides insights on cybersecurity risks and potential mitigation actions to support risk assessment and resource allocation. Among applications on electric power systems, Ciapessoni et al. [14] propose a methodology to assess the security of such systems by analyzing the vulnerabilities to natural and human threats. On the other hand, Shelar and Amin [15] formulate a game theoretic framework to assess the security of an electricity distribution network, based on which the defender optimizes the security strategy of the network nodes.

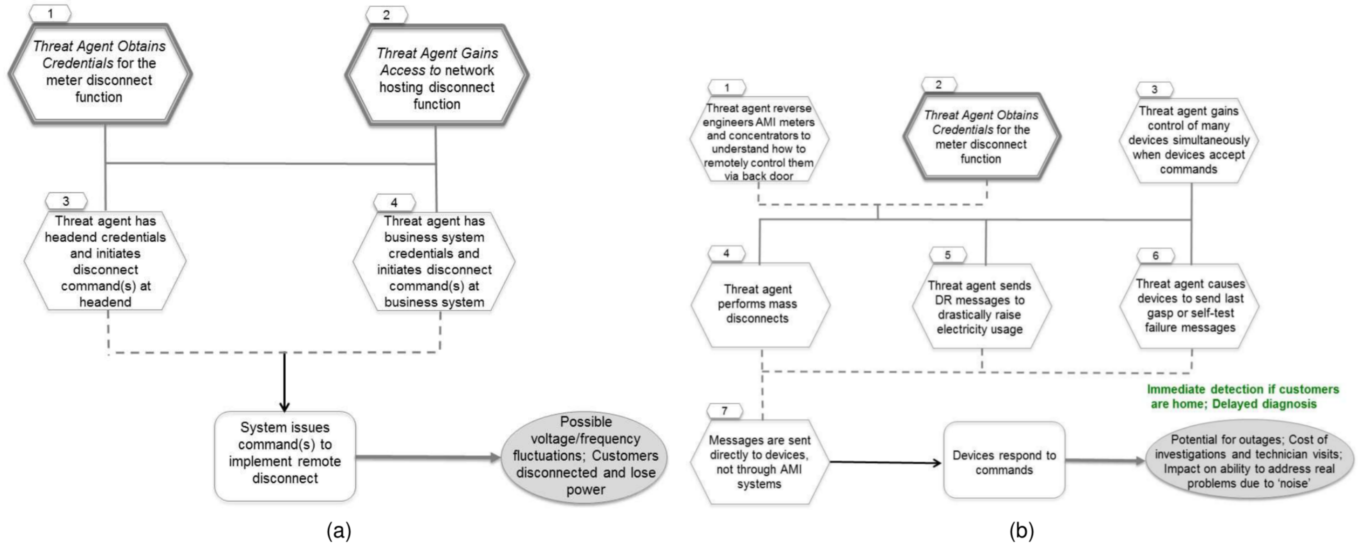
In this paper, Section 2 reviews the practice proposed by the EPRI to select appropriate mitigation actions for the electric power system. Specifically, we present a critical analysis of the ranking procedure of individual cyber threats, which could lead to inefficient or unfeasible allocations for the system. This problem is addressed in Section 3, which provides an alternative to the EPRI practice by evaluating portfolios of mitigation actions to protect the cyber physical system against multiple cyber threat scenarios. In addition, an optimization model supports the selection of the mitigation strategies that minimize the expected impacts of cyber attacks, based on financial and technical constraints. Section 4 illustrates the methodology by analyzing the cyber threat scenarios concerning the Advanced Metering Infrastructure (AMI) of an electric power system. Section 5 discusses the potential and limits of the proposed framework, suggesting possible ways to overcome some inconveniences. Finally, Section 6 concludes the paper and outlines extensions for future research.

2 Analysis of the EPRI practice

Cybersecurity management calls for an extensive analysis of the system vulnerabilities, which leads to an efficient allocation of resources to protect the electric power system. In particular, the EPRI proposes the analysis of individual cyber threat scenarios based on attack graphs, multi-leveled diagrams describing threats on cyber physical systems and possible attacks to realize such threats [16]. Attack graphs are increasingly being applied to computer control systems, especially related to electric power systems, but they have also been used to analyze threats to physical systems [17]. Figure 1 illustrates the graphical notation of two attack graphs, where a cyber threat scenario is represented through sequences of events (shown as diamonds) leading to the possible impacts of the cyber attack (shown as ellipses). The impacts of the cyber attack occur if a combination of events of the cyber threat scenario has proven to be successful, based on the binary representation of AND and OR gates (shown as solid and dashed lines, respectively).

Attack graphs represents cyber threat scenarios, which are evaluated based on the *likelihood* of occurrence and *impact*. According to the EPRI analyses, the likelihood depends on 5 criteria whereas the impact depends on 15 criteria which are reported in Tables 3 and 4, respectively. These tables also report the EPRI scoring system for quantifying the likelihood and impact of cyber threat scenarios. Each score is an integer value in the range $0 - 9$, thus the likelihood and impact are computed by summing the scores over the respective criteria. However, this scoring system can be questioned on the meaningfulness of the $0 - 9$ scale. For instance, is a public accessible asset three times more accessible than a fenced asset with standard locks and nine times more accessible than a guarded/monitored asset? Furthermore, the additive model can be questioned on the sum of scores across different criteria. For instance, is "Public safety concern" comparable to "Long term economic damage"?

Mapping the likelihood and impact of all cyber threat scenarios in a risk matrix [18] makes it possible to rank the priority of individual cyber threats. This procedure clusters each cyber threat into *High*, *Medium* or *Low* likelihood and *High*, *Medium* or *Low* impact in order to prioritize the selection of mitigation actions. Specifically, cyber threats with *High* likelihood and *High* impact deserve the highest priority in the choice of mitigation actions, whereas priority


1
Threat Agent Obtains Credentials for the meter disconnect function
2
Threat Agent Gains Access to network hosting disconnect function
3
Threat agent has headend credentials and initiates disconnect command(s) at headend
4
Threat agent has business system credentials and initiates disconnect command(s) at business system
System issues command(s) to implement remote disconnect
Possible voltage/frequency fluctuations; Customers disconnected and lose power
1
Threat agent reverse engineers AMI meters and concentrators to understand how to remotely control them via back door
2
Threat Agent Obtains Credentials for the meter disconnect function
3
Threat agent gains control of many devices simultaneously when devices accept commands
4
Threat agent performs mass disconnects
5
Threat agent sends DR messages to drastically raise electricity usage
6
Threat agent causes devices to send last gasp or self-test failure messages
Immediate detection if customers are home; Delayed diagnosis
7
Messages are sent directly to devices, not through AMI systems
Devices respond to commands
Potential for outages; Cost of investigations and technician visits; Impact on ability to address real problems due to 'noise'

(a) (b)

Fig. 1. Attack graphs for (a) "Invalid disconnect messages to meters impact customers and utility" and (b) "Reverse engineering of AMI equipment allows unauthorized mass control" [13].

decreases for cyber threats with lower likelihood and/or impact until the budget is depleted.

Despite the intuitive appeal and simplicity, risk matrices do not necessarily recommend effective risk management decisions, instead they may lead to incorrect risk prioritization [19], [20]. Thus, the sequential choices of mitigation actions may result in a sub-optimal resource allocation because they are based on an incorrect prioritization of cyber threats. Furthermore, this procedure does not consider technical and budget constraints across different scenarios. In conclusion, the EPRI practice presents several inconsistencies in assessing the risk of cyber threats and supporting the selection of mitigation actions.

3 Bayesian framework

We propose a Bayesian framework, which provides an alternative to the EPRI practice for the risk assessment of cyber threats and the risk-based selection of mitigation strategies. In particular, the proposed risk assessment is based on a comprehensive analysis of multiple cyber threats that can affect the cyber physical system [21]. The framework also includes an optimization model for determining non-dominated mitigation strategies in order to protect the system from cyber threats. Specifically, an optimization algorithm computes the portfolios of mitigation actions that minimize the expected impacts of cyber attacks, considering budget and technical constraints.

3.1 From attack graphs to Bayesian network

In contrast to the EPRI analysis of individual cyber threat scenarios, the Bayesian framework relies on a comprehensive analysis of multiple attack graphs [22]. Each attack graph represents a single cyber threat scenario, however some events could be equivalent among different attack graphs. For instance, in Figures 1a and 1b the event "Threat agent obtains credentials for the meter disconnect function" is equivalent among both attack graphs. For this reason, multiple attack graphs can be integrated into a directed acyclic graph by combining the corresponding events into single nodes. This integration leads to a comprehensive representation of cyber threat scenarios that overviews the alternative opportunities to attack the system.

This directed acyclic graph can be converted into a Bayesian network [23], a probabilistic graphical model that consists of:

- *chance nodes* (shown as circles) representing the random events of cyber threat scenarios;
- *value nodes* (shown as diamonds) representing the possible impacts of the cyber attacks;
- *arcs* (shown as directed edges) indicating causal dependencies between nodes.

Specifically, chance nodes are connected by arcs to represent combinations of events leading to the respective final impacts [24]. In this framework, the combinations of events indicate the stages of cyber threat scenarios, whereas the final impacts indicate the possible outcomes of the cyber attacks. Arcs connects the nodes to represent the causal dependencies between the events of the attack graph. Figure 2 illustrates a Bayesian network, where each chance node represents a random event that encodes a finite set of discrete states, including a state of *No occurrence* of the event. Bayesian networks typically consider discrete states, nevertheless it is possible to include continuous variables under specific conditions [23].

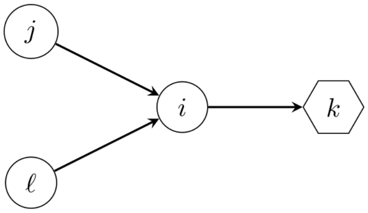


Fig. 2. Example of a Bayesian network.

Statistical analyses and expert judgment provide information to define the probability distributions of events that do not depend on any other chance node (nodes j and ℓ in Figure 2). For events that show causal dependencies on other chance nodes through directed arcs (node i in Figure 2), the probabilistic representation is based on the state of the events they are depending on. Thus, it is necessary to define conditional probability tables for such nodes. Following the binary representation of attack graphs, conditional probability tables are derived from the information provided by AND and OR gates. Specifically, if the event i depends on the events j and ℓ through AND(OR) gates, then the occurrence probability of the event i is 1 if the events j and(or) ℓ occur as well, and 0 otherwise. For illustrative purposes, we consider the event "Threat agent performs mass disconnects" in Figure 1b as an example throughout the paper. Table 1 displays the conditional probability table of the event, based on the binary representation of *Occurrence* and *No occurrence* of the dependent events. The conditional probability table is derived from the information provided by AND and OR gates of the attack graph in Figure 1b.

Bayesian networks represent the events such that the occurrence probability is not necessarily limited to 0 and 1, but it is a real value in the set $[0, 1]$. This model leads to a more realistic representation of the stages of cyber threat scenarios, in contrast to the binary representation. Table 2 displays the conditional probability table of the event "Threat agent performs mass disconnects", based on the multiple states of the events. This conditional probability table is not meant to represent any actual electric power system. According to the EPRI analyses, the occurrence probability of each event depends on (i) skill required, (ii) physical accessibility, (iii) logical accessibility and (iv) attack vector. In particular, the occurrence probability increases by enhancing the accessibility to equipment and information, while it decreases by requiring specialized knowledge and technical means to pursue the cyber threat.

The cascading events of the cyber threat scenarios finally lead to the possible impacts, assessed according to a set of criteria represented by the set K of value nodes [25]. The EPRI lists 14 possible impact criteria of cyber attacks on electric power systems, including financial, safety and service impacts. As a result, each value node of the Bayesian network represents a single impact criterion k, whose score depends on the state of events leading to that specific out-

TABLE 1
Conditional Probability Table Based on Binary States.

			Threat agent performs mass disconnects	
Threat agent reverse engineers AMI equipment	Threat agent obtains credentials	Threat agent gains control of devices	*Occurrence*	*No occurrence*
Occurrence	*Occurrence*	*Occurrence*	1	0
		No occurrence	0	1
	No occurrence	*Occurrence*	1	0
		No occurrence	0	1
No occurrence	*Occurrence*	*Occurrence*	1	0
		No occurrence	0	1
	No occurrence	*Occurrence*	0	1
		No occurrence	0	1

TABLE 2
Conditional Probability Table Based on Multiple States.

			Threat agent performs mass disconnects [MW]			
Threat agent reverse engineers AMI equipment	Threat agent obtains credentials	Threat agent gains control of devices	*No occurrence*	(0 50]	(50 100]	> 100
Occurrence	*Occurrence*	*None*	1	0	0	0
		Few	0.6	0.4	0	0
		Moderate	0.4	0.2	0.4	0
		High	0.3	0.1	0.2	0.4
	No occurrence	*None*	1	0	0	0
		Few	0.6	0.4	0	0
		Moderate	0.4	0.2	0.4	0
		High	0.3	0.1	0.2	0.4
No occurrence	*Occurrence*	*None*	1	0	0	0
		Few	0.6	0.4	0	0
		Moderate	0.4	0.2	0.4	0
		High	0.3	0.1	0.2	0.4
	No occurrence	*None*	1	0	0	0
		Few	1	0	0	0
		Moderate	1	0	0	0
		High	1	0	0	0

come. Ideally, the scores should be evaluated by a specific scale that reflects its unit of measure.

3.2 Probabilistic risk assessment

The probabilistic risk assessment of cyber threats is based on the computation of the expected impact for every impact criteria. Each chance node i represents a random event that encodes a finite set $\mathbb{S}_i$ of discrete states, including a state of *No occurrence* of the event. In particular, the occurrence probability of events that show causal dependencies relies on the occurrence probability of the events they depend on. For this reason, we define Δ_i as the set of all possible combinations of states of the chance nodes affecting the event i, such that

$$\Delta_i = \prod_{j|(j,i)\in E} \mathbb{S}_j, \tag{1}$$

where E denotes the set of all arcs.

Let the random variable X_i represent the probability distribution of event i over the states $s_i \in \mathbb{S}_i$. Then, $\hat{\mathbf{X}}_i$ is a vector of random events on which X_i directly depends, meaning the vector of random variables X_j for all nodes j such that $(j, i) \in E$. For the d-separation property of Bayesian networks [23], the probability that the events affecting the event i meets a specific combination of states $\delta_i \in \Delta_i$ is

$$\mathbb{P}[\hat{\mathbf{X}}_i = \delta_i] = \prod_{s_j \in \delta_i} \mathbb{P}[X_j = s_j] \quad \forall \delta_i \in \Delta_i. \tag{2}$$

Thus, the occurrence probability of the event i is computed by the *law of total probability* as the weighted average of the posterior probabilities across all $\delta_i \in \Delta_i$, such that

$$\mathbb{P}[X_i = s_i] = \sum_{\delta_i \in \Delta_i} \mathbb{P}[X_i = s_i | \hat{\mathbf{X}}_i = \delta_i] \; \mathbb{P}[\hat{\mathbf{X}}_i = \delta_i]. \tag{3}$$

Because the occurrence probabilities are computed recursively, it is necessary to start the computation from the initial events throughout the dependent events of the cyber threat scenarios. The risk of cyber threats is then evaluated as the expected impact of the scenarios for each criterion $k \in K$, such that

$$\mathbb{E}[V_k] = \sum_{\delta_k \in \Delta_k} \mathbb{P}[\hat{\mathbf{X}}_k = \delta_k] \; V_k[\hat{\mathbf{X}}_k = \delta_k], \tag{4}$$

where $V_k[\hat{\mathbf{X}}_k = \delta_k]$ is the score of the impact criterion k depending on the combination of states δ_k of the events leading to that specific impact.

The expected impacts can be significantly reduced by deploying mitigation actions on the cyber physical system. Specifically, the mitigation actions affect the occurrence probability of one or multiple events in the cyber threat scenarios. In Bayesian networks, decision nodes (shown as squares) represent the choice of mitigation actions, as illustrated in Figure 3. Each arc directed from a decision node to a chance node indicates that the deployment of the mitigation action affects the occurrence probability of the event represented by the chance node. Because this paper

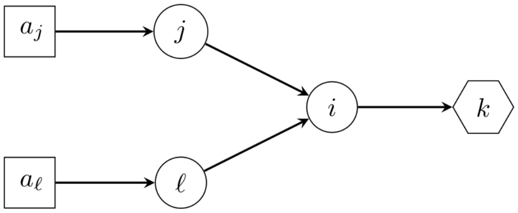


Fig. 3. Example of a Bayesian network with decision nodes.

focuses on system design, the decision nodes do not depend on any event (no incoming arcs). Future research will focus on system control with decision nodes depending on other events.

Mitigation actions are numbered $a \in \{1, 2, ..., N\}$, such that the binary variable z_a indicates the deployment of the mitigation action a. Specifically, the binary variable is $z_a = 1$ for the deployment of the mitigation action a and $z_a = 0$ otherwise. Thus, a portfolio is defined by the binary vector $\mathbf{z}$ as a combination of binary variables z_a for all the possible mitigation actions. With no loss of generality, the vector $\mathbf{z}$ lists binary variables such that

$$\mathbf{z} = [z_1,\, z_2,\, \dots,\, z_N]. \tag{5}$$

The deployment of mitigation actions reduces the occurrence probabilities of affected events. Bayesian networks compute probability updates of the cascading events throughout the cyber threat scenarios by the *law of total probability*, such that

$$\mathbb{P}[X_i = s_i|\mathbf{z}] = \sum_{\delta_i \in \Delta_i} \mathbb{P}[X_i = s_i|\hat{\mathbf{X}}_i = \delta_i]\; \mathbb{P}[\hat{\mathbf{X}}_i = \delta_i|\mathbf{z}]. \tag{6}$$

Thus, the risk of cyber threats depends on the portfolio $\mathbf{z}$ so that the expected impact of each criterion $k \in K$ is

$$\mathbb{E}[V_k](\mathbf{z}) = \sum_{\delta_k \in \Delta_k} \mathbb{P}[\hat{\mathbf{X}}_k = \delta_k|\mathbf{z}]\; V_k[\hat{\mathbf{X}}_k = \delta_k]. \tag{7}$$

This framework aims to compute the risk of cyber threats for each impact criterion, making it possible to select mitigation strategies based on the minimization of the expected impacts.

3.3 Optimization model

The risk-based selection of mitigation strategies is performed through a multi-objective optimization model. Unlike the EPRI practice, the selection of mitigation actions is not based on the additive model of scores across different impact criteria. Instead, our optimization model determines the portfolios of mitigation actions that minimize the risk of cyber threats for every impact criteria. The selection is based on the analysis of expected impacts derived from the deployment of different mitigation strategies, so that the optimization model determines the portfolios that fulfill the Pareto condition

$$\mathbf{z}^* \succ \mathbf{z} \Longleftrightarrow \left\{ \begin{array}{ll} \mathbb{E}[V_k](\mathbf{z}^*) \leq \mathbb{E}[V_k](\mathbf{z}) & \text{for all } k \\ \mathbb{E}[V_k](\mathbf{z}^*) < \mathbb{E}[V_k](\mathbf{z}) & \text{for some } k \end{array} \right. . \tag{8}$$

This condition indicates that portfolio $\mathbf{z}^*$ dominates $\mathbf{z}$ if it reduces the risk of cyber threats for any impact criterion without increasing the risk for other impact criteria.

In addition to the Pareto condition, the optimal mitigation strategies need to fulfill budget and technical constraints. Budget constraints specify the financial feasibility of the deployment of a mitigation strategy. Each mitigation action a is associated to a cost c_a, thus the overall cost of portfolio $\mathbf{z}$ must not exceed the budget B such that

$$\sum_a z_a\; c_a \leq B. \tag{9}$$

Technical constraints specify the properties of the system, such as mutually exclusive or mutually inclusive conditions of mitigation actions. For instance in Figure 3, the linear constraints

$$z_{a_j} + z_{a_\ell} \leq 1 \tag{10}$$

$$z_{a_j} - z_{a_\ell} = 0 \tag{11}$$

indicate that mitigation actions a_j and a_ℓ cannot be deployed together or they must be deployed together, respectively.

Technical constraints also include risk acceptability limits that are represented by non-linear inequalities. In particular, specific regulatory conditions may apply to some events of the cyber threat scenarios. For such event i, the subset $\tilde{\mathbb{S}}_i \subset \mathbb{S}_i$ includes the critical states whose occurrence probability must not exceed a risk acceptability threshold ϵ_i, such that

$$\sum_{s_i \in \tilde{\mathbb{S}}^i} \mathbb{P}[X_i = s_i|\mathbf{z}] \leq \epsilon_i. \tag{12}$$

Risk acceptability thresholds are usually provided by regulatory offices or internal company policies.

Feasible portfolios belong to the set $\mathbf{Z}_F$, which includes all binary vector $\mathbf{z}$ that fulfill linear and non-linear constraints. Then, the set of non-dominated solutions consists of the feasible portfolios that fulfill the Pareto condition for any other feasible portfolio, meaning that

$$\mathbf{Z}_{ND} = \{\mathbf{z}^* \in \mathbf{Z}_F | \nexists\; \mathbf{z} \in Z_F \text{ such that } \mathbf{z} \succ \mathbf{z}^*\}. \tag{13}$$

Generally, the set of non-dominated portfolios can include multiple alternative solutions, so the selection of a single mitigation strategy is not straightforward. For this reason, it is necessary to support the selection of the optimal mitigation strategy through additional analyses. A possible approach is the computation of the *core index* of each mitigation action. Analogously to Liesiö et al. [26], the core index $CI(a)$ is defined as the fraction of non-dominated portfolios that include the mitigation action a, such that

$$CI(a) = \frac{|\{\mathbf{z}^* \in \mathbf{Z}_{ND} | z_a = 1\}|}{|\mathbf{Z}_{ND}|}. \tag{14}$$

The analysis of the core indexes helps determine the mitigation actions that should be selected or rejected. If the core index of a mitigation action is 1, that measure is included in all non-dominated portfolios; on the other hand, if the core index is 0, that measure is not included in any non-dominated portfolio. Finally, mitigation actions whose core index is in the range $(0, 1)$ require further analyses in order to be selected or rejected.

An implicit enumeration algorithm computes the set of non-dominated portfolios that minimize the risk of cyber threats over the impact criteria. The algorithm is an adaptation of

Liesiö [27] and has been proposed by Mancuso et al. [28] for multi-objective optimization. This optimization algorithm is computationally efficient but it may be time consuming for a large amount of mitigation actions (over 40). In this case, evolutionary algorithms are a possible alternative to approximate non-dominated solutions for a lower computational time [29].

4 Case study

We illustrate the potential of the Bayesian framework by optimizing the selection of mitigation strategies for the Advanced Metering Infrastructure (AMI) of an electric power system. AMI systems have raised many security concerns since they connect traditionally self-contained power system operations with unreliable customer sites that are widely dispersed. The deployment of AMI systems is introducing millions of components to the electric grid that support two-way communication for next-generation grid applications. Although these systems can increase operational efficiency and enable new capabilities such as demand-response, they also increase the attack opportunity for potential adversaries. For this reason, electric power companies must address these new cybersecurity risks as part of their risk management strategy.

Information about AMI systems is provided by the National Electric Sector Cybersecurity Organization Resource (NESCOR), a program funded by the U.S. Department of Energy to protect electric power systems from cybersecurity incidents, both malicious and non-malicious. The NESCOR document "Electric Sector Failure Scenarios and Impact Analyses" [16] provides short descriptions of approximately 125 cyber threat scenarios in seven domains of the electric sector: *Advanced Metering Infrastructure*, *Distributed Energy Resources*, *Wide Area Monitoring*, *Protection and Control*, *Electric Transportation*, *Demand Response* and *Distribution Grid Management*. Furthermore, the NESCOR document "Analysis of Selected Electric Sector High Risk Failure Scenarios" [13] presents the analyses of a selection of these cyber threat scenarios. Specifically, each analysis includes an attack graph that details the logical dependencies of events leading to a successful cyber attack. In addition to the attack graph, several of the analyses also provide a detailed description of each scenario.

Based on the NESCOR analyses, we select 8 cyber threats with the highest priority for AMI systems, in particular:

- Invalid disconnect messages to meters impact customers and utility;
- Reverse engineering of AMI equipment allows unau-

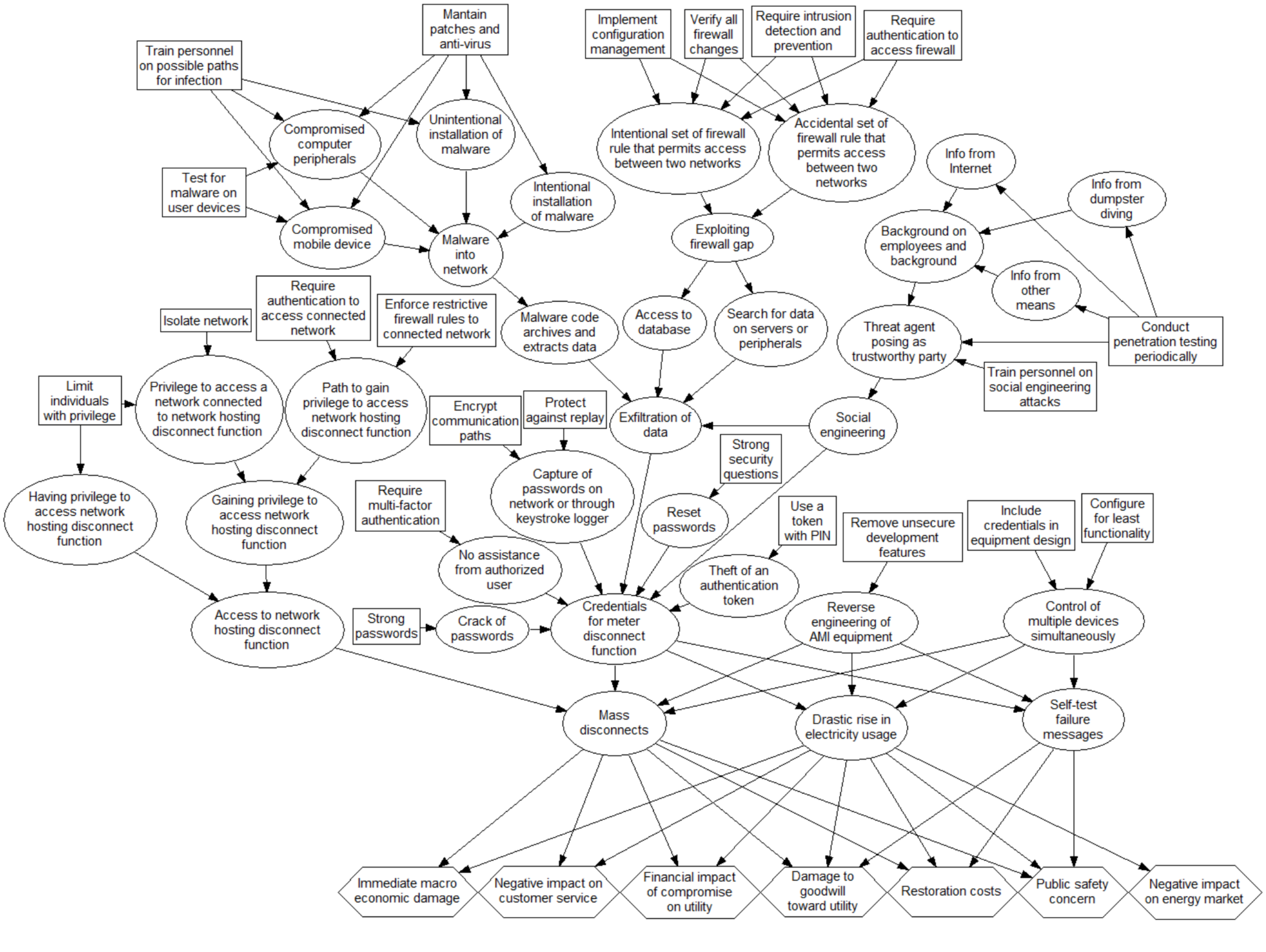

Mantain patches and anti-virus
Train personnel on possible paths for infection
Implement configuration management
Verify all firewall changes
Require intrusion detection and prevention
Require authentication to access firewall
Compromised computer peripherals
Unintentional installation of malware
Intentional set of firewall rule that permits access between two networks
Accidental set of firewall rule that permits access between two networks
Info from Internet
Test for malware on user devices
Intentional installation of malware
Info from dumpster diving
Compromised mobile device
Malware into network
Exploiting firewall gap
Background on employees and background
Info from other means
Require authentication to access connected network
Isolate network
Enforce restrictive firewall rules to connected network
Malware code archives and extracts data
Access to database
Search for data on servers or peripherals
Threat agent posing as trustworthy party
Conduct penetration testing periodically
Train personnel on social engineering attacks
Limit individuals with privilege
Privilege to access a network connected to network hosting disconnect function
Path to gain privilege to access network hosting disconnect function
Encrypt communication paths
Protect against replay
Exfiltration of data
Social engineering
Strong security questions
Having privilege to access network hosting disconnect function
Gaining privilege to access network hosting disconnect function
Require multi-factor authentication
Capture of passwords on network or through keystroke logger
Reset passwords
Use a token with PIN
Remove unsecure development features
Include credentials in equipment design
Configure for least functionality
No assistance from authorized user
Theft of an authentication token
Access to network hosting disconnect function
Strong passwords
Crack of passwords
Credentials for meter disconnect function
Reverse engineering of AMI equipment
Control of multiple devices simultaneously
Mass disconnects
Drastic rise in electricity usage
Self-test failure messages
Immediate macro economic damage
Negative impact on customer service
Financial impact of compromise on utility
Damage to goodwill toward utility
Restoration costs
Public safety concern
Negative impact on energy market

Fig. 4. Bayesian network for selected cyber threat scenarios to the Advanced Metering Infrastructure of an electric power system.

thorized mass control;

- Threat agent obtains credentials for system or function;
- Threat agent uses social engineering;
- Threat agent gains access to network;
- Threat agent exfiltrates data;
- Authorized employee brings malware into system or network;
- Threat agent exploits firewall gap.

These cyber threats potentially lead to "Threat agent performs mass disconnects", "Threat agent sends demand-response messages to drastically raise electricity usage" and "Threat agent causes devices to send last gasps or self-test failure messages", which indicate the possible outcomes of cyber attacks to the AMI systems. In Figure 4, the Bayesian network is based on the attack graphs of the 8 cyber threat scenarios to represent the alternative opportunities to attack the system. In particular, the circles represent the events of the cyber threat scenarios, the diamonds indicate the possible impacts of cyber attacks whereas the squares show mitigation actions that could be deployed for protecting the AMI system from cyber threats. Note that the event "Threat agent obtains credentials for the meter disconnect function" is equivalent among both cyber threat scenarios in Figures 1a and 1b. For this reason, this event has been represented by one chance node named "Credentials for meter disconnect function" in the Bayesian network. Reducing redundancies of equivalent events in multiple cyber threat scenarios facilitates the comprehensive analysis of cyber threats as a Bayesian network. In addition, the events "Threat agent has headend credentials and initiates disconnect(s) at headend" and "Threat agent has business system credentials and initiates disconnect(s) at business system" in Figure 1a are not considered in the Bayesian network because it is sufficient that the threat agent gains access to the network hosting the meter disconnect function and obtains the relative credentials to cause "Possible voltage/frequency fluctuations with disconnected customers".

In the Bayesian network, probability distributions of the chance nodes have been set according to information provided by the NESCOR documents. For instance, the psychological manipulation (social engineering) of an employee may be expensive and it could lead to a public disclosure if the attempt fails, which summarizes to a low occurrence probability. However, such information is not sufficient to specifically quantify the occurrence probability. For this reason, the occurrence probabilities of this example are not meant to be representative of any existing AMI system, but they are illustrative values to prove the viability of the Bayesian framework. The value nodes list the impact criteria of Table 4, in particular the ones affected by a possible cyber attack to the AMI systems. For illustrative purposes, impacts V_k have been quantified based on the scoring system of Table 4 due to the lack of detailed information in literature. For instance, the event "Threat agent performs mass disconnects" is quantified in different states of mass disconnects: *No occurrence*, $(0\ 50]$MW, $(50\ 100]$MW, $>$ 100MW. Thus, each value node maps the impact score depending on the states of that event, as illustrated in Figure 5 for the impact criterion "Restoration costs". Note that the impacts of cyber threats are not necessarily evaluated by every criteria of Table 4. For instance, the event "Threat agent performs mass disconnects" does not affect the impact criterion "Loss of privacy" for any state.

Assuming that event i is "Threat agent performs mass disconnects" and the value node represents the impact criterion "Restoration costs" in Figure 2, the expected impact of "Restoration costs" $[k = RC]$ is the weighted average of the impact scores for every state in Figure 5, such that

$$\begin{aligned}\mathbb{E}[V_{RC}] = \mathbb{P}[X_i \leq 50MW]\ V_{RC}[X_i \leq 50MW] + ...\\ + \mathbb{P}[X_i \geq 100MW]\ V_{RC}[X_i \geq 100MW]. \quad (15)\end{aligned}$$

The NESCOR documents also list possible mitigation actions that could be deployed to protect the AMI systems from cyber threats, specifying the events affected by each mitigation action. The deployment of a mitigation action af-

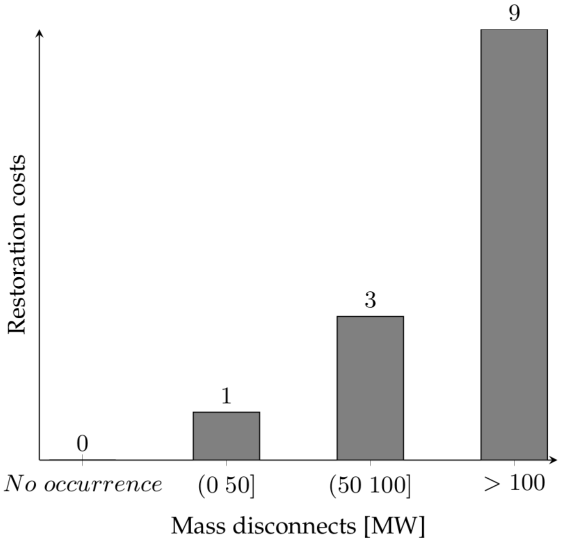


Fig. 5. Illustrative impact scores for "Restoration costs".

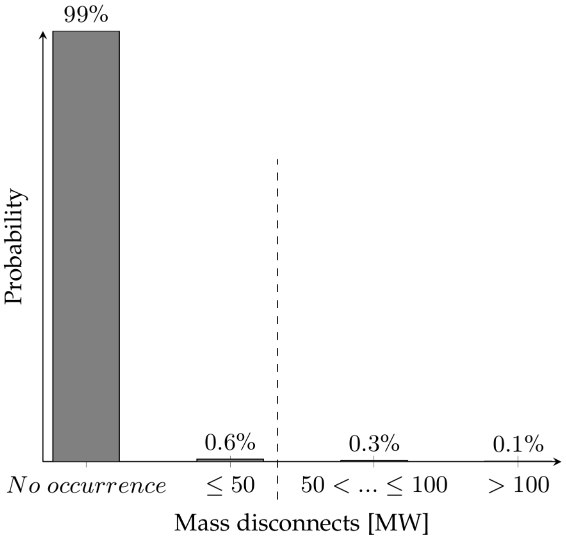


Fig. 6. Illustrative probability distribution for mass disconnects.

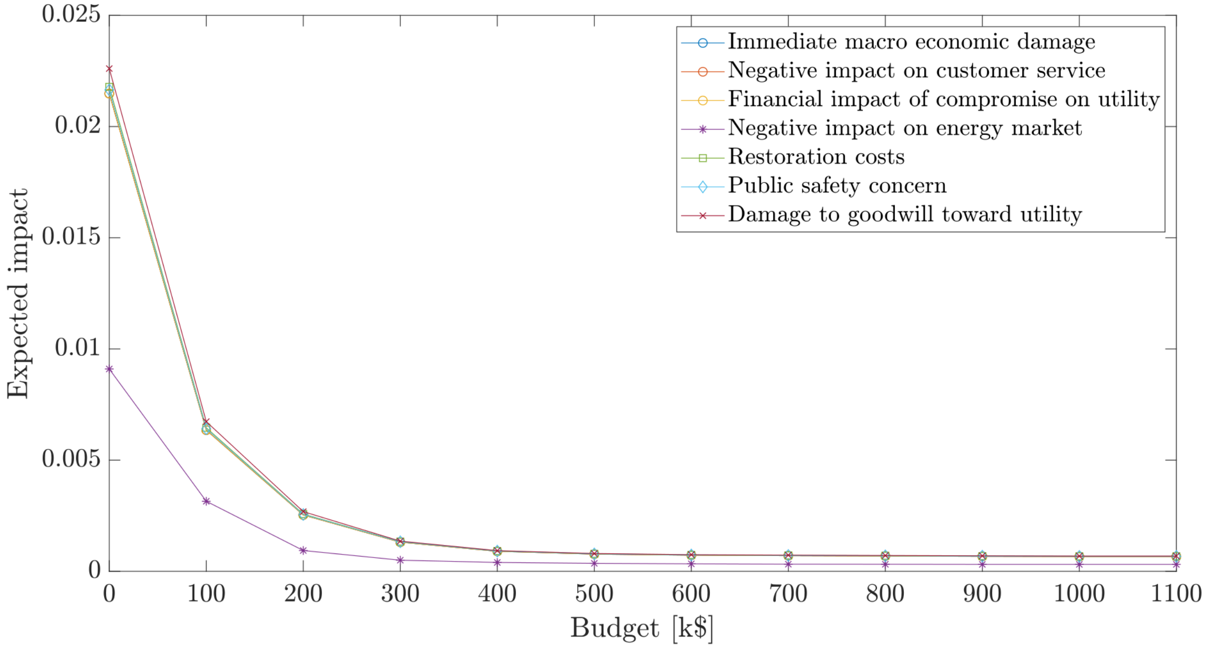


Fig. 7. Expected impact of each impact criterion for different budget levels.

fects the occurrence probability of the cyber threats according to the effect of the action. In particular, this case study accounts for 22 possible mitigation actions, which lead to 2^{22} mitigation strategies. Tables 5-10 list the 22 mitigation actions for the selected cyber threat scenarios, specifying the affected events based on the NESCOR analyses. The first column of the tables lists the index of the action in the portfolio $\mathbf{z}$, whereas the third column lists the cost of each mitigation action. The illustrative costs of mitigation actions aim to include a budget constraint to the optimization model. In addition, the optimization model includes a technical constraint on the risk acceptability of mass disconnects above 50MW. Figure 6 illustrates the probability distribution of the event “Threat agent performs mass disconnects” deriving from the deployment of a generic portfolio $\mathbf{z}$. Assuming that experts set the risk acceptability threshold to 0.5%, then the occurrence probability of the critical states must fulfill the constraint

$$\mathbb{P}[X_i > 50MW|\mathbf{z}] \leq 0.5\%. \tag{16}$$

The results of the multi-objective optimization show a decrease of the risk of every impact criteria by increasing the budget level. Figure 7 shows that larger budgets lead to more effective mitigation strategies to reduce the risk of every impact criteria. In this case study, the risk profiles of some impact criteria are overlapping because the impact scores are based on the same $0-9$ scale that limit the quantification of the impacts. The analysis of the risk profiles supports the definition of the optimal budget by selecting the budget level above which the risk converges for every impact criteria, such as $B \geq 400$ in this example. Computational time is around one hour on a regular laptop, however it depends on the constraints limiting the set of feasible portfolios. For instance, relaxing the budget constraint leads to higher computational time because the algorithm considers a larger set of feasible portfolios.

In Figure 7, the risk profiles consider all the non-dominated portfolios selected by the optimization algorithm for each budget level. Then, the core index of each mitigation action is computed to support the choice of actions that should be selected or rejected. Figure 8 maps the core index of each mitigation action through a gray scale. Specifically, a black square indicates that the action is included in every non-dominated portfolio, whereas a white square indicates that the action is not included in any non-dominated portfolio. Gray squares indicate a core index in the range $(0, 1)$, meaning that the mitigation action is included in some non-dominated portfolios, but not all.

As a result, the black-squared actions should be selected whereas the white-squared actions should be rejected. On the other hand, gray-squared actions need additional analyses to support the selection or rejection of the deployment on the AMI system. In this case study, the additional analyses would be necessary only for a limited number of mitigation actions for some budget levels. For instance, for budget $B = 500$k$ the mitigation actions z_5, z_6 and z_{16} belong to 50% of the non-dominated portfolios. The other mitigation actions belongs to either all or none of the non-dominated portfolios, so they do not require any additional analysis.

5 Discussion

The case study shows the potential of a comprehensive analysis of multiple cyber threats. Integrating the cyber threat scenarios into a Bayesian network facilitates the detection of system vulnerabilities and the definition of appropriate mitigation actions for protecting the cyber physical system. In this respect, actions affecting multiple cyber threats and synergies of actions affecting the same event(s) can be easily represented in a single model. This model results in a

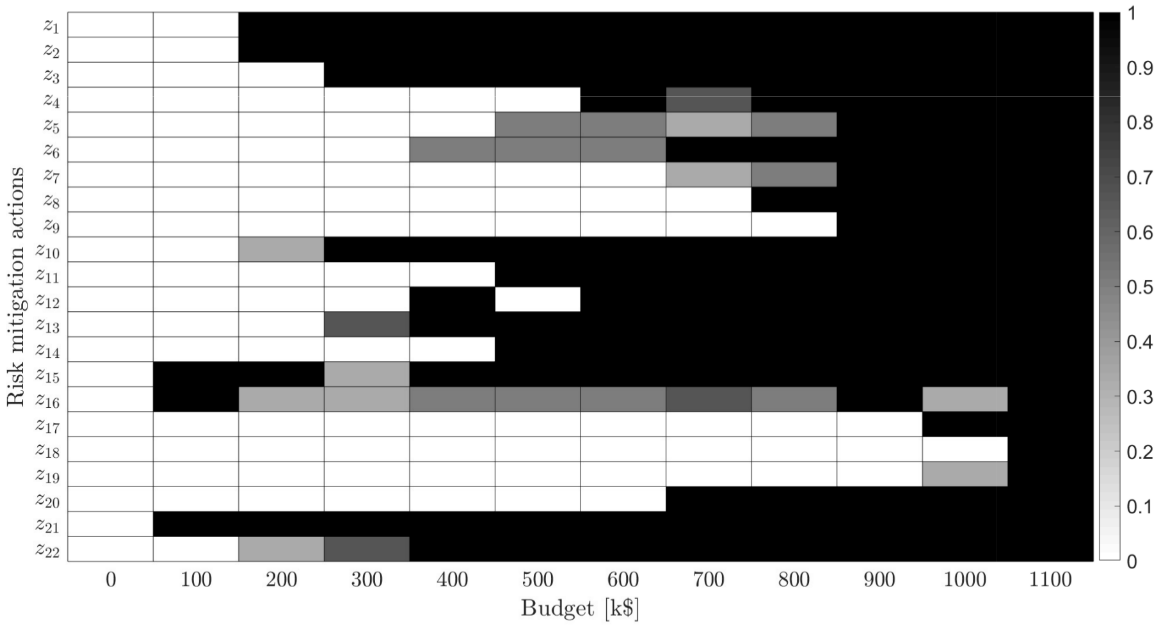


Fig. 8. Core index map of mitigation actions for different budget levels.

clear graphical representation of the possible cyber threats to the system by erasing the redundancies deriving from equivalent events in multiple scenarios.

The model relies on the definition of the occurrence probability of cyber threats, which could be a troublesome task. However, the decomposition of the cyber threat scenario into cascading events facilitates the definition of the occurrence probabilities of the single events. In addition, the collection of information on successful and unsuccessful cyber attacks could provide valuable data to estimate the occurrence probability of specific events [30]. These statistical analyses are not sufficient because the threat agents would exploit system vulnerabilities that were not necessarily available in past attacks, which by definition are not included in the existing data [31]. Specifically, a cyber threat may not be recognized until it manifests, thus it may be missed in threat scenarios that are examined as part of the risk assessment [32]. For this reason, it is necessary to integrate statistical analyses with information provided by experts based on investigations on possible system vulnerabilities.

The probabilistic representation of cyber threat scenarios provides a solid framework for the risk assessment of cyber physical system. It also enhances detailed analyses for risk management, in contrast to the binary representation through the attack graphs. Moreover, Bayesian networks make it possible to update the probability of the cascading events of cyber threat scenarios. As a result, the model represents the effect of the deployment of mitigation actions on the system, even considering intrusion detectors to tackle cyber threats that have not been examined for the risk assessment [33]. The evaluation of the risk for each impact criteria provides additional insights into risk management, which would not be possible with the additive model of scores proposed by the EPRI.

In the case study, the impacts of the cyber threats have been set according to the scoring system in Table 4. However, it is advisable to set different numeric scales based on the specificity of the impact criterion, for instance the criterion “Restoration costs” should be evaluated through a monetary scale. Note that the choice of the scale could lead to different solutions of the optimization model [34].

Finally, the Bayesian framework has broader applications than electric power systems to consider cyber threats on any cyber physical system. For instance, the National Vulnerability Database provides information about vulnerabilities of IT systems through the Common Vulnerability Scoring System [35].

6 Conclusions

In this paper, we have developed a Bayesian framework to analyze the vulnerabilities of cyber physical systems and optimize the resource allocation to protect the system from cyber threats. In particular, the selection of mitigation actions is based on the analysis of multiple outcomes of cyber attacks, including financial, safety and service impacts. Cyber threat scenarios are modeled through Bayesian networks to overview the alternative opportunities to pursue a cyber attack leading to such impacts. Thus, the minimization of the expected impacts supports the choice of mitigation strategies based on a multi-objective optimization model.

The optimization model integrates budget and technical constraints that limit the set of feasible portfolios in order to select the optimal mitigation strategies. Specifically, the optimal mitigation strategies correspond to the portfolios that reduce the risk of cyber threats for any impact criterion without increasing the risk for other impact criteria. As a result, we have showed that a comprehensive analysis of the cyber threat scenarios leads to an optimal mitigation strategy for the system. The viability of the Bayesian framework has been illustrated through a case study concerning the Advanced Metering Infrastructure of an electric power system, which have raised several security concerns.

In conclusion, this framework can be introduced as a novel practice for assessing the risks of cyber threats and for supporting risk-based decisions on resource allocation to cyber physical systems. Possible extensions need to be investigated, such as modeling the objectives of the threat agent(s) through Adversarial Risk Analysis [36]. Future research will focus on the analysis of the cyber resilience [37], meaning the ability of the cyber physical system to continuously deliver the intended outcome despite adverse cyber events.

References

[1] Lee, J., Bagheri, B. and Kao, H.A., 2015. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manufacturing Letters, 3, pp.18-23.

[2] Smith, M.D. and Paté-Cornell, M.E., 2018. Cyber Risk Analysis for a Smart Grid: How Smart is Smart Enough? A Multiarmed Bandit Approach to Cyber Security Investment. IEEE Transactions on Engineering Management.

[3] Whitehead, D.E., Owens, K., Gammel, D. and Smith, J., 2017, April. Ukraine cyber-induced power outage: Analysis and practical mitigation strategies. In Protective Relay Engineers (CPRE), 2017 70th Annual Conference for (pp. 1-8). IEEE.

[4] Yaqoob, I., Ahmed, E., ur Rehman, M.H., Ahmed, A.I.A., Al-garadi, M.A., Imran, M. and Guizani, M., 2017. The rise of ransomware and emerging security challenges in the Internet of Things. Computer Networks, 129, pp.444-458.

[5] Nourian, A. and Madnick, S., 2018. A systems theoretic approach to the security threats in cyber physical systems applied to stuxnet. IEEE Transactions on Dependable and Secure Computing, 15(1), pp.2-13.

[6] Lee, R.M., Assante, M.J. and Conway, T., 2014. German steel mill cyber attack. Industrial Control Systems, 30, p.62.

[7] Kshetri, N., 2010. The global cybercrime industry: economic, institutional and strategic perspectives. Springer Science and Business Media.

[8] Poolsappasit, N., Dewri, R. and Ray, I., 2012. Dynamic security risk management using bayesian attack graphs. IEEE Transactions on Dependable and Secure Computing, 9(1), pp.61-74.

[9] Shameli-Sendi, A., Louafi, H., He, W. and Cheriet, M., 2018. Dynamic optimal countermeasure selection for intrusion response system. IEEE Transactions on Dependable and Secure Computing, 15(5), pp.755-770.

[10] Mancuso, A., Compare, M., Salo, A. and Zio, E., 2017. Portfolio optimization of safety measures for reducing risks in nuclear systems. Reliability Engineering and System Safety, 167, pp.20-29.

[11] Barrett, M.P., 2018. Framework for Improving Critical Infrastructure Cybersecurity Version 1.1 (No. NIST Cybersecurity Framework).

[12] Zio, E., 2009. Computational methods for reliability and risk analysis (Vol. 14). World Scientific Publishing Company.

[13] Lee, A., 2015. Analysis of selected electric sector high risk failure scenarios. National Electric Sector Cybersecurity Organization Resource (NESCOR) Technical Working Group 1.

[14] Ciapessoni, E., Cirio, D., Kjølle, G., Massucco, S., Pitto, A. and Sforna, M., 2016. Probabilistic risk-based security assessment of power systems considering incumbent threats and uncertainties. IEEE Transactions on Smart Grid, 7(6), pp.2890-2903.

[15] Shelar, D. and Amin, S., 2017. Security assessment of electricity distribution networks under DER node compromises. IEEE Transactions on Control of Network Systems, 4(1), pp.23-36.

[16] Lee, A., 2015. Electric sector failure scenarios and impact analyses. National Electric Sector Cybersecurity Organization Resource (NESCOR) Technical Working Group 1.

[17] Johnson, P., Vernotte, A., Gorton, D., Ekstedt, M. and Lagerström, R., 2016, October. Quantitative information security risk estimation using probabilistic attack graphs. In International Workshop on Risk Assessment and Risk-driven Testing (pp. 37-52). Springer, Cham.

[18] Ni, H., Chen, A. and Chen, N., 2010. Some extensions on risk matrix approach. Safety Science, 48(10), pp.1269-1278.

[19] Duijm, N.J., 2015. Recommendations on the use and design of risk matrices. Safety science, 76, pp.21-31.

[20] Allodi, L. and Massacci, F., 2017. Security events and vulnerability data for cybersecurity risk estimation. Risk Analysis, 37(8), pp.1606-1627.

[21] Liu, Y. and Man, H., 2005, March. Network vulnerability assessment using Bayesian networks. In Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2005 (Vol. 5812, pp. 61-72). International Society for Optics and Photonics.

[22] Frigault, M. and Wang, L., 2008, July. Measuring network security using bayesian network-based attack graphs. In Annual IEEE International Computer Software and Applications Conference (pp. 698-703). IEEE.

[23] Nielsen, T.D. and Jensen, F.V., 2009. Bayesian networks and decision graphs. Springer Science and Business Media.

[24] Xie, P., Li, J.H., Ou, X., Liu, P. and Levy, R., 2010, June. Using Bayesian networks for cyber security analysis. In Dependable Systems and Networks (DSN), 2010 IEEE/IFIP international conference on (pp. 211-220). IEEE.

[25] Couce-Vieira, A., Houmb, S.H. and Ros-Insua, D., 2017, August. CSIRA: A Method for Analysing the Risk of Cybersecurity Incidents. In International Workshop on Graphical Models for Security (pp. 57-74). Springer, Cham.

[26] Liesiö, J., Mild, P. and Salo, A., 2008. Robust portfolio modeling with incomplete cost information and project interdependencies. European Journal of Operational Research, 190(3), pp.679-695.

[27] Liesiö, J., 2014. Measurable multiattribute value functions for portfolio decision analysis. Decision Analysis, 11(1), pp.1-20.

[28] Mancuso, A., Compare, M., Salo, A. and Zio, E., 2019. Portfolio optimization of preventive safety measures for time-dependent accident scenarios. Reliability Engineering and System Safety, conditionally accepted.

[29] Coello, C.A.C., Lamont, G.B. and Van Veldhuizen, D.A., 2007. Evolutionary algorithms for solving multi-objective problems (Vol. 5). New York: Springer.

[30] Holm, H., 2014. A large-scale study of the time required to compromise a computer system. IEEE Transactions on Dependable and Secure Computing, 11(1), pp.2-15.

[31] Paté-Cornell, M.E., Kuypers, M., Smith, M. and Keller, P., 2018. Cyber risk management for critical infrastructure: a risk analysis model and three case studies. Risk Analysis, 38(2), pp.226-241.

[32] Linkov, I., Eisenberg, D.A., Plourde, K., Seager, T.P., Allen, J. and Kott, A., 2013. Resilience metrics for cyber systems. Environment Systems and Decisions, 33(4), pp.471-476.

[33] Modelo-Howard, G., Bagchi, S. and Lebanon, G., 2008, September. Determining placement of intrusion detectors for a distributed application through bayesian network modeling. In International Workshop on Recent Advances in Intrusion Detection (pp. 271-290). Springer, Berlin, Heidelberg.

[34] Hämäläinen, R.P. and Lahtinen, T.J., 2016. Path dependence in Operational ResearchHow the modeling process can influence the results. Operations Research Perspectives, 3, pp.14-20.

[35] Zhang, Y., Wang, L., Xiang, Y. and Ten, C.W., 2015. Power system reliability evaluation with SCADA cybersecurity considerations. IEEE Transactions on Smart Grid, 6(4), pp.1707-1721.

[36] Banks, D.L., Aliaga, J.M.R. and Insua, D.R., 2015. Adversarial risk analysis. Chapman and Hall/CRC.

[37] Gisladottir, V., Ganin, A.A., Keisler, J.M., Kepner, J. and Linkov, I., 2017. Resilience of cyber systems with overand Underregulation. Risk Analysis, 37(9), pp.1644-1651.

TABLE 3
Likelihood Criteria With Scoring System [16].

Likelihood criterion	Scoring system
Skill required	0: Deep domain/insider knowledge and ability to build custom attack tools; 1: Domain knowledge and cyber attack techniques; 3: Special insider knowledge needed; 9: Basic domain understanding and computer skills.
Accessibility (physical)	0: Inaccessible; 1: Guarded, monitored; Fence, standard locks; 9: Publicly accessible.
Accessibility (logical, assume have physical access)	0: High expertise to gain access; 1: Not readily accessible; 3: Publicly accessible but not common knowledge; 9: Common knowledge or none needed.
Attack vector (assume have physical and logical access)	0: Theoretical; 1: Similar attack has been described; 3: Similar attack has occurred; 9: Straightforward, for example script or tools available.
Common vulnerability among others	0: Isolated occurrence; 1: More than one utility; 3: Half or more of power infrastructure; 9: Nearly all utilities.

TABLE 4
Impact Criteria With Scoring System [16].

Impact criterion	Scoring system
Public safety concern	0: none; 1: 10-20 injuries possible; 3: 100 injured possible; 9: one death possible.
Workforce safety concern	0: none; 3: any possible injury; 9: any possible death.
Ecological concern	0: none; 1: logical ecological damage such as localized fire or spill, repairable; 3: permanent local ecological damage; 9: widespread temporary or permanent damage to one or more ecosystems.
Financial impact of compromise on utility	0: petty cash or less; 1: up to 2% of utility revenue; 3: up to 5 %; 9: greater than 5 %.
Restoration costs	0: petty cash or less; 1: up to 1% of utility organization O&M budget; 3: up to 10%; 9: greater than 10%.
Negative impact on generation capacity	0: no effect; 1: small generation facility off-line or degraded operation of large facility; 3: more than 10% loss of generation capacity for 8 hours or less; 9: more than 10% loss of generation capacity for more than 8 hours.
Negative impact on the energy market	0: no effect; 1: localized price manipulation, lost transactions, loss of market participation; 3: price manipulation. lost transactions, loss of market participation impacting a large metro area; 9: market or key aspects of market non operational.
Negative impact on the bulk transmission system	0: no; 1: loss of transmission capability to meet peak demand or isolate problem areas; 3: major transmission system interruption; 9: complete operational failure or shut down of the transmission system.
Negative impact on customer service	0: no; 1: up to 4 hour delay in customer ability to contact utility and gain resolution, lasting one day; 3: up to 4 hour delay in customer ability to contact utility and gain resolution, lasting a week; 9: complete operational failure or shut-down of the transmission system.
Negative impact on billing functions	0: none; 1: isolated recoverable errors in customer bills; 3: widespread but correctible errors in bills; 9: widespread loss of accurate power usage data.
Damage to goodwill toward utility	0: no effect; 1: negative publicity but this does not cause financial loss to utility; 3: negative publicity causing up to 20% less interest in programs; 9: negative publicity causing more than 20% less interest in programs.
Immediate macro economic damage	0: none; 1: local businesses down for a week; 3: regional infrastructure damage; 9: widespread runs on banks.
Long term economic damage	0: none; 3: several years of local recession; 9: several years of national recession.
Loss of privacy	0: none; 1: 1000 or less individuals; 3: thousands of individuals; 9: millions of individuals.

TABLE 5
Mitigation Actions for Scenario “Authorized Employee Brings Malware Into System or Network”.

Index	Mitigation actions	Cost [k$]	Affected event(s)
1	Train personnel on possible paths for infection	30	Compromised mobile device
			Compromised computer peripherals
			Unintentional installation of malware
2	Maintain patches and anti-virus	70	Compromised mobile device
			Compromised computer peripherals
			Unintentional installation of malware
			Intentional installation of malware
3	Test for malware before connection	50	Compromised mobile device
			Compromised computer peripherals

TABLE 6
Mitigation Actions for Scenario "Threat Agent Exploits Firewall Gap".

Index	Mitigation actions	Cost [k$]	Affected event(s)
4	Implement configuration management	40	Intentional set of firewall rule that permits access between two networks
			Accidental set of firewall rule that permits access between two networks
5	Verify all firewall changes	60	Intentional set of firewall rule that permits access between two networks
			Accidental set of firewall rule that permits access between two networks
6	Require intrusion detection	30	Intentional set of firewall rule that permits access between two networks
			Accidental set of firewall rule that permits access between two networks
7	Require authentication to access firewall	50	Intentional set of firewall rule that permits access between two networks
			Accidental set of firewall rule that permits access between two networks

TABLE 7
Mitigation Actions for Scenario "Threat Agent Uses Social Engineering".

Index	Mitigation actions	Cost [k$]	Affected event(s)
8	Conduct penetration testing periodically	70	Info from Internet
			Info from dumpster diving
			Info from other means
			Threat agent posing as trustworthy party
9	Train personnel on social engineering attacks	40	Threat agent posing as trustworthy party

TABLE 8
Mitigation Actions for Scenario "Threat Agent Obtains Credentials for System or Function".

Index	Mitigation actions	Cost [k$]	Affected event(s)
10	Strong passwords	30	Crack of passwords
11	Encrypt communication paths	80	Capture of passwords on network or through keystroke logger
12	Protect against replay	60	Capture of passwords on network or through keystroke logger
13	Strong security questions	30	Reset passwords
14	Require multi-factor authentication	50	No assistance from authorized user
15	Use a token with PIN	20	Theft of an authentication token

TABLE 9
Mitigation Actions for Scenario "Threat Agent Gains Access to Network".

Index	Mitigation actions	Cost [k$]	Affected event(s)
16	Limit individuals with privilege	30	Having privilege to access network hosting disconnect function
			Privilege to access a network connected to network hosting disconnect function
17	Isolate network	90	Privilege to access a network connected to network hosting disconnect function
18	Enforce restrictive firewall rules to access connected network	70	Path to gain privilege to access network hosting disconnect function
19	Require authentication to access connected network	40	Path to gain privilege to access network hosting disconnect function

TABLE 10
Mitigation Actions for Scenario "Reverse Engineering of AMI Equipment Allows Unauthorized Mass Control".

Index	Mitigation actions	Cost [k$]	Affected event(s)
20	Remove unsecure development features	80	Reverse engineering of AMI meters
21	Include credentials in equipment design	50	Control of many devices simultaneously
22	Configure for least functionality	30	Control of many devices simultaneously

Annex 8: Paper: Adversarial Risk Analysis for Structured Expert Judgement Modelling

Adversarial Risk Analysis for Structured Expert Judgement Modelling

Abstract

We argue that adversarial risk analysis may be incorporated into the structured expert judgement modelling toolkit for cases in which we need to forecast the actions of competitors or adversaries. This is relevant in areas such as cybersecurity, security, defense and competitive business. We also illustrate key concepts and modelling strategies in ARA.

Keywords: Structured expert judgement, adversarial risk analysis, decomposition, security, cybersecurity.

1 Introduction

Structured Expert Judgement (SEJ) elicitation has a long history of successes, both in methodology and application, many of which stem from Roger Cooke's work, e.g. Cooke (1991) and Goossens et al. (1998). Hence, it has become a major ingredient within the risk and decision analysis practice (Bedford and Cooke, 2001). A significant feature of these disciplines, as acknowledged in the classic book by Raiffa (1968), is their emphasis in decomposing complex problems into smaller pieces that are easier to handle and then recombining the piecewise solutions to tackle the global problem. Examples of such decomposition principles include:

- The exercise of decision analysis, as in French and Ríos Insua (2000) or Clemen and Reilly (2013). This methodology seeks to solve complex decision making problems by the principle of maximising expected utility. In doing so, one avoids direct comparison of alternatives which, in the context of uncertainty and multiple objectives, may be cognitively intricate and prone to bias. Instead, one first structures the problem by identifying decisions, uncertainties and objectives, assessing beliefs and preferences and then finding the alternative with maximum expected utility.

Problem structuring is typically presented as the first stage of a decision or risk analysis cycle. The value of such a decomposition is assessed in Watson and Brown (1978).

- Preference assessment also uses decomposition. It may be difficult to compare consequences of alternatives without determining a utility function, specially in presence of multiple conflicting attributes. A typical approach is to search for a decomposable functional form (often additive, linear or multilinear, e.g. González-Ortega et al. (2018)), and then assess the component utilities and weights in order to recompose the global utility function whose expected value must be maximised. Ravinder and Kleinmuntz (1991) provide theory showing the advantages of undertaking such decompositions in utility assessment, given certain conditions.

- Belief assessment also benefits from decomposition, typically through the argument of *extending the conversation*. Rather than directly assessing the probability of an outcome, one finds a conditioning partition and assesses the probabilities of the outcome given the conditioning events. From these, and the probabilities of the conditioning events, the law of total probabilities enables calculation of the unconditional probability of the outcome. Ravinder et al. (1988) and Andradottir and Bier (1997, 1998) provide a methodological framework to validate the advantages of this approach, evaluated empirically in e.g. MacGregor and Kleinmuntz (1994) and MacGregor (2001). Tetlock and Gardner (2015) call this approach *Fermitisation* and present it as a key strategy for the success of their super-forecasters.

To sum up, various forms of decomposition pervade risk and decision analysis. They simplify the complex cognitive tasks and mitigate expert reliance on heuristics that can introduce bias, ensuring that experts and decision makers actually analyse their decision making problems (Montibeller and von Winterfeldt, 2015). The decomposition typically entails more assessments, though these tend to be simpler and more meaningful, leading to better decisions.

In this paper, we present Adversarial Risk Analysis (ARA) (Ríos Insua et al., 2009; Banks et al., 2015) as a decomposition strategy for game theoretic problems studied from a Bayesian perspective. ARA stems from the observation that common knowledge assumptions in standard game theoretic approaches, based on Nash equilibria and their refinements, do not hold in many applications, such as counterterrorism or cybersecurity, since competitors try to conceal information. We formulate the problem as a Bayesian game, as described in Kadane and Larkey (1982) and Raiffa (1982), and operationalise the approach by providing procedures to forecast the actions of the adversary.

ARA can be complex because the opponent strategises intelligently. But ARA is a powerful decomposition tool in finding solutions. ARA provides prescriptive support to one of the agents (D, she), based on a subjective expected utility model for a probability distribution of the decisions of the adversary (A, he). D models A's decision problem and, assuming that he is an expected utility maximiser (or has some other criterion, as in prospect theory), tries to assess his probabilities and utilities. If these were known, she could identify his optimal action. However, her uncertainty about A's probabilities and utilities is propagated to his decisions, leading to a probability distribution over his actions. ARA can be framed as a tool for SEJ elicitation when we need to deal with probabilities referring to actions by opponents. As an example, in Chen et al. (2016) nearly 30% of the questions posed to experts somehow involved adversaries (e.g. *Will Syria use chemical or biological weapons before January 2013?*).

After sketching the ARA approach (Section 2), we show how this strategy can actually improve non-structured expert assessment of the opponent's actions (Section 3). We then propose several ways to implement ARA in practice (Section 4), include a numerical example (Section 5), and end with a discussion (Section 6).

2 ARA in Sequential Games

To simplify the discussion, we focus on sequential Defend-Attack games: agent D first makes her decision $d \in \mathcal{D}$, then agent A observes d and chooses his alternative $a \in \mathcal{A}$. The outcome is usually a random variable whose distribution depends upon a and d. As an example, imagine that a company deploys cybersecurity countermeasures and then, having observed them, a cybercriminal decides whether to launch an attack. The cost to the company is a random variable that is conditioned upon both decisions.

The corresponding bi-agent influence diagram (Banks et al., 2015) is shown in Figure 1. The dashed arc reflects the fact that D's decision is observed by A, before he makes his decision. The consequences for both players relate to the outcome $s \in \mathcal{S}$ of the interaction. Each decision maker conducts their own assessment of the probability of outcome s, $p_D(s \,|\, d, a)$ and $p_A(s \,|\, d, a)$ for D and A, respectively, which depends on both their decisions d and a. The utility function $u_D(d, s)$ of D is subject to her choice d and the result s. Similarly, A's utility function is $u_A(a, s)$. In some situations an agent's utility function may depend upon the choices of both opponents, as well as the outcome.

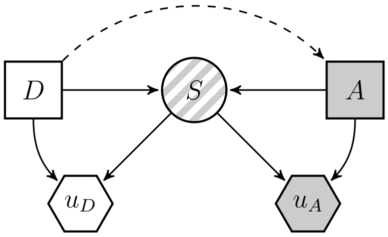


Figure 1: The two player sequential decision game.

The ARA approach weakens the standard, but unrealistic, common knowledge assumption in game theoretic approaches (Hargreaves-Heap and Varoufakis, 1995), according to which the agents share information about their probabilities and utilities. In our case, not having common knowledge means that D does not know (p_A, u_A). The problem she faces is depicted in the influence diagram in Figure 2.

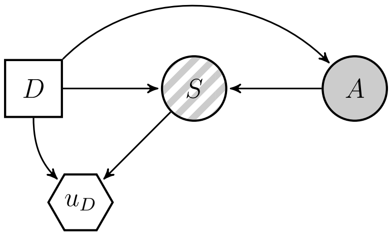


Figure 2: The decision problem as seen by D.

To solve this problem D requires, besides $p_D(s \mid d, a)$ and $u_D(d, s)$ mentioned above, the distribution $p_D(a \mid d)$, which is her assessment of the probability that A will choose action a after having observed her choice d. Once D has completed these assessments, she can compute the expected utility of d as

$$\psi_D(d) = \int \left[\int u_D(d, s)\, p_D(s \mid d, a)\, \mathrm{d}s \right] p_D(a \mid d)\, \mathrm{d}a,$$

and seek the optimal decision $d^*_{ARA} = \arg\max_{d \in \mathcal{D}} \psi_D(d)$.

Eliciting $p_D(a \mid d)$ is complicated since it entails strategic elements. D could try to assess it from a standard SEJ perspective, as in Cooke (1991) or O'Hagan et al. (2006), but ARA usefully suggests a decomposition approach that requires her to analyse the problem from A's perspective, as shown in the influence diagram in Figure 3.

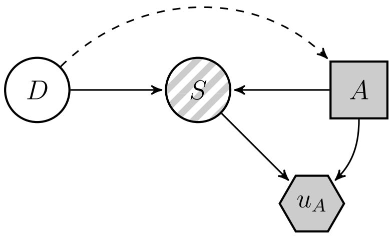


Figure 3: D's analysis of the decision problem as seen by A.

First, D puts herself in A's shoes. She would use all the information she can obtain about A's probabilities, utilities and strategy, assuming he is an expected utility maximiser. Then, instead of using point estimates for p_A and u_A in order to find A's optimal response for a given d, her uncertainty about A's decision should derive from her uncertainty about (p_A, u_A), through a distribution F on the space of probabilities and utilities. This induces a distribution over A's expected utility, where his random expected utility is

$$\Psi_A(d, a) = \int U_A(a, s)\, P_A(s \,|\, d, a)\, \mathrm{d}s,$$

for $(P_A, U_A) \sim F$. Then D finds

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

in the discrete case, and analogously in the continuous one. She can use Monte Carlo simulation to approximate $p_D(a \,|\, d)$, as shown in Sections 3 and 5.

The above approach extends to simultaneous decision making problems, general interactions between both agents, multiple agents, agents who employ principles different than maximum expected utility, as well as to other situations presented in Banks et al. (2015). Here we exclusively explore the relevance of ARA as part of the SEJ toolkit.

3 ARA as a Decomposition Approach

We study ARA as a decomposition approach within the sequential Defend-Attack model described above. There are two possible ways to assess distribution $p_D(a \,|\, d)$ needed for D's decision:

- One could do it directly with standard SEJ procedures (Cooke, 1991). Denote such assessment by $p_D^{SEJ}(a \,|\, d)$.
- Otherwise, one could assess it indirectly through ARA as in Section 2. D would assess A's beliefs and preferences from her own uncertain perspective, represented

by $(P_A, U_A) \sim F$, and then solve A's decision making problem using these random probabilities and utilities by computing

$$p_D^{ARA}(a \,|\, d) = \mathbb{P}_F \left[a = \arg\max_{x \in \mathcal{A}} \int U_A(x, s)\, P_A(s \,|\, d, x)\, \mathrm{d}s \right].$$

To compare both the SEJ and ARA approaches, it is convenient to make three simplifying assumptions: (i) D, whom we support, has only two options, defend (d_1) or not (d_0); (ii) A can solely choose to attack (a_1) or not (a_0); and (iii) if A decides to attack, the only two outcomes are success (s_1) or failure (s_0). For A, the problem can be viewed as the decision tree in Figure 4, with $d \in \{d_0, d_1\}$, which parallels the influence diagram in Figure 3. The ARA approach obtains the conditional probabilities $p_D^{ARA}(a \,|\, d)$ by solving the decision tree using D's (random) assessments for A's inputs.

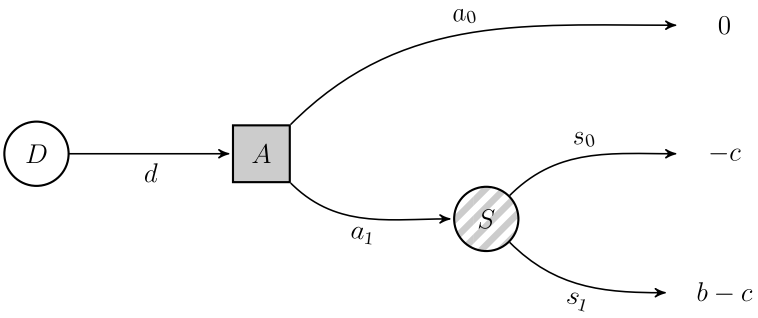


Figure 4: Decision tree representing A's problem.

For example, D may believe that A is an expected utility maximiser. She would then elicit her beliefs about A's probabilities and utilities (random variables to her) and use these to solve his problem from her perspective, which entails computing the (random) action that maximises his random expected utility. Thus, D should model both A's preferences for the consequences in Figure 4 as well as his beliefs about the likelihood of success.

Suppose D thinks A bases his decision on a cost-benefit analysis. In that case, the consequences for A are described in Table 1. Moreover, D may use a multi-attribute value model to decompose her assessment about A's valuation of consequences into simpler assessments regarding the costs and benefits associated with each of his consequences. Later, she can aggregate these assessments as shown in the row *Profit* in Table 1 and reflected in Figure 4.

	(Attack, Outcome) - $(\boldsymbol{a}, \boldsymbol{s})$		
	$(\boldsymbol{a_0}, \boldsymbol{s_0})$	$(\boldsymbol{a_1}, \boldsymbol{s_0})$	$(\boldsymbol{a_1}, \boldsymbol{s_1})$
Cost	0	c	c
Benefit	0	0	b
Profit	0	$-c$	$b-c$

Table 1: Cost-benefit analysis of A's consequences.

Specifically, this requires D to assess two quantities: c and b, A's cost of undertaking an attack and his benefit if successful, respectively. We assume that $0 < c < b$, implying that an attack (a_1) is a more costly action for A than not attacking (a_0), but potentially more beneficial; and that a successful attack (s_1) is better for A than an unsuccessful one (s_0). Since D is generally uncertain about these quantities, she will provide probability distributions to model her beliefs about them. Suppose her self-elicitations correspond to the uniform distributions:

- A's cost of an attack: $c \sim \mathcal{U}\left(c_{\min}, c_{\max}\right)$ (uniform).
- A's economic benefit from a successful attack: $b \sim \mathcal{U}\left(b_{\min}, b_{\max}\right)$.

These allow D to compute the random values related to A's consequences in Table 1. We have assumed that D believes that A's costs and benefits are uniformly distributed and, quite importantly, independent. However, in many cases there is dependence; e.g. a more costly attack is most likely correlated with larger gains for A. In that case, one needs to model c and b jointly. For simplicity, this example assumes independence.

If D believes that A is risk neutral (i.e. seeking to maximise his expected profit), she would now elicit her beliefs about A's beliefs for his probability of success. Otherwise, beforehand, she would have to model A's risk preferences over the random consequences. She could do that by eliciting a utility function over profits for him and model his risk attitude as shown in Section 4.2 and exemplified in Section 5, where her uncertainty about A's risk attitude is captured through a probability distribution over the risk aversion coefficient of a parametric utility function.

Alternatively, because there are just three possible outcomes for A (no attack, failed attack, successful attack), D may directly assess her belief about his utility for each of them. Without loss of generality, a utility of 0 and 1 can be respectively assigned to the worst and best consequences for A. Since D believes that $-c < 0 < b - c$, $u_A(-c) = 0$ and $u_A(b - c) = 1$, even when she does not know the exact values of b and c. Thus, she only needs to elicit her distribution for $u_A(0) = u$, knowing that $0 < u < 1$, though being uncertain of A's exact value of u. Recall that this could be elicited as the probability at which A is indifferent between getting profit 0 for sure and a lottery ticket in which he wins $b - c$ with probability u and looses c with probability $1 - u$. So D could elicit a distribution for the random variable U_A that represents her full uncertainty over A's utility u.

D would also need to assess A's beliefs about his chance of success, determined by $p_A(s_1 \,|\, d_0, a_1) = \pi_{d_0}$ and $p_A(s_1 \,|\, d_1, a_1) = \pi_{d_1}$. She should model her uncertainty about these with random probabilities $\pi_{d_0} \sim P_A^{d_0}$ and $\pi_{d_1} \sim P_A^{d_1}$, with $\pi_{d_1} < \pi_{d_0}$ to ensure that defending (d_1) reduces the chance of a successful attack. Then for each of her actions $d \in \{d_0, d_1\}$, D can compute A's random expected utilities as

$$\Psi(d, a_0) = u \sim U_A, \qquad \Psi(d, a_1) = \pi_d \sim P_A^d,$$

and the ARA probabilities of attack as

$$p_D^{ARA}(a_1 \,|\, d) = \mathbb{P}_{(U_A, P_A^d)}\left(u < \pi_d\right). \tag{1}$$

Once she has self-elicited her distributions for U_A, $P_A^{d_0}$ and $P_A^{d_1}$, she may compute the attack probabilities as in (1), which represent D's ARA probabilistic predictions of how A will respond to each of her possible choices. For example, assuming that these distributions are $U_A \sim \mathcal{B}e(1,2)$ (beta) and $P_A^{d_0} \sim \mathcal{U}(0.5, 1)$ and $P_A^{d_1} \sim \mathcal{U}(0.1, 0.4)$, then Monte Carlo (MC) approximation estimates the attack probabilities as $\hat{p}_D^{ARA}(a_1 \,|\, d_0) \approx 0.92$ and $\hat{p}_D^{ARA}(a_1 \,|\, d_1) \approx 0.43$ (based on an MC sample size of 10^6). In this case, choosing to defend (d_1) acts as a deterrent for A to attack (a_1).

We now address whether this ARA decomposition approach leads to better attack probability estimates than those obtained by standard SEJ methods. Adopting a normative viewpoint, we show through simulation that under certain conditions the variance of the ARA estimates are smaller than those of the SEJ estimates.

In our case, due to the assumptions behind expression (1), we have no reason to believe that D finds one attack distribution more (or less) likely than another, except that an attack is more likely when no defence is attempted. That is, $p_{d_0}^{SEJ} \geq p_{d_1}^{SEJ}$ where $p_{d_0}^{SEJ} = p_D^{SEJ}(a_1 \,|\, d_0)$ and $p_{d_1}^{SEJ} = p_D^{SEJ}(a_1 \,|\, d_1)$. Thus, as a high-entropy benchmark, we assume that $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ are uniformly distributed over the set $\{0 \leq p_{d_1}^{SEJ} \leq p_{d_0}^{SEJ} \leq 1\}$, whose variance-covariance matrix is

$$\begin{pmatrix} \frac{1}{18} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{18} \end{pmatrix} \approx \begin{pmatrix} 5.56 & 2.78 \\ 2.78 & 5.56 \end{pmatrix} \cdot 10^{-2}. \tag{2}$$

As before, D's assessment of the ARA attack probabilities involves eliciting distributions $(U_A, P_A^{d_0}, P_A^{d_1})$. It is reasonable to assume that u is independent of π_{d_0} and π_{d_1}.

Since the support of all three random variables is $[0,1]$, an equitable framework for the benchmark may assume that $U_A \sim \mathcal{U}(0,1)$ and $(P_A^{d_0}, P_A^{d_1})$ are uniformly distributed over the set $\{0 \leq \pi_{d_1} \leq \pi_{d_0} \leq 1\}$. We conducted 10^4 MC estimates of the attack probabilities using these distributions, each based on an MC sample size of 10^4, leading to a variance-covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ of

$$\begin{pmatrix} 2.24 & 1.10 \\ 1.10 & 2.22 \end{pmatrix} \cdot 10^{-5}. \tag{3}$$

Thus, as a result of the decomposition approach inherent to the ARA methodology, both variances and the covariance in the ARA approach (3) are significantly smaller than those in the SEJ benchmark (2), providing a more precise assessment.

Typically, one would have more information about $(U_A, P_A^{d_0}, P_A^{d_1})$. For example, suppose D believes that the mean values of the three random variables are $E[U_A] = \frac{2}{5}$, $E[P_A^{d_0}] = \frac{2}{3}$ and $E[P_A^{d_1}] = \frac{1}{3}$. If she assumes they all are uniformly distributed with maximum variance, then $U_A \sim \mathcal{U}(0, \frac{4}{5})$, $P_A^{d_0} \sim \mathcal{U}(\frac{1}{3}, 1)$ and $P_A^{d_1} \sim \mathcal{U}(0, \frac{2}{3})$ (and $\pi_{d_1} \leq \pi_{d_0}$), and the variance-covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ is

$$\begin{pmatrix} 1.42 & 0.65 \\ 0.65 & 2.35 \end{pmatrix} \cdot 10^{-5}.$$

Compared to (3), these assumptions reduce the variance for $p_{d_0}^{ARA}$ and the covariance, although slightly increase the variance of $p_{d_1}^{ARA}$.

Finally, if the random variables followed beta distributions with common variance $\frac{1}{10}$, then $U_A \sim \mathcal{B}e(0.56, 0.84)$, $P_A^{d_0} \sim \mathcal{B}e(0.81, 0.41)$ and $P_A^{d_1} \sim \mathcal{B}e(0.41, 0.81)$ (and $\pi_{d_1} \leq \pi_{d_0}$), and the variance-covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ is

$$\begin{pmatrix} 1.52 & 0.64 \\ 0.64 & 2.25 \end{pmatrix} \cdot 10^{-5}.$$

Again, the covariance matrix is significantly more precise than the SEJ benchmark.

For further comparison of the SEJ and ARA benchmarks, assume that the SEJ elicitation process incorporates additional information, so that $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ are now uniformly distributed over the set $\{\varepsilon \leq p_{d_1}^{SEJ} \leq p_{d_0}^{SEJ} \leq 1 - \varepsilon\}$ with $0 \leq \varepsilon \leq \frac{1}{2}$. Then the variance-covariance matrix for $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ is

$$\begin{pmatrix} \frac{(1-2\varepsilon)^2}{18} & \frac{(1-2\varepsilon)^2}{36} \\ \frac{(1-2\varepsilon)^2}{36} & \frac{(1-2\varepsilon)^2}{18} \end{pmatrix}. \tag{4}$$

From (3) and (4), we see that one must take $\varepsilon > 0.49$, a very precise assessment, in order for the corresponding variance-covariance matrix of $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ to be less variable than $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$.

All these comparisons indicate that although the ARA approach requires more assessments to obtain the relevant probabilities of the adversarial actions, ARA tends to provide more precise ones. However, if the SEJ information is sufficiently specific, then SEJ can outperform ARA in terms of reduced variance for the relevant random variables.

4 ARA Modelling Strategies

We have shown that the ARA decomposition can have advantages over the plain SEJ approach. We now describe strategies to implement ARA through random probabilities and utilities.

4.1 Random probabilities

We focus first on D's assessments over A's perspective of the different random events involved in the decision making problem, that is, the random probabilities. To fix ideas, assume we have a single chance node to model, say S, which depends on both D's and A's choices. We designate D's probability model for S by $p_D(s \mid d, a)$. Our task is to develop a (random) model $P_A(s \mid d, a)$ that reflects D's uncertainty about A's prospect of S. We distinguish three cases. In all of them, Bayesian updating could be used to dynamically adjust the assessed prior distributions as data accumulates, thus attaining subsequent random posterior distributions that better reflect D's perspective over A's uncertainty, as shown in Section 4.1.1.

4.1.1 Probability of a single event

Suppose first that the chance node S just consists of a single event s which may $(s = 1)$ or not $(s = 0)$ happen. Then $p_A(s \mid d, a)$ is completely determined by $p_A(s = 1 \mid d, a)$, for each of the possible combinations of D's and A's decisions, as $p_A(s = 0 \mid d, a) = 1 - p_A(s = 1 \mid d, a)$.

One possibility would be to base $P_A(s = 1 \mid d, a)$ on an estimate π_D of $p_A(s = 1 \mid d, a)$, with some uncertainty around it. This may be accomplished in several ways. We could do it through a uniform distribution $\mathcal{U}(\pi_D - \mu, \pi_D + \mu)$ centred around π_D in which the parameter μ would also have to be assessed, e.g. in terms of the expected variance ν of the distribution so that $\mu = \sqrt{3\nu}$. Another option would be to use a beta distribution $\mathcal{B}e(\alpha, \beta)$ in which π_D may be regarded as the mean (or the median or mode) of the distribution and we would have to assess the parameters α and β to shape the distribution, e.g. based on a further assessment of the variance ν. This would lead, when

π_D is the mean, to:

$$\alpha = \frac{\pi_D}{\nu}\left(\pi_D\left(1-\pi_D\right)-\nu\right), \qquad \beta = \frac{1-\pi_D}{\nu}\left(\pi_D\left(1-\pi_D\right)-\nu\right)$$

Note that when D thinks that A has information similar to hers, an adequate best guess for π_D could be based on her own assessment of $p_D(s = 1 \,|\, d, a)$.

If the possible occurrence of single event s were to be repeated over time, random prior distributions could be reassessed by means of Bayesian updating. Consider, for example, the second case in which a beta distribution $\mathcal{B}e(\alpha, \beta)$ is used to approximate $P_A(s = 1 \,|\, d, a)$. If event s has had y chances to happen and has only materialised z times (and $y - z$ has not), our new random posterior would be $\mathcal{B}e(\alpha + z, \beta + y - z)$.

4.1.2 Probabilities of multiple events

We assume now that the chance node S includes N events $\{s_1, \ldots, s_N\}$. In this case, probabilities $p_A(s = s_1 \,|\, d, a), \ldots, p_A(s = s_{N-1} \,|\, d, a)$ would determine $p_A(s \,|\, d, a)$ completely, for each pair of D's and A's decisions, as $p_A(s = s_N \,|\, d, a) = 1 - \sum_{n=1}^{N-1} p_A(s = s_n \,|\, d, a)$. Therefore, we only need to model $P_A(s = s_1 \,|\, d, a), \ldots, P_A(s = s_{N-1} \,|\, d, a)$, which we jointly designate $P_A(s \,|\, d, a)$.

In line with the previous case, we could base $P_A(s \,|\, d, a)$ on a best guess $\pi_D(s)$, for example $p_D(s \,|\, d, a)$ when D believes that A has similar information, with some uncertainty around it. We could use a parametric probability distribution, randomising each of its parameters much as we have done in the preceding section. In this manner, for each pair of decisions d and a, we could estimate $\pi_{D,n}$ of $p_A(s = s_n \,|\, d, a)$ $\forall n \in \{1, \ldots, N-1\}$ and then incorporate the uncertainty through a uniform $\mathcal{U}(\pi_{D,n} - \mu_n, \pi_{D,n} + \mu_n)$ or beta distribution $\mathcal{B}e(\alpha_n, \beta_n)$ centred around $\pi_{D,n}$, making sure that their sum does not exceed 1.

A more effective way would be modelling $P_A(s \,|\, d, a)$ as a Dirichlet distribution with mean $\pi_D(s)$ and parameters assessed based on one further judgement concerning, e.g. the variance of one of them. To do this, for each pair of decisions (d, a), we would obtain from D an estimate $\pi_{D,n}$ of $p_A(s = s_n \,|\, d, a)$ $\forall n \in \{1, \ldots, N\}$ and associate random variables S_n such that $E\left[S_n\right] = \pi_{D,n}$. Their joint distribution could then be described as Dirichlet, $(S_1, \ldots, S_N) \sim \mathcal{D}ir(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_N)$. If $\hat{\alpha} = \sum_{n=1}^{N} \alpha_n$, it follows that

$$E\left[S_n\right] = \frac{\alpha_n}{\hat{\alpha}}, \qquad Var\left[S_n\right] = \frac{\alpha_n(\hat{\alpha} - \alpha_n)}{\hat{\alpha}^2(\hat{\alpha} + 1)};$$

and it suffices to fix a value, e.g. $Var\left[S_1\right]$, to calculate the required α_n parameters.

4.1.3 The continuous case

We consider now the case in which the chance node S involves a continuous set of events. Trying to determine a random probability for each realisation s is no longer possible, so we need to model a distribution over the various probability distributions that A might entertain.

Techniques are similar to those described to assess the probabilities of multiple events. We could base $P_A(s \,|\, d, a)$ on a guess $\pi_D(s)$, say $p_D(s \,|\, d, a)$, with some uncertainty around it. For example, this may be achieved by means of a Dirichlet process, with base distribution $\pi_D(s)$ and concentration parameter ρ as perceived by D, which allows to sample approximate distributions of $P_A(s \,|\, d, a)$. Other non-parametric approaches such as hidden Markov models or hierarchical Pitman-Yor processes (Teh and Jordan, 2010) could be used with reference to the above guess.

4.2 Random utilities

We draw now attention over D's perspective on A's preference assessments concerning the consequences of the decision making problem, that is, the random utilities. We shall usually have some information about A's multiple interests, e.g. when dealing with terrorism problems, Keeney (2007) and Keeney and von Winterfeldt (2010) present extensive classifications of criteria amongst which to choose. Keeney (2007) then advocates that standard utility methods may be adopted by interviewing experts in the problem at hand, therefore developing utility functions modelling A's preferences. However, note that such preferences are not directly elicited from A, but rather through a surrogate. Thus, intrinsically, there is uncertainty about A's preferences.

An alternative approach, illustrated in Banks et al. (2015), is to aggregate the objectives with a weighted measurable value function, as in Dyer and Sarin (1979). As an example, we could consider an additive value function for A in which his objectives $v_1, \ldots, v_R$ may be aggregated using weights $w_1, \ldots, w_R \geq 0$, $\sum_{r=1}^R w_r = 1$ as $v_A = \sum_{r=1}^R w_r \, v_r$. The uncertainty about the weights could be modelled using a Dirichlet distribution, just as in Section 4.1.2, so that we may estimate their value and then associate random variables W_r such that $E\left[W_r\right] = w_r$, their joint distribution being Dirichlet, $(W_1, \ldots, W_R) \sim \mathcal{D}ir(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_R)$ with one further judgement, e.g. fixing one of the parameter's variance. Finally, using the relative risk aversion concept (Dyer and Sarin, 1982), we could assume different risk attitudes when modelling A's utility function. Continuing the example and assuming an exponential utility function, we may transform the (random) value function $V_A = \sum_{r=1}^R W_r \, v_r$ into one of the three following utilities depending on A's risk attitude:

- *Risk aversion.* $U_A = 1 - \exp(-\lambda\, V_A + c), \quad \lambda > 0.$
- *Risk neutrality.* $U_A = V_A + c.$
- *Risk proneness.* $U_A = \exp(\lambda\, V_A + c), \quad \lambda > 0.$

Further uncertainty about the risk coefficient λ and the adjusting constant c may be modelled e.g. through uniform distributions $\Lambda \sim \mathcal{U}(\lambda_1, \lambda_2)$ and $C \sim \mathcal{U}(c_1, c_2)$. In any case, to determine all the required distributions, we may ask experts to directly elaborate such distributions or request them to provide point estimates of the weights and coefficients and build the distributions from these.

An alternative approach for building a distribution over A's preferences is described in Wang and Bier (2013). As before, we suppose that they are represented through a multi-attribute utility function, which may involve the above attributes $v_1, \ldots, v_R$ as well as an unobserved one v_0. For simplicity, consider A's utility to be linear in each of the attributes which are additively independent. Then we ask several experts to provide rank orders of A's action valuations and derive probability distributions that can match those orderings to obtain the (random) weights $(W_0, W_1, \ldots, W_R)$ for his utility function. To do this, we consider as input such rankings and as output a distribution over A's preferences (expected utilities) for which two methods are suggested. One is an adaptation of probabilistic inversion (Neslo et al., 2008); essentially, it identifies a probability distribution Q over the space of all possible attribute weights $(W_0, W_1, \ldots, W_R)$ that can match the empirical distribution matrix of expert rankings with minimum Kullback-Leibler divergence to a predetermined (e.g. non-informative, Dirichlet) starting probability measure Q_0. The other one uses Bayesian density estimation (Müller et al., 2015) based on a prior distribution Q_p (e.g. chosen in accordance to a Dirichlet process with base distribution Q_0) over the space of attribute weights $(W_0, W_1, \ldots, W_R)$ and treating the expert rankings as observations to update that prior leading to a posterior distribution Q, obtained through a Gibbs sampling scheme.

5 A Numerical Example

As an illustration, we consider a sequential defend-attack cybersecurity problem. A user (D, defender) needs to make a connection to a site, either through a safe, but costly, route (d_0) or through a cheaper, but more dangerous protocol. In the latter case, she may use a security key, rendering the protocol less dangerous. While using the dangerous protocol, whether unprotected (d_1) or protected by a security key (d_2), the defender may be the target of a cybercriminal who may decide to attack (a_1) or not (a_0). Table

2 (respectively, Table 3) display the defender's (respectively, attacker's) consequences, expressed as costs, for the various defend and attack possibilities.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	h	—
	$\boldsymbol{d_1}$	0	$c\,\theta_1$
	$\boldsymbol{d_2}$	k	$k + c\,\theta_2$

Table 2: Defender's loss function.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	0	—
	$\boldsymbol{d_1}$	0	$L - G\,\theta_1$
	$\boldsymbol{d_2}$	0	$L - G\,\theta_2$

Table 3: Attacker's loss function.

The following parameters are used: (i) h is the cost of using the expensive protocol; (ii) θ_1 is the fraction of assets lost by the defender when attacked and unprotected; (iii) θ_2 is the fraction of assets lost by the defender when attacked but protected; (iv) k is the security key's cost; (v) c is the defender's scaling cost relative to the fraction of assets lost; (vi) L is the uncertain cost of an attack; and (vii) G is the uncertain cybercriminal's scaling gain relative to the fraction of assets lost by the defender. The global problem may be viewed through the game tree in Figure 5.

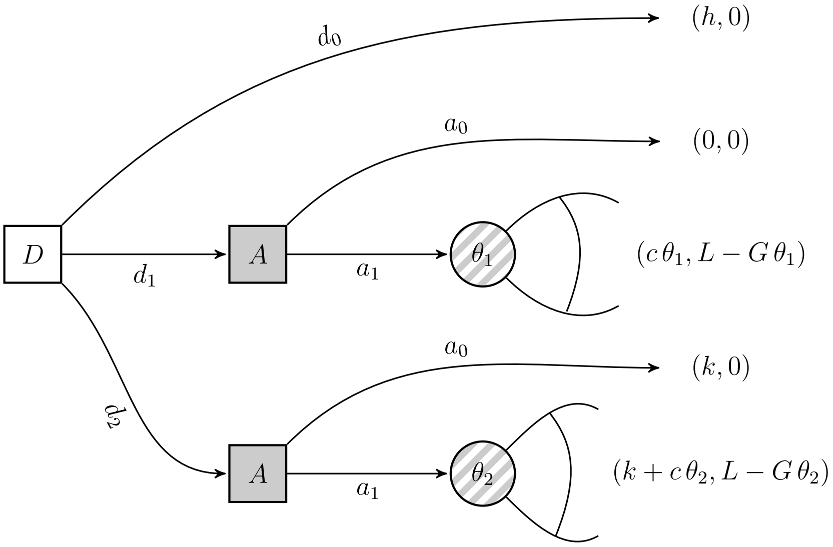


Figure 5: Game tree for the cybersecurity routing problem (losses).

The defender believes that the asset fractions θ_i follow distributions $p_D(\theta_i \,|\, d_i, a_1)$ with $\theta_i \sim \mathcal{B}e(\alpha_i^D, \beta_i^D)$, $i = 1, 2$. She is risk averse and her utility function is strategically equivalent to $1 - e^{\lambda_D\, x}$, where x is her cost and $\lambda_D > 0$ her risk aversion coefficient.

The attacker has different beliefs about θ_i, $p_A(\theta_i \,|\, d_i, a_1)$, with $\theta_i \sim \mathcal{B}e(\alpha_i^A, \beta_i^A)$, $i = 1, 2$; the defender's uncertainty about α_i^A and β_i^A inducing its randomness. He is risk prone and his utility function is strategically equivalent to $e^{-\Lambda_A\, x} - 1$, where x is his cost and $\Lambda_A > 0$ his uncertain risk proneness coefficient. Both agents expect θ_1 to be greater than θ_2, but not necessarily. This may be reflected in the choice of the beta parameters, for example with $E\,[\theta_1] = \frac{\alpha_1^D}{\alpha_1^D + \beta_1^D} > \frac{\alpha_2^D}{\alpha_2^D + \beta_2^D} = E\,[\theta_2]$, in the case of the defender. Table 4 (respectively, Table 5) provides the defender's (respectively, attacker's random) expected utilities u_D (respectively, U_A) under the various interaction scenarios.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	$1 - e^{\lambda_D\, h}$	—
	$\boldsymbol{d_1}$	0	$1 - \int e^{\lambda_D\, c\,\theta_1}\, p_D(\theta_1)\, \mathrm{d}\theta_1$
	$\boldsymbol{d_2}$	$1 - e^{\lambda_D\, k}$	$1 - \int e^{\lambda_D (k + c\,\theta_2)}\, p_D(\theta_2)\, \mathrm{d}\theta_2$

Table 4: Defender's expected utility.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	0	—
	$\boldsymbol{d_1}$	0	$\int e^{\Lambda_A (G\,\theta_1 - L)}\, P_A(\theta_1)\, \mathrm{d}\theta_1 - 1$
	$\boldsymbol{d_2}$	0	$\int e^{\Lambda_A (G\,\theta_2 - L)}\, P_A(\theta_2)\, \mathrm{d}\theta_2 - 1$

Table 5: Attacker's random expected utility.

Suppose we assess from the defender: (i) a cost of $h = 150,000$ € for implementing the expensive protocol; (ii) a security key's cost of $k = 50,000$ € to be protected when using the dangerous protocol; (iii) a scaling cost of $c = 200,000$ € relative to the fraction of assets lost; (iv) a risk aversion coefficient of $\lambda_D = 3 \cdot 10^{-5}$; (v) the distribution $\theta_1 \sim \mathcal{B}e(\alpha_1^D, \beta_1^D)$ with a expected fraction (mean) of 0.6 of the assets lost and standard deviation 0.15 when attacked and unprotected, leading to $\alpha_1^D = 0.36$ and $\beta_1^D = 0.24$; and (vi) the distribution $\theta_2 \sim \mathcal{B}e(\alpha_2^D, \beta_2^D)$ with a expected fraction (mean) of 0.3 of the assets lost and standard deviation 0.07 when attacked but protected, leading to $\alpha_2^D = 0.6$ and

$\beta_2^D = 1.4$. These are standard decision analytic assessments and the resulting problem faced by her is described in the decision tree in Figure 6.

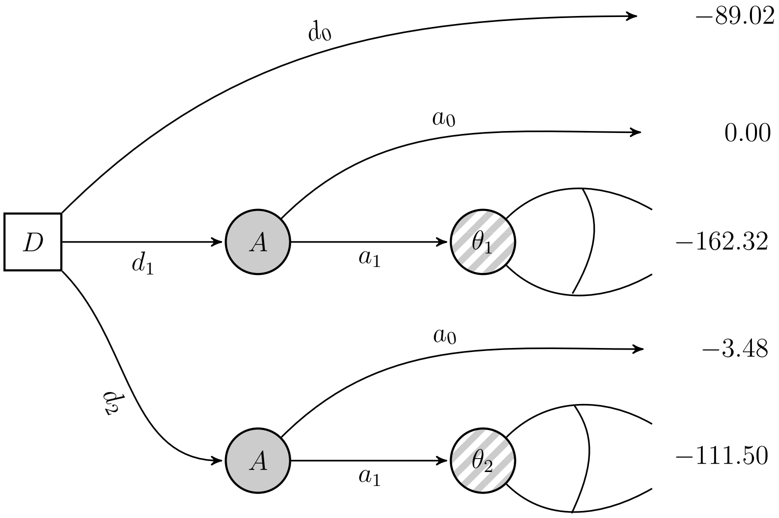


Figure 6: Decision tree representing the defender's problem (expected utilities).

The expected utility of the first alternative (d_0, use the expensive protocol) may be directly estimated as

$$\psi_D(d_0) = 1 - e^{\lambda_D\, h} \approx -89.02,$$

given the fact that there is no chance of attack in this scenario. However, those of the other two alternatives have the form

$$\psi_D(d_i) = \sum_{j=0}^{1} p_D(a_j \,|\, d_i)\, u_D(d_i, a_j), \quad i = 1, 2;$$

where $u_D(d_i, a_j)$ may be obtained from Table 4 with the specific values indicated in Figure 6. Thus, we need to assess the attack probabilities $p_D(a_1 \,|\, d_i)$ (and $p_D(a_0 \,|\, d_i) = 1 - p_D(a_1 \,|\, d_i)$) given the implemented defence d_i.

Suppose that, in line with Section 4, we assess that: (i) the uncertain cost of an attack is $L \sim \mathcal{U}(10^4, 2 \cdot 10^4)$ with an expected cost of $15,000$ €; (ii) the uncertain cybercriminals scaling gain relative to the fraction of assets lost by the defender is $G \sim \mathcal{U}(10^4, 5 \cdot 10^4)$ with an expected scaling gain of $30,000$ €; (iii) his risk proneness coefficient is $\Lambda_A \sim \mathcal{U}(10^{-4}, 2 \cdot 10^{-4})$ with an expectation of $1.5 \cdot 10^{-4}$; (iv) the distribution $\theta_1 \sim \mathcal{B}e(\alpha_1^A, \beta_1^A)$ has a expected fraction (mean) of 0.6 assets lost when the defender is attacked but protected, with $\alpha_1^A \sim \mathcal{U}(5, 7)$ and $\beta_1^A \sim \mathcal{U}(3, 5)$; and (v) the distribution $\theta_2 \sim \mathcal{B}e(\alpha_2^A, \beta_2^A)$ has a expected fraction (mean) of 0.3 assets lost when the defender is attacked but protected, with $\alpha_2^A \sim \mathcal{U}(2, 4)$ and $\beta_2^A \sim \mathcal{U}(6, 8)$. We may then use

Algorithm 1 to estimate the required probabilities $\hat{p}_D(a_1 \mid d)$, where $\Psi_A^n(d_i, a)$ designates the expected utility that the cybercriminal reaches when the defender implements d, he chooses action a and the sampled parameters are $l^n, g^n, \lambda_A^n, \alpha_i^{A,n}, \beta_i^{A,n}$.

Algorithm 1 Numerical example: Simulation of $\hat{p}_D(a_1 \mid d)$

Data: Number of iterations N.

```
Set $p_1, p_2 = 0$.
For $n = 1$ to $N$ do
    Draw $l^n$ from $\mathcal{U}(10^4, 2 \cdot 10^4)$ and $g^n$ from $\mathcal{U}(10^4, 5 \cdot 10^4)$.
    Draw $\lambda_A^n$ from $\mathcal{U}(10^{-4}, 2 \cdot 10^{-4})$.
    Draw $\alpha_1^{A,n}$ from $\mathcal{U}(2, 7)$ and $\beta_1^{A,n}$ from $\mathcal{U}(1, 5)$.
    Draw $\alpha_2^{A,n}$ from $\mathcal{U}(0, 3)$ and $\beta_2^{A,n}$ from $\mathcal{U}(1, 6)$.
    For $i = 1$ to 2 do
        $\Psi_A^n(d_i, a_0) = 0$.
        $\Psi_A^n(d_i, a_1) = \int e^{\lambda_A^n (g^n \theta_i - l^n)} \frac{\theta_i^{\alpha_i^{A,n}-1} (1-\theta_i)^{\beta_i^{A,n}-1}}{\mathrm{Beta}(\alpha_i^{A,n}, \beta_i^{A,n})} \, \mathrm{d}\theta_i - 1$.
        If $\Psi_A^n(d_i, a_1) \geq \Psi_A^n(d_i, a_0)$ then
            $p_i = p_i + 1$.
        End If
    End For
End For
For $i = 1$ to 2 do
    $\hat{p}(a_1 \mid d_i) = p_i / N$.
End For
```

In our case, with $N = 10^6$, we obtain $\hat{p}(a_1 \mid d_1) = 0.66$ (and, consequently, $\hat{p}(a_0 \mid d_1) = 0.34$). Similarly, $\hat{p}(a_1 \mid d_2) = 0.23$ (and $\hat{p}(a_0 \mid d_2) = 0.77$). Then, we have $\psi_D(d_0) = -89.02$, $\psi_D(d_1) = -107.13$ and $\psi_D(d_2) = -28.32$ being the optimal cyberdefense $d^*_{ARA} = d_2$, that is using the dangerous protocol protected by the security key.

6 Discussion

Adversarial Risk Analysis is an emergent paradigm when supporting a decision maker who faces adversaries and such that the consequences are random and depend on the actions of all participating agents. We have illustrated the relevance of such approach as a decomposition technique to forecast adversarial actions in game theoretic contexts, which could be added to the SEJ toolkit. We have also presented key implementation strategies. We have limited the analysis to the simpler sequential case, but ideas extend to simultaneous problems, albeit with technical difficulties, due to the belief recursions typical of level-k thinking.

Throughout the examples expert judgement has been assessed assuming just one single expert. However, in practice, several experts might be available and aggregation techniques such as Cooke's classical method (Cooke, 1991) would be needed. Diverse adversarial rationalities, such as non-strategic or prospect-maximising players, could be handled by means of mixtures.

We have seen how the ARA decomposition strategy breaks down an attack probability assessment into multi-attribute utility and probability assessments for the adversary. For the ARA approach to be worthwhile, it is expected that the resulting probabilities are more accurate than the ones that would have been directly obtained and, also, that the corresponding increased number of necessary judgements are cognitively easier. Experiments should be conducted to validate these ideas.

Last, we have shown that ARA might improve the results of direct SEJ. As this is not always the case, we would study the combination of both techniques to provide a broader view on the prescriptive optimal decisions.

Acknowledgements

The work of DRI is supported by the Spanish Ministry of Economy and Innovation programs MTM2014-56949-C3-1-R and MTM2017-86875-C3-1-R, the ESF-COST Action IS1304 on Expert Judgement and the AXA-ICMAT Chair on Adversarial Risk Analysis. Besides, JGO's research is financed by the Spanish Ministry of Economy and Competitiveness under FPI SO grant agreement BES-2015-072892. This work has also been partially supported by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Program for Centers of Excellence in R&D (SEV-2015-0554) and project MTM2015-72907-EXP, and has received funding from the European Union's H2020 Program for Research, Technological Development and Demonstration, under grant agreement no. 740920 (CYBECO).

7 An Experiment

We have seen how the ARA decomposition strategy breaks down an attack probability assessment into multi-attribute utility and probability assessments for A. This strategy increases the number of judgements necessary to determine attack probabilities, although these might be simpler. For the ARA approach to be worthwhile, it is expected that the resulting probabilities are more accurate than the ones that would have been directly obtained. Also, for the ARA approach to be practical it is expected that the corresponding assessments will be cognitively easier so that all the necessary inputs can be obtained to compute the ARA attack probabilities. We describe here an experiment performed to validate these ideas.

The essence of the experiment. We have a group of participants. We randomly split them into two groups: I and II. Each group makes a few assessments, say five. Both groups have to assess the same probabilities in an adversarial situation.

- *Group I receives just the event to be assessed.*
- *Group II receives the event but also the decomposition and they are asked to assess the random probabilities and utilities.*

References

S. Andradottir and V. M. Bier. Choosing the number of conditioning events in judgemental forecasting. *Journal of Forecasting*, 16(4): 255–286, 1997.

S. Andradottir and V. M. Bier. An analysis of decomposition for subjective estimation in decision analysis. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 28(4): 443–453, 1998.

D. Banks, J. Ríos and D. Ríos Insua. *Adversarial Risk Analysis* (2016 ed.). CRC Press, Boca Raton, FL, 2015.

T. Bedford and R. M. Cooke. *Probabilistic Risk Analysis: Foundations and Methods* (2011 ed.). Cambridge University Press, Cambridge, United Kingdom, 2001.

E. Chen, D. V. Budescu, S. K. Lakshmikanth, B. A. Mellers and P. E. Tetlock. Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, 13(3): 128-152, 2016.

R. T. Clemen and T. Reilly. *Making Hard Decisions with DecisionTools* (2013 ed.). Cengage Learning, Mason, OH, 2013.

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

J. S. Dyer and R. K. Sarin. Group preference aggregation rules based on strength of preference. *Management Science*, 25(9): 822–832, 1979.

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

S. French and D. Ríos Insua. *Statistical Decision Theory* (2000 ed.). Wiley, New York, NY, 2000.

J. González-Ortega, V. Radovic and D. Ríos Insua. Utility elicitation. In *Elicitation: The Science and Art of Structuring Judgement* (pp. 241–264), Springer, New York, NY, 2018.

L. H. J. Goossens, R. M. Cooke and B. C. P. Kraan. Evaluation of weighting schemes for expert judgement studies. *Proceedings of the Fourth International Conference on Probabilistic Safety Assessment and Management*: 1937–1942, 1998.

S. Hargreaves-Heap and Y. Varoufakis. *Game Theory: A Critical Introduction* (2004 ed.). Routledge, New York, NY, 1995.

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.

D. G. MacGregor. Decomposition for judgmental forecasting and estimation. In *Principles of Forecasting* (pp. 107–123), Springer, Boston, MA, 2001.

D. G. MacGregor and J. S. Armstrong. Judgmental decomposition: When does it work? *International Journal of Forecasting*, 10(4): 495–506, 1994.

G. Montibeller and D. von Winterfeldt. Biases and debiasing in multi-criteria decision analysis. *IEEE 2015 48th Hawaii International Conference on System Sciences*: 1218–1226, 2015.

P. Müller, F.A. Quintana, A. Jara and T. Hanson. *Bayesian Nonparametric Data Analysis* (2015 ed.). Springer, Cham, Switzerland, 2015.

R. Neslo, F. Micheli, C. V. Kappel, K. A. Selkoe, B. S. Halpern and R. M. Cooke. Modeling stakeholder preferences with probabilistic inversion: Application to prioritizing marine ecosystem vulnerabilities. In *Real-Time and Deliberative Decision Making* (pp. 265–284), Springer, Dordrecht, Netherlands, 2008.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. arthwaite, D. J. Jenkinson, J. E. Oakley and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities* (2006 ed.). Wiley, Chichester, UK, 2006.

H. Raiffa. *Decision Analysis: Introductory Lectures on Choices Under Undertainty* (1968 ed.). Addison-Wesley, Menlo Park, CA, 1968.

H. Raiffa. *The Art and Science of Negotiation* (2003 ed.). Harvard University Press, Cambridge, MA, 1982.

H. V. Ravinder and D. N. Kleinmuntz. Random error in additive decompositions of multiattribute utility. *Journal of Behavioral Decision Making*, 4(2): 83–97, 1991.

H. V. Ravinder, D. N. Kleinmuntz and J. S. Dyer. The reliability of subjective probabilities obtained through decomposition. *Management Science*, 34(2): 186–199, 1988.

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

Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In *Cambridge Series in Statistical and Probabilistic Mathematics: Bayesian nonparametrics* (pp. 158–207), Cambridge University Press, New York, NY, 2010.

P. E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction* (2015 ed.). Broadway Books, New York, NY, 2015.

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

S. R. Watson and R. V. Brown. The valuation of decision analysis. *Journal of the Royal Statistical Society. Series A (General)*, 141(1): 69–78, 1978.

Annex 9: Skeleton and examples of the new R Routines

Annex 9: Skeleton and examples of new R routines

Basic schema of the R script

The script

1. Configures the R environment that runs the simulation (e.g., check if required libraries are in the system)
2. Reads the input from the CYBECO Toolbox:
    1. *Front-end input* with the parameters selected by the end user.
    2. *Back-end input* with parameters defined by the admin user.
3. Checks whether the input is valid and loads it to the R environment.
4. Runs the simulations.
    1. Simulates the hacktivists and cybercriminals problems to forecast their attacks. For each agent:
        1. Definition of decisions and portfolios of decisions.
        2. Configuration of simulation (needs the decision portfolios).
        3. Load the functions for each node, which has its own script containing its functions (some nodes have the same functions in both the attacker and defender problems, other nodes might have different functions).
        4. Problem-solving through simulation and optimization. Implementation of the algorithm for the attacker (see annex 1).
    2. Simulates the defender problem to find optimal portfolio.
        1. Definition of decisions and portfolios of decisions.
        2. Configuration of simulation (needs the decision portfolios).
        3. Load the functions for each node, which has its own script containing its functions (some nodes have the same functions in both the attacker and defender problems, other nodes might have different functions).
        4. Problem-solving through optimization. Implementation of the algorithm for the defender (see annex 1).
5. Generates the results of the analysis.
6. Writes the output for the CYBECO Toolbox.

Files of the algorithm

- **toolbox_model.R** – R script that executes the algorithm.
- **\input** – Folder with input files.
  - **frontend_input.R** – input from CYBECO Toolbox user interface.
  - **backend_input.R** – input from CYBECO Toolbox admin user.
- **\output** – Folder with output files
  - **results_csv.R** – output for CYBECO Toolbox user interface.
- **\config** – Configuration files of algorithm (eg, check libraries or valid input).
- **\analysis** – Folder with script containing the model.
  - **\constraints** – Folder with script modelling the constraints (e.g., budget).
  - **\nodes** – Folder with script modelling the different nodes of the model (e.g., environmental threats).
  - **\problems** – Folder with script defining the problems for the different agents involved (e.g., defender, hacktivists).
  - **\simulations** – Folder with script modelling the heuristics and iterations to run the simulation.
  - **\summaries** – Folder with script that generates the results of the analysis.

Example scripts

Environmental threat

```
#! Rscript N03_envthreats.R --

#### ENVIRONMENTAL THREAT: FIRE ----
envThreatFire <- function () {
  if (envthreat_fire_included == FALSE) {0}
  else if (envthreat_fire_included == TRUE) {
    min(1,rpois(1,.022))
  } else {cat("Error in N03_envthreats.R")}
}

d_envThreatFire <- function () {
  if (envthreat_fire_included == FALSE) {0}
  else if (envthreat_fire_included == TRUE) {
    .022
  } else {cat("Error in N03_envthreats.R")}
}
```

Defender utility

```
#! Rscript N15H_defender_utility.R --

#### DEFENDER UTILITY ----

# Function that models the defender utility
defenderUtility <- function(defender_mon_results,
                            iimp_pii_records_exposed) {
 monetary_result <- feature_turnover_money-defender_mon_results
 defender_utility <- 1-utility_defender_coef_exp*(1-exp(utility_defender_rho*(monetary_result +
pii_personal_value*iimp_pii_records_exposed)))
 defender_utility
}
```

Hacktivist problem

```
#! Rscript hacktivists.R --

####  CONFIGURES THE DECISION-MAKING ----

#  generates the hacktivists decisions

ifelse(tarthreat_dataexf_included == TRUE,
                                N06H1_targexf_options <-c(1,0),
                                N06H1_targexf_options <-c(0))

ifelse(tarthreat_dataman_included == TRUE,
                                N06H1_targman_options <-c(1,0),
                                N06H1_targman_options <-c(0))

ifelse(tarthreat_dos_included == TRUE,
                                N06H1_targdos_options <- c(1,0),
                                N06H1_targdos_options <-  c(0))


#  generates the decision portfolio of the hacktivists,
# which is a table with all the possible combinations of decisions
H_decision_portfolios <- expand.grid(H_targexf_decision = N06H1_targexf_options,
                                   H_targman_decision = N06H1_targman_options,
                                   H_targdos_decision = N06H1_targdos_options)

####  CONFIGURES THE SIMULATION OF THE HACKTIVISTS PROBLEM ----

#  defines the number of simulations per decision portfolio
H_portfolio_simsize <- input_portfolio_simsize*5

# s defines the number of portfolios
H_portfolios_numberof <- portfolioSize(H_decision_portfolios)
```

```
#  assigns a numeration to each of the individual simulations
H_portfolio_num <- portfolioNumeration(H_portfolios_numberof)

#  generates a table with the individual simulations
# and their corresponding decision portfolio
H_portfolios_table <- portfolioTable(H_portfolio_num,
                                     H_decision_portfolios)

####  LOADS THE FUNCTIONS OF EACH NODE ----

source("analysis/nodes/hacktivists/N02H_security_portfolio_options.R", echo = echoing)
source("analysis/nodes/hacktivists/N06H2_targatt_hacktivists_results.R", echo = echoing)
source("analysis/nodes/hacktivists/N09H_impacts.R", echo = echoing)
source("analysis/nodes/hacktivists/N11H_hacktivists_detection.R", echo = echoing)
source("analysis/nodes/hacktivists/N15H_hacktivists_utility.R", echo = echoing)

####  SOLVES THE HACKTIVISTS PROBLEM THROUGH SIMULATION ----

# Simulation:
# For each security control observed,
# we perform an individual simulation [hacktivists_simulation.R script]
# to obtain the optimal decision portfolio of the hacktivists
# in that individual simulation [which.max(...)], i.e.,
# the one that maximises expected utility.
# We repeat this process a number of times
# based on the size of the simulation defined by the user [H_security_portfolio_simsize].
# The H_fullsim_table stores all the simulated data.
# The H_optimal_portfolio_sim table contains
# the optimal portfolio in each individual simulation.
#H_fullsim_table <- NULL
H_optimal_portfolio_sim <- NULL
jj <- 0
for (j in H_security_portfolio_options) {
  H_security_portfolio_obs  <- j
  jj <- jj+1
  for (i in 1:H_portfolio_simsize) {
    cat('\r',floor(100*(jj-1+i/H_portfolio_simsize)/(length(H_security_portfolio_options))), "% of
hacktivists problem completed ...")
    source("analysis/simulations/hacktivists_simulation.R", echo = echoing)
#    H_fullsim_table <- dplyr::bind_rows(H_fullsim_table, H_portfolio_simulation)
    H_optimal_portfolio_sim <-
      dplyr::bind_rows(H_optimal_portfolio_sim,
                       dplyr::slice(H_portfolio_simulation,
                                    which.max(H_portfolio_simulation$hacktivists_utility)))
  }
}

# The warning messages are not relevant,
# they reffer to automatic changes in formating made by R.
# Solution:
# From the simulated data [H_optimal_portfolio_sim],
# we generate the distribution [H_random_optimal_portfolio]
# of hacktivists decisions [H_targef_decision, H_targman_decision, ...],
# given the security control observation [H_security_portfolio_obs].
H_random_optimal_portfolio <- NULL
for (j in H_security_portfolio_options) {
  H_random_optimal_portfolio <-
    dplyr::bind_rows(H_random_optimal_portfolio,
                     data.frame(H_security_portfolio_obs = j,
                                H_random_optimal_targexf_attack =
                                  mean(dplyr::filter(H_optimal_portfolio_sim,
                                                     H_security_portfolio_obs == j
)$H_targexf_decision),
                                H_random_optimal_targman_attack =
                                  mean(dplyr::filter(H_optimal_portfolio_sim,
                                                     H_security_portfolio_obs == j
)$H_targman_decision),
                                H_random_optimal_targdos_attack =
                                  mean(dplyr::filter(H_optimal_portfolio_sim,
                                                     H_security_portfolio_obs == j
)$H_targdos_decision)
                                ))
}
```

Annex 10: Recommendations from focus groups, advisory board and reviewers

Annex 10

Recommendations from focus groups, advisory board and reviewers

This annex contains the recommendations from the focus groups, the advisory board and the reviewers. These comments motivate the proposed improvements summarised in Section 2 of the document.

Recommendations from focus groups

We only include here the recommendations to be addressed in the modelling part, leaving aside the interface aspects dealt with in such groups or in the experiments. The core information comes from the experts in AXA Group Security.

Global comments:

1. There is a conflict between "adversarial risk" mentioned several times in the text of the pages and considering "non-intentional" threats, such as the way Fire or Computer Virus are described. From an information security perspective, this is wrong.

This was mainly due to our wording. An effort has been made in this phase to conform to the ISF terminology which covers intentional and non-intentional threats. Adversarial risk takes care of the intentional part, whereas standard risk analysis takes care of the non-adversarial part.

2. At the moment, the way it works feels like Each Threat/Risk has a single control mapped against it, which won't ever be the case; Many controls can support many risks and as such there would need to be a weighting behind one. For example, a Virus threat would be reduced by
   a. Patching
   b. HIPS/HIDS
   c. Firewall
   d. Anti-Malware
   e. Etc.

The underlying model is generic and of course may be used for the case in which there are several controls for a threat and controls for several threats. The same with the tool.

3. There's huge gaps here in security controls even taking into account basic control recommendations like UK Cyber Hygiene, Australian Signals Directorate, CIS Top 20, etc.

This was just an initial example. The model is generic and can accommodate the catalogues as wished. The new version is extended.

4. The number of security controls is extremely reduced and doesn't include several ones that would impact the security posture of the organisation used as an example

Same comment as before, to mitigate we shall include an Other measures option.

5. It's unclear how the profile of the organisation (document management company) affects the tool and the results. If the profile of the organization is to imply that some security controls are implemented, then it should say so. Otherwise, this risk analysis is only partial and most likely to be inaccurate.

Again this was an initial example. The method can accommodate the full case. In the example is just reflected in the size of the company, the budget available,…. An updated model will include generic features affecting various elements in the model.

6. Its mixing up availability and confidentiality/integrity by having the physical fire stuff; Would it be better to have that separate?

The supporting paper separates availability, etc. We have actually separated the impacts over various assets distinguishing between insurable and non insurable impacts. We have worked to improve the distinction taking into account the new preference model. There is a trend on combining physical and logical security and safety issues.

1. Assets
   - Are these the only Assets you consider? With the description given of the SME, I'm expecting the organization to list its most important assets, which should include, as document management company, the documents it's managing.
   - What about information assets?

*As we said this is an initial example, which needs to be completed. This has been complemented and we also added the Other * option when required.*

2. Threats
   - Are these the only threat considered? A document management company must be concerned about insider threat, data leakage, ransomware and many other threats.

This is an initial conceptual example. We have completed methodologically as requested.

- I don't get the fact that you (the SME) can uncheck threats. If this is the case, then it kind of breaks the point I think. Unless the Uncheck is meant that I don't want insurance in that area. A complete risk analysis needs to consider all the threats.

In some instances, it could be the case that the risk analysts do not consider relevant a certain threat. For example, a nation state may be relevant for a reduced set of organisations, even SMEs, of certain characteristics or from some countries. Implicitly this means that the insurance won't cover against such threat. This requires some kind of parametrization of insurance products.

- You write 'CYBECO looks at adversaries in the cybersecurity context', yet the threat of Fire excludes 'sabotage' and Computer virus is 'non-intentional'. Then, they are not 'adversarial' threats, so I don't really understand.

Not only adversaries. Some threats may be intentional, some not. We could view a fire as intentional if performed by an adversary, as well as an intentional computer virus. But is not the case in the example. Care has been taken as far as wording is concerned.

3. Threat / Fire
   - Makes sense and allows flexibility, however what about Earthquakes, Floods, why just fires?

This is an example and not full. The system is open to extensions.

- The home page mentions that CYBECO looks at adversaries in the cybersecurity context. But the popup on Threats/Fire says "We assume that a fire can occur only by accident, not considering the possibility of sabotage". So why is it considered at all?

We consider both adversarial and non adversarial threats. Wording has been improved

4. Threat / Computer Virus
   - You consider computer virus among the threat but do not include anti-virus in the security controls. Is this because some security controls are expected to be implemented and not documented here?

We include the firewall and risk mitigation procedures as examples. As we said this is not the final system just a demo prototype version. But it could be the case that some controls are already implemented and this needs to be contemplated. We have improved definition and implementation of controls.

   - I would change this to Malware Infection, better language.

The whole vocabulary has been updated.

   - The Information popup is pretty bad as it talks about unintentional. Most virus outbreaks are malicious with malicious intent. Yes, employees are probably doing this accidentally but the wording describing this is pretty bad.

To be modified. However, most virus are not targeted (although intentional). Language adapted to ISF.

   - What difference does "non-intentional" make? As a document management organization, I'd be more concerned by intentional virus attacks, trying to compromise the availability or integrity of my customers' documents rather than the non-intentional ones that an antivirus will pick up.

This is a simplified example. In such case, you would include targeted virus within the intentional attacks. The model has been enriched to take account of this. Antiviruses are not as excellent as you seem to imply.

   - If it's "non-intentional", then it's "accidental risk" and not "adversarial risk", so, hence the introduction to the site, why is it considered?

The wording has been modified.

   - Cost of repairing an infected computer, this at least needs to be split into Workstations and Servers as the cost for either will be very different

Indeed. But this is a simplified example!!!! We would differentiate in assets (right now just computer equipment, better workstations and servers). The asset definition has been improved.

   - Furthermore, assuming no controls, then most attacks would impact all the computers unless you've got good network segmentation (i.e., other security controls). It's a generalisation but it would be a good way to simplify it and protect the insurance company from losing money. E.g. assuming worse case.

Same comment.

5. Threat / Competitor Attack
   - This is really weird, because none of the others imply a threat actor but this one does.

Because this is targeted. Wording has been improved.

   - I doubt that most competitors in this space attack each other. The most likely threats are DDOS where the SME would be impacting by being on the same network or a blackmail type attempt to illicit money.

Again this is based in the paper where everything is explained. On the other hand many attacks come from competitors… Again the list of targeted attacks is simplified. We have enriched the targeted and non targeted panels.

6. Security controls
    - Are these the only security controls you are considering? What about antivirus and data leakage prevention that are quite expected in an organization that manages documents.

Same comment as above. This is a simplified demo system. We have enriched the list this year.

- There's huge gaps here in security controls even taking into account basic control recommendations like UK Cyber Hygiene, Australian Signals Directorate, CIS Top 20, etc.

Same comment and action undertaken.

- Also, it doesn't talk about operational and maintenance costs just the cost of a Network Firewall. There would be ongoing maintenance and operational costs to manage it (Depending on the size of company) This might be intentional as it's just a model but it's too basic to figure out if the model is correct.

Such costs would be sunk within the full costs. This is a yearly planning example. We have separated between both types of costs.

7. Firewall – Because its next to anti-fire my mind immediately jumps to a physical firewall in a building. Consider Network Firewall.

The whole concepts has been revised and reworded when necessary.

8. I'm not sure I understand why Traditional insurance is mentioned here. The tool is about cyber insurance against adversarial threat. If the traditional insurance is to mitigate the threat of fire, sure, but fire is not considered an adversarial threat, so what does it do here in the first place?

More and more, there is a movement towards combining physical and cyber security. Standards like ISO , ISF etc… combine both types of threats. The main split is between targeted and non targeted. Someone could buy 'standard' insurance but no cyber or vice versa. Wording has been carefully revised.

9. The introduction of the page for the risk analysis from an insurance company perspective refers to a paper titled An Adversarial Risk Analysis Framework for Cybersecurity. Yet, the form below considers the risk of Fire that is considered non-intentional, i.e., non-adversarial.[1]

We have improved and simplified the wording.

10. The different prices for cyber insurance is surprising. The price should vary depending on the number of security control selected (except anti-fire), and the combination of a Firewall ***and*** DDoS protection is more effective than either of them individually at protecting the organization against cyber threat, so the price should be further reduced.

[1] *Note: the paper was not made available to the participants..*

We have revised thoroughly the model assumptions and revised computations.

11. The tool doesn't say if the "risk mitigation procedures" are expected when either the firewall or DDoS protection are implemented. From a security point of view, it makes a significant difference, as it reflects a difference of maturity. Was the firewall and/or DDoS protection installed as a result of a "risk mitigation procedure" or in an ad-hoc way?

The "risk mitigation procedures" refers to the adoption of secure practices and procedures by the workers when they are using their computers. Actually this could be one of the UK Cyber essentials. In the example case of the model, they can be the sole security measure implemented but they can also be implemented in conjunction with the firewall and the DDoS protection. The wording has been revised.

The following comments were made on the output of the risk analysis.

1. "Risk analysis of impact" is odd. I would rather call it "impact analysis".

Impact analysis is other type of analysis but we can change the text. There is somehow a debate on terminology but we have made an effort to adapt terminology to ISO 27005.

2. From a security standpoint, you are listing all the possible combination – that fit in the budget. But have you checked that you can get a cyber insurance without the most basic controls such as a firewall? Without these basic security controls, if cyber insurance was available, the price difference would not be just 20% or so.

The model covers this type of constraints. The included example is a simple version. The cyberinsurance product defined is very simple. We allow for more constraints and more realistic cyberinsurance products.

3. The number of expected DDoS attempts seems to be related to the type of Cloud-Based DDoS protection (1tbps: 0; 10gbps: 22; 5gpbs: 27; 2gpbs: 28; no: 28). What are those numbers based on?

In the example case, the attacker can test the site to know the DDoS protection implemented and thus adapt the number of attack attempts. Basically the stronger the protection, the less attack attempts that you shall suffer. 0, 22, 27, 28 etc… comes from the underlying model. We revise text and explanations.

4. Besides, the number of "attempts" cannot be related to the type of protection you have. The number of successful attacks or disruptions are, not the number of attempts.

It actually can as the attacker may observe some of the implemented defences, through tools like network scanners. We have revised text and improved explanation.

5. Are you assuming that the selection of the anti-fire system reduces the impact on the facility of €0? What type of anti-fire are we talking about?

We don't see where this comment comes from. There is always a possibility of fire and, as a consequence, an expected impact on the facility in all the results. Anti-fire detects a fire so that it can be mitigated earlier and thus fewer contents are affected, therefore reducing the damage. We have revised thoroughly the database.

6. Once again, I don't understand the connection with the impact on the facilities. It's a cyber risk and cyber insurance. What does fire, impact on facilities and traditional insurance have to do with this scenario.

As mentioned, we are providing an integrated security-cybersecurity approach. We have a global budget for security which needs to be invested for standard and cyber threats, for targeted and non-targeted threats. The wording has been improved.

7. I have a hard time understanding how the risk analysis is performed. It seems that for every line the variables (e.g., number of viruses, likelihood of fire, DDoS attempts) are changing. How can you compare the effectiveness and relevance of a controls if each lines are based on a different situation?
8. To take an example, I got two lines that only differ by their insurance – 1: comprehensive and 2: Traditional. Yet, the Probability of Fire changes from 1: 1.5% to 2: 2.58%; the expected number of Virus infections from 1: 5.39 to 2: 5.43. Why are these probabilities are changing? It makes the comparison between comprehensive and traditional insurance impossible. The probability of fire is not affected by whether or not you have an insurance or a cyber insurance and the expected number of virus cannot be affected by whether or not you have a cyber insurance. Or, the wording is completely wrong.

The calculations are based on MC simulation. For the prototype we performed a small simulation but big enough for calculating valid results for evaluating the different portfolios. A final version of this model would provide results based on a more strong simulation (having therefore less variation). We have also worked on improving the involved algorithms. We have revised text and improved the underlying algorithms.

The following comments were specific to the Risk Analysis for an SME (Simulation)

1. The price selection for the available insurance is ambiguous.
   a. Why is there a specific price for the traditional insurance with firewall or DDoS protection?
   b. Are the benefits cumulative? Does the cost with procedure expect all the above (i.e., anti-fire and firewall or DDoS)?

Indeed, this needs to be corrected. It's correct in the model/paper:

DDoS protection: why the only variable considered is the attack capability in gbps? What about the probability of being attacked? Is it a probability of 1?

The probability of attack is not one, but we have not included in the output. We shall include it. We also include other variables such as costs and gains for attacker, detection probability. We shall explore other technical relevant variables. We have also revised text.

Recommendations from advisory board

As reflected in D2.5, the advisory board members reflected the following needs in relation with the developments in WP·:

- Further reflect about threats coming from insiders.

We have devoted a paper to the problem of cyber insiders available as Annex 3.*

- There should be some kind of segmentation of attackers, as well as their objectives, much as you have done for the Defenders.

We have developed several templates for various types of attackers (specifically, hacktivists and cybercriminals) and described the computational strategies when there are several attackers. Described in detail in Annex 8. We have also sketched the preference models for attackers.

- The need to consider in further detail third party liabilities.

We have thoroughly updated our Defender model document, available as Annex 2 and worked in the related issue of supply chain cyber risk, as reflected in the Annexes 4 and 5.

- Translate the usual A, I, C criteria more into business related attributes.

We have reflected and discuss this in detail in Annex 2.

- The objectives of attacker and defender do not be aligned/related.

Indeed, and we have further clarify this point in Appendix 1 and the examples there included.

- The preference list needs further refinement e.g. for possible overlaps.

We have thoroughly revised the document in Appendix 2 in relation with this idea.

Recommendations from reviewers

- REVIEW RECOMMENDATION 1: Dissemination & exploitation activities fall short of what could be expected from the project, in particular in regard to the relatively short duration (24 months) of the project. While neither SEJ nor ARA are particularly new approaches, their inclusion into utility-oriented risk modelling for cyber security is – according to the claims of the DoA. This claim should be better substantiated by scientific publication and dissemination activities, of which there are precious little at the time of the review. The reviewers suggest that the validity and benefits of the mathematical approach is corroborated by at least one new and original high profile scientific publication from WP3.

Our paper An Adversarial Risk Analysis Framework for Cybersecurity (appendix 1 of D3.1) has been accepted by the journal Risk Analysis on April 9, 2019. The journal Risk Analysis is a Q1 journal and one of the main references in the risk analysis field. The computational paper has been published in European Journal of Operational Research, also a Q1 journal.

- REVIEW RECOMMENDATION 4: The project designed two experiments to evaluate the attitude and the behaviour of people with respect to security, protection, and insurance. However, no test or validation of the toolbox with real end-users has been foreseen. In order to evaluate the effectiveness of the developed tool-kit, we require the consortium to perform an experimental evaluation with at least two types of end-users: insurance providers and potential customers evaluating the option of buying a cyber insurance product. The evaluation should be performed at M18 and cover both usability of the tool and effectiveness for the specific scope important for the user. Results should then be analysed to improve the tool towards its final version released at the end of the project.

As detailed in Sect. 2.2, we evaluated the toolbox with end users, in this case potential customers, to provide inputs to both the Toolbox and the underlying WP3 models. We assessed them, and incorporate most of their suggestions as well as solutions to their concerns in the improved versions of the WP3 models.

- REVIEW RECOMMENDATION 5: The presentation of the software prototype left the reviewers somewhat confused, which was mainly due to its generic form (supporting different stakeholders) and its representation of general ontology data. Future versions of this prototype should give much better indications about (a) the stakeholder type currently supported by the software, (b) the stakeholder's decision the system offers to support (c) motivation why specific types of data have to be provided for decision support, and (d) intelligible output stating and explaining the lines of action available to the user.

Most of these improvements took place in the Toolbox itself (better indications, explanations and intelligibility). When it comes to WP3 models, we focused on simplifying the type of data the users have to provide to our models through the Toolbox, specifically they are general business or IT information about their organization (number of employees or computers) that it is relatively easy to obtain or estimate - compared to cybersecurity information.

- REVIEW RECOMMENDATION 6: The project might very much benefit by a comprehensive overview of existing cyber insurance products (and insurance elements linked to Service Level Agreement at IaaS, PaaS and SaaS level) from different providers – in regard to:

  - variations of coverage,
  - data points required for defining specific policies, and
  - data points required to create more targeted or more comprehensive ones,
  - types of coverage and pricing customers would like to see on offer, and
  - measures customers and their users have to take in order to become 'insurable'.

  Such kind of overview could be of particular use for guiding and validating the modelling activities. It also appears of the essence for any serious exploitation planning, e.g. for motivating the shape insurance products, potential discontinuations or discouragements of old ones, or for evaluating advantages and disadvantages of separating resp. combining cyber-specific risks from other insurable risks to business continuity. It would also help to clarify the types and probabilities of risks currently not covered by SLA agreement between IT service providers and their customers and which may consequently be targeted by insurance products.

Such overviews were available in Marotta et al (2017) and Romanosky et al (2018) which we have followed for the purpose considered.

- REVIEW RECOMMENDATION 7: In principle, CYBECO's utility-based approach is amenable for designing insurance products, and the model designers are confident that they may be able to capture multi-variant and drifting strategies of potential defenders, let alone those of potential attackers of IT systems. We stress again that this is a hypothesis whose validity we expect to be thoroughly checked and documented.

Besides the publication of our paper "An Adversarial Risk Analysis Framework for Cybersecurity", we also worked on a more specific paper dealing with the elicitation of preferences and utilities of the potential defenders (annex 2), specifically we describe a methodology that consists in (1) helping the defender define their goals regarding cybersecurity through a comprehensive tree of cybersecurity objectives, (2) the integration of these objectives into a multi-attribute utility function and (3) the integration of their risk attitude into their utility function. We have included in them a discussion on market segmentation and insurance design.

Annex 11: Validation of ARA for belief formation

Annex 11. Validation of ARA for belief formations

Rationale

Structured Expert Judgement (SEJ) elicitation has become a major ingredient within the risk analysis practice (Bedford and Cooke, 2001). A significant feature of this discipline is its emphasis in decomposing complex problems into smaller pieces that are easier to handle and recombining the piecewise solutions to tackle the global problem. This simplifies the complex cognitive tasks and mitigates expert reliance on heuristics that can introduce bias, ensuring that experts and decision makers actually analyze their decision making problems. The decomposition typically entails more assessments, though these tend to be simpler and more meaningful, leading to better decisions. One example refers to belief assessment through the argument of *extending the conversation*. Rather than directly assessing the probability of an outcome, one finds a conditioning partition and assesses the probabilities of the outcome given the conditioning events. From these, and the probabilities of the conditioning events, the law of total probabilities enables calculation of the unconditional probability of the outcome

During CYBECO we have presented and studied Adversarial Risk Analysis (ARA) as a decomposition strategy for game theoretic problems from a Bayesian perspective applying it in the cybersecurity domain. ARA can be framed as a tool for SEJ elicitation when we need to deal with probabilities referring to forecasting actions to be performed by opponents or, more generally, by other agents not available for elicitation, like cyberattackers. As an example, in Chen et al. (2016) nearly 30% of the questions posed to experts somehow involved adversaries (e.g. Will Syria use chemical or biological weapons before January 2013?, asked in 2011).

One of the most interesting issues uncovered in the first CYBECO experiment referred to the relevance of properly assessing adversarial probabilities. We have also exposed theoretically through simulation how the indirect ARA approach may lead to more accurate forecasts than if such assessments are performed directly with standard SEJ techniques. With this final group of experiments we aim at confirming such issue.

Experiment implementation

The face-to-face experiment was carried out in an experimental lab in Barcelona (Spain), during April 2019. It was designed by DEVSTAT with the aid of the CYBECO advisory board and the CSIC team. The experiment was run using an experimental software designed ad-hoc (developed in PHP) and tested before the sessions to guarantee its usability and understandability.

The experiment embraced four sessions with 96 participants in total. At the beginning of each session, the subjects were randomly scattered among the semi-cubicles in a room. To avoid presentation bias, the same lecturer led all sessions. The lecturer read aloud the instructions and accompanied his speech with a slide projector to explain the kind of decisions that the subjects would have to take through three examples, and the benefits that they would obtain in the experiment depending of the performance. They then undertook the actual tasks and completed two simple questionnaires (socio-demographic and level of familiarity with the context of the decisions) The economic experiment included a variable payment to each subject, depending on the probabilities assigned to a series of events and its actual realisation during the days following the experimental sessions.

The distribution by profile of the respondents is shown in Table 1. Three quarters of them were public/private workers and were almost equally distributed between both genders. Regarding the education of participants, most of them had either some years of university or a university degree.

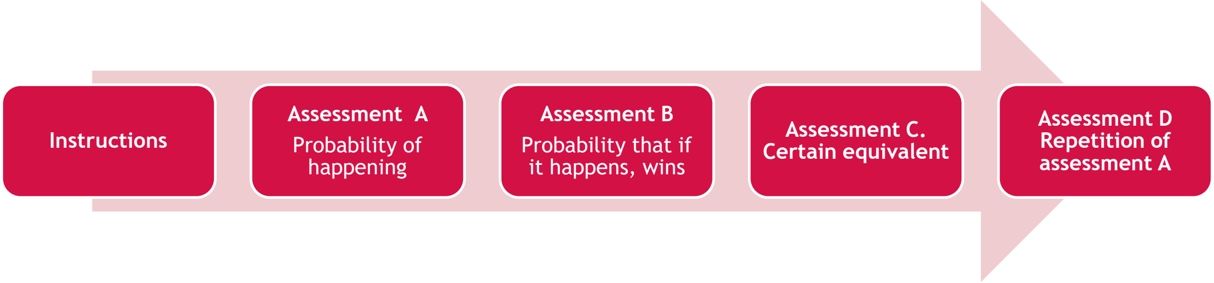
Profile	*N*	*%*
18-35	26	27,08
36-50	39	40,63
51-74	31	32,29
Male	50	52,08
Female	46	47,92
0-11 years of education	1	1,04
12 years of education	18	18,75
Some years of university (not completed)	28	29,17
University degree	35	36,46
Post-graduate degree	14	14,58
Freelance	8	8,33
Public/private worker	72	75,00
Unemployed	7	7,29
Housemaker	2	2,08
Student	3	3,13
Disabled	3	3,13
Retired	0	0,00
Other	1	1,04

Table 1. Distribution of participants by age, gender, education and employment.

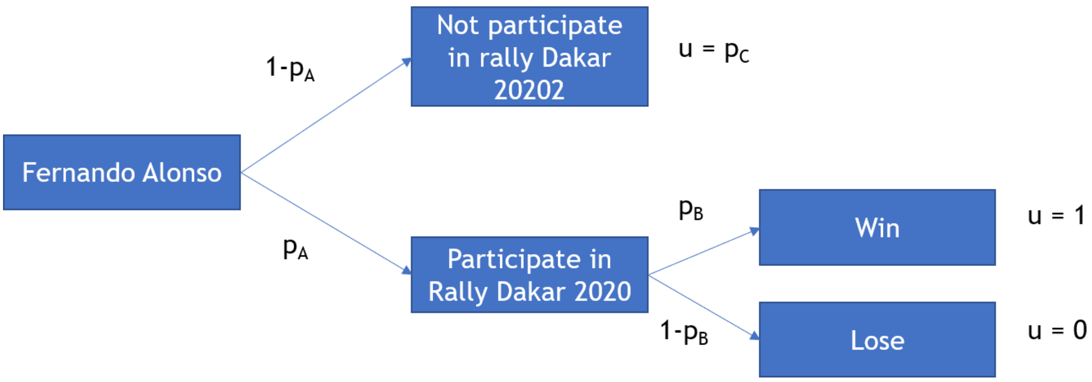
The median duration of the experiment was 30 minutes.

Results

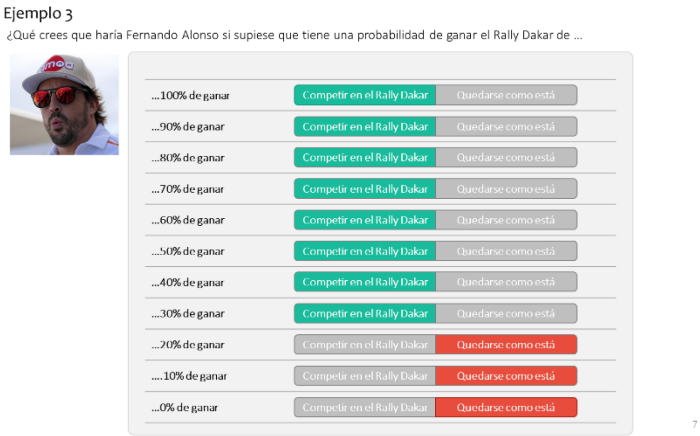
We describe here the design of the four cases presented to the subjects.



After providing the instructions, we then ask about the probability pA that some person makes a decision which may improve her status or not. We then assess our belief about her beliefs of improving her position as well assess indirectly his evaluation of the status quo. Finally, we repeat the first assessment. The interrelations of all the tasks are presented in the following scheme for a motivating example.



The assessment p_C is suggested by the perceived moment in which the subject decides to take part in the event rather than remaining in the status quo. For example in this case it would be p_C=70%


Ejemplo 3

¿Qué crees que haría Fernando Alonso si supiese que tiene una probabilidad de ganar el Rally Dakar de ...

...100% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...90% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...80% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...70% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...60% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...50% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...40% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...30% de ganar	[✓ Competir en el Rally Dakar]	Quedarse como está
...20% de ganar	Competir en el Rally Dakar	[✓ Quedarse como está]
....10% de ganar	Competir en el Rally Dakar	[✓ Quedarse como está]
...0% de ganar	Competir en el Rally Dakar	[✓ Quedarse como está]

7

The comparisons that we are undertaking refer to:

- Checking whether after reflections in tasks B and C, there is a change of opinion concerning p_A vs $p_{D.}$ In other words whether further reflection about the problem induces a change of beliefs.

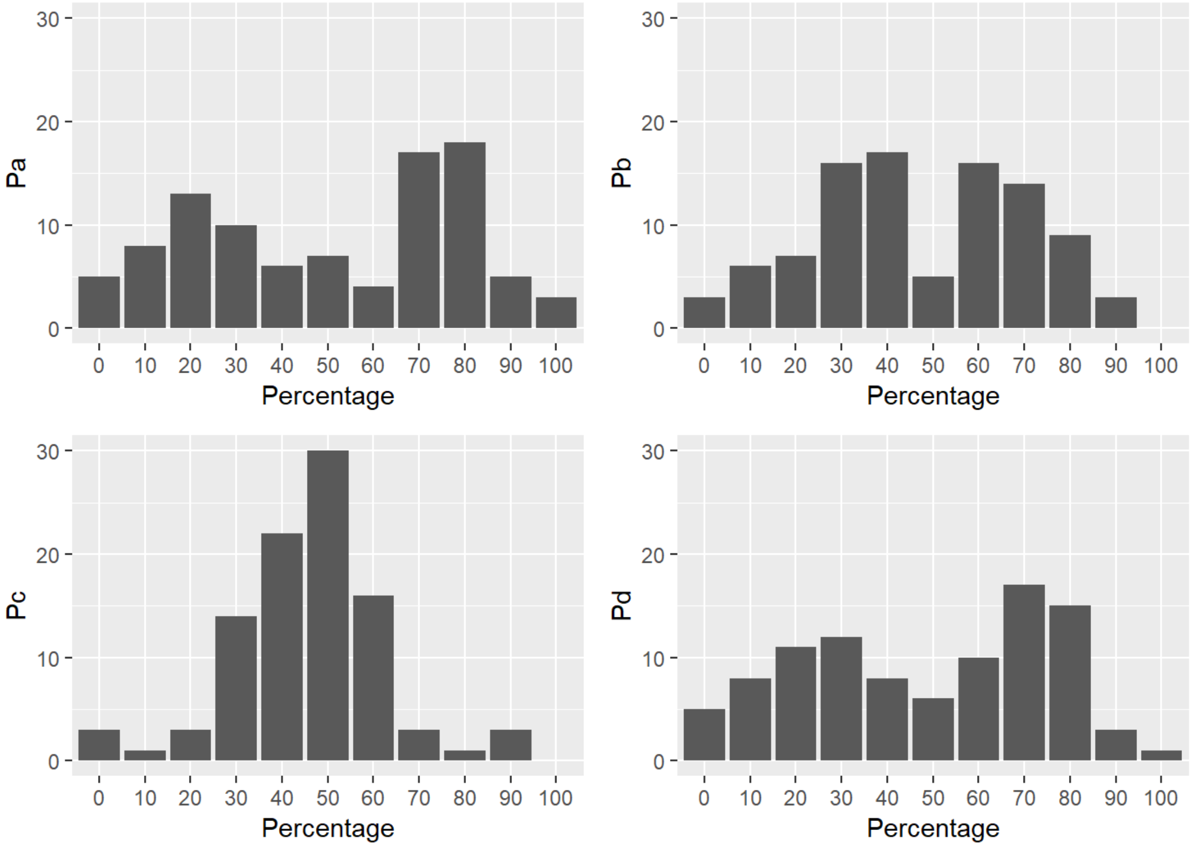
- Estimate $p_{A,}$ p_D based on p_B and p_C. Following a simple random choice model, we may use the form $q=1-p_C/(p_B+p_C)$ and compare with p_A and $p_{D,}$ to check a similar phenomenon and analyse the ARA decomposition.

- Score with a proper scoring rule the performance of both forecasting approaches, once the events are realized by the end of May.

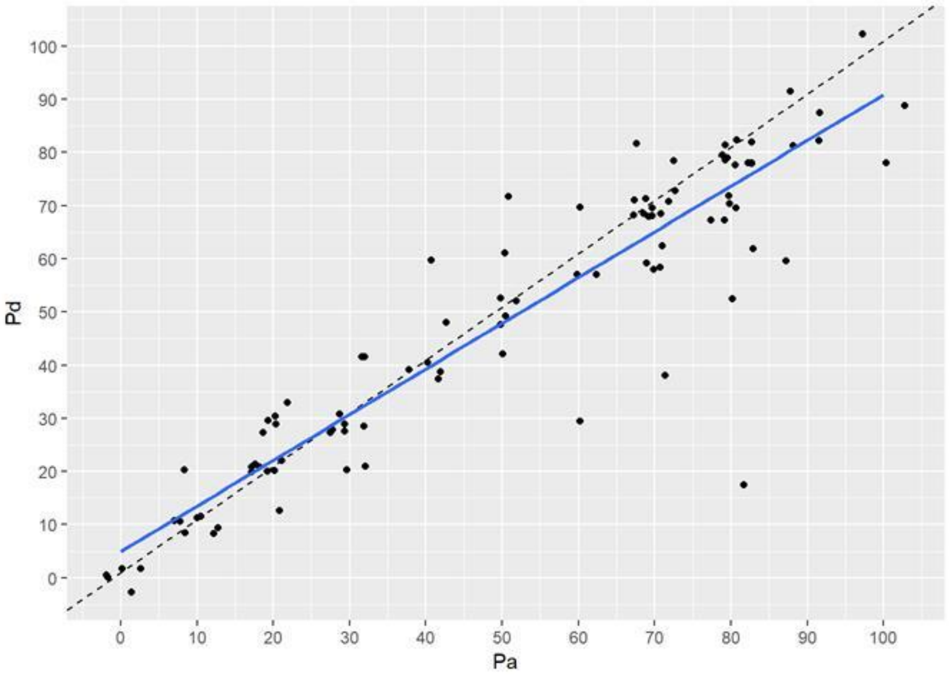
- Check the effects of knowledge level and other demographic factors (age, gender, education).

Here we present descriptive statistics concerning the first of the four questions posed which we refer to the decision of Theresa May to call for elections in the next 30 days.

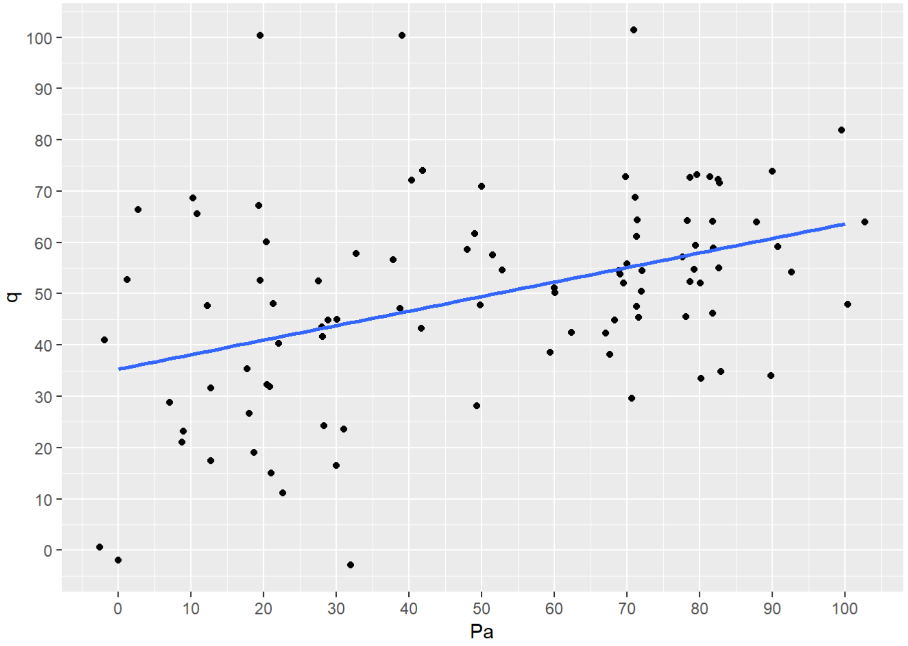
Will Theresa May call for elections within the next 30 days



Comparison of the probability for Theresa May to call for elections within the next 30 days before (pA) and after (pD) the reflection.



Comparison of the probability for Theresa May to call for elections within the next 30 days before (pA) the reflection and estimation with a simple random choice model $q=1-p_C/(p_B+p_C)$.



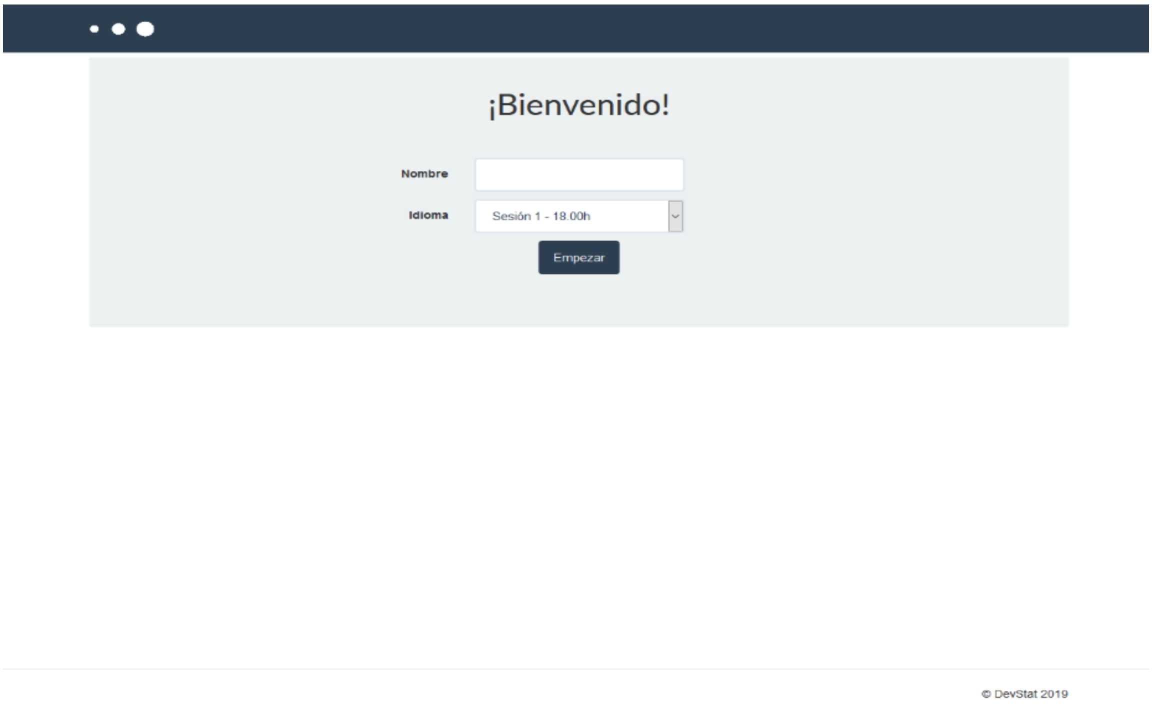
Similar results hold for the other three other questions (concerning a new product release, the participation in a TV show and the participation in a sport event).

The results will be fully analysed when the deadline for the events takes place. Initial exploratory analyses seem to suggest the relevance of the ARA decomposition in belief formation.

We end up by showing the screenshots of the experiment for the first of the four decisions considered in the experiment (Theresa’s May decision of calling for elections). The screenshots are in Spanish, since the experiment was run in Spain and no version of the experiment in English was developed.

Some screens from the experimental software

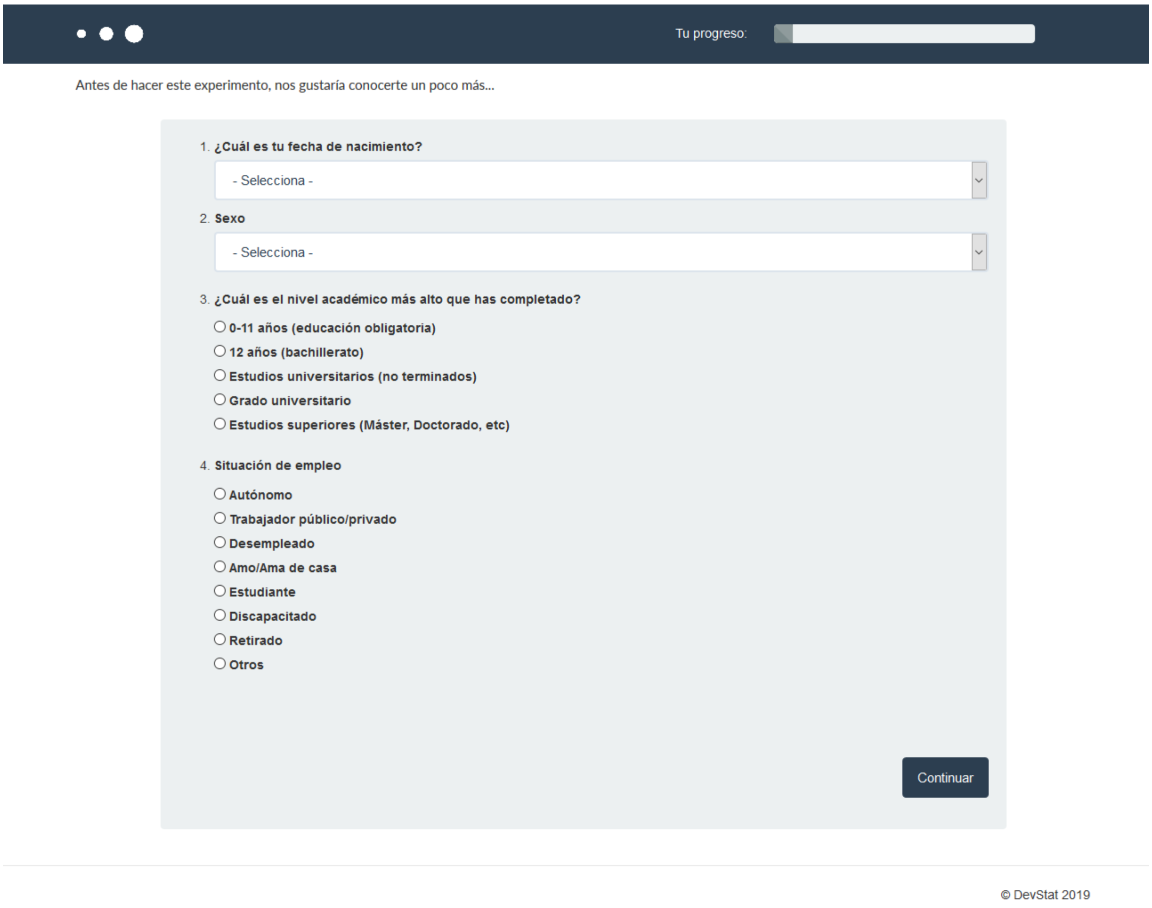
Welcome and basic demographic data



¡Bienvenido!

Nombre

Idioma: Sesión 1 - 18.00h



Antes de hacer este experimento, nos gustaría conocerte un poco más...

1. ¿Cuál es tu fecha de nacimiento?

   - Selecciona -

2. Sexo

   - Selecciona -

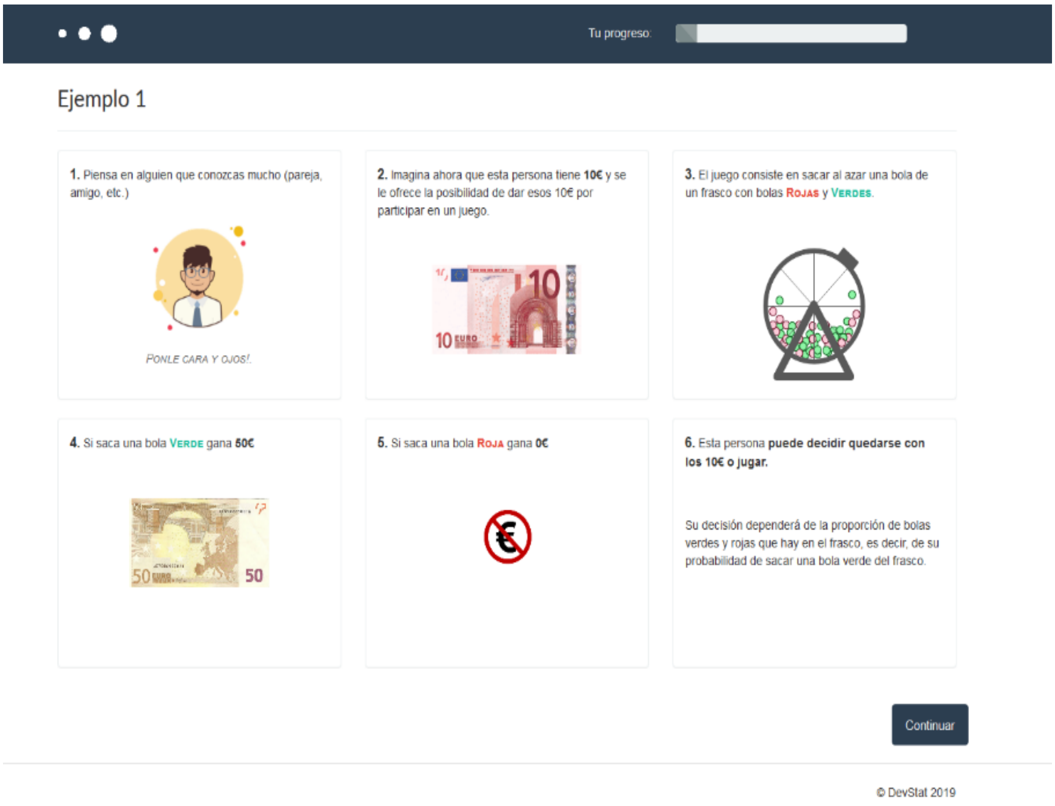
3. ¿Cuál es el nivel académico más alto que has completado?

   ○ 0-11 años (educación obligatoria)
   ○ 12 años (bachillerato)
   ○ Estudios universitarios (no terminados)
   ○ Grado universitario
   ○ Estudios superiores (Máster, Doctorado, etc)

4. Situación de empleo

   ○ Autónomo
   ○ Trabajador público/privado
   ○ Desempleado
   ○ Amo/Ama de casa
   ○ Estudiante
   ○ Discapacitado
   ○ Retirado
   ○ Otros

Becoming familiar with adversarial probabilities. Ficticious case



Ejemplo 1

1. Piensa en alguien que conozcas mucho (pareja, amigo, etc.)

Ponle cara y ojos!.

2. Imagina ahora que esta persona tiene **10€** y se le ofrece la posibilidad de dar esos 10€ por participar en un juego.

3. El juego consiste en sacar al azar una bola de un frasco con bolas **Rojas** y **Verdes**.

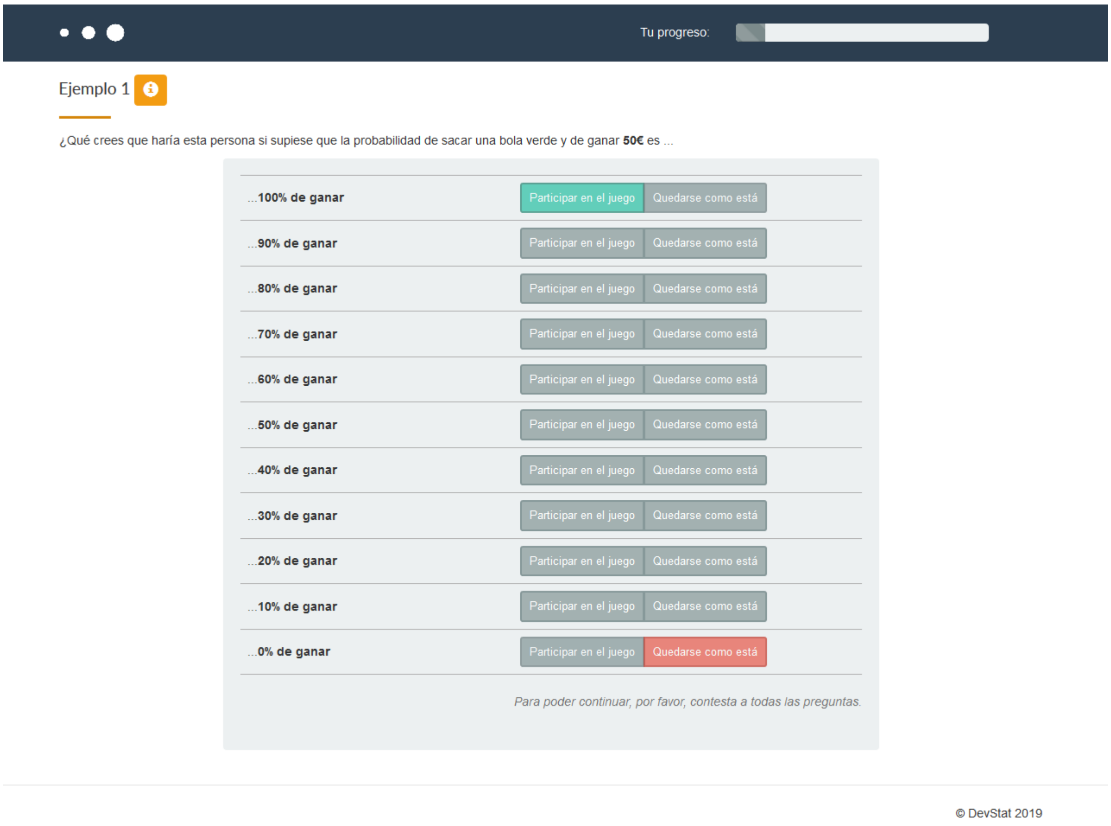
4. Si saca una bola **Verde** gana **50€**

5. Si saca una bola **Roja** gana **0€**

6. Esta persona **puede decidir quedarse con los 10€ o jugar.**

Su decisión dependerá de la proporción de bolas verdes y rojas que hay en el frasco, es decir, de su probabilidad de sacar una bola verde del frasco.

© DevStat 2019



Ejemplo 1

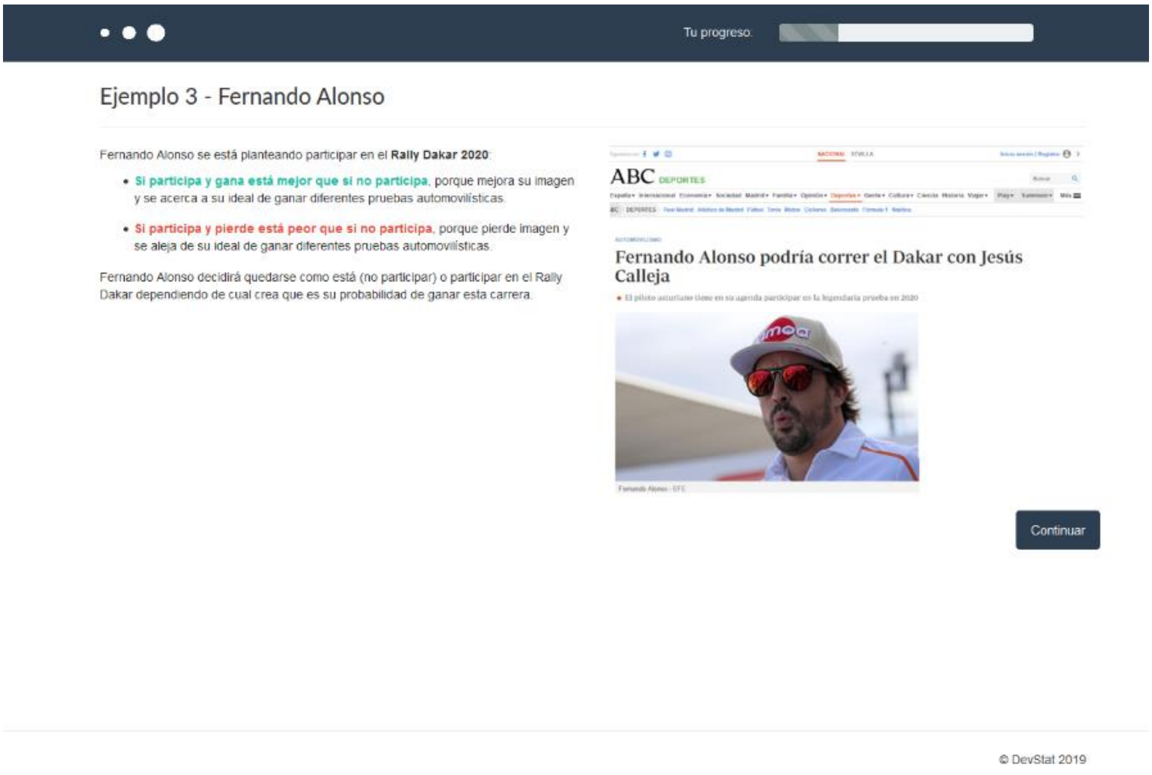
¿Qué crees que haría esta persona si supiese que la probabilidad de sacar una bola verde y de ganar **50€** es ...

...**100% de ganar**	[✓ Participar en el juego]	Quedarse como está
...**90% de ganar**	Participar en el juego	Quedarse como está
...**80% de ganar**	Participar en el juego	Quedarse como está
...**70% de ganar**	Participar en el juego	Quedarse como está
...**60% de ganar**	Participar en el juego	Quedarse como está
...**50% de ganar**	Participar en el juego	Quedarse como está
...**40% de ganar**	Participar en el juego	Quedarse como está
...**30% de ganar**	Participar en el juego	Quedarse como está
...**20% de ganar**	Participar en el juego	Quedarse como está
...**10% de ganar**	Participar en el juego	Quedarse como está
...**0% de ganar**	Participar en el juego	[✓ Quedarse como está]

Para poder continuar, por favor, contesta a todas las preguntas.

© DevStat 2019

Becoming familiar with adversarial probabilities. Realistic case



Ejemplo 3 - Fernando Alonso

Fernando Alonso se está planteando participar en el **Rally Dakar 2020**:

- **Si participa y gana está mejor que si no participa**, porque mejora su imagen y se acerca a su ideal de ganar diferentes pruebas automovilísticas.
- **Si participa y pierde está peor que si no participa**, porque pierde imagen y se aleja de su ideal de ganar diferentes pruebas automovilísticas.

Fernando Alonso decidirá quedarse como está (no participar) o participar en el Rally Dakar dependiendo de cual crea que es su probabilidad de ganar esta carrera.

ABC DEPORTES

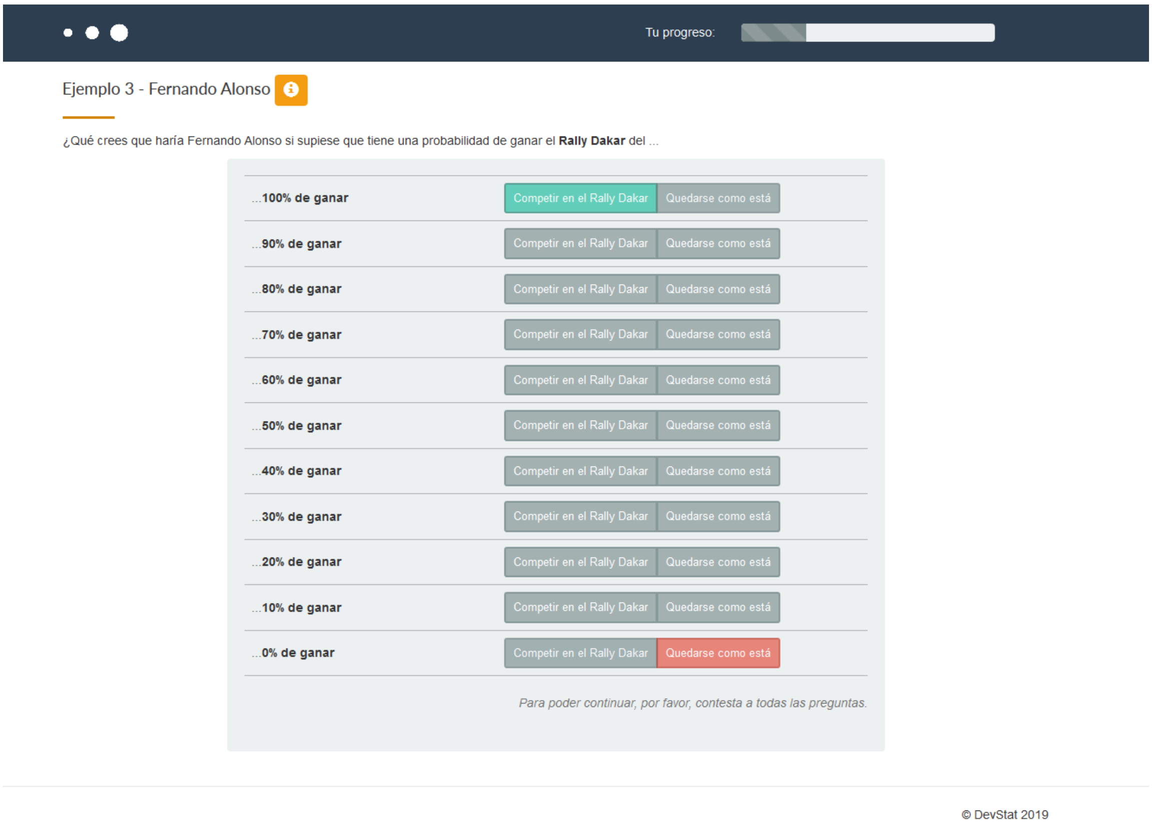
AUTOMOVILISMO

Fernando Alonso podría correr el Dakar con Jesús Calleja

- El piloto asturiano tiene en su agenda participar en la legendaria prueba en 2020

Fernando Alonso - EFE

© DevStat 2019



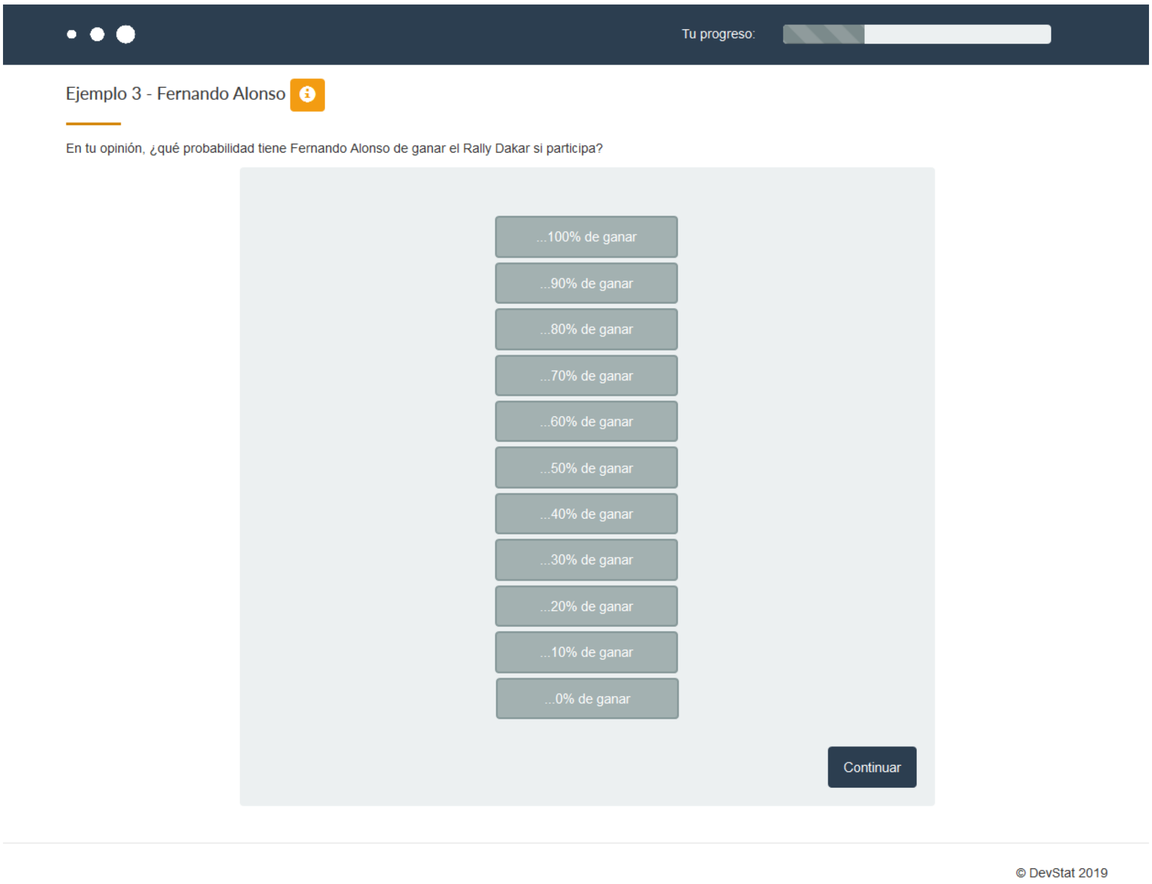
Ejemplo 3 - Fernando Alonso

¿Qué crees que haría Fernando Alonso si supiese que tiene una probabilidad de ganar el **Rally Dakar** del ...

...**100% de ganar**	[✓ Competir en el Rally Dakar]	Quedarse como está
...**90% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**80% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**70% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**60% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**50% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**40% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**30% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**20% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**10% de ganar**	Competir en el Rally Dakar	Quedarse como está
...**0% de ganar**	Competir en el Rally Dakar	[✓ Quedarse como está]

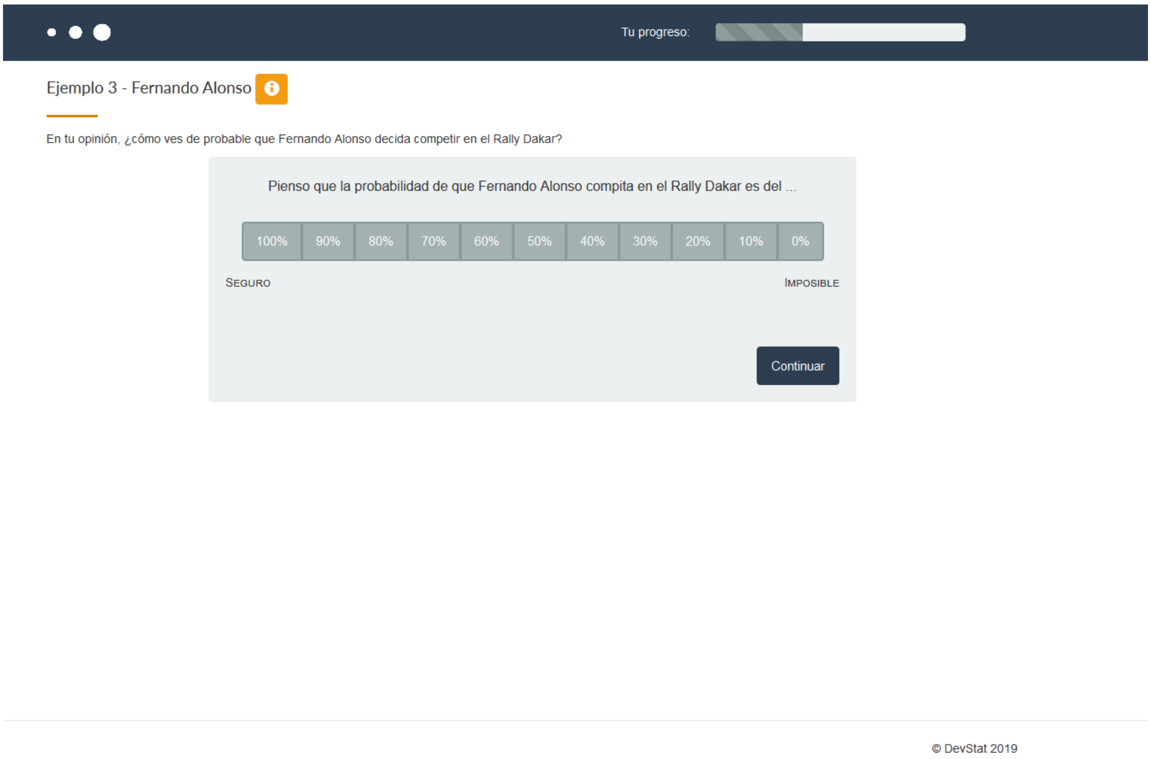
Para poder continuar, por favor, contesta a todas las preguntas.

© DevStat 2019



Ejemplo 3 - Fernando Alonso

En tu opinión, ¿qué probabilidad tiene Fernando Alonso de ganar el Rally Dakar si participa?

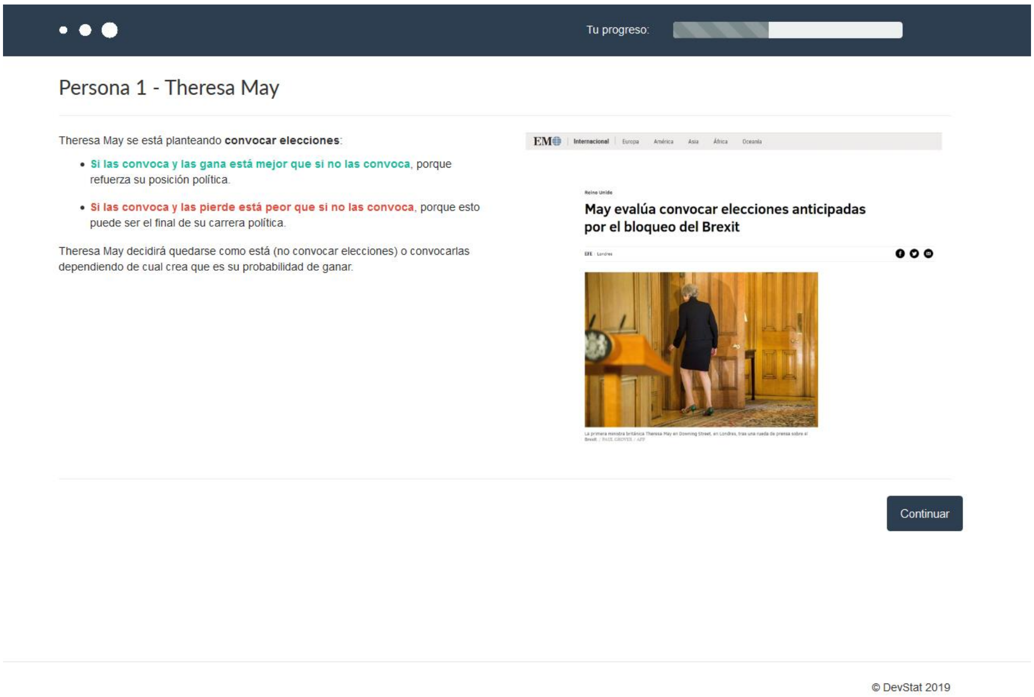


Ejemplo 3 - Fernando Alonso

En tu opinión, ¿cómo ves de probable que Fernando Alonso decida competir en el Rally Dakar?

Pienso que la probabilidad de que Fernando Alonso compita en el Rally Dakar es del ...

Instructions for actual experiment nad first case concerning Theresa May



Persona 1 - Theresa May

Theresa May se está planteando **convocar elecciones**:

- **Si las convoca y las gana está mejor que si no las convoca**, porque refuerza su posición política.
- **Si las convoca y las pierde está peor que si no las convoca**, porque esto puede ser el final de su carrera política.

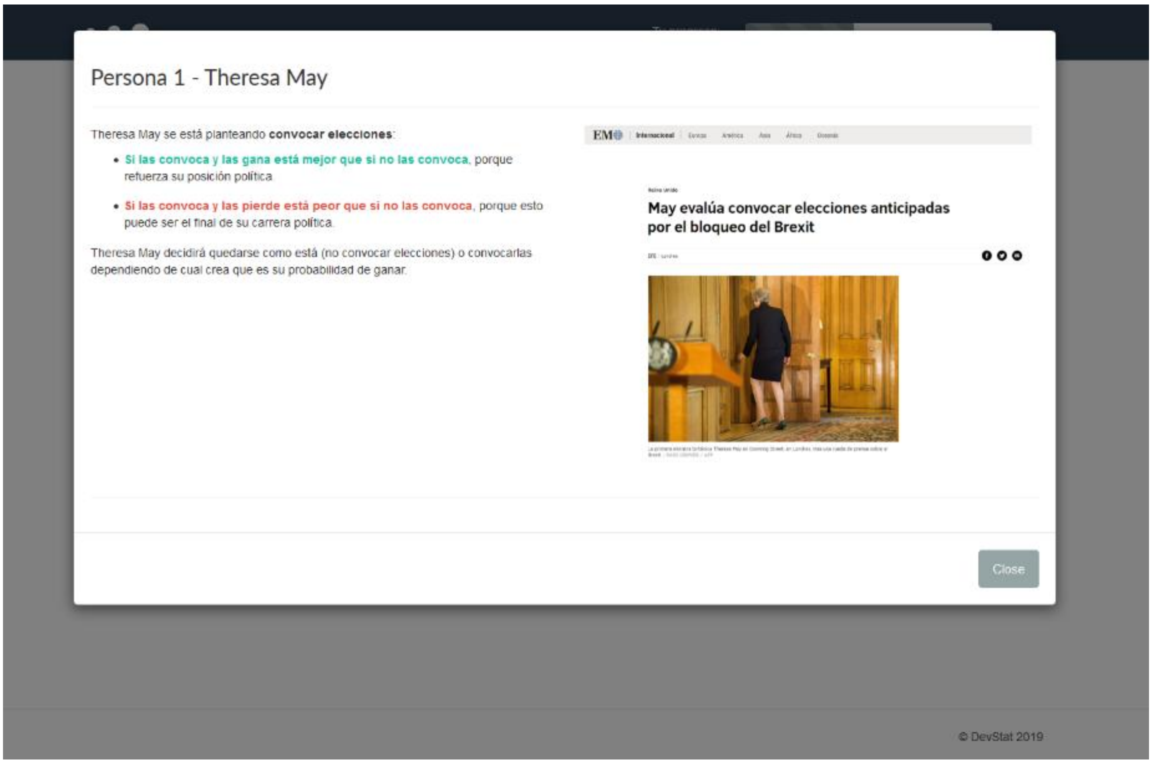
Theresa May decidirá quedarse como está (no convocar elecciones) o convocarlas dependiendo de cual crea que es su probabilidad de ganar.

Reino Unido

May evalúa convocar elecciones anticipadas por el bloqueo del Brexit

EFE · Londres

La primera ministra británica Theresa May en Downing Street, en Londres, tras una rueda de prensa sobre el Brexit / PAUL GROVER / AFP



Persona 1 - Theresa May

Theresa May se está planteando **convocar elecciones**:

- **Si las convoca y las gana está mejor que si no las convoca**, porque refuerza su posición política.
- **Si las convoca y las pierde está peor que si no las convoca**, porque esto puede ser el final de su carrera política.

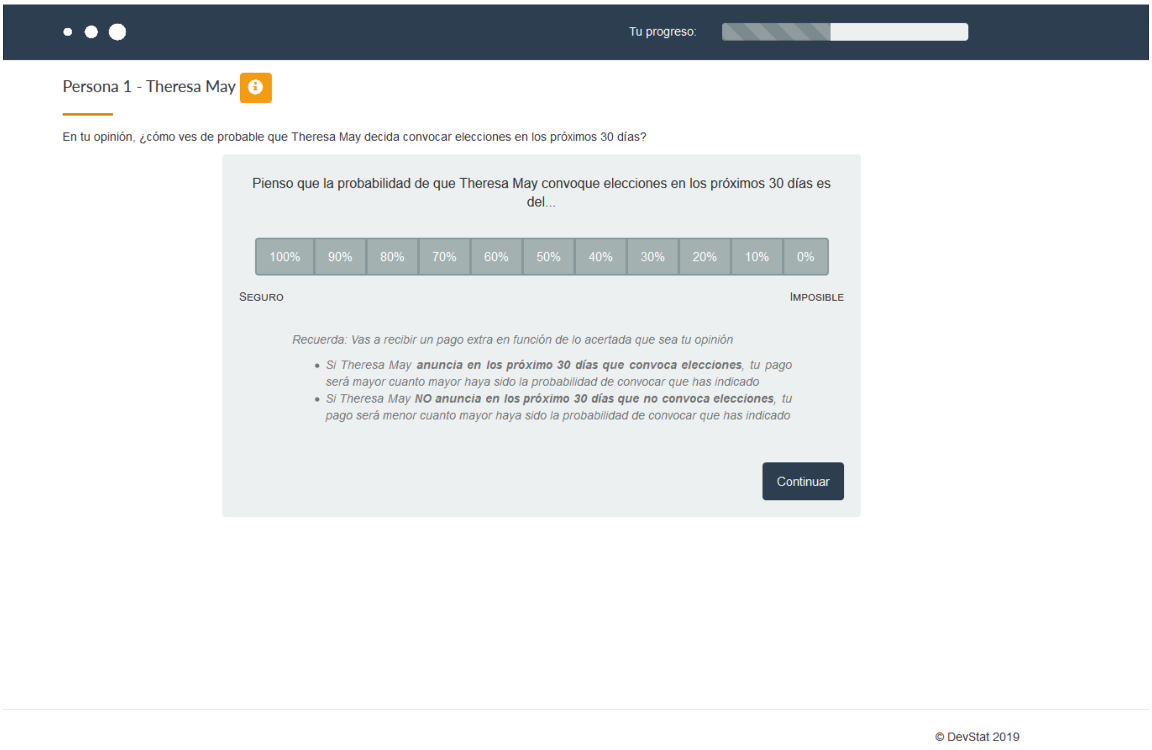
Theresa May decidirá quedarse como está (no convocar elecciones) o convocarlas dependiendo de cual crea que es su probabilidad de ganar.

Reino Unido

May evalúa convocar elecciones anticipadas por el bloqueo del Brexit

EFE · Londres

La primera ministra británica Theresa May en Downing Street, en Londres, tras una rueda de prensa sobre el Brexit / PAUL GROVER / AFP



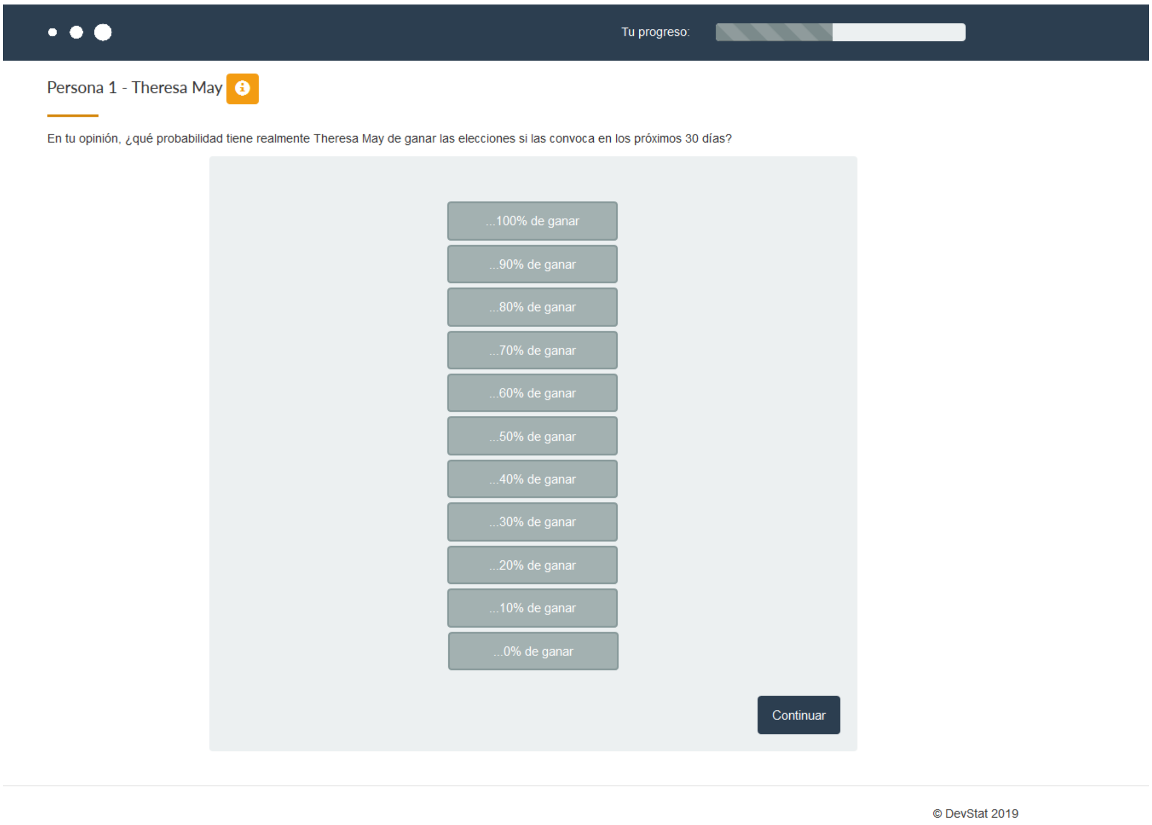
Persona 1 - Theresa May

En tu opinión, ¿cómo ves de probable que Theresa May decida convocar elecciones en los próximos 30 días?

Pienso que la probabilidad de que Theresa May convoque elecciones en los próximos 30 días es del...

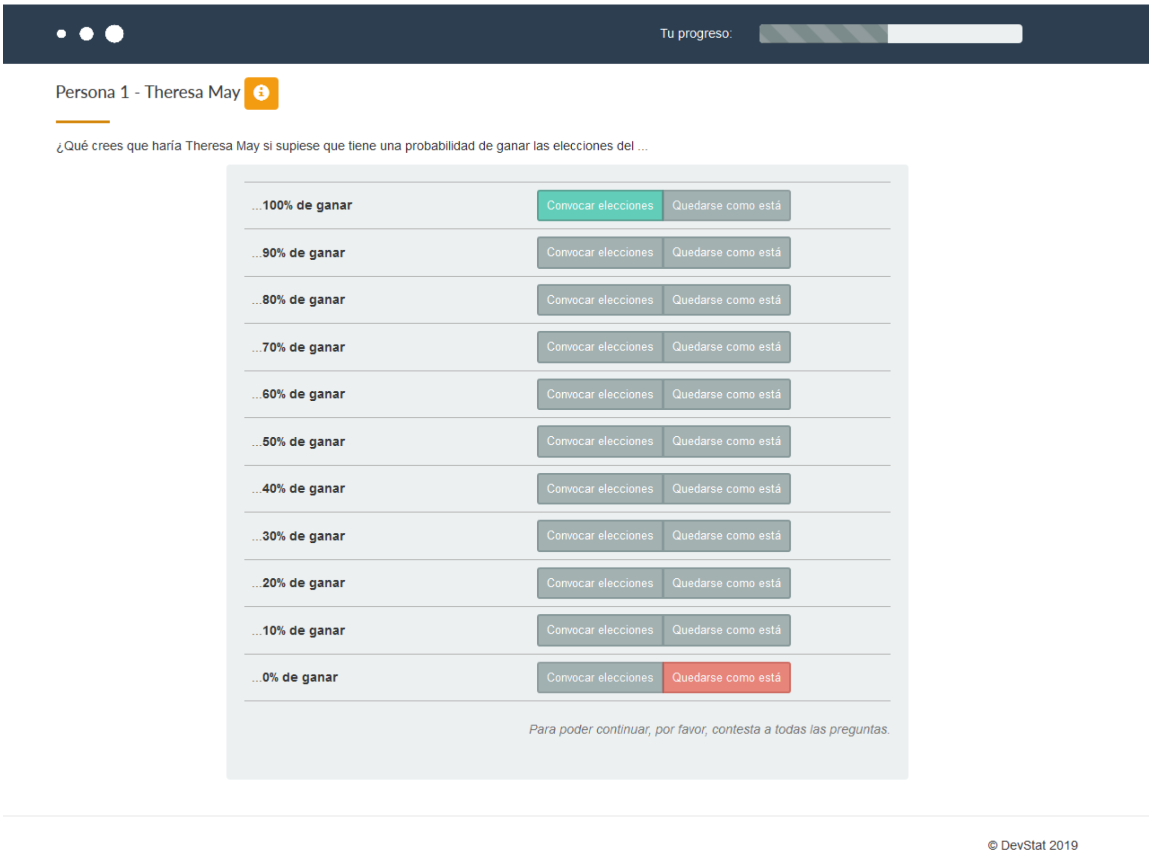
Recuerda: Vas a recibir un pago extra en función de lo acertada que sea tu opinión

- *Si Theresa May **anuncia en los próximo 30 días que convoca elecciones**, tu pago será mayor cuanto mayor haya sido la probabilidad de convocar que has indicado*
- *Si Theresa May **NO anuncia en los próximo 30 días que no convoca elecciones**, tu pago será menor cuanto mayor haya sido la probabilidad de convocar que has indicado*



Persona 1 - Theresa May

En tu opinión, ¿qué probabilidad tiene realmente Theresa May de ganar las elecciones si las convoca en los próximos 30 días?



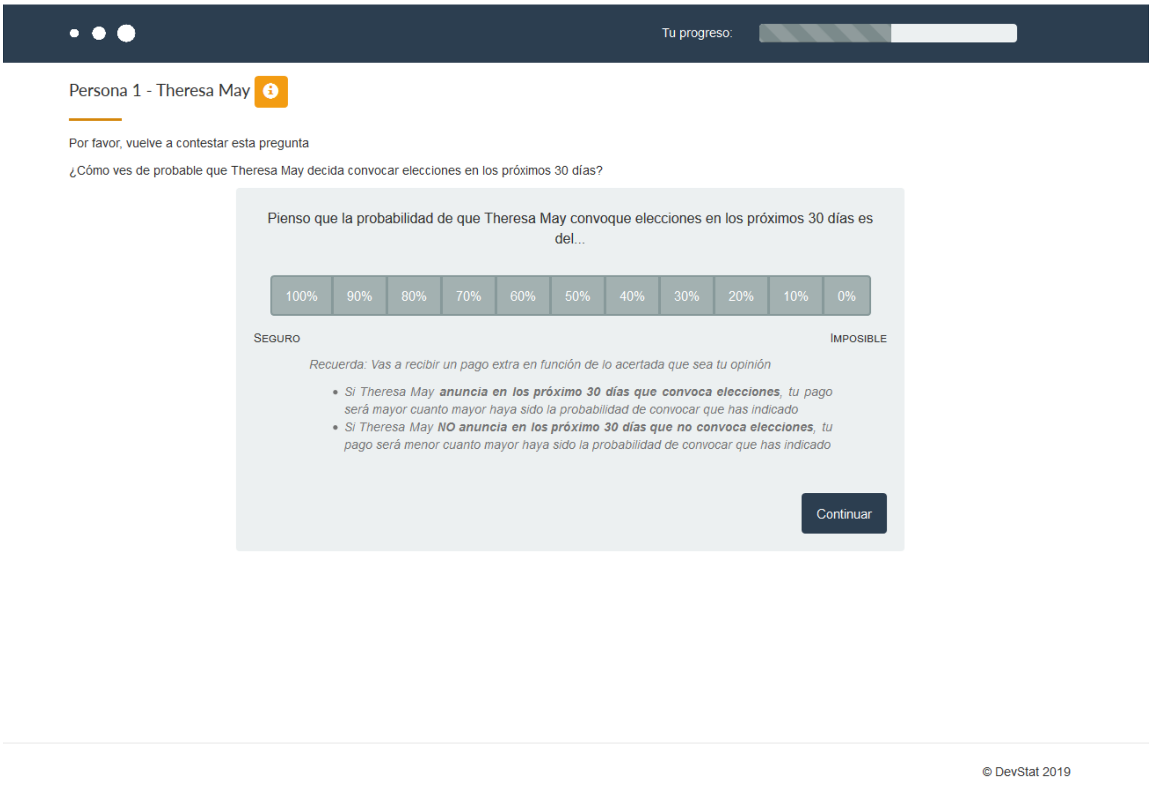
Persona 1 - Theresa May

¿Qué crees que haría Theresa May si supiese que tiene una probabilidad de ganar las elecciones del ...

...**100% de ganar**	[✓ Convocar elecciones] Quedarse como está
...**90% de ganar**	Convocar elecciones Quedarse como está
...**80% de ganar**	Convocar elecciones Quedarse como está
...**70% de ganar**	Convocar elecciones Quedarse como está
...**60% de ganar**	Convocar elecciones Quedarse como está
...**50% de ganar**	Convocar elecciones Quedarse como está
...**40% de ganar**	Convocar elecciones Quedarse como está
...**30% de ganar**	Convocar elecciones Quedarse como está
...**20% de ganar**	Convocar elecciones Quedarse como está
...**10% de ganar**	Convocar elecciones Quedarse como está
...**0% de ganar**	Convocar elecciones [✓ Quedarse como está]

Para poder continuar, por favor, contesta a todas las preguntas.

© DevStat 2019



Persona 1 - Theresa May

Por favor, vuelve a contestar esta pregunta

¿Cómo ves de probable que Theresa May decida convocar elecciones en los próximos 30 días?

Pienso que la probabilidad de que Theresa May convoque elecciones en los próximos 30 días es del...

100% 90% 80% 70% 60% 50% 40% 30% 20% 10% 0%

SEGURO IMPOSIBLE

Recuerda: Vas a recibir un pago extra en función de lo acertada que sea tu opinión

- *Si Theresa May **anuncia en los próximo 30 días que convoca elecciones**, tu pago será mayor cuanto mayor haya sido la probabilidad de convocar que has indicado*
- *Si Theresa May **NO anuncia en los próximo 30 días que no convoca elecciones**, tu pago será menor cuanto mayor haya sido la probabilidad de convocar que has indicado*

© DevStat 2019