D6.3 - Report on Experimental Analysis

Woohyun Shim, Fabio Massacci, Martina De Gramatica, Luca Allodi, Viet Hung Nguyen (UNITN), Julian Williams (UNIABDN/UDUR), Raminder Ruprai (NGRID).

Pending of approval from the Research Executive Agency - EC

Document Number	D6.3
Document Title	Report on Experimental Analysis
Version	10.0
Status	Draft
Work Package	WP 6
Deliverable Type	Report
Contractual Date of Delivery	30.04.2014
Actual Date of Delivery	30.04.2014
Responsible Unit	UNITN
Contributors	UNIABDN, UDUR, NGRID, URJC
Keyword List	Experimental Analysis, Comparative Statics, Workshops, Controlled Experiment, Case-Control Studies, Media Analysis
Dissemination level	PU

SECONOMICS Consortium

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

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

Document change record

Version	Date	Status	Author (Unit)	Description
1.0	25/09/2013	Draft	W. Shim, F. Massacci, L. Allodi, V. Nguyen (UNITN), J. Williams (UNIABDN), R. Ruprai (NGRID)	Draft table of contents.
2.0	29/01/2014	Draft	W. Shim (UNITN), J. Williams (UDUR)	Draft Sections 2 and 3.
3.0	20/02/2014	Draft	W. Shim, M. Gramatica (UNITN)	Revise Sections 2 and 3.
			A. Massa (UNITN)	Consolidate comments and suggestions w.r.t mathematical optimization models and comparative statics.
4.0	25/02/2014	Draft	W. Shim (UNITN)	Draft Sections 4 and 5.
5.0	04/03/2014	Draft	W. Shim (UNITN)	Draft Executive Summary.
6.0	14/03/2014	Draft	D. Rios (URJC)	Scientific Review.
7.0	18/03/2014	Draft	W. Shim, F. Massacci, L. Allodi, (UNITN), J. Williams (UDUR)	Reflect comments from the review and GA.
8.0	25/03/2014	Draft	W. Shim (UNITN)	Reimplementation in LaTex.
9.0	04/04/2014	Draft	E. Chiarani (UNITN)	Quality Check.
10.0	04/04/2014	Final	W. Shim, F. Massacci (UNITN)	Final version ready for submission.

Index

Executive summary

Technical Work Packages (WPs) in the SECONOMICS project have developed a series of theoretical frameworks and models applicable to various infrastructure sectors. While this theoretical approach can provide stakeholders with a general idea for designing and developing appropriate security strategies and policies, it might not be able to offer sufficiently practical information on the effectiveness of a security measure, strategy or policy, and to guarantee the validity. In order to attest validity and practicality of the theoretical models developed in the technical WPs, it is necessary to consider ways of complementing the models based on empirical approach.

D6.3 therefore aims at providing a discussion on an array of methodologies able to overcome the shortcomings in the theoretical models and circumvent issues related to the lack of market-driven data. In particular, we discuss the experimental methodologies for the systematization of exploration of the security measures, strategies and policies, and provide detailed information on experimental frameworks used for evaluating various security measures and policies tackled in the three case study WPs.

In Section 2, we discuss various quantitative and qualitative experimental methods that have been and can be used to complement the models and frameworks developed in the technical WPs. The proposed experimental methodologies include comparative statics, workshops (semi-structured interviews and focus groups), controlled experiments, case-control studies and media analysis.

Section 3 illustrates how different experimental methodologies have been and can be applied to complement theoretical models developed for the three SECONOMICS case studies. We focus particularly on presenting various applications of experimental methods for a specific case study. Furthermore, the section explores how an experimental method has been implemented for analysing regulatory frameworks from a cross-domain perspective.

Section 4 presents a way to employ an experimental method for evaluating impact of future and emerging threats and developing desirable policy strategies and instruments for such threats. We propose in particular to use sensitivity analysis to assess the impacts caused by emerging threats and compare resilience and sustainability of different security system designs.

Section 5 concludes the document. Particularly, we propose to employ a wide set of experimental methods to complement and evaluate security frameworks and models, since a single experimental method cannot explore and analyse all research frameworks and models developed in the SECONOMICS project.

The annexes have three research papers that focus on the application of the experimental methods. ANNEX1 shows how socially optimal regulatory rules can be identified using comparative statics. ANNEX2 attempts to analyse emerging cyber-threats using cross-domain application explained in Section 3. ANNEX3 compares different methods for security risk assessment using a controlled experiment.

1. Introduction

In the context of the SECONOMICS project, several theoretical models and frameworks applicable to information ecosystems for ensuring security have been developed and designed in the technical WPs 4, 5 and 6. The range of the models and frameworks that the WPs tackled is relatively large; they cover a broad spectrum of interests ranging from low level strategies for helping managers, consultants, and practitioners in delivering the security to the systems (e.g., D5.1 & D5.2), to high level policies for assisting policy-makers and industry/business leaders to ensure the overall sustainability and resilience of security ecosystems (e.g., D4.2, D4.3 & D6.1). Furthermore, D6.2 reviews current legal instruments in EU to address Pan-European exercises on security. However, depending sorely on such theoretical models might raise the question of validity – the degree with which the results are an accurate representation of stakeholders' views: to what extent are the findings of a given model valid?

For the models to move beyond a theoretical analysis of the effectiveness of a particular level of a security measure or a specific strategy or policy, it is necessary to consider some way of identifying functional relationships between inputs and outputs and estimating the values for providing different levels and types of security measures, strategies and policies. Furthermore, we need to obtain the information on current status and on changes in the current situation that is substantially outside of the range of current experience. These values can be obtained by using "observed" or "stated" information from research subjects.

While many methodologies have been developed and used to examine a theoretical model empirically, the underlying principle is to test the relationship between an input and an output by examining data-driven information. For example, in valuing a particular policy, economists have tried to estimate what are the costs and benefits for a particular policy under the situation where markets exist. Under this situation, the prices for marketed goods would contain sufficient information to ascertain what market players will gain and lose from the interaction.

However, since a situation like this does not exist for a public good such as security, costs and benefits cannot be estimated directly. For example, determining the effect of security provision is driven by difficult-to-value costs and benefits, and security that is considered as a public good do not have a market price for many cases. As a result, economists and social scientists have been forced to develop techniques to infer the values of public goods, including security, and various theories and models developed through the work of them helped formalize the notion of an equilibrium set of public goods and how to obtain them (e.g., contingent valuation and hedonic pricing).

The aim of this document is to present a set of methodologies which can create experiment spaces to perform valuation of the security models we have developed. Indeed, during the project period, sufficient experience has been acquired to devise methodologies for the systematization of exploration of the experiment space of policies for specified security architectures. This document details experimental methodologies that are and can be applied to analyse security-related issues (e.g., security rules and regulations) of various types of stakeholders. In the case studies of WPs 1, 2 and 3, various experimental evaluation methodologies are and will be used to determine and measure the performance of different security mechanisms.

2. Experimental Methods

As explained in D6.1 and [1], given the popularity of strategic behaviour and decision in an information ecosystem, designing appropriate theoretical models and employing proper experimental methods is crucial in studying such behaviour and decision. In economics, for example, many researchers have tried to develop a more rigorous and testable theory to identify principal–agent relationship addressed in D6.1, with the focus on linking it to experimental tools. Furthermore, in many research fields, researchers have started to combine multiple experimental methods, using both quantitative and qualitative approaches, in a study to realize the advantages of both and lessen their weaknesses. For example, in D4.2, ISASCR shows how a qualitative analysis on risk perception can be incorporated in a quantitative analysis. Accordingly, this report explains how we can employ various quantitative and qualitative experimental tools for evaluating and validating the theoretical models developed by the technical WPs. Table 1 illustrates the methods proposed as experimental tools and the examples of technical deliverables that are appropriate for the application of such methods.

Table 1: Applicable Experimental Methods

	Data/Formula-Driven	Human Perspective-Driven
Qualitative	Comparative Statics (e.g., D6.1, D6.2, D5.1 & D5.2)	Workshops (e.g., D6.1, D5.1, D4.2 & D4.3)
Quantitative	Case-Control Studies (e.g., D6.1 & D6.2)	Media Analysis / Controlled Experiments (e.g., D4.2 & D4.3)

Each experimental method is driven by different approaches and data collection procedures, and aims at exploring different types of research hypotheses. For example, let us consider the following illustrative hypotheses:

- Germany worries more about STUXNET than Italy.
- Patching vulnerabilities reduce risk by 45%.
- Vulnerabilities follow a linear law.
- A One-size-fits-all regulatory rule is more appropriate to U.K than Italy.

These are examples of hypotheses that are better tested by quantitative experimental methods. In contrast, certain hypotheses should be evaluated by qualitative methods. Some illustrative examples are:

- An increase in passenger profiling activities leads to a higher level of security.
- An increase in inspection capacity induces increased security.

- An increase in the terrorists' cost of carrying out an attack results in greater security.
- An increase in security regulation induces increased security.

In the subsections below, each experimental method is described in more detail. More specifically, we define the experimental methods, provide various application examples previously implemented by the SECONOMICS partners, and show how the tools can line up with the theoretical models and frameworks developed in technical WPs.

2.1 Qualitative Methods

2.1.1 Data/Formula-Driven: Comparative Statics

Comparative statics has been widely employed in various policy and economic models to show whether the variables are related to each other by some functional form and whether a change in an input variable induces a change of any output variable. More specifically, it is used to compare a set of equilibrium conditions that are related to different values of decision variables of the policy maker and parameters not governed by his decision (i.e., exogenous to the model). For purposes of a comparison, comparative statics always starts with the assumption that an initial equilibrium state exists. It then shows how a deviation in the model – in the form of change in the value of a parameter or an exogenous variable – alters the initial equilibrium state.

A comparative static analysis is commonly conducted in various ways. The analysis can be designed to show the direction of change in the equilibrium state with respect to the change in a decision variable or a parameter. For example, using an economic model, we can design a qualitative comparative statics analysis that shows the impact of an increase in security investment on the direction of change in the current equilibrium social welfare. On the other hand, a researcher might be interested in the magnitude of the change in the equilibrium state due to a given change in a parameter or an exogenous variable. For example, we can design a comparative static analysis which measures the magnitude of social welfare change resulting from a change in security investment as shown in ANNEX1. It should be noted, however, that the latter analysis always embraces the former since the direction can be obtained from the algebraic sign identified by the latter analysis [2].

Using a comparative statics analysis, many testable implications of models can be examined. The common approach to test such implications is to use simulation or regression analysis. Simulation provides economic agents with an array of possible outcomes (and the probabilities they will occur) in response to any choice of action. For example, in ANNEX 2 of D6.2 and [1], the authors explained how changes in a variable affects the equilibrium state (e.g., with respect to the expected loss from an attack). Other examples can be found in D5.1 and D5.2 which show how changes in a security investment portfolio affect the outcomes (e.g., expected utilities or estimated attack probabilities).

Regression analysis in contrast furnishes coefficients for given input and output variables. Examples can be found in [3]. In this study, the authors studied software vulnerabilities as an emerging threat, and estimated the direction and the magnitude of security vulnerability disclosure based on time-based vulnerability discovery models, see Figure 1 as an example. They showed how a logistic-based model can be applied to predict the discovery process of

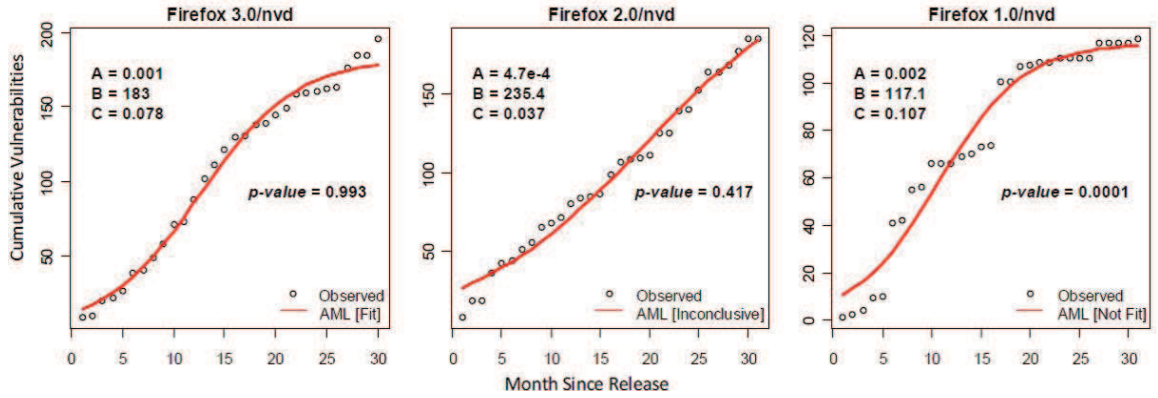


Figure 1: Fitting Vulnerability Data Sets

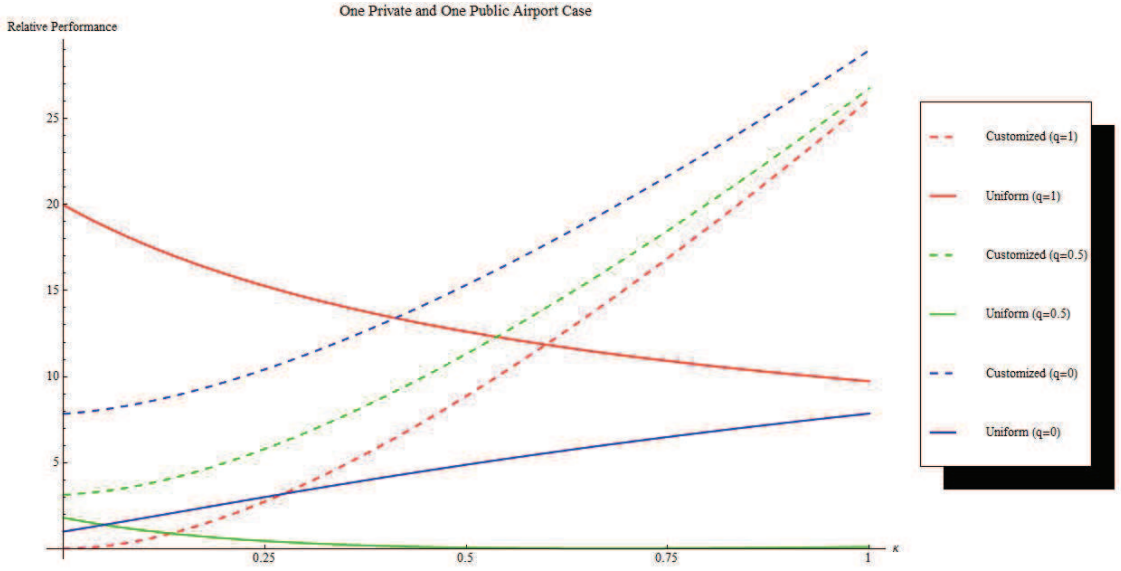


Figure 2: Effect of changes in state security charges on expected loss

vulnerabilities. The study further tested the goodness-of-fit in order to measure the explanatory power of the model.

In the context of SECONOMICS, we use a comparative statics to characterize the strategic decision and behaviour of economic agents. In detail, we use a comparative statics method to verify the models developed in the project and get a more detailed insight on dynamic processes of an equilibrium change for the alternation of a given variable. This study, therefore, will not only provide a simple change in equilibrium given the alteration of an input variable, but also disclose profound information on the whole structure of the systems. For example, in the models for regulatory structure of airport and national grid, an extensive comparative statics analysis will be conducted to study the relationship between different regulatory and financing systems (e.g., customized or one-size-fits-all regulations, and centralized or decentralized financing systems) and social outcomes (For example, see Figure 2 and the figures in ANNEX1).

In summary, a comparative statics analysis is a powerful experimental tool, as it can provide policy makers with better ideas on the relationships between various inputs and outcomes. It should be, however, mentioned that, for the models based on a game theoretic framework addressed in D6.1 and D6.2, comparative statics should be used with care since a change in an input variable results in a feedback effect between the strategies of different players: when an input is altered, not only does one player's equilibrium state change in response, but also other players' reaction is altered as a result of a change in the former [4].

2.1.2 Human Perspective-Driven: Workshops

The main objective of workshops as an experimental method is to capture in-depth meaning and interpretation of the information that arithmetic formulas cannot convey. While workshops can employ various experimental methods, we use two experimental tools: semi-structured interviews and focus groups. These tools are particularly known to be particularly useful for getting in-depth knowledge and insights from the respondents.

Interviews aim at understanding a state or a phenomenon from the interviewees' perspectives and at uncovering its underlying meaning from their experience and knowledge. Even if survey-based methods might be cost- and time-effective compared to interviews, they are only able to use closed-ended questions that might prevent a researcher from getting more detailed feedback that can potentially be obtained from respondents. In contrast, interviews allow respondents to express their opinions from their own perspectives with their own words. Since an interview is conducted based on structured conversations designed and controlled by an interviewer, it is particularly useful for revealing the in-depth story of interviewees. In conducting a series of interviews, we use a semi-structured procedure, since it allows an interviewer to pursue new ideas that emerge during the interview.

The other tool used in workshops is focus groups. Similarly to interviews, focus groups can provide deeper insights into how people feel and think about a phenomenon under study. Although interviews allow a researcher to capture detailed information from respondents, conducting a series of interviews may be highly time-consuming and expensive. A focus group makes it possible to overcome this issue: it can be conducted in a time-saving and economical manner. In detail, focus groups are conducted by organizing group interviews. In the procedure, a researcher, as a moderator, induces participants to interact with each other and to express their experience and opinions. As a result, focus groups can capture detailed information boosted by group interaction and non-verbal communication: by participating in the discussions in a focus group, participants might be able to link diverse concepts and develop collective thoughts that cannot be obtained from individual interviews.

In the context of SECONOMICS, workshops are therefore particularly suited to obtain in-depth information from and knowledge of the participants and to link them with theoretical models and frameworks developed by the technical WPs. While interviews and focus groups can be conducted separately for a specific topic being explored, we integrate these two tools into a workshop. This gives us a better chance to get both individual and group knowledge. Particularly, in the first half phase of the project, we organized a series of small workshops with experts to better understand the security environment in the corresponding field, and to feed the information obtained from the interview back to technical WPs to identify appropriate

research design. The information has been reflected to the technical deliverables (e.g., D4.2, D5.1, and D6.1) as well as case study deliverables.

From the second half phase of the project, we started to organize workshops to evaluate modelling approaches developed in the technical WPs. In detail, the workshops aimed at discussing and validating the models for security decision making developed in the technical WPs, together with obtaining various practical perspectives from stakeholders. For example, SECONOMICS project partners presented the models and the results from the project, and asked the participants to hold a focus group discussion for evaluating them. This made it possible for us to have the participants' collective views. In addition, a series of interviews was also conducted during the workshops. It was found that there are actually diverse views on different security issues depending on the participants' job rank, knowledge and experience. The participants' views gathered in the workshops have been (and will further be) reflected and consolidated in case study and technical deliverables (see, for example, D1.3, D1.4, D3.3 and D3.4 as well as ANNEX1).

2.2 Quantitative Methods

2.2.1 Data/Formula-Driven: Case-Control Studies

A common obstacle in studying information ecosystems is that events of interest are rare. While rare events such as terrorist attacks and security incidents attract substantial interest in the fields of economics and sociology, these events might be difficult to be analysed by a conventional statistical model since they are too sparse across time and agents. Consequently, studying rare events with traditional experimental research methods might be neither effective nor desirable. Such problems in analysing rare events occur frequently in the security research field.

The main objective of a case-control study is to overcome the problem related to studying rare events and provide an efficient solution to the problem. It can be defined as an investigation of the association between a specific event and potentially affecting factors by taking separate samples of event cases and of controls that have a chance of event occurrence. While this method provides substantial efficiencies, it has not been employed frequently in the field of security research.

The defining feature of a case-control study is that the sample is stratified on a discrete dependent variable, not, as in conventional research practice, on one or more independent variables. For example, in ANNEX2 and the study of vulnerability exploitation [5], UNITN enumerated all the vulnerabilities exploited in the wild ("the cases") during a certain period of time, and the population at risk – all the vulnerabilities disclosed during that time period ("the controls"). The studies then estimated the effect of vulnerability characteristics on the risk of vulnerability exploitation in the wild.

The advantage of case-control analysis is that it would yield a valid estimate of the relative effect of input variables on the hazard rate for event occurrence, similarly with the results that can be obtained from a large panel data analysis. Consequently, a case-control study would make it possible for a researcher to gain similar results at a small fraction of the time and costs for conducting a longitudinal study.

The rarer an event is, the greater such efficiency will likely be. In the field of information security, for example, an economist might wish to explore factors that cause organizations

to convert to sustainable security systems by following a panel of organizations, year after year, until a sufficient number converts to the sustainable system. These hypothetical studies could be done relatively inexpensively and quickly using case-control designs. The feasibility of such a study, and that of case-control studies in general, depends on how easy it is to obtain some sort of listing of units which experienced the rare event.

Consequently, in the context of SECONOMICS, case-control design might offer economy of effort without loss of validity, and without resort to highly specialized or unusual methods of analysis.

2.2.2 Human Perspective-Driven: Controlled Experiment & Media Analysis

Controlled Experiment Debates on various security regulations and demands for an appropriate evaluation have grown over the last years. For example, 3D body scanner and CCTV have caused serious privacy concerns and their effectiveness has been questioned. Furthermore, many security measures, policies and strategies implemented by an organization or mandated by a government agency have been asked whether they truly improve the effectiveness and efficiency of the security environment. One of the viable experimental tools for evaluating the performance of these new measures, policies and strategies is a controlled experiment.

A controlled experiment can be defined as an assessment or a series of assessments in which intentional alterations are made to input variables of a system so that a researcher can identify the effects of alterations that may be detected from an outcome [6]. In detail, a controlled experiment is conducted by researchers who control input variables in a laboratory setting, in order to observe any change in the output variable that can be attributed to the inputs [7]. If a researcher changes only one input with all the other inputs held constant, he might be able to estimate the effect of changes in the input on the outcome. Using this experimental method, a researcher can obtain a high degree of confidence that any changes in the outcome result from the input altered by him [7]. Therefore, the main advantage of a controlled experiment is that a researcher can manipulate the inputs being studied to analyse the potential impacts on the outcome without any confounding effects.

For example, in the field of security and risk management, Massacci et al. [8] conducted a controlled experiment to evaluate whether the effectiveness of the approach in modelling requirements evolution in the aviation domain depends on the analyst's level of knowledge on the approach and on the domain. Other examples are the study using a Smart Grid application scenario presented in ANNEX3 and [9]. By using a randomized block design, the authors investigated the effect of perception of two different risk-based methods, visual and textual, on the effectiveness of the methods.

As for the SECONOMICS project, a controlled experiment can be used in various ways. For high-level security regulations, it can provide policy makers with a better ideas of how newly proposed regulations can improve the overall performance of the system (WP6). For low-level operational cases, it can give managers a clearer understanding of how well proposed security procedures and measures achieve security goals (WP5).

Media Analysis This is a specialized subset of content analysis and a well-established research methodology. It is widely used to analyse a broad range of texts from media. Media

analysis can be defined as a technique that focuses on analyzing the contents of text, where "'contents' refer to words, meanings, pictures, symbols, ideas, themes, or any message that can be communicated and the 'text' is anything written, visual, or spoken that serves as a medium for communication" [10]. The benefits of a media analysis are twofold: it provides a deep insight both on the creative process of public opinion's attitudes and perceptions and, at the same time, reflects existing attitudes, perceptions and cultural patterns.

In order to conduct a media analysis, a coding method should be chosen: the coding system provides the setting of a codebook containing a list of codes considered to be relevant and crucial to the comprehension of the text. The choice of the text strongly influences the results of the analysis; media forms, genre, time-frame and key-words should be carefully selected to obtain a set of useful data that depict a realistic picture of the situation.

A media analysis can be regarded as a suitable research method for evaluating various security mechanisms since it can unearth the actual interrelationship between security-related issues and public attitudes towards these issues unveiled by the analysis of the media.

In the context of SECONOMICS, a media analysis is employed by WP4 partner to examine the effect of security-related issues (i.e., 3D body scanner employment, Stuxnet incident and CCTV camera implementation) on the general public. The main objectives of the analysis are therefore to obtain information on longitudinal changes and cultural heterogeneity of risk perception with respect to a certain security issues from a sociological perspective, and to gain a better knowledge on risk phenomena by investigating attitudes, perception and behaviour of citizens.

In D4.2, extensive reviews of the literature have identified that: (1) risk perception, as well as risk tolerance, depends on different cultural and socio-economic conditions and therefore are strictly culturally defined; (2) varying degrees of acceptance level of security measures are related to social context alike, changes according to cross-cultural differences and are affected by time changes; and (3) communication on risk and security topics strongly affects citizens' attitude and position towards policies and political decision-makers. As a result, a media analysis is conducted based on a cross-cultural and cross-national perspective.

The results of the analysis are fed into theoretical models and used to evaluate and validate them. Furthermore, the results are used to: (1) identify effective channels and patterns of communication, as well as risk perception related issues; (2) conceptualize and assess public acceptance of security measures; and (3) elaborate recommendations for practitioners, communication strategies between policy makers, stakeholders and citizens in the area of security and risk (see D4.3 and D4.4).

3. Case Study Analysis

In this section, we illustrate how the previous experimental methods can be linked with various security scenarios identified in the case study WPs. Here, this section focuses particularly on explaining specific experimental methods that are employed for analysing different scenarios in the case studies. Furthermore, we explore how an experimental method can be employed for studying a cross-domain issue and an emerging threat.

3.1 Grid Case Study

3.1.1 Comparative Statics

As presented in D6.1 and D6.2, NGRID has employed very complex Industrial control systems (ICS) / Supervisory Control and Data Acquisition (SCADA) systems to deal with its systems and networks, and incorporated various standardized communication technologies such as internet protocols and mobile networks in order to obtain high flexibility as well as low cost operation for its networks and systems. Furthermore, various entities have started to share corporate networks for exchanging on-going information of the normal business activities. This implies that, while the systems and networks used today in NGRID can have high level of interoperability, they are also likely to have high level of systematic security risks. As a result, current NGRID systems and networks might provide attackers with a more effective way of penetrating attacks.

In WP6, we analyse how security investment decisions across various operators lead to changes in an attacker's attacking intensity and how externality can be internalized by making coordination across the operators. More specifically, in ANNEX 2 of D6.2, using comparative statics, we illustrate various cases with different security environments, for example, (1) where operators are unregulated and make investment decision based on the Nash equilibrium; (2) where the policy-maker is fully informed and imposes a mandatory investment level on the operators; and (3) where the policy-maker cannot observe operators' shift of assets (i.e., reducing the policy-maker's abilities) and cannot mandate security investment any longer. The results provide detailed information on how shocks resulting in different levels of attacking intensity cause operators' shift of assets and why operators might prefer an unregulated environment to a regulated environment.

3.1.2 Controlled Experiment

While there have been various proposed methods to identify and evaluate security threats and vulnerabilities, little experimental evaluation has been conducted in practice. Furthermore, studies for experimental evaluation have been conducted by the same researchers who have proposed the methods. As a result, security practitioners are not confident in adopting the proposed security methods, or even criticize the effectiveness of the methods in practice. In order to address this problem, there have been constant demands for conducting empirical evaluations to analyse which methods work better to assess security threats and vulnerabilities and why [5].

In WP2, we intend to conduct a controlled experiment to compare and evaluate the performance of different methods for risk analysis (that identifies security assets, unwanted

incidents, threats and vulnerabilities) and treatment. For example, in ANNEX3 and [9], UNITN and NGRID conducted controlled experiments that compare the performance of two different classes of security methods for a Smart Grid application scenario (ranging from security management to database security): visual methods and textual methods. The goal of the experiment was to evaluate the effectiveness of the methods, and the participants' perception of them. By employing and extending the similar experimental design with this study, we will be able to identify an effective risk-based security method used or potentially used in NGRID, and to provide decision-makers with helpful insights about the pros and cons of different security methods for treating and mitigating security risks.

3.1.3 Media Analysis

For the NGRID case, the media analysis is centered on the Stuxnet incident, as an example of a possible threat endangering the functionality of a critical system. Topics, issues, quotations and correlation among codes are analysed for the selected topic to have insights about how they are framed and debated in the selected media. It allows us to identify involved actors, and to assess the discourses and justifications of security and risk appeared in the domestic and international media.

In order to proceed with the analysis, we investigated a parameter which can be used as a proxy for assessing the social acceptance related to security measures designed to prevent cyber incidents, such as Stuxnet attacks. We use media salience as the parameter reveals the degree to which common citizens are exposed to and familiar with these issues. With respect to the topic, we investigate frequency, coverage and characterization of the debates. Furthermore, we conducted a comparative assessment to identify cross-cultural and cross-national differences and common points among different countries.

We designed the analysis to investigate two different aspects: the degree of social acceptance on the Stuxnet incident and the framework of the discussion itself. Therefore, we consider the following:

- By correlating the representativeness of certain codes (salience rate) with the argumentative direction used to refer to the code itself (inclination), the study identifies the degree of social acceptance of security measures
- By examining co-occurrence among different actors, topics, argumentative strategies and justifications, together with a cultural interpretation of the social and economic context, the study provide relevant elements allowing depicting the framework of the discussion itself, how it is framed and debated by the public opinion.

As a result, a media analysis on the Stuxnet incident makes it possible to evaluate the perception of the public opinion related to cyber attacks on critical infrastructure sectors, and to provide a comprehensive conceptualization as contribution to a suitable and usable knowledge for decision makers and stakeholders on security risks.

3.2 Airport Case Study

3.2.1 Comparative Statics

After 9/11 events, a large number of security regulations that require increased expenditure have been established in the aviation industry. These regulations are often very complex and deal with a wide spectrum of security policies and measures affecting various aspects of the system in question. These regulations have different rules (e.g., customized and one-size-fits-all) and financing mechanisms (e.g., centralization and decentralization).

While different European States use different financing and regulatory systems, the trade-offs between these mechanisms have not been analysed well in the field of aviation security research. For example, while we found various industry reports arguing that the government should avoid setting and imposing a one-size-fits-all regulation, and try to mirror particularities and characteristics of different airports in designing a security regulation, we presently do not have much theoretical and empirical evidence supporting this argument.

Therefore, we develop a model based on political economics and conduct an experimental analysis using comparative statics. We take a fresh look at different financing and regulatory structures in the field of aviation security and examine their pros and cons. More specifically, we develop a model with an airport network that encompasses airports with different characteristics (e.g., sizes and externalities) and evaluate the performance of centralized and decentralized financing, and customized and one-size-fits-all regulatory approaches, and compare the trade-offs between these approaches.

Using comparative statics, we illustrate how the relative performance of these approaches changes with different characteristics of the aviation network and externalities. The detailed analysis and results are attached in ANNEX1.

3.2.2 Workshops

In order to better develop narrative scenarios and to design and evaluate the theoretical models, a series of workshops were organized by the case study partners. In the workshops, we conducted individual interviews and focus group discussions with stakeholders, experts and well-informed airport security representatives at national and international levels. These exchanges allowed us to identify and interpret the main issues in airport security, to feed the obtained information into the models, and to evaluate the performance of the developed models.

A series of interviews and discussions provided an alternative solution for gathering data that is not available from traditional sources (statistical data, cross-national surveys). Moreover, given its subjective nature, these tools allow to deepen and enrich the quality of the answers catching the personal and collective points of view of respondents that may not always coincide with the official discourse provided by formal documents and speeches. Furthermore, interviews and discussions guarantee a better understanding of the meanings that different actors attribute to the context in which they live in, the motivations leading their behavior and the social patterns they follow.

In the first phase of the SECONOMICS project, we conducted a series of individual interviews and focus group discussions with stakeholders to identify the main security issues in the aviation field. More specifically, we had the chance to speak with high level representatives of the Italian State Airports Authority (ENAV) and the main Italian airport operator, which gave us interesting insights for reflection on the economic impacts affecting the air-

port side regarding the implementation of different aviation security policies and regulations. They made points in relation to the importance of establishing and supporting a security culture involving all actors within the airport domain. It was also identified that they consider training as a cost-efficient way to improve security. We used the results of the interviews and discussions to design theoretical models, and further developed more detailed questions.

In the subsequent workshops, we conducted interviews and discussions in two medium-small size airports with stakeholders and experts at different levels, and dealt mainly with airport management issues with a reference to the national and international regulatory structure. Data collected through the interviews and discussions were used to populate and calibrate - in a later phase - the theoretical models, examining the existing trade-offs between implementing security measures and costs and efficiency issues in relation with the current regulatory setting.

Additional interviews and discussions will be conducted in order to investigate and model the contractual relationship between airports and the outsourced service companies, with particular attention to private security agencies. This might lead us to identifying, developing and validating theoretical models based on the principal-agent theory, outlining the interplay among different actors in the arena (see D6.1 for more details about principal-agent theory).

3.2.3 Media Analysis

A media analysis is applied to investigate the case study for airport security, focusing on the implementation of a 3D body scanner in an airport. The main focus is on analyzing the tradeoffs between security and possible restrictions on personal freedom and privacy, as well as health and cost concerns. The analysis investigates how the proposal of introducing the 3D body scanner technology in the airport domain affects citizens' perception on security and risks. It allows us to understand how the topic is framed and communicated to the general public in the selected countries with different degrees of social acceptance on this security technology. The salience concept, moreover, is used as a relevant measure providing insights for identifying the perception on security issues among the citizens and passengers.

Looking into details, a discussion to introduce the 3D body scanner technology started in the U.S., just after the failed terrorist attempt in 2009 on a flight from Amsterdam to Detroit. From that moment, the topic became a priority in the the political agenda at an international level. Some EU countries reacted to the events happening in the U.S. supporting and implementing this technology in their airports, while other countries expressed negative opinions mainly due to privacy and health concerns. One common tendency ties all countries: the debate is led by two well defined groups such as the Transport Security Agency and politicians strongly supporting the implementation, whereas passengers and experts arguing against it. The main argument used to criticize the adoption of 3D body scanners is supported by the concerns about legality, privacy and health issues, while supporting opinions are summarized as the enhancement of the efficiency in the counter-terrorism security measures.

The analysis on the countries' opinion reveals that the social acceptance of this technology depends on several factors: countries with previous experience of national and international terrorism are more inclined to the introduction of a 3D body scanner as they feel to be generally more exposed to potential terrorist attacks. Furthermore, not only past experience

but also cultural patterns and structures of thought affect the perception of risk and threat, as well as the positive/negative inclination towards security measures applied.

3.3 Cross-Domain Applications

In the field of security, there has been a fundamental question regarding which regulatory frameworks can produce a socially better outcome. According to our study (see, for example, Deliverables 2.1, 2.2, 2.3 and 6.1), security regulatory frameworks that are implemented in various industry sectors and countries are fundamentally different. These frameworks can be broadly classified into two approaches: principles-based and rules-based approaches.

While a principles-based framework defines high-level general statements that define a goal or objective of the entity adhering to the principle, a rules-based framework commonly provides a wide set of compulsory instructions and requirements. As a result, it is likely that a regulatory system based on a rules-based framework has a wide array of requirements. In the cases of information security or cyber-security, the main constituent of a principles-based approach is a risk-based approach (See D6.1 for more details). Risk mitigation is therefore built into the principle.

The main advantage of a rules-based regulatory system is that it can ensure that all parties that need to adhere to it are applying the same set of security controls, and may even specify the details of how the controls are to be implemented. On the other hand, a principles- or risks-based regulatory approach can have flexibility encompassing a wider range of scenarios than a rules-based approach. Both approaches can be considered as a "double-edged sword": some argue that a rules-based approach is undesirable since it might not be able to foster development and innovations in security practices and might entail a high systematic risk. The others claim that, while a risks-based approach might ensure a low systematic risk, it cannot guarantee the certainty of achieving a higher level of consistency in the system.

Many evidences are likely to suggest that there has been an incremental shift along the spectrum of principles to rules in security regulations and standards towards the direction of more rules. We believe that this trend is due to a focus on compliance with the detail of the requirement, which may or may not be in alignment with its spirit. However, a shift in the opposite direction would require more attention to defining the purpose of the standard and regulation and articulating an assurance and enforcement framework.

In order to assess the adequacy of rules- versus principles-based approaches in various domains, a series of experimental methodologies can be applied. One example is to conduct a comparative statics analysis in the context of the incentive structures of various agents within a system, as explained in D6.1. If a public policy maker imposes certain behavioural constraints on targets regarding security, a natural issue of aligning incentives, which is a standard principal-agent problem in economics, appears. A natural question is then where to place the specific constraints on behaviour and what regulatory framework should be used to enforce those constraints. In a principles-based system, a set of idealized outcomes is specified. Alternatively, if a public policy maker sets a series of rules, then these rules may conflict with the risk targets of the agents.

A generalization from the economics literature is that the alignment of incentives tends to favour principles-based approaches (i.e. aligning risk preferences and the inter-temporal

substitution between short and long term risks via discount rates). However, this is not always the case. For situations whereby incentives cannot be, or are too difficult to be, aligned, contractual requirements with dichotomous adherence are sometimes favoured. On the other hand, setting a penalty structure based on violations of rules does not always result in the correct internalization of externalities at both the level of the agents and the wider economy.

Using comparative statics, we can analyse the level at which behavioural restrictions need to be enforced contractually and potential regulatory frameworks and policy mechanisms that affect the evolution of these arrangements.

Another example for assessing the adequacy of regulatory frameworks is to use a case-controlled experiment as shown in the study of vulnerability exploitation in ANNEX2. In the study, the authors at UNITN evaluated the efficacy of a rule-based framework in mitigating security risk. In detail, the U.S. federal government uses a rule-based cyber-security policy by stipulating that software vulnerabilities marked as “high risk” by the Common Vulnerability Scoring System (CVSS) should be fixed with high priority. In the analysis, using a case-controlled methodology, the authors tried to identify whether a rule-based approach based on CVSS is an effective strategy for mitigating security risks. The results found that, while a shift towards the direction of increased rules has been witnessed in the field of information security, a major adjustment may be require in security regulations and policies to be more effective in mitigating security risks.

4. Future and Emerging Threats

We have witnessed that many critical infrastructure sectors have become more technologically and managerially complex and tightly linked. Most of the systems interact through networks and form complex information ecosystems. While the high level of connectivity and complexity might increase efficiency, it is also known to be more vulnerable. The performance of the whole ecosystem is affected by the performance of each system since the disruption of one system (e.g., due to terrorism or catastrophic natural disasters) can result in transboundary effects on other systems. For example, many of the blackout events in U.K. and U.S. were caused by a malfunction of one of the components. This implies that tight linkage and complex structure of the information architectures might amplify the impact of a small shock.

It seems therefore obvious that, while many national and supranational governmental bodies try to avoid a failure in the critical infrastructure sectors, such failures might not be avoidable. If they need to deal with future and emerging threats (e.g., cyberterrorism and bioterrorisms), the problem may become even more severe since the impact of an incident resulting from an emerging threat is characterized by high degree of uncertainty and unpredictability: the known unknowns [11][1].

As a result, this section illustrates how to evaluate impacts of emerging threats and develop desirable policy strategies and instruments. Addressing the problems caused by future and emerging threats in the infrastructure sectors involves two procedures. The first is to understand the consequences of system failures caused by these threats. As information ecosystems become more complex, infrastructure sectors come to confront more diverse and unpredictable risks: a disruption of one infrastructure may directly and indirectly influence other infrastructure sectors, and affect large spatial areas. Furthermore, as witnessed from a series of Stuxnet incidents, it came out into the open that a disruption caused by an unknown threat can result in not only financial losses but also non-pecuniary damage on the economy and society. While it is difficult to predict what emerging and future threats to security can actually cause impacts on the economy and society, one thing is clear: without a reliable structure, a desirable information ecosystem cannot be realized.

The second issue is to identify a way to design a system that can improve its sustainability and resilience to a shock. Since unprepared systems facing ad-hoc crises caused by emerging threats can face drastic consequences, a public policy-maker or a security manager need to design and implement security requirements or preserve resources that enable the information ecosystem to renew and restructure itself after shocks. As a result, concepts of system sustainability and resilience, which are recognized as two possible ways of tackling the impacts after shocks are being more and more used both at a governmental and supragovernmental level. Whereas sustainability refers to the competence that makes the system to maintain itself within acceptable bounds of operating state despite the disturbance in the system, resilience refers to the capacity to restore the system to an acceptable operating state after a shock.

As explained in D6.1, in the field of information security, many researchers have developed a model designed to interpret the dynamics of complex information ecosystems with a shock and to increase resilience and sustainability of the ecosystem. They mostly focused

[1] See also http://www.cost.eu/domains_actions/isch/Actions/IS1304 for more details.

on studying the efficiency of various systems and the time the systems require restoring to an operating state after a shock, since having a sustainable and resilient system is a key element for achieving a more steady-state information ecosystem that can deliver essential functions for economy and society, particularly in times of crises. The idea of sustainability and resilience might therefore be able to provide us with new stimuli for facing the challenges due to emerging threats.

However, our knowledge about the systems and structures that improve resilience and sustainability, and practical information on how a security manager and a policy-maker can design a strategy and a policy that can cope with the challenges from emerging threats are still rather limited. Since impacts caused by emerging and future threats might be characterized by high degrees of complexity, uncertainty and unpredictability, policy-makers and security managers are facing more dynamic and complex problems for which they need to find economically and socially suitable answers. As a result, coping successfully with emerging threats needs more than traditional approaches.

In order to conduct an assessment on the impacts caused by emerging threats and to compare resilience and sustainability of different system designs, sensitivity analysis can be used [12]. In detail, sensitivity analysis is an approach to study how "sensitive" a system is to changes in the values of the elements in the system and to changes in the system structure. Sensitivity analysis is commonly conducted as a series of tests in which the investigator sets different values for the elements or different scenarios for the system to see how the changes in these result in alterations in system performance. By showing how the system behaves in response to changes, sensitivity analysis helps an investigator not only to evaluate the system performance with a shock but also to design a sustainable and resilient system structure (see ANNEX 2 in D6.2 as an example).

In addition, sensitivity analysis is a useful tool in figuring out system dynamics. Running an analysis with diverse values of the elements can provide an investigator with deep insights into the behavior of a system and possible improvements in the process being modeled. Sensitivity tests for highly uncertain emerging threats will therefore make it possible for us to identify the important links between the system and parameters, and a sustainable and resilient system structure together with optimal rules for security regulations and strategies, and resource allocation. In WP6, we have designed and developed a framework for security investment decision-making, and conducted a series of sensitivity tests by changing the values of the variables in the model as well as by making changes in premises and assumptions of a scenario (see the annex in D6.2). We believe that, knowing the values used in the analysis are based on some available information and heuristic choice as well as the subjective judgments, the characteristics of sensitivity analysis can provide information on a more realistic and reliable decision-making for providing against emerging and future threats.

5. Conclusions

In this deliverable, we have outlined experimental methodologies that are or could be employed for analysing models developed in technical WPs. We discuss how we can create experimental spaces to perform qualitative and quantitative analyses, and how we can evaluate different security measures and rules in relevant economic systems. In detail, we include discussion of stakeholders, preferences and norms in the information ecosystem as well as methodologies for the systemization of exploration of the space of policies for specified security architectures.

We propose to employ a wide set of experimental methods to evaluate security mechanisms and frameworks in sociological and economic settings, since a single experimental method cannot explore and analyse all research frameworks and models developed in the SECONOMICS project. We therefore focus particularly on understanding and assessing the experimental methodologies that can complement each other. In addition, we include a discussion of applications of experimental methods on the case studies delineating security policies in the context of economics, law and society. It is important to note that, while we do not explicitly include, similar experimental methods illustrated above they have been also integrated to analyse the cases provided in WP3 (e.g., workshops and media analysis). Further work will focus on refining the experimental frameworks, and producing a set of policy papers based on the proposed experimental methodologies in conjunction with WPs 1, 2, 3, 4, 5 and 6.

BIBLIOGRAPHY

[1] David Pym, Joe Swierzbinski, and Julian Williams. The need for public policy interventions in information security. Available at http://www0.cs.ucl.ac.uk/staff/D.Pym/InfoSecPubPol.pdf [Last accessed: 14 February 2014], .

[2] Jacco Thijssen. *Investment under uncertainty, coalition spillovers and market evolution in a game theoretic perspective*. Kluwer Academic Publishers, 2004.

[3] Viet Hung Nguyen and Fabio Massacci. The (un) reliability of nvd vulnerable versions data: an empirical experiment on google chrome vulnerabilities. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, pages 493–498. ACM, 2013.

[4] Eric Rasmussen. *Games and information: an introduction to game theory*. Oxford : Blackwell, 2002.

[5] Luca Allodi and Fabio Massacci. My software has a vulnerability, should i worry? *arXiv preprint arXiv:1301.1275*, 2013.

[6] Douglas C Montgomery. *Design and analysis of experiments*. John Wiley & Sons, 7th edition, 2008.

[7] Cary Coglianese. Empirical analysis and administrative law. *Univ. of Illinois Law Rev.*, pages 1111–37, 2002.

[8] Fabio Massacci, Federica Paci, Le Minh Sang Tran, and Alessandra Tedeschi. Assessing a requirements evolution approach: Empirical studies in the air traffic management domain. *Journal of Systems and Software*, 2013.

[9] Katsiaryna Labunets, Fabio Massacci, Federica Paci, and Le Minh Sang Tran. An experimental comparison of two risk-based security methods. In *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on*, pages 163–172. IEEE, 2013.

[10] William Lawrence Neuman. *Social research methods: Quantitative and qualitative approaches*. Allyn and Bacon, 2005.

[11] LeMinhSang Tran and Fabio Massacci. Dealing with known unknowns: Towards a game-theoretic foundation for software requirement evolution. In Haralambos Mouratidis and Colette Rolland, editors, *Advanced Information Systems Engineering*, volume 6741 of *Lecture Notes in Computer Science*, pages 62–76. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-21639-8. doi: 10.1007/978-3-642-21640-4_7. URL http://dx.doi.org/10.1007/978-3-642-21640-4_7.

[12] Andres Redchuk and David Rios Insua. Sensitivity analysis. In SaulI. Gass and MichaelC. Fu, editors, *Encyclopedia of Operations Research and Management Science*, pages 1379–1384. Springer US, 2013. ISBN 978-1-4419-1137-7. doi: 10.1007/

978-1-4419-1153-7_1155. URL http://dx.doi.org/10.1007/978-1-4419-1153-7_1155.

[13] European Commission. Report from the commission on financing aviation security. European Commission, 2009.

[14] Michael Carney and Keith Mew. Airport governance reform: a strategic management perspective. *Journal of Air Transport Management*, 9(4):221 – 232, 2003. ISSN 0969-6997. doi: http://dx.doi.org/10.1016/S0969-6997(03)00003-6. URL http://www.sciencedirect.com/science/article/pii/S0969699703000036.

[15] Vicki M Bier, Naraphorn Haphuriwat, Jaime Menoyo, Rae Zimmerman, and Alison M Culpen. Optimal resource allocation for defense of targets based on differing measures of attractiveness. *Risk Analysis*, 28(3):763–770, 2008.

[16] Oates Wallace. *Fiscal federalism*. New York: Harcourt Brace Jovanovich, 1972.

[17] Wallace E Oates. An essay on fiscal federalism. *Journal of economic literature*, 37: 1120–1149, 1999.

[18] Robert Dur and Hein Roelfsema. Why does centralisation fail to internalise policy externalities? *Public Choice*, 122(3-4):395–416, 2005.

[19] Timothy Besley and Stephen Coate. Centralized versus decentralized provision of local public goods: a political economy approach. *Journal of public economics*, 87(12):2611–2637, 2003.

[20] European Commission. Directive of the european parliament and of the council on airport charges. European Commission, 2007.

[21] Martin Thelle, Torben Pedersen, and Frederik Harhoff. Airport competition in europe. Copenhagen Economics, 2004.

[22] Irish Aviation Authority & Aviasolutions. Study on civil aviation security financing. Irish Aviation Authority & Aviasolutions, 2004.

[23] SH&E Limited. Capital needs and regulatory oversight arrangements. SH&E Limited, 2006.

[24] Stefano Baronci. Aci europe position on airport charges. Airports Council International, 2007.

[25] Christos Ioannidis, David Pym, and Julian Williams. Sustainability in information stewardship: Time preferences, externalities, and social co-ordination. In *The Twelfth Workshop on the Economics of Inforamtion Security (WEIS 2013)*, 2013.

[26] David Pym, Julian Williams, and Iffat Gheyas. Resilience in information stewardship. 2013, in preparation, .

[27] Lawrence A. Gordon and Martin P. Loeb. The economics of information security investment. *ACM Trans. Inf. Syst. Secur.*, 5(4):438–457, November 2002. ISSN 1094-9224. doi: 10.1145/581271.581274. URL http://doi.acm.org/10.1145/581271.581274.

[28] Eurostat. Nearly 830 million air passengers in 2012. Eurostat, the statistical office of the European Union, 2013.

[29] Sheldon H Jacobson, Tamana Karnani, and John E Kobza. Assessing the impact of deterrence on aviation checked baggage screening strategies. *International Journal of Risk Assessment and Management*, 5(1):1–15, 2005.

[30] Sheldon H Jacobson, Tamana Karnani, John E Kobza, and Lynsey Ritchie. A cost-benefit analysis of alternative device configurations for aviation-checked baggage security screening. *Risk Analysis*, 26(2):297–310, 2006.

[31] James Chow, James Chiesa, Paul Dreyer, Mel Eisman, Theodore W Karasik, Joel Kvitky, Sherrill Lingel, David Ochmanek, and Chad Shirley. Protecting commercial aviation against the shoulder-fired missile threat. Technical report, RAND Corporation, 2005.

[32] ACI Europe. Revision of the 2005 ec guidelines on financing of airports and start-up aids to airlines departing from regional airports. Airports Council International Europe, 2012.

[33] ACI Europe. Aci europe position on the proposal for a directive on security charges (com (2009) 217). Airports Council International Europe, 2009.

[34] Peter Mell and Karen Scarfone. *A Complete Guide to the Common Vulnerability Scoring System Version 2.0*. 2007.

[35] Stephen D. Quinn, Karen A. Scarfone, Matthew Barrett, and Christopher S. Johnson. Sp 800-117. guide to adopting and using the security content automation protocol (scap) version 1.0. Technical report, 2010.

[36] PCI Council. Pci dss requirements and security assessment procedures, version 2.0., 2010. URL https://www.pcisecuritystandards.org/documents/pci_dss_v2.pdf.

[37] Stefan Frei, Martin May, Ulrich Fiedler, and Bernhard Plattner. Large-scale vulnerability analysis. In *Proceedings of the 2006 SIGCOMM workshop on Large-scale attack defense*, pages 131–138. ACM, 2006.

[38] Muhammad Shahzad, Muhammad Zubair Shafiq, and Alex X. Liu. A large scale exploratory analysis of software vulnerability life cycles. In *Proceedings of the 34th International Conference on Software Engineering*, pages 771–781. IEEE Press, 2012.

[39] Siv Hilde Houmb, Virginia NL Franqueira, and Erlend A Engum. Quantifying security risk level from cvss estimates of frequency and impact. 83(9):1622–1634, 2010.

[40] C. Miller. The legitimate vulnerability market: Inside the secretive world of 0-day exploit sales. In *Proceedings of the 6th Workshop on Economics and Information Security*, 2007.

[41] Guido Schryen. A comprehensive and comparative analysis of the patching behavior of open source and closed source software vendors. In *Proceedings of the 2009 Fifth International Conference on IT Security Incident Management and IT Forensics*, IMF '09, pages 153–168, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3807-5. doi: 10.1109/IMF.2009.15. URL http://dx.doi.org/10.1109/IMF.2009.15.

[42] Mehran Bozorgi, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Beyond heuristics: learning to classify vulnerabilities and predict exploits. In *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining*, pages 105–114. ACM, 2010.

[43] Fabio Massacci, Stephan Neuhaus, and Viet Nguyen. After-life vulnerabilities: A study on firefox evolution, its vulnerabilities, and fixes. In *Proceedings of the 2011 Engineering Secure Software and Systems Conference (ESSoS'11)*, Lecture Notes in Computer Science, pages 195–208, 2011.

[44] Karen Scarfone and Peter Mell. An analysis of cvss version 2 vulnerability scoring. In *Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 516–525, 2009.

[45] Vadim Kotov and Fabio Massacci. Anatomy of exploit kits. preliminary analysis of exploit kits as software artefacts. In *Proc. of ESSoS 2013*, 2013.

[46] Chris Grier, Lucas Ballard, Juan Caballero, Neha Chachra, Christian J. Dietrich, Kirill Levchenko, Panayiotis Mavrommatis, Damon McCoy, Antonio Nappa, Andreas Pitsillidis, Niels Provos, M. Zubair Rafique, Moheeb Abu Rajab, Christian Rossow, Kurt Thomas, Vern Paxson, Stefan Savage, and Geoffrey M. Voelker. Manufacturing compromise: the emergence of exploit-as-a-service. In *Proceedings of the 19th ACM Conference on Computer and Communications Security*, pages 821–832. ACM, 2012.

[47] Tudor Dumitras and Darren Shou. Toward a standard benchmark for computer security research: The worldwide intelligence network environment (wine). In *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pages 89–96. ACM, 2011.

[48] Branden R Williams and Anton Chuvakin. *PCI Compliance: Understand and implement effective PCI data security standard compliance*. Syngress Elsevier, 2012.

[49] Richard Doll and A Bradford Hill. Smoking and carcinoma of the lung. *British Medical Journal*, 2(4682):739–748, 1950.

[50] Tudor Dumitras and Petros Efstathopoulos. Ask wine: are we safer today? evaluating operating system security through big data analysis. In *Proceeding of the 2012 USENIX Workshop on Large-Scale Exploits and Emergent Threats*, LEET'12, pages 11–11, 2012.

[51] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org. ISBN 3-900051-07-0.

[52] L. Evans. The effectiveness of safety belts in preventing fatalities. *Accident Analysis & Prevention*, 18(3):229–241, 1986.

[53] J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. 310(6973):170, 1995.

[54] Dewayne E. Perry, Adam A. Porter, and Lawrence G. Votta. Empirical studies of software engineering: a roadmap. In *Proceedings of the 22nd Conference on The Future of Software Engineering*, pages 345–355. ACM, 2000.

[55] Luca Allodi, Vadim Kotov, and Fabio Massacci. Malwarelab: Experimentation with cybercrime attack tools. In *Proceedings of the 2013 6th Workshop on Cybersecurity Security and Test*, 2013.

[56] Steve Christey and Brian Martin. Buying into the bias: why vulnerability statistics suck. https://www.blackhat.com/us-13/archives.html#Martin, July 2013.

[57] Luca Allodi and Fabio Massacci. A preliminary analysis of vulnerability scores for attacks in wild. In *Proceedings of the 2012 ACM CCS Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 2012.

[58] Pratyusa K. Manadhata and Jeannette M. Wing. An attack surface metric. *IEEE Transactions on Software Engineering*, 37:371–386, 2011. ISSN 0098-5589. doi: http://doi.ieeecomputersociety.org/10.1109/TSE.2010.60.

[59] Lingyu Wang, Tania Islam, Tao Long, Anoop Singhal, and Sushil Jajodia. An attack graph-based probabilistic security metric. In *Proceedings of the 22nd IFIP WG 11.3 Working Conference on Data and Applications Security*, volume 5094 of *Lecture Notes in Computer Science*, pages 283–296. Springer Berlin / Heidelberg, 2008.

[60] L. Gallon. Vulnerability discrimination using cvss framework. In *Proceedings of the 4th IFIP International Conference on New Technologies, Mobility and Security*, pages 1–6, 2011.

[61] Yonghee Shin and Laurie Williams. Can traditional fault prediction models be used for vulnerability prediction? *Empirical Software Engineering*, 18(1):25–59, 2013. ISSN 1382-3256. doi: 10.1007/s10664-011-9190-8. URL http://dx.doi.org/10.1007/s10664-011-9190-8.

[62] Michael Gegick, Pete Rotella, and Laurie A. Williams. Predicting attack-prone components. In *Proceedings of the 2nd International Conference on Software Testing Verification and Validation (ICST'09)*, pages 181–190, 2009.

[63] Stephan Neuhaus, Thomas Zimmermann, Christian Holler, and Andreas Zeller. Predicting vulnerable software components. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, pages 529–540, 2007.

[64] O.H. Alhazmi and Y.K. Malaiya. Application of vulnerability discovery models to major operating systems. *IEEE Transactions on Reliability*, 57(1):14 –22, march 2008. ISSN 0018-9529. doi: 10.1109/TR.2008.916872.

[65] Andy Ozment. Improving vulnerability discovery models. In *Proceedings of the 3rd Workshop on Quality of Protection*, pages 6–11, 2007.

[66] Fabio Massacci and Viet Nguyen. An independent validation of vulnerability discovery models. In *Proceeding of the 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS'12)*, 2012.

[67] A. Ozment. The likelihood of vulnerability rediscovery and the social utility of vulnerability hunting. In *Proceedings of the 4th Workshop on Economics and Information Security*, 2005.

[68] Sandy Clark, Stefan Frei, Matt Blaze, and Jonathan Smith. Familiarity breeds contempt: the honeymoon effect and the role of legacy code in zero-day vulnerabilities. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 251–260, 2010. URL http://doi.acm.org/10.1145/1920261.1920299.

[69] C. Herley and D. Florencio. Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy. *Economics of Information Security and Privacy*, 2010.

[70] K. Labunets, F. Massacci, F. Paci, and L. M. Tran. An experimental comparison of two risk-based security methods. In *Proc. of ESEM '13*, pages 163–172, 2013.

[71] Andreas L. Opdahl and Guttorm Sindre. Experimental comparison of attack trees and misuse cases for security threat identification. *Inf. Soft. Technology*, 51(5):916–932, 2009.

[72] Riccardo Scandariato, Kim Wuyts, and Wouter Joosen. A descriptive study of microsoft's threat modeling technique. *REJ*, pages 1–18, 2014.

[73] Mass Soldal Lund, Bjornar Solhaug, and Ketil Stolen. A guided tour of the coras method. In *Model-Driven Risk Analysis*, pages 23–43. Springer, 2011.

[74] Daniel Mellado, Eduardo Fernández-Medina, and Mario Piattini. Applying a security requirements engineering process. In *Proc. of ESORICS '06*, pages 192–206. Springer, 2006.

[75] Tor Stålhane and Guttorm Sindre. Identifying safety hazards: An experimental comparison of system diagrams and textual use cases. In *Proc. BPMDS '12*, volume 113, pages 378–392, 2012.

[76] Tor Stålhane and Guttorm Sindre. Safety hazard identification by misuse cases: Experimental comparison of text and diagrams. In *Proc. MODELS '08*, pages 721–735, 2008. URL http://dx.doi.org/10.1007/978-3-540-87875-9_50.

[77] Tor Stålhane, Guttorm Sindre, and Lydie Bousquet. Comparing safety analysis based on sequence diagrams and textual use cases. In *Proc. CAISE '10*, volume 6051, pages 165–179, 2010.

[78] Peter Karpati, Yonathan Redda, Andreas L. Opdahl, and Guttorm Sindre. Comparing attack trees and misuse cases in an industrial setting. *Inf. Soft. Technology*, 56(3):294 – 308, 2014.

[79] Daniel L. Moody. The method evaluation model: a theoretical model for validating information systems design methods. In *Proc. of ECIS '03*, pages 1327–1336, 2003.

[80] Tor Stålhane and Guttorm Sindre. A comparison of two approaches to safety analysis based on use cases. In *Proc. of ER '07*, volume 4801, pages 423–437, 2007.

[81] Fabio Massacci and Federica Paci. How to select a security requirements method? A comparative study with students and practitioners. In *Proc. of NordSec '12*, pages 89–104. Springer, 2012.

[82] Claes Wohlin, Per Runeson, Martin Hst, Magnus C Ohlsson, Bjrn Regnell, and Anders Wessln. *Experimentation in software engineering*. Springer, 2012.

[83] ISO/IEC. *31000:2009 – Risk Management*. 2009.

[84] *EATM, ATM Security Risk Assessment Methodology, Edition 1.0*. EUROCONTROL, May 2008.

[85] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, pages 319–340, 1989.

[86] UNITN. Experiment website. http://securitylab.disi.unitn.it/doku.php?id=seceng-course-exp-2013.

[87] Thom Baguley. Calculating and graphing within-subject confidence intervals for anova. *Behavior research methods*, 44(1):158–175, 2012.

[88] Anselm L. Strauss and Juliet M. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, 1998. ISBN 0803959400.

ANNEX1. Regulations and Security Expenditure in the Aviation Network

Following the September 11 attacks, various security regulations that require increased expenditures have been enacted in the aviation industry. These regulations have different regulatory mechanisms (i.e., customized vs. one-size-fits-all) and financing rules (i.e., centralization vs. decentralization). While different countries use different financing and regulatory rules, the trade-off between these rules has not been studied well in the field of aviation security research. For example, while we found various industry reports arguing that the government should avoid setting and imposing one-size-fits-all regulation, and try to mirror the particularities and characteristics of different airports in setting a regulation, we presently do not have much theoretical and empirical evidence supporting this argument. Therefore, in this study, we take a fresh look at the different financing and regulatory structures in the field of aviation security. Particularly, a model with two airports, their security expenditures and externality is developed. With the model, we first theoretically characterize and evaluate the performance of centralized and decentralized financing systems, and customized and one-size-fits-all regulatory approaches, and compare the trade-offs between these mechanisms. We then graphically illustrate how the relative performance of these mechanisms changes with the different characteristics of the aviation network and externality.

ANNEX1.1 Introduction

Regulators designing and implementing security regulations on airports always face difficult problems. They believe that strong measures which may require significant investments are essential to address the threats posed by international terrorism and to restore public confidence in the aviation security [13]. In order to prevent and reduce security risks, they try to develop regulatory rules that can induce the appropriate level of security expenditures by airport operators. These might be customized rules that reflect the characteristics of each airport or uniform rules that can be generally applied to all airports. It seems however that these rules might not align well with the operators' incentive since necessary action imposed by these security rules might only favor some of the airports or might not reflect the differences in each airport.

Accordingly, as security regulations on the aviation industry have intensively tightened in recent years, various questions regarding the effectiveness of the regulations in reducing security risks have arisen. For example, in the previous literature, some authors have questioned regarding the effectiveness of the reform on regulatory rules, and generally concluded that the regulator's passion for making regulatory rules tighter does not align well with the interest of airport operators [14].

On the other hand, for airport operators, determining the optimal level of security investment has become a major task. After the events of September 11, 2001, security costs now represent up to 35% of overall airport operating costs [13]. They need to decide the best mechanism for their resource allocation in compliance with regulatory standards. For example, airport operators want to avoid a costly security accident with minimum resources while according with regulatory requirements. However, some authors have recently pointed out that the optimal investment and resource allocation are likely to vary across airports (e.g.,

[15]). Each airport might have different security preference, and would want to allocate its resources differently from other airports.

While a fully tailored security rule might maximize social surplus, the regulator might not be able to explore all the possible regulatory options since it will be too costly to him. Even if he can investigate all the possible options, it would be very difficult for them to determine the correct enforcement level due to limited information. As a result, the regulator faces to choose a regulation, including setting the level of security expenditures and imposing different types of security charges, based on his limited knowledge and it might result in suboptimal outcome. Furthermore, the externality between airports might make this problem more severe.

Our study therefore aims at addressing directly to this issue and investigating how best the regulator can design security regulations. In detail, we consider different regulatory and financing rules and analyze the likely consequences of such rules. As for the financing rules, we consider how the security expenditures imposed by the government are financed: centralized and decentralized financing systems. Regarding the regulatory rules, we take into account different types of enforcement on airports' security spending: customized and uniform levels. By using the combinations of financing and regulatory rules, we investigate whether a specific rule is appropriate for minimizing the expected losses from a terrorist's successful attack and derive the conditions that the rules can improve the social surplus. We show that, without a proper combination of the rules and prudent consideration on the characteristics of airports, the regulator might in fact generate a worse global outcome.

The questions to be addressed in this study therefore are:

1. What are the optimal security expenditures for different financing and regulatory rules?
2. How to design a rule that is close to the social optimum for airport security?
3. Is a current security rule effective for induce efficient security expenditures? Does it incur a better global outcome than do other rules?
4. What are the impacts of different forms of security regulations on different types of airports (e.g., big and small, hub & spoke, and point to point)?

The remainder of the study is organized as follows. As background, Section 2 reviews the previous literature. Section 3 discusses a theoretic model that shows the optimal security expenditures and the impact of a regulation on such expenditures. Based on the model, Section 4 provides graphical illustration for different regulatory and financing settings. Finally, Section 5 offers concluding remarks and further discussion.

ANNEX1.2 Literature Review

There has been a wide array of studies that investigate the pros and cons of different financing systems. Since Oates' work on fiscal decentralization [16], many studies have tried to explain why decentralization can be more efficient than centralization, or vice versa (e.g., [17, 18, 19]). The crucial feature of these studies is that the optimal financing structure for public good provision brings about a potential tradeoff: for example, while a decentralized system can enjoy the benefits from reflecting diverse preferences for public good provision, it would induce the costs since it cannot enjoy economies of scale and internalize

spillovers. Similarly, while a centralized provision of a uniform supply of public goods might be able to internalize spillovers, it would experience the coordination failure that causes costs for those whose preferences are not taken into account. As a result, they concluded that (de)centralization should only take place once the costs imposed by a (de)centralized provision of public goods are outweighed by the benefits coming from a (de)centralized provision.

While these studies are well suited to explain the costs of (de)centralization in policy domains where public goods cannot be differentiated according to the preferences of localities, in many cases, it is possible to decide centrally on geographically differentiated levels of public good in line with the diverse regional preferences and cultures. This opens up to redistribution games among regions to gather in a larger share of central spending. However, these studies only considered public good provisions in policy domains, and assumed that all players in the model as social welfare maximizers. While security can be considered as a public good, the models developed in these studies cannot be directly applied to a security study for several reasons. First, many organizations, including airports, which provide security are profit (or utility) maximizers rather than social welfare maximizers. Particularly, in the aviation industry, as noted in [20, 21], many airports are public-owned or corporatised. Second, the models in the previous studies has mainly focused on financing mechanisms for public good provisions without considering regulatory systems. In this study, we therefore expand the model developed in the previous literature and examine the effect of different regulatory mechanisms as well as financing mechanisms in security provisions.

In the field of aviation security which is the main focus of this study, while relatively little studies have been conducted regarding the effectiveness of particular financing and regulatory rules, there have been extensive industry and government reports that explain the current financing and regulatory structures in various countries and try to offer some best direction forward. One example is the report published by Irish Aviation Authority & Aviasolutions [22]. The report illustrated the current status of regulatory and financing structure related to aviation security in European countries. It provided the detailed information on how aviation security is administered and funded in each European country. Another example is the reports published by ACI including [23, 24]. The reports argued that one-size-fits-all regulatory approach should be avoided since airports are affected by different factors including size, role and needs. The reports further argued that a regulation should be sufficiently flexible to make it possible to consider different types of airports.

However, all of these reports did not provide concrete evidence why some regulatory and financing mechanisms are better than others and how the country uses the combination of the regulatory and financing mechanisms to obtain a better outcome. The main contribution of this study is therefore to provide a way to identify optimal combination of the regulatory and financing mechanisms which can produce the best outcome under the given condition.

ANNEX1.3 The Model

In this section, a simple model is introduced with two distinct airports, indexed by $i \in \{1, 2\}$. Each airport makes a particular security spending, x_i, per passenger. This can be thought of as costs for security training or maintaining security devices. When the number of passengers in airport i is n_i, therefore, the total security expenditures can be given as $n_i x_i$. Each

airport is characterized by a subjective security preference, η_i; The larger is an airport's η, the stronger its preference for security protection.[2] The parameter $\kappa \in [0, 1]$ denotes the degree of externality; security spending in one airport has externality effects on the expected loss in the other airport. If $\kappa = 0$, the externality effect is absent: airport i is not affected by the security condition in airport $-i$. The larger is κ, the larger is the externality effect. If $\kappa = 1$, airport security is equally determined by security spending in both airports.[3]

It is considered that security expenditures are financed by head taxes assessed against passengers.[4] These taxes might be levied by two different ways: centralized and decentralized financing systems. A decentralized system refers that security expenditures are funded by airport charges from passengers (hereinafter referred to as "airport charges"). On the other hand, a centralized system means that security spending is funded by the government which charges from passengers (hereinafter referred to as "state charges"). Different countries use different systems; countries use airport charges, state charges, or both [22]. In the case where there is only "airport charges", the government determines the head tax rate, p, and airport i can charge the total of pn_ix_i from the passengers to cover some of its security expenses. If the government uses only "state charges", the security expenditures are being financed by a tax charged from all passengers in both airports and shared among the airports through a common budget, i.e., $p(\sum_i n_i x_i)/\sum_i 1$. We further assume that, if the government uses both "airport charges" and "state charges", q portion of the security expenditures is paid by "airport charges" and $(1-q)$ portion is paid by "state charges". In this case, airport i's total charges can be denoted as $qpn_ix_i + (1-q)p(\sum_i n_i x_i)/\sum_i 1$.

The probability of a successful attack, $\tilde{\sigma}$, is assumed to be determined by security expenditures per passenger x_i, i.e., $\tilde{\sigma}_i(x_i)$.[5] As is common with most of the economic models, we presume that the function $\tilde{\sigma}_i$ is continuously twice differentiable. Particularly, for all x_i, $\partial\tilde{\sigma}_i/\partial x_i < 0$ and $\partial^2\tilde{\sigma}_i/\partial x_i^2 > 0$ are assumed. The notations mentioned above are summarized in Table 2.

Table 2: Description of Variables

Variable	Description
x_i	Security expenditure per passenger in airport i.
n_i	Number of passengers of airport i.
$\tilde{\sigma}_i$	Probability of a successful attack for airport i.
η_i	Airport i's security preference.
κ	The degree of security externality.
L_i	Airport i's loss from a successful attack.
p	Head tax rate
q	The portion of airport charges

Since there are several configurations regarding airport ownership in the aviation net-

[2] This preference can be regarded as attacking intensity used in [25] and [26], or vulnerability used in [27].

[3] In public economics, a public good with $\kappa = 1$ is referred to as a "global" public goods [18].

[4] To finance airport security expenditures, the airport or the national authorities can levy a tax, a fee or a charge on airlines, passengers and cargo shippers. However, the major funder of these expenditures is passengers [13].

[5] For simplicity, in this section, we limit ourselves to this parameter. Other parameters including the effectiveness of a security technology can be easily included in the probability function as shown in [27].

work, we need to consider different expected loss functions that are affected by these ownerships. We first consider the case where there are only government-owned airports (hereinafter referred to as "public airports"). We then expand our consideration to the cases where there are privately-owned airports (hereinafter referred to as "private airports"), and airports with the mixed ownership (i.e., one private airport and one public airport).

In each of the following subsections, we first derive socially optimal security expenditures for a specific airport network configuration. The socially optimal expenditures can be defined as the outcomes that minimizes the sum of expected losses from a successful attack in both airports. These optimal expenditures serve as criteria against which to examine the performance of different regulatory and funding structures. While these can be obtained from a social surplus maximizing (or social loss minimizing) function, in reality, the regulator cannot set a regulatory rule which coincides perfectly with the solution of this function due to cost reasons and the lack of information. Therefore, we investigate a set of solutions of various expected functions derived from different regulatory and financing systems.

ANNEX1.3.1 Two Public Airport Case

According to the document published by European Commission [20], the majority of EU airports are still owned by the governments. Therefore, as a starting point, we begin our analysis by assuming the airport network with two public airports.

The Social Optimum

When there are only public airports, the socially optimal security expenditures can be identified by maximizing:

$$V_G = \sum_i \left[C_i + n_i x_i + T_i^a + T_i^s\right] \tag{1}$$

where

$$\begin{aligned} C_i &= L_i \eta_i \left[\tilde{\sigma}_i + \kappa \tilde{\sigma}_{-i}\right] \\ T_i^a &= qpn_i x_i \\ T_i^s &= (1-q)p(n_i x_i + n_{-i} x_{-i})/2. \end{aligned}$$

The social loss minimizing levels of security expenditures are described by the following first-order condition:

$$\tilde{\sigma}'(x_i^*) = \frac{-(1+p)n_i}{\left(L_i \eta_i + L_{-i} \eta_{-i} \kappa\right)}, \quad i \in 1, 2. \tag{2}$$

where x_i^* denotes the socially optimal level of security expenditures. It is clear from (2) that the socially optimal security expenditures x_i increases in airport i's losses from a successful attack $L_{\{i,-i\}}$, security preference $\eta_{\{i,-i\}}$ and the externality κ, and decrease in the number of passengers n_i and the head tax rate p.

Customized Imposition of Security Expenditures

We now consider a case where the government enforces different levels of security expenditures on the airports. Under this type of decision making, the regulator first collects necessary information from the different airport operators and, based on this information,

determines the expenditure level for each airport (hereinafter referred to as a "*customized approach*"). However, we assume that he has only limited information and chooses the expenditure level based on each airport's expected loss function rather than an appropriately determined social loss function: that is, the regulator makes separate decisions on security expenditures for each airport. This can be considered as customized imposition of security expenditures with minimal government intervention. In this case, optimal security expenditure for airport i is found by minimizing:

$$V_{i,g}^{c} = C_i + n_i x_i + qpn_i x_i + (1-q)p(n_i x_i + n_{-i} x_{-i})/2 \quad for \quad i \in 1,2. \tag{3}$$

We first consider decentralized financing. Under a decentralized financing system, security expenditures are financed locally (i.e., $q = 1$). As a result, the first-order condition of (3) can be given by:

$$\tilde{\sigma}'(x_i) = \frac{-(1+p)n_i}{(L_i \eta_i)}, \quad i \in 1,2. \tag{4}$$

From Eq (4), it can be seen that security expenditures under decentralized financing increase in L_i and η_i, and decrease in n_i and p. In contrast to the socially optimal case, the security expenditures are independent of the externalities κ. Comparing Eq (2) with Eq (4), it can be identified that the customized expenditure under decentralized financing the socially optimal level in (2) when $\kappa = 0$. For $\kappa > 0$, however, decentralized financing makes airports underspend on security and this underspending is increasing in the extent of the externality. This implies that, if the regulator does not consider the externality effect of security expenditures on the airports, it makes the regulatory impose too low levels of security expenditures to the airports.

If the government uses a centralized financing system, security expenditures are financed through a common central budget (i.e., $q = 0$). The levels of security expenditures for minimizing the expected losses under a centralized financing system can therefore be found by the following first-order conditions:

$$\tilde{\sigma}'(x_i) = \frac{-(2+p)n_i}{2L_i \eta_i}, \quad i \in 1,2. \tag{5}$$

Comparing (5) with (2), it can be identified that when the airports are identical and the externality effect is maximal, the security expenditure in airport i is too low from the social perspective. In contrast, when the airports are identical and the externality effect is minimal, the security expenditure in airport i is too high from the social perspective. This implies that there is a critical level for the externality that can make the regulator impose the socially optimal security expenditures. As for a heterogeneous case, both airports are enforced to overspend on security if the externality is minimal. However, the effect of a centralized financing system becomes unclear if $\kappa > 0$.

Uniform Imposition of Security Expenditures

In order to avoid the problems caused by a customized approach, the government may want to use a different regulatory mechanism. If the government has enough information and resources, it might be able to identify the sum of expected losses identical to V_G given in (1), and to resolve the problem found in the customized approach.

However, since it would be very difficult and costly for the government to identify the sum of expected losses and set the levels of security expenditures based on this, the government

might have to use a different approach. We consider here a case where the government set the uniform level of security expenditures for both airports. In the field of public economics, it has been referred to as an "one-size-fits-all" approach. The objective function of the regulator using this approach is:

$$V_g^u = \sum_i [L_i \eta_i [\tilde{\sigma} + \kappa \tilde{\sigma}] + n_i x + qpn_i x + (1-q)p(n_i x + n_{-i} x)/2]. \quad (6)$$

Note that both centralized and decentralized financing systems have the identical objective function in this case. Therefore, the first-order condition for both centralized and decentralized financing systems can be defined as:

$$\tilde{\sigma}'(x) = \frac{-(1+p)(n_1+n_2)}{(L_1\eta_1+L_2\eta_2)(1+\kappa)}. \quad (7)$$

Unlike (4) and (5), the solution is now dependent on the level of the externality and equals the socially optimal level in (2) when the airports are homogeneous. However, when the airports are not identical, $\kappa = 1$ makes both airports 1 & 2 underspend on security. If $\kappa < 1$, it is undeterminable whether the imposed level results in overspending or underspending.

airport 1 underspends on security if $\kappa = 1$, but Therefore, it is clear from the results that, when the airports are identical, an "one-size-fits-all" approach dominates a "customized" approach whenever externality exists, whereas a "customized" approach with a decentralized financing system is superior when there is no externality and the airports are heterogeneous.

ANNEX1.3.2 Two Private Airport Case

We have witnessed that the number of private airports has increased [20]. Furthermore, even when an airport is owned by the government, it has begun to focus more on profitability similarly with their private counterparts.[6] Therefore, we also consider the case where there are two private airports. Since the results are very similar with the two public airport case, we limit ourselves to present the results briefly.

The Social Optimum

If an airport is privately-owned (or corporatised), it can be considered as a profit-maximizer and regards security charges as a part of its income. Therefore, the expected loss function of the airport has negative signs for security charges, T_i^a and T_i^s. The socially optimal security expenditure for private airports can be identified by minimizing:

$$V_P = \sum_i [C_i + n_i x_i - T_i^a - T_i^s] \quad (8)$$

yielding the following first-order conditions:

$$\tilde{\sigma}'(x_i) = \frac{-(1-p)n_i}{(L_i\eta_i + L_{-i}\eta_{-i}\kappa)}, \quad i \in 1, 2. \quad (9)$$

Similar interpretation with (2) can be given to (9) except for the head tax rate p: the increase in p now raises the level of security expenditure. Comparing (9) with (2), it is clear that the socially optimal expenditure level for the private airport case is higher than that for the public airport case.

[6] According to [21], in Europe, 80% of airports have been corporatised and they now operate as commercial entities.

Customized Imposition of Security Expenditures

A customized approach makes the regulator enforce security expenditure for each airport separately, and the levels of security expenditure of the airports are determined by their own expected loss functions. Therefore, optimal security expenditure for airport i under a customized approach is found by minimizing:

$$V_{i,p}^{c} = C_i + n_i x_i - qpn_i x_i - (1-q)p(n_i x_i + n_{-i} x_{-i})/2. \tag{10}$$

If the government uses a decentralized financing system, the security expenditure is financed locally (i.e., q is set to 1). Therefore, the first-order conditions of (10) under decentralized financing is given by:

$$\tilde{\sigma}'(x_i) = \frac{-(1-p)n_i}{(L_i \eta_i)}, \quad i \in 1, 2. \tag{11}$$

Eq (11) implies that, no matter whether the airports are identical or not, the imposed security expenditure levels for the airports coincide with the socially optimal expenditure levels when there is no interdependence in security (i.e., $\kappa = 0$). However, if $\kappa > 0$, a customized approach under decentralized financing always makes the regulator mandate security expenditure lower than the socially optimal level, and the gap increases as the externality rises.

If we assume that the government uses centralized financing (i.e., q is set to 0), the expected loss minimizing level for security expenditure can be found by the following first-order conditions:

$$\tilde{\sigma}'(x_i) = \frac{-(2-p)n_i}{2L_i \eta_i}, \quad i \in 1, 2. \tag{12}$$

By comparing (12) with (9), it is clear that the compulsory security expenditure based on a customized approach always makes both airports underspend on security from a social perspective for both identical and heterogeneous airport cases, and this underspending becomes more severe as the externality increases.

Furthermore, from (12) and (11), it can be identified that the customized approach with a decentralized financing system performs better than that with a centralized financing system regardless of homogeneity or heterogeneity of the airports.

Uniform Imposition of Security Expenditures

As in the public airport case, we analyse the case where the government uses an "one-size-fits-all" approach. The optimal solution can be identified by minimizing:

$$V_p^u = \textstyle\sum_i \left[L_i \eta_i \left[\tilde{\sigma} + \kappa \tilde{\sigma}\right] + n_i x - qpn_i x - (1-q)p(n_i x + n_{-i} x)/2\right]. \tag{13}$$

The first-order condition for both centralized and decentralized financing systems can be defined as:

$$\tilde{\sigma}'(x) = \frac{-(1-p)(n_1+n_2)}{(L_1\eta_1 + L_2\eta_2)(1+\kappa)}. \tag{14}$$

The above solution implies that, if the airports are identical, an "one-size-fits-all" approach can always impose socially optimal security expenditure regardless of the existence of interdependence. As a result, an "one-size-fits-all" approach dominates a "customized" approach when the airports are identical. If we assume that the airports are heterogeneous, airport 1 overspends and airport 2 underspends if $\kappa = 1$. However, the effect of the rule becomes unclear if $\kappa < 1$.

ANNEX1.3.3 One Private and One Public Airport Case

While the number of private (or corporatised) airports is increasing, they are likely to be relatively big size airports [20]. In contrast, there are still many public airports particularly when they are small. In this subsection, we consider a case where there are both private and public airports in the aviation network.

The Social Optimum

As mentioned before, we assume that a public airport is a social surplus maximizer and a private airport is a profit maximizer. Therefore, security charges T_i^a and T_i^s in the expected loss functions have the opposite signs: a private airport has a negative sign for security charges whereas a public airport has a positive sign for those. If we assume that airport 1 is privately-owned and airport 2 is publicly-owned, the social optimum can be found by minimizing:

$$V_{PG} = [C_1 + n_1 x_1 - T_1^a - T_1^s] + [C_2 + n_2 x_2 + T_2^a + T_2^s] \tag{15}$$

yielding the following first-order conditions:

$$\begin{aligned} \tilde{\sigma}'(x_1) &= \frac{-(1-qp)n_1}{(L_1\eta_1 + L_2\eta_2\kappa)} \\ \tilde{\sigma}'(x_2) &= \frac{-(1+qp)n_2}{(L_2\eta_2 + L_1\eta_1\kappa)} \end{aligned} \tag{16}$$

Eq (16) indicates that the optimal expenditure levels are now affected by the portion of the security expenditures, q. Similar interpretation with the previous solutions can be made: the increase in q or p raises the level of security expenditure for airport 1, whereas the increase in such values results in the decrease in the security expenditure level for airport 2.

Imposing Different Levels of Security Expenditures

By considering two different types of airport ownerships, we now need to consider two different expected loss functions in analyzing a customized approach. Optimal security expenditure for private airport 1 and public airport 2 under a customized approach can be identified by minimizing:

$$\begin{aligned} V_{1,p}^c &= C_1 + n_1 x_1 - qpn_1 x_1 - (1-q)p(n_1 x_1 + n_2 x_2)/2 \\ V_{2,g}^c &= C_2 + n_2 x_2 + qpn_2 x_2 + (1-q)p(n_2 x_2 + n_1 x_1)/2 \end{aligned} \tag{17}$$

We first assume that the government uses a decentralized financing system (i.e., $q = 1$). The first-order conditions for (17) under decentralized financing are given by:

$$\begin{aligned} \tilde{\sigma}'(x_1) &= \frac{-(1-p)n_1}{(L_1\eta_1)} \\ \tilde{\sigma}'(x_2) &= \frac{-(1+p)n_2}{(L_2\eta_2)} \end{aligned} \tag{18}$$

From (16) and (18), it can be found that if the airports are homogeneous, the imposition makes the airports spend the socially optimal levels for security if the externality effect is zero. However, if the externality effect exists, customized imposition with decentralized financing makes both airport underspend on security, and the underspending increases as the externality level rises. Even for the heterogeneous airport case, the same result can be

identified: while the regulator can impose the socially optimal expenditure levels to airports 1 and 2 if there is no externality, the imposed expenditure levels for airports 1 and 2 enforce them to underspend on security as long as the externality exists.

If the government uses centralized financing (i.e., $q = 0$), the expected loss minimizing levels for security expenditure can be found by the following first-order conditions:

$$\begin{aligned} \tilde{\sigma}'(x_1) &= \frac{-(2-p)n_1}{2L_1\eta_1} \\ \tilde{\sigma}'(x_2) &= \frac{-(2+p)n_2}{2L_2\eta_2} \end{aligned} \tag{19}$$

The conditions implies that, for the identical airport case, this type of regulatory and financing rules makes airport 1 overspend if $\kappa = 0$ and underspend if $\kappa = 1$. Therefore, there is a critical level for the externality that makes the rule socially optimal. However, for airport 2, the imposition makes the airport always underspend on security and the level becomes higher as the externality increases. If we consider the heterogeneous airport case, airport 1 is enforced to overspend on security if $\kappa = 0$, while the effect of the imposition is unclear if $\kappa > 0$. Airport 2 in this case is always forced to overspend on security and the overspending level increases as the externality rises.

Imposing a Uniform Level of Security Expenditure

If the government employs a "one-size-fits-all" regulation, the optimal solution can be identified by setting $x = x_i = x_{-i}$ in (15). If should be noted that in this case the expected loss functions for centralized and decentralized financing are not identical. The first-order condition under decentralized financing (i.e., $q = 1$) can be given by:

$$\tilde{\sigma}'(x) = \frac{-(n_1+n_2)+p(n_1-n_2)}{(L_1\eta_1+L_2\eta_2)(1+\kappa)}. \tag{20}$$

For the identical airport case, this type of rules makes airport 1 always underspend and airport 2 always overspend regardless of the level of externality. If we assume that there are heterogeneous airports, the effect of the rule becomes unclear for both airports 1 and 2.

On the other hand, the first-order condition for centralized financing (i.e., $q = 0$) is:

$$\tilde{\sigma}'(x) = \frac{-(n_1+n_2)}{(L_1\eta_1+L_2\eta_2)(1+\kappa)}. \tag{21}$$

If the airports are identical, this rule can produce the socially optimal levels of security expenditure for both airports 1 and 2 irrespective of the externality level. As for the heterogeneous airport case, airport 1 is always forced to overspend on security and the level of overspending increases as the externality rises. On the other hand, airport 2 is made to underspend on security if $\kappa = 1$, whereas the effect of the imposition cannot be determined if $\kappa < 1$.

ANNEX1.3.4 Optimal Financing and Regulatory Rules

In the previous sections, we showed that, while some combination of financing and regulatory structures can induce the socially optimal levels of security expenditure, others might cause an underspending or overspending problem. These results are summarized in Tables 3 and 4. As can be seen in the tables, if there is externality, none of the solutions can induce socially optimal outcomes. Furthermore, as the externality increases, some of the models

start to produce the worse outcomes. As a result, to bring these financing and regulatory structures closer to the social optimal values, the regulator might want to use the combination of these structures. For example, the regulator can use the combination of centralized and decentralized financing systems by setting q between 0 and 1. On the other hand, he might change a regulatory rule from a uniform approach to a customized approach, or vice versa, with q fixed to some value.

While the relative performance of these combinations might provide socially better outcomes, it would be difficult to theoretically compare the performance of the different combinations with different aviation network settings. As a result, in the following sections, we present graphical illustration using simple functional forms with specific scenarios.

Table 3: Comparison of Optimal Security Expenditure Levels: Private or Public Airports

		Customized & Decentralized	Customized & Centralized	Uniform
Two Public Airports	Identical	$\kappa = 0$: Socially optimal. $\kappa > 0$: Underspending & becomes severe as κ increases	Overspending to underspending as κ increases	Always socially optimal.
	Not Identical	$\kappa = 0$: Socially optimal. $\kappa > 0$: Underspending & becomes severe as κ increases	$\kappa = 0$: Overspending. Unclear if $\kappa > 0$	$\kappa = 1$: Underspending. Unclear if $\kappa < 1$
Two Private Airports	Identical	$\kappa = 0$: Socially optimal. $\kappa > 0$: Underspending & becomes severe as κ increases	Always underspending	Always socially optimal
	Not Identical	$\kappa = 0$: Socially optimal. $\kappa > 0$: Underspending & becomes severe as κ increases	Always underspending	$\kappa = 1$: Airport1: Overspending. Airport2: Underspending. $\kappa < 1$: Unclear

ANNEX1.4 Graphical Illustration

While useful, the analysis in the previous sections cannot sufficiently illustrate the interaction between the externality, regulatory rules and financing systems. For example, the analysis cannot provide information on how the proportion of airport charges (i.e., q) affects the levels of mandated security expenditure. Moreover, it cannot show the relative performance and outcomes from different regulatory and financing structures. Therefore, in this section, we explore the results in a fully specified setup. For an illustrative purpose, we use the simple

Table 4: Comparison of Optimal Security Expenditure Levels: Mixed Airports

		Customized & Decentralized	Customized & Centralized	Uniform & Dencentralized	Uniform & Centralized
One Private and One Public Airports	Identical	κ = 0: Socially optimal. $\kappa > 0$: Underspending & becomes severe as κ increases	κ = 0: Airport 1: Overspending κ = 1: Airport 1: Underspending. Airport 2: Always underspending	Airport 1: Always underspend. Airport 2: Always overspend.	Always socially optimal
	Not Identical	κ = 0: Socially optimal. $\kappa > 0$: Underspending & becomes severe as κ increases	κ = 0: Airport 1: Overspending $\kappa > 0$: Airport 1: Unclear. Airport 2: Always overspending	Unclear (depends on *p*)	Airport 1: Always overspending & becomes more severe as κ increases. Airport2: Underspending if κ = 1, unclear if $\kappa < 1$

functional forms of $\tilde{\sigma}_i$ as e^{-x_i} and η_i as $(1-e^{-n_i})$. Table 5 shows optimal solutions for different regulatory settings. It can be clearly seen that the levels are affected by p, q, n_i and κ. In the followings, we compare the relative performance of different settings by giving numerical values to parameters for specific scenarios.

Table 5: Optimal Solutions for Different Regulatory Settings

	Social Optimum
$x^*_{(i,g)}$	$Log\left[\frac{\left(1-e^{-n_i}\right)L_i+\left(1-e^{-n_{-i}}\right)\kappa L_{-i}}{(1+p)n_i}\right], i\in\{1,2\}$
$x^*_{(i,p)}$	$Log\left[\frac{\left(1-e^{-n_i}\right)L_i+\left(1-e^{-n_{-i}}\right)\kappa L_{-i}}{(1-p)n_i}\right], i\in\{1,2\}$
$\left(x^*_{(1,p)}, x^*_{(2,g)}\right)$	$Log\left[\frac{\left(1-e^{-n_1}\right)L_1+\left(1-e^{-n_2}\right)\kappa L_2}{(1-pq)n_1}\right]$, $Log\left[\frac{\left(1-e^{-n_2}\right)L_2+\left(1-e^{-n_1}\right)\kappa L_1}{(1+pq)n_2}\right]$
	Customized Intervention
$x^c_{(i,g)}$	$Log\left[\frac{2e^{-n_i}\left(-1+e^{n_i}\right)L_i}{(2+p+pq)n_i}\right], i\in\{1,2\}$
$x^c_{(i,p)}$	$Log\left[\frac{2e^{-n_i}\left(-1+e^{n_i}\right)L_i}{(2-p-pq)n_i}\right], i\in\{1,2\}$
$\left(x^c_{(1,p)}, x^c_{(2,g)}\right)$	$Log\left[\frac{2e^{-n_i}\left(-1+e^{n_i}\right)L_i}{(2-p-pq)n_i}\right]$, $Log\left[\frac{2e^{-n_i}\left(-1+e^{n_i}\right)L_i}{(2+p+pq)n_i}\right]$
	One-Size-Fits-All Intervention
$\left(x^u_g\right)$	$\left(Log\left[\frac{(1+\kappa)\left[e^{n_2}\left(-1+e^{n_1}\right)L_1+e^{n_1}\left(-1+e^{n_2}\right)L_2\right]}{(1+p)(n_1+n_2)}\right]-n_1-n_2\right)$
$\left(x^u_p\right)$	$\left(Log\left[\frac{(1+\kappa)\left[e^{n_2}L_1+e^{n_1}\left(L_2-e^{n_2}(L_1+L_2)\right)\right]}{(-1+p)(n_1+n_2)}\right]-n_1-n_2\right)$
$\left(x^u_{pg}\right)$	$\left(Log\left[\frac{(1+\kappa)\left(e^{n_2}\left(-1+e^{n_1}\right)L_1+e^{n_1}\left(-1+e^{n_2}\right)L_2\right)}{n_1-pqn_1+n_2+pqn_2}\right]\right)$

ANNEX1.4.1 Value Assumptions

In order to illustrate the effect of different financing and regulatory rules, we start the analysis by giving the likely values for the parameters so that the models developed in the previous section can be calibrated. The assumptions for the values of the parameters will provide numerical results which make it possible for us to compare the optimal regulatory settings and social welfare under different conditions. Before providing any assumptions regarding the values, it should be noted that we consider three different ownership cases in this section: the first case considers two public airports with different sizes. We can think of this as a case where there are one big and one medium-sized public airports operating in a country. The second is the counterpart of the first case: two private airports with different sizes. Lastly, the third case presumes that there are one big private airport and one medium-sized public airport. We particularly select the third case since this is a dominant feature in the aviation industry in Europe [21]. The assumptions we made for the values of the parameters are as follows:

- **Big private/pubic airport:** We set the number of passengers n_i in a big airport as 45

million. This number comes from the calculation of average number of passengers in the 10 biggest European airports in 2012 [28]. Regarding the loss from a successful attack, the estimations of it were very different depending on the studies. For example, in the studies of Jacobson and his colleagues, they used $1.4 billion [29] and $30 billion [30]. Chow et al. [31] estimated that the loss can range from $1.4 billion up to $70.7 billion or more since a successful attack can cause not only damage on an airport and an airplane but also loss of life, damage on infrastructure and huge undesirable impacts on the society and economy. We conservatively set the loss to 30 billion.

- **Medium-sized private/public airport:** We define a medium-sized airport as a regional (or international) airport which is used mostly by small airplanes that go to the national hubs or the international hubs. Due to the rapid growth of low-cost airlines, the presence of this type of airports is likely to become more important. While there is various available information on the busiest and largest airports, however, such information cannot be found easily for medium-sized airports. Therefore, we assume that a medium-sized airport serves 15 million passengers. Furthermore, we assume that the loss caused by a successful attack through a small airport result in 15 billion. This is because of the fact that damage on the economy and society in this case would be smaller than the case when there is a successful attack on a big airport.

- **Other parameter values** We further assume the value of the head tax rate, p. If the head tax rate is 1.0, it means an airport recover all of its security expenditures from the security charges. According to [32, 33], however, it is unlikely that an airport can recover its security expenses from the security charges. As a result, we set p to 0.9.

We believe that the changes in the parameters only affects the responsiveness of the relationships, but not the main conclusions. Therefore, the illustrations can provide one with general ideas on the relative performance of different regulatory and financing settings and the relationship between the parameters and the social losses.

ANNEX1.4.2 Case 1: Different Financing Structures

As indicated in [22, 33], countries uses either airport charges or state charges, or both charges. However, there is likely to be no clear evidence why a country selects different charges. Therefore, we now consider a scenario that the regulator with different regulatory mechanisms faces to determine the level of state security charges. We provide a graphical illustration that shows whether the regulator's attempt to increase state charges can improve the social welfare. More specifically, we analyze the case where the regulator tries to determine the appropriate level of tax quota, q, based on the country's regulatory approach and the level of externality. For example, the government using a one-size-fits-all or customized security regulation might change q in order to minimize the social loss and maximize the social surplus that depends on its regulatory rule. Therefore, the investigation in this section will answer the question "what would be the best q when the government uses a specific regulatory setting?"

For our starting graphical illustration, we first assume that there are two heterogeneous public airports. Figure 3 represents the relative performance of q with different levels of externality. Relative performance is measured by calculating the difference between aggregated

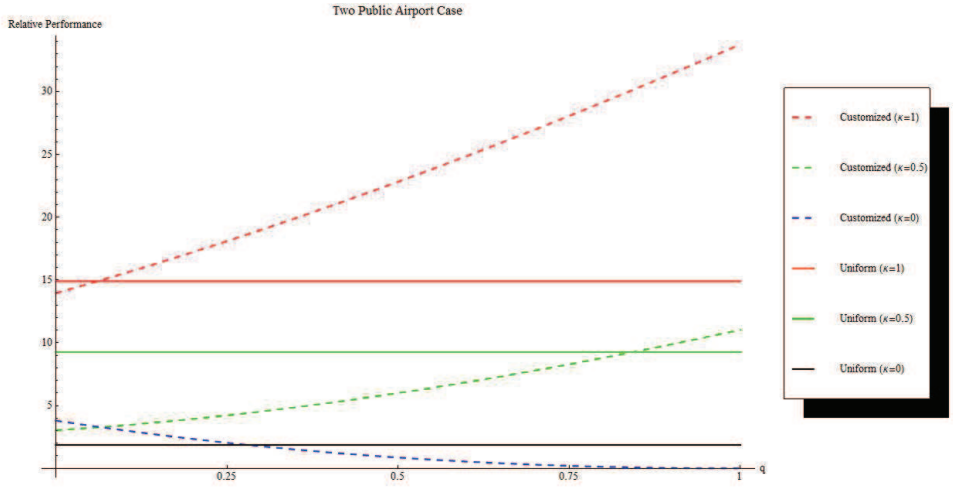


Figure 3: Relative Performance of Two Heterogeneous Public Airports with Changing q

expected loss from a specific regulatory and financing setting and socially optimal expected loss. For example, for a customized approach, the relative performance is calculated by:

$$V_G(x^c_{(1,g)}, x^c_{(2,g)}) - V_G(x^*_{(1,g)}, x^*_{(2,g)}). \tag{22}$$

From Figure 3, several interesting points can be identified. First, if the aviation network is comprised of two heterogeneous airports and the government uses a one-size-fits-all approach, it cannot improve the regulatory performance by altering the level of q. In addition, the performance of a one-size-fits-all approach becomes worse as the externality effect increases. Second, if the government relies on a customized regulatory approach, lowering the level of q gives a better outcome when the externality is between medium to high (i.e., $k = 0.5$ or 1), whereas raising q performs better when the externality is low (i.e., $k = 0$). This implies that there might be a critical value of externality (e.g., $0 < \kappa < 0.5$) that changes the effect of q.

We now consider the second case where there are two heterogeneous private airports. As can be seen in Figure 4, it can be identified that the performance of a one-size-fits-all approach for the different levels of externality is almost identical. Furthermore, a rule based on a one-size-fits-all approach can provide the outcome very close to the socially optimal level of expected loss. As for the customized regulatory rules, increasing q always results in better performance regardless of the externality levels. In addition, when q is set to maximal, the outcomes based on customization approach to the socially optimal outcomes: a financing system based on decentralization always performs better than centralization.

Figure 5 displays the case where the aviation network is composed of two heterogeneous private and public airports. Similarly with the previous case for two private airports, raising q always provides better performance for a regulation based on a customized approach: a decentralized financing system outperforms a centralized financing system. On the other hand, for a one-size-fits-all regulatory rule, increasing q might provide a worse outcome: there is a specific value of q for each externality level that approaches to a socially optimal level. If q is set to the values close to these, the social optimum can be achieved. As q

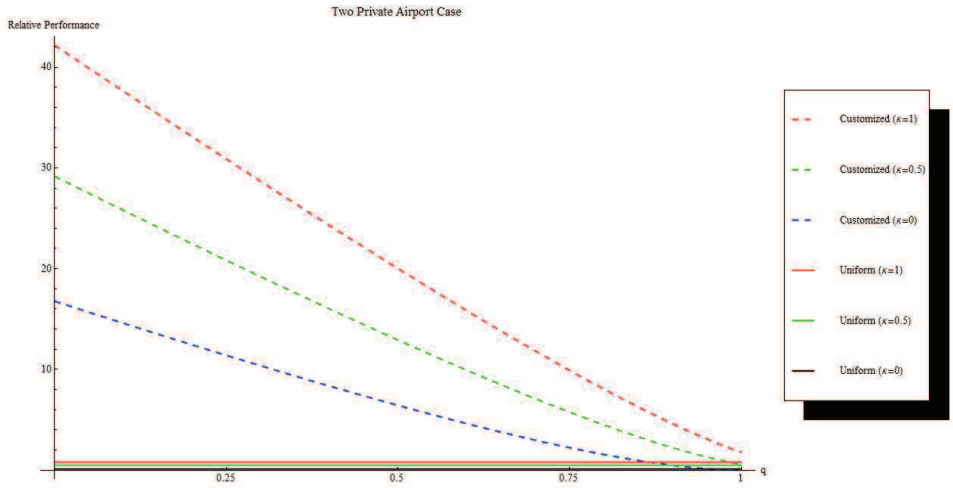


Figure 4: Relative Performance of Two Heterogeneous Private Airports with Changing q

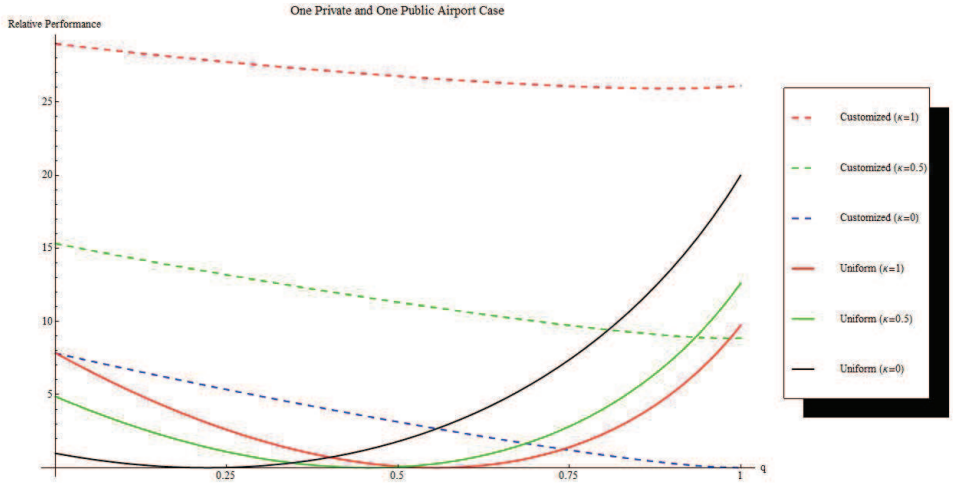


Figure 5: Relative Performance of Heterogeneous Private and Public Private Airports with Changing q

diversifies from these critical values, however, the performance becomes worse. In addition, as the externality level becomes higher, these critical q values also increase.

Table 6 summarizes the above results. It shows that if the externality is minimal and the government uses a customized approach, the decentralized financing (i.e., $q = 1$) is a dominating strategy and makes it possible to obtain a socially optimal outcome for all airport network settings. However, as the externality increases, centralized financing dominates in the case for two public airports whereas decentralized financing still dominates in the other cases. For the government which relies on a uniform approach, q does not affect the performance for the first two cases, while the different level of q can produce a socially optimal outcome in the last case.

Table 6: Optimal q with a Specific Regulatory Rule and κ

		κ		
		0	0.5	1
Two Public Airports	Customized	Decentralized*	Centralized	Centralized
	Uniform	-	-	-
Two Private Airports	Customized	Decentralized*	Decentralized	Decentralized
	Uniform	-*	-	-
One Private One Public	Customized	Decentralized*	Decentralized	Decentralized
	Uniform	$q = 0.25$*	$q = 0.47$*	$q = 0.61$*

Note: * indicates that the setting can produce a socially optimal outcome and - indicates q does not affect the result.

ANNEX1.4.3 Case 2: Different Regulatory Rules

Since the study conducted by Oates [16], various authors argued that one-size-fits-all regulations impose significant costs on economic actors. The report published by ACI [33] also pointed out that one-size-fits-all approaches should be avoided. Even in a series of the interviews we conducted with various airport operators, we noticed that some of them concerned about these types of airport regulations. However, it is unclear whether one-size-fits-all regulations undermine security and produce a socially worse outcome, and tailored regulations outperforms one-size-fits-all regulations. Indeed, Besley & Coate [19] indicated that, with externality and non-identical players, the performance of uniform and customized systems depends on the levels of externality and heterogeneity. However, their study considered not private entities but public entities. In this subsection, we therefore explore the performance of different regulatory structures in a setting with different financing systems.

As discussed in the previous sections, it is obvious that the effect of externality differs, depending on which financing rule the government is using. This view introduces a new dimension in the view of the trade-off between the performance of centralized and decentralized financing mechanisms to which analysis we now turn. Here, we assume the case where q is set to a certain level (i.e., $q = \{0, 0.5, 1\}$) and the regulator determine an appropriate regulatory rule based on the externality level, k. Therefore, we aim at answering the question: "When q is fixed to a certain level, what would be the best regulatory rule to produce a socially desirable outcome?"

As before, we begin our discussion by considering two heterogeneous public airports. Figure 6 presents the performance of the regulatory approaches with different levels of q and κ. Several points should be mentioned. Similarly with the previous scenario, the outcomes based on a one-size-fits-all approach does not depend on the level of q. If the government uses a decentralized financing system (i.e., $q = 1$), a customized approach performs better than a one-size-fits-all approach when the externality is low. However, after a certain level of the externality, a one-size-fits-all approach starts to incur a better outcome. Similar interpretation can be applied to the case when q equals 0.5, except the critical point for q which makes a uniform approach outperform a customized approach is higher than the critical point with $q = 1$. If the government employs a centralized financing system (i.e., $q = 0$), the result is opposite to the previous cases: a rule based on an uniform approach performs better if the externality is minimal, but a rule based on customization becomes better as the

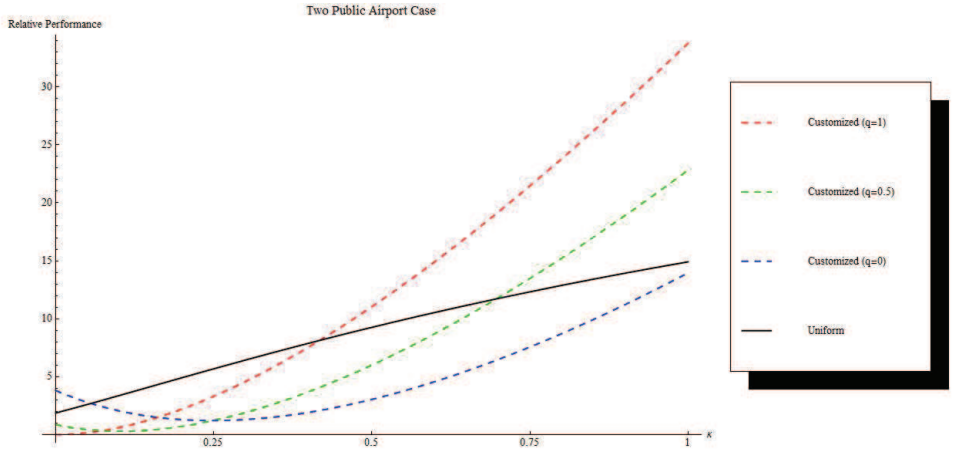


Figure 6: Relative Performance of Two Heterogeneous Public Airports with Changing κ

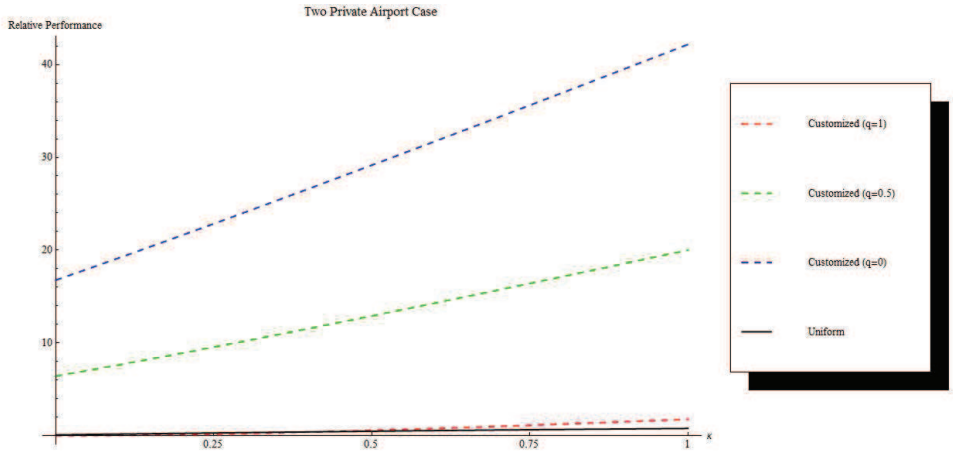


Figure 7: Relative Performance of Two Heterogeneous Private Airports with Changing κ

externality rises.

If the airports are privately-owned and heterogeneous, regardless of the level of externality, the imposition of uniform security expenditure performs better than that of customized security expenditure if $q < 1$ (i.e., a centralized financing system or a combination of the financing systems). In contrast, if the government uses a decentralized financing system, the customized imposition performs slightly better when the externality is low whereas the uniform imposition becomes better as the externality increases (see Figure 7).

Figure 8 illustrates the relative performance of regulatory rules when there are heterogeneous privately-owned and government-owned airports. The figure indicates that, if the regulator set q to 1 (i.e., decentralized financing), the customized imposition of security expenditures dominates when the externality is small and the uniform imposition dominates as the externality becomes high. If the government solely depends on a centralized financing system, the security expenditure based on the uniform imposition always outperforms the expenditure imposed by a customized approach, even if both regulatory mechanisms produce

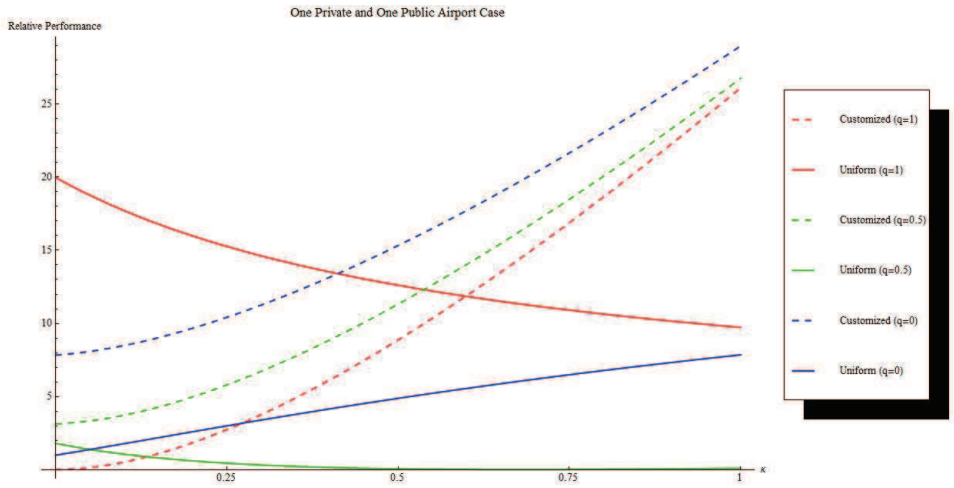


Figure 8: Relative Performance of Heterogeneous Private and Public Private Airports with Changing κ

worse outcome as the externality increases. Lastly, it can be seen that, if the government uses the combination of centralized and decentralized financing systems (i.e., $q = 0.5$), the security expenditure with the uniform imposition always dominates. Moreover, as the externality increases, the uniform regulatory rule produces a socially optimal outcome whereas the performance of the customized regulatory rule becomes worse.

Table 7 displays the summary of the results. From this table, it can be seen that, in many cases, a rule based on a uniform approach begins to perform better than a rule based on a customized approach as the externality increases.

ANNEX1.5 Conclusion

This study showed how different policy structures lead to different levels of security expenditures and expected losses with various industry settings. We identified that some financing and regulatory mechanisms can result in overspending on security, whereas with other mechanisms security spending might be too low. As shown in Table 8, we found that if the externality is low, the mix of decentralized financing and customized regulation can always produce a better outcome regardless of the industry settings. However, this result does not hold as the externality increases. Particularly, unlike the argument of researchers and practitioners that a customized regulation might perform better than a uniform regulation, we identified that a customized regulation might sometimes produce a worse outcome than a uniform regulation. For example, as shown in Table 8, while a customized regulation with two public airport case can produce the best outcome with an appropriate financing mechanisms, the other cases indicate that a customized regulation can produce the best outcome only when the externality is minimal. Since the other cases are more realistic aviation network settings, we believe that the regulator might be able to obtain a socially desirable outcome by developing a well-designed uniform regulation when the aviation network has high externality.

Table 7: Optimal Rules with Specific q and κ

			κ		
			0	0.5	1
Two Public Airports	q	Centralized	Uniform	Customized*	Customized*
		Combined	Customized	Customized	Uniform
		Decentralized	Customized*	Uniform	Uniform
Two Private Airports	q	Centralized	Uniform	Uniform*	Uniform*
		Combined	Uniform	Uniform*	Uniform*
		Decentralized	Customized*	Uniform*	Uniform*
Private & Public Airports	q	Centralized	Uniform	Uniform	Uniform
		Combined	Uniform	Uniform*	Uniform*
		Decentralized	Customized*	Customized	Uniform

Note: * indicates the setting that can produce the best outcome with the specific level of the externality. For example, in the two public airport case, a rule based on the decentralized financing and customized regulatory mechanism can produce the best outcome if $\kappa = 0$.

Table 8: Optimal Financing and Regulatory Rules with Different Network Settings

	κ		
	Low	Moderate	High
Two Public Airports	Decentralized & Customized	Centralized & Customized	Centralized & Customized
Two Private Airports	Decentralized & Customized	Any & Uniform	Any & Uniform
Public & Private Airports	Decentralized & Customized	Combined & Uniform	Combined & Uniform

Note: "Any" indicates that a financing system does not affect the overall performance.

The results also provided further implications. While the results from the theoretical base model with two public airport case showed that either centralized or decentralized financing mechanism can induce a better outcome, the results from the more realistic models indicated that combined financing mechanism might produce a better outcome than relying sorely on either centralized or decentralized financing mechanism.

Taken these results together, we might be able to conclude that the determination of optimal financing and regulatory rules should be based on the level of the externality and the aviation network setting.

We need to mention that much is left to be done. First of all, our future analysis will be extended by incorporate "strategic delegation" by the regulator [18]. Several scholars have argued that the regulator sometimes has an incentive to misrepresent his policy preference [18, 19]. For example, in the course of interviews supported by SECONOMICS project, we identified that, in setting regulatory rules for airport security, many of them are led and distorted by only opinions from big and dominant airports rather than reflect the opinions and perspectives of all airports. In this case, the regulator might set the security expenditure based on the expected losses of the dominant airports, and the other airports are forced to spend the security expenditure that only benefits these dominant airports. This analysis may add to our understanding of how the strategic delegation affects policy outcomes. Another interesting extension of the model would be to allow expenditures for different security measures. This seems particularly relevant in the light of the mandatory expenditures for different security measures.

Our analysis has shed light on problems that arise with determining an optimal security rule, an issue that has become more important over time. While there has been a wide array of discussions about optimal security rules in the aviation industry, there has been a lack of research efforts in conducting rigorous theoretical investigation for this matter. We believe that our model is well placed to explain the efficiencies and inefficiencies of various security rules with different settings for the aviation network, and to resolve the misconception of researchers and practitioners that a customized rule is the key to everything.

ANNEX2. Risk-based vs. Rule-based Approaches in Information Security

(U.S) Rule-based policies to mitigate software risk suggest using the CVSS score to measure the risk of an individual vulnerability and "act" accordingly. A key issue is whether the "danger" score does actually match the risk of exploitation in the wild, and if and how such score can be improved.

To address this question we propose to use a case-control study methodology similar to the procedure used to link lung cancer and smoking in the 1950s. A case-control study allows the researcher to draw conclusions on the relation between some *risk factor* (e.g. smoking) and an effect (e.g. cancer) by looking backward at the *cases* (e.g. patients) and comparing them with *controls* (e.g. randomly selected patients with similar characteristics). The methodology allows to quantify (statistically) the *risk reduction* achievable by acting on the risk factor.

We illustrate the methodology by using publicly available data on vulnerabilities, exploits

and exploits in the wild to (1) evaluate the performances of the current risk factor in the industry, the CVSS base score; (2) determine whether it can be improved by considering additional factors such the existence of a proof-of-concept exploit, or of an exploit in the black markets. In this example our analysis shows that (a) fixing a vulnerability just because it was assigned a high CVSS score only yields negligible risk reduction; (b) the additional existence of proof of concept exploits yields a relevant risk reduction; (c) fixing in response to exploit presence in black markets yields the largest risk reduction.

ANNEX2.1 Introduction

Software security configuration managers (e.g. Tripwire Enterprise, HP SCAP Scanner, QualysGuard FDCC Module, Rapid 7 Nexpose) usually rely on the National (US) Vulnerability Database[7] (NVD for short). Each vulnerability is reported alongside a "technical assessment" given by the Common Vulnerability Scoring System[8] (CVSS), which evaluates different technical aspects of the vulnerability [34].

Despite not being designed to be a metric for risk, the CVSS score is often used as such. For example, the US Federal government with QTA0-08-HC-B-0003 reference notice requires all IT products for the US Government to manage and assess the security of IT configurations with the NIST certified S-CAP protocol [35], which explicitly says: "*Organizations should use CVSS base scores to assist in prioritizing the remediation of known security-related software flaws based on the relative severity of the flaws.*". Another notable example is PCI DSS, the standard for security of credit card data, that states a similar rule: "*Risk rankings should be based on industry best practices. For example, criteria for ranking High risk vulnerabilities may include a CVSS base score of 4.0 or above [..].*" [36]. As a result, using the CVSS (base) score as-is to identify "high risk" vulnerabilities that must be fixed with the highest priority has become common practice in industry. However, at the present moment, no scientifically sound method is available to assess the effectiveness of those "risk metrics" to actually correlate with attacks in the wild. The absence of a sound methodology to treat the data at hand is also reflected by the variety of different approaches employed by security assessment tools such as those of Rapid7, Symantec, Qualys etc. Each assessment tool available on the market supplement it with different information and combine those with different functions.

In order to address this unsatisfactory state of practice we need to address the following questions:

1. *Do we have a methodology to assess the actual risk reduction that acting on a risk factor would entail?*
2. *Can we use the methodology to assess whether the CVSS base score is suitable in identifying high risk vulnerabilities?*
3. *If not, how can it be improved by considering other risk factors?*

[7] http://nvd.nist.gov
[8] http://www.first.org/cvss

A major problem to answer these questions is the nature of the data at hand. Vulnerability-database information is rife with problems and its use to classify risks of exploit is often inappropriate (see e.g [37, 38, 39] as some examples). An egregious example is present in [38] where a large majority of "exploits" are reported as zero-days[9] by subtracting two dates in the database. Unfortunately, the exploit time reported in databases such as OSVDB (Open Sourced Vulnerability Data Base[10]) only measures the time when a proof-of-concept exploit becomes known. Security researchers normally submit proof-of-concept exploits to vendors and vulnerability white markets in order to prove that the vulnerability is worth the bounty [40]: the zero-gap between exploit and vulnerability disclosure date is created by the disclosure reporting process! Timing data on vulnerabilities is often unreliable as well; quoting from [41], "*[using NVD] the computation of patch times and exploit times would contain errors of unknown size*". A major problem is therefore to identify a methodology and data sources that can correctly find and use data of actual exploitation (whether attempted or successful).

A positive example, is the paper by Bozorgi et al. [42] who have tried to use CVSS data to predict exploitation. Their results showed that the Exploitability CVSS subscore distribution does not correlate well with existence of known exploits from the ExploitDB. There are two ways to interpret this result (possibly both correct): the exploitability of CVSS is the wrong metric, or Bozorgi and his co-authors used the wrong DB to measure exploits. ExploitDB could be just used by security researchers to show off their skills (and obtain more contracts as penetration testers) but might not have a correlation with actual attacks by hackers.

To address these problems, in this paper we:

1. Introduce the case-control study as a fully-replicable methodology to soundly analyze vulnerability and exploit data;

2. Check the suitability of the current use of the CVSS score as a risk metric by comparing it against *actual exploits recorded in the wild* and by performing a break-down analysis of its characteristics and values

3. We use our case-control study methodology to show how one can improve the current practice of "Base CVSS" by considering other risk factors and quantitatively assess their performance in terms of risk reduction. The risk factors considered in our study are:

   (a) The CVSS base score as reported by the National Vulnerability Database.

   (b) Existence of a public proof-of-concept exploit.

   (c) Existence of an exploit traded in the cybercrime black markets.

   Any other risk factors, like software popularity, CVSS subscores, or other measurable values may be considered when replicating our methodology.

We stress our presentation on the reproducibility of our methodology and the possibility to integrate it into any practical scenario. To this aim we provide (1) within this manuscript

[9] A zero-day exploit is present when the exploit is reported before or on the date that the vulnerability is disclosed.

[10] http://osvdb.org

an exhaustive description of the methodology and of the rational behind our decisions; (2) making available our datasets for replication or comparative purposes.

In the rest of the paper we introduce our four datasets. An overview of the datasets illustrating the problem is then presented. We then draw a first, observational analysis of the performance of current practices by analysing the CVSS Impact and Exploitability submetrics. In the core of the paper we perform a randomized case-control analysis, and discuss our observations from the findings and threats to validity. We finally review related work and conclude.

ANNEX2.2 Datasets

We base our analysis on four datasets:

- NVD (National Vulnerability Database): the "universe" of vulnerabilities. NVD is the reference database for disclosed vulnerabilities held by NIST. It has been widely used and analyzed in previous vulnerability studies [43, 38, 44]. Our copy of the NVD dataset contains data on 49599 vulnerabilities reported until June 2012.
- EDB (Exploit-db[11]): proof-of-concept exploits. EDB includes information on proof-of-concept exploits and references the respective CVE. Our EDB copy contains data on 8122 *proof-of-concept* exploits and affected CVEs.
- EKITS: black-marketed exploits. EKITS is our dataset of vulnerabilities bundled in Exploit Kits. Exploit Kits are malicious web sites that the attacker deploys on some public webserver he/she controls. Their purpose is to attack and infect systems that connect to them. For further details refer to [45, 46]. EKITS is based on Contagio's Exploit Pack Table[12] and, at the time of writing, represents a substantial expansion over it in terms of reported exploit kits. EKITS reports exploits for 103 unique CVEs bundled in 90+ exploit kits. A sample of notable names of those are: *Elenonore, Blackhole, Crimepack, Fragus, Sakura, Icepack*.
- SYM: vulnerabilities exploited in the wild. SYM reports vulnerabilities that have been exploited in the wild as documented in Symantec's AttackSignature[13] and ThreatExplorer[14] public datasets. SYM contains 1277 CVEs identified in viruses (local threats) and remote attacks (network threats) by Symantec's commercial products. This has of course some limitation as direct attacks by individual motivated hackers against specific companies are not considered in this metric. The SYM dataset can be seen as an "index" of the wider WINE dataset [47] where actual volumes of attacks are reported. We do not use it here as we are trying to characterize a worst case scenario where "one exploit is too many". We are currently using the full WINE information to help improving the CVSS v3 upcoming standard.

[11] http://www.exploit-db.com/
[12] http://contagiodump.blogspot.it/2010/06/overview-of-exploit-packs-update.html
[13] http://www.symantec.com/security_response/attacksignatures/
[14] http://www.symantec.com/security_response/threatexplorer/

Table 9: Summary of our datasets

DB	Content	Collection method	#Entries
NVD	CVEs	XML parsing	49599
EDB	Publicly exploited CVEs	Download and web parsing to correlate with CVEs	8122
SYM	CVEs exploited in the wild	Web parsing to correlate with CVEs	1277
EKITS	CVEs in the black market	ad-hoc analysis + Contagio's Exploit table	103

Table 9 summarizes the content of each dataset and the collection methodology. All datasets used in this study are available to the community upon request[15].

ANNEX2.2.1 A coarse-grained overview of the datasets

We report in Figure 9 the histogram distribution of the CVSS base scores. Three clusters of vulnerabilities are visually identifiable throughout our datasets:

1. HIGH: CVSS ≥ 9
2. MEDIUM: $6 \leq$ CVSS < 9
3. LOW: CVSS < 6

The role of the CVSS score is, in the context of our analysis, to discern dangerous vulnerabilities from non-dangerous ones. Therefore an important analysis at this stage is to understand the overlap between the datasets, in order to grasp whether they and the CVSS score are capturing the same phenomenon.

In Figure 10 we report a Venn diagram of our datasets. Area size is proportional to the number of vulnerabilities that belong to it; the color is an indication of the CVSS score. Red, orange and cyan areas represent HIGH, MEDIUM and LOW score vulnerabilities respectively. NVD reports a large number of HIGH CVSS vulnerabilities that are not included in SYM. Similarly, while most of the intersection between EDB and SYM is covered by HIGH score CVEs, much of the red area for EDB is not included in SYM. A similar conclusion can be drawn for MEDIUM score vulnerabilities. This map gives a first intuition of the problem one may encounter by using the CVSS base score as a metric for risk of exploitation: much of the "red area" is located *outside* of SYM (false positives), while *within* SYM about half the vulnerabilities are of LOW or MEDIUM score (false negatives).

Table 10 reports the likelihood of a vulnerability being in SYM if it is contained in one of our datasets.

[15] http://securitylab.disi.unitn.it/doku.php?id=datasets

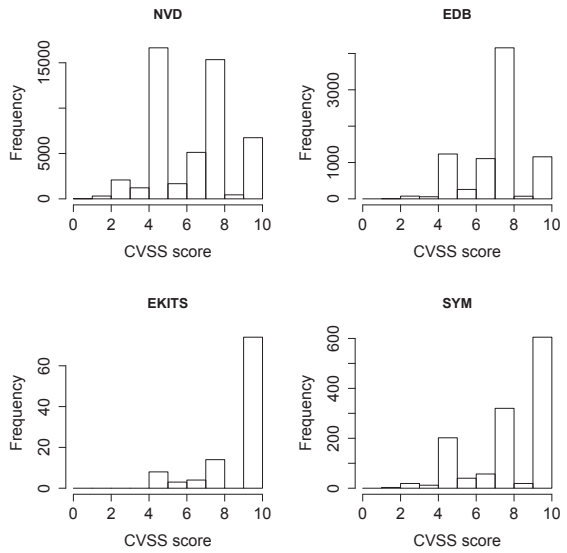


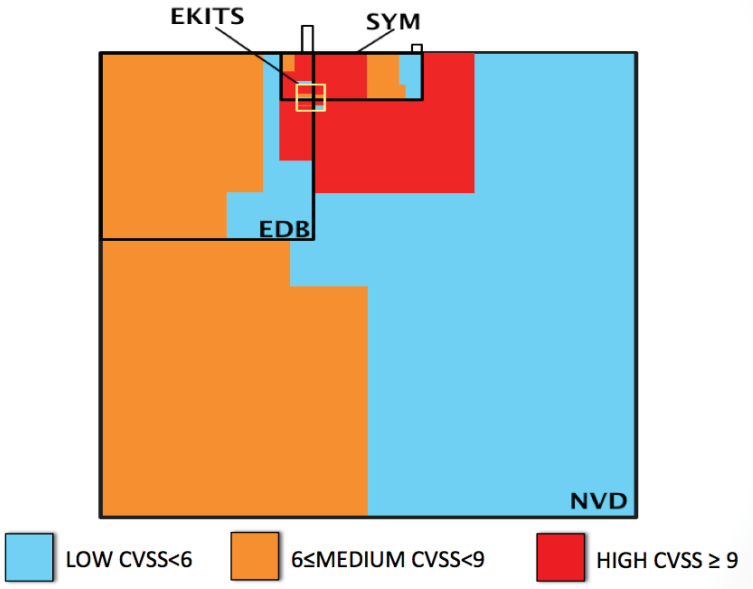
Figure 9: Distribution of CVSS scores per dataset.

Table 10: A (potentially erroneous) conditional probability of vulnerability being a threat

	vuln in SYM	vuln not in SYM
EKITS	75.73%	24.27%
EDB	4.81%	95.19%
NVD	2.57%	97.43%

Note: Conditional probability that a vulnerability v is listed by Symantec as threat knowing that it is contained in a dataset, i.e. $P(v \in SYM \mid v \in dataset)$. This is a rushing computation because datasets might be constructed with different criteria.

A rushing conclusion might be that, if one sees a vulnerability affecting his/her software in the black market, there is roughly a 75% chance that it is exploited in the wild. For NVD and EDB the likelihood would be measured to be less than 5%. However, this conclusion can be *grossly* incorrect. For example SYM might report only vulnerabilities of interest to Symantec's costumers. Suppose they mostly use Windows; then all Linux vulnerabilities listed in EDB would not be mentioned in SYM not because they are not exploited in the wild, but simply because they are not interesting for Symantec. Another possible example might be that Symantec only focuses on vulnerabilities exploited over the network while NVD might report also lots of vulnerabilities exploitable by social engineering. We would have in other words a "selection bias" problem. In order to offer more scientifically sound conclusions we need to (a) better understand the internals of the CVSS base score (which we do in the next subsection) and (b) propose a methodology to make sure we are comparing apples with apples (which we do in the case-control methodology section).



Dimensions are proportional to data size. In red vulnerabilities with CVSS≥9 score. Medium score vulnerabilities are orange, and cyan represents vulnerability with CVSS lower than 6. The two small rectangles outside of NVDspace are vulnerabilities whose CVEs were not present in NVD at the time of sampling.

Figure 10: Relative Map of vulnerabilities per dataset

ANNEX2.3 CVSS score breakdown

The Common Vulnerability Scoring System identifies three scores: the *base score*, the *temporal score*, and the *environmental score* [34]. The base score identifies "fundamental characteristics of a vulnerability that are constant over time and user environments"; the temporal score considers assessments like existence of a patch for the vulnerability, or the presence of an exploit in the wild; the environmental score considers further assessments tailored around the particular system implementation in which the vulnerability is present. However, of the three only the *base score* is identified, by standards and best practices alike, as the metric to rely upon for vulnerability management [48, 35]. The base score is also the only one commonly available in vulnerability bulletins and public datasets. We will therefore follow these guidelines and use the base score as a factor of risk.

The CVSS base score is computed as a product of two submetrics: the Impact submetric and the Exploitability submetric. The CVSS base score $CVSS_b$ assumes therefore the form:

$$CVSS_b = Impact \times Exploitability \tag{23}$$

which closely recalls the traditional definition of risk as "impact $\times$ likelihood". The Impact submetric is an assessment of the impact the exploitation of the vulnerability has on the system. The Exploitability subscore is defined by factors such as the difficulty of the exploitation and reachability of the vulnerability (e.g. from the network or physical access only). For this reason it can be interpreted as a measure of "likelihood of exploit" [42], even if it not defined as such.

Table 11: Possible values for the Exploitability and Impact subscores.

Impact subscore		
Confidentiality	Integrity	Availability
Undefined	Undefined	Undefined
None	None	None
Partial	Partial	Partial
Complete	Complete	Complete

Exploitability subscore		
Access Vector	Access complexity	Authentication
Undefined	Undefined	Undefined
Local	High	Multiple
Adjacent Net.	Medium	Single
Network	Low	None

ANNEX2.3.1 The Impact and Exploitability Subscores

The Impact and Exploitability subscores are calculated on the basis of additional variables, reported in Table 11. The Impact submetric is identified by three separate assessments on the Confidentiality, Integrity and Availability impacts on a victim system This triple will be identified as the "CIA" impact in this manuscript. Each variable can assume three values: Complete (C), Partial (P), None (N).

The Exploitability submetric is as well identified by three variables:

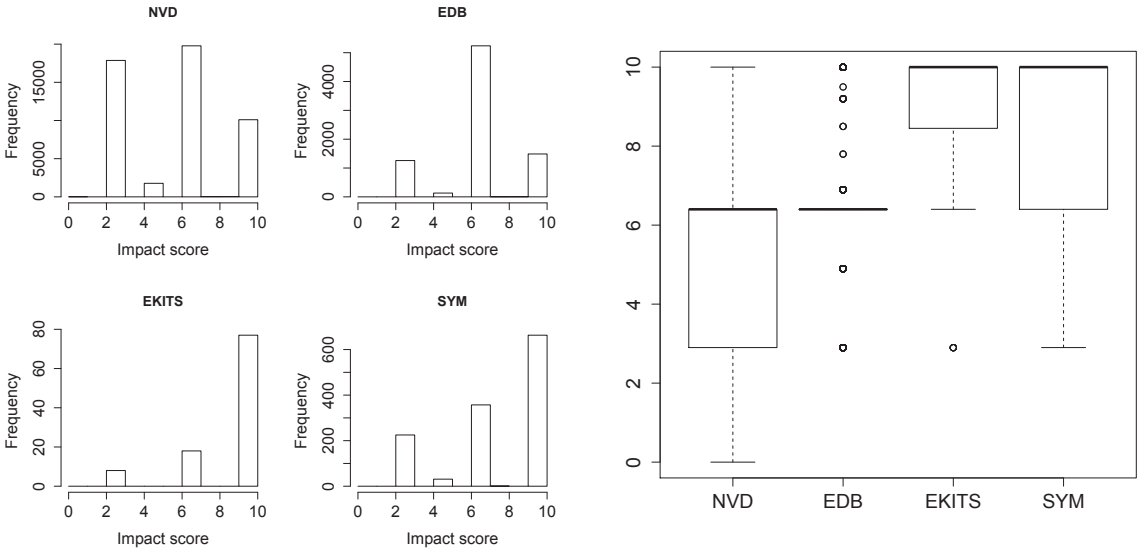
- Access Vector gives information on the accessibility of the vulnerability by distinguishing the case when the attacker can exploit it remotely from the Network, (N); from an Adjacent Network (A); by Locally (L) accessing the vulnerable component.
- Access Complexity provides information on the difficulty the attacker may encounter in recreating the conditions for the vulnerability to be exploited. This assessment can assume three values: High (H), Medium (M), or Low (L).
- Authentication represents the number of steps of authentication the attacker has to go through to trigger the vulnerability. The levels of the assessment can be: None (N), Single (S), Multiple (M).

Table 11 reports a summary of the CVSS base score's variables and respective possible values.

ANNEX2.3.2 Breakdown of the Impact subscore

The histogram distribution of the Impact subscore is depicted in Figure 11. By looking at all vulnerabilities (NVD), it is apparent how the subscore does not assume all possible values, and as a result the subscore distribution has gaps below score 2, between 3 and 6, and between 7 and 9. Still, the Impact subscore shows some variability throughout all datasets. In EDB scores between 6 and 7 characterize the great majority of vulnerabilities. This distribution may be an effect of the nature of the dataset: EDB features proof-of-concepts for vulnerabilities discovered by security researchers, likely with the intent of selling them to the software vendors; lower score vulnerabilities may be of too little value to be worth the bounty [40]; medium-score ones may instead represent the "low-hanging fruits" that maximize the researchers' return-on-investment.

The majority of vulnerabilities in SYM and EKITS have a high Impact subscore; unsurprisingly, vulnerabilities exploited in real attacks in the wild tend to yield a higher Impact



The histogram on the left represents the frequency distribution of CVSS Impact values among the datasets. The boxplot on the right reports the distribution of values around the median (represented by a thick horizontal line). Outliers are represented by dots.

Figure 11: Histogram and boxplot of CVSS Impact subscores per dataset.

on the system victim of the attack than the average vulnerability. Vulnerabilities bundled in exploit kits also tend to have a higher impact than the average vulnerability; this is again unsurprising as the whole purpose of an exploit kit is to "drop" malware on the victim system and executing it (i.e. the vulnerability should allow for arbitrary code execution). About 20% of the vulnerabilities in SYM have an impact score lower than 6.

The different distribution of the CVSS Impact subscore among the datasets is apparent in the boxplot reported in Figure 11. NVD results distributed in the whole range [0..10], with median just above 6 (6.4). The distribution of impact values in EDB is highly dense around the median (6.4). The distribution of the Impact subscore for SYM and EKITS are clearly different from the other two datasets; their median impact score of 10 is also significantly higher than those of NVD and EDB.

To explain the gaps in the histogram in Figure 11, we drilled down the distribution of Impact subscores throughout our datasets. To simplify discussion, in Table 12 we report the incidence of the existing values for the CIA assessments in NVD only. It is immediate to see that only few values are actually used. For example there is only one vulnerability whose CIA impact is "PCP" (i.e. partial impact on confidentiality, complete on integrity and partial on availability).

Availability almost always assumes the same value of Integrity, apart from the case where there is no impact on Confidentiality, and looks therefore of limited importance for a descriptive discussion.

For the sake of readability, we therefore exclude the Availability assessment from the analysis, and proceed by looking at the two remaining Impact variables in the four datasets. This analysis is reported in Table 13. Even with this aggregation on place many possible values of the CIA assessment result unused. "PP" vulnerabilities characterize the majority

Table 12: Incidence of values of CIA triad within NVD.

Confidentiality	Integrity	Availability	Absolute no.	Incidence
C	C	C	9972	20%
C	C	P	0	-
C	C	N	43	<1%
C	P	C	2	<1%
C	P	P	13	<1%
C	P	N	3	<1%
C	N	C	15	<1%
C	N	P	2	<1%
C	N	N	417	1%
P	C	C	5	<1%
P	C	P	1	<1%
P	C	N	0	-
P	P	C	22	-
P	P	P	17550	35%
P	P	N	1196	2%
P	N	C	9	<1%
P	N	P	110	<1%
P	N	N	5147	10%
N	C	C	64	<1%
N	C	P	1	<1%
N	C	N	43	<1%
N	P	C	17	<1%
N	P	P	465	1%
N	P	N	7714	16%
N	N	C	1769	4%
N	N	P	5003	10%
N	N	N	16	<1%

Table 13: Combinations of Confidentiality and Integrity values per dataset.

Confidentiality	Integrity	SYM	EKITS	EDB	NVD
C	C	51.61%	74.76%	18.11%	20.20%
C	P	0.00%	0.00%	0.02%	0.04%
C	N	0.31%	0.97%	0.71%	0.87%
P	C	0.00%	0.00%	0.01%	0.01%
P	P	27.80%	16.50%	63.52%	37.83%
P	N	7.83%	0.97%	5.61%	10.62%
N	C	0.23%	0.00%	0.18%	0.22%
N	P	4.39%	2.91%	5.07%	16.52%
N	N	7.83%	3.88%	6.75%	13.69%

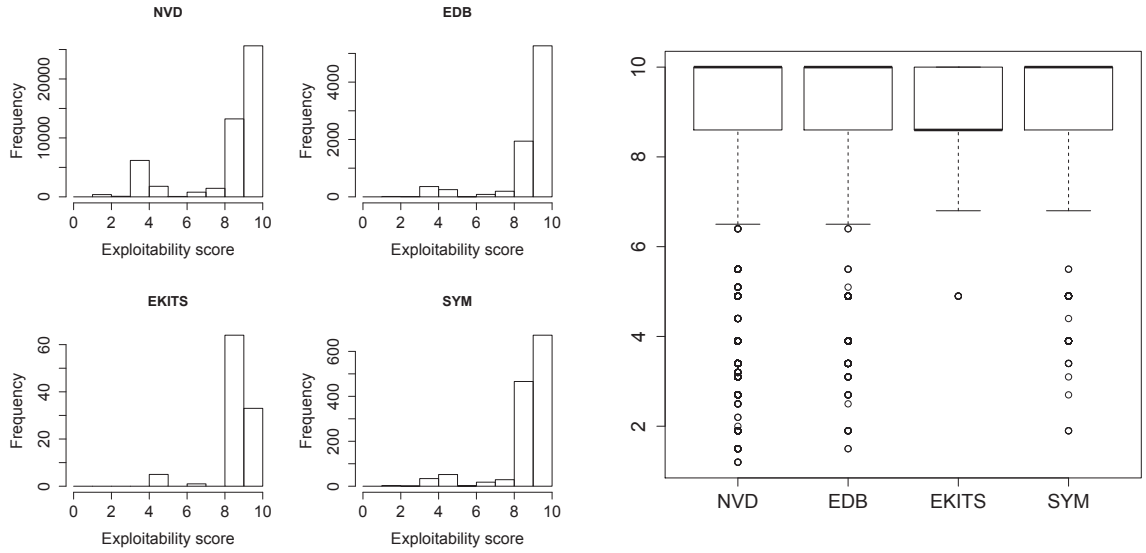


Figure 12: Distribution of CVSS Exploitability subscores.

of disclosed vulnerabilities (NVD) and vulnerabilities with a proof-of-concept exploit (EDB). This observation changes completely when looking at the SYM and EKITS datasets, for which most vulnerabilities (50%, 75%) score "CC". This shift alone can be considered responsible for the different distribution of scores depicted in Figure 11.

Observation

The "gaps" in the impact score distribution visible in the histogram are explained by the negligible influence of the Availability assessment on the overall score, and by the observed little variance in the values assumed by Confidentiality and Integrity. Two out of 27 CIA configuration ("CC*" and "PP*") describe almost 60% of the assessments.

ANNEX2.3.3 Breakdown of the Exploitability subscore

Figure 12 shows the distribution of the Exploitability subscore for each dataset. The Exploitability score shows little variability among the datasets. Almost all vulnerabilities score between 8 and 10, and from the boxplot it is evident that the distribution of exploitability subscores is indistinguishable among the datasets.

Bozorgi et al. [42] did not find any correlation between Exploitability subscore and existence of a proof-of-concept exploit in EDB; we confirm their conclusion by finding no relation between the Exploitability score and *actual exploitation in the wild*: the CVSS Exploitability subscore resembles more a constant than a variable: everything is exploitable. This means that it has little to no influence on the variance of the final CVSS score, which may in turn affect the suitability of the CVSS as a *risk* metric.

If we drill down the Exploitability subscores we find that most vulnerabilities do not require any authentication (Authentication = (N)one, 95%), and most are accessible from remote (Access Vector = (N)etwork, 87%). This observation is even more extreme in datasets other

Table 14: Exploitability Subfactors for each dataset.

	metric	value	SYM	EKITS	EDB	NVD
9*Exploitability	3*Acc. Vec.	local	2.98%	0%	4.57%	13.07%
		adj.	0.23%	0%	0.12%	0.35%
		net	96.79%	100%	95.31%	86.58%
	3*Acc. Com.	high	4.23%	4.85%	3.37%	4.70%
		medium	38.53%	63.11%	25.49%	30.17%
		low	57.24%	32.04%	71.14%	65.13%
	3*Auth.	multiple	0%	0%	0.02%	0.05%
		single	3.92%	0.97%	3.71%	5.30%
		none	96.08%	99.03%	96.27%	94.65%

than NVD, that we show in Table 14.

Observation

The Exploitability subscore's low variability depicted in Figure 12 is explained by (1) Authentication being set essentially to one value only; (2) Access Complexity having only two possible values ("Local", "Network") with an overwhelming prevalence of the latter. Thus Exploitability is essentially a constant and therefore cannot predict exploits as indeed observed by [42].

To assess how these characteristics of the CVSS assessment affect the use of the score as a risk metric, we now introduce our case-control study.

ANNEX2.4 Randomized case-control study

Randomized case-control studies are often set up in the medical domain to assess the effectiveness of a medical test or of a medical treatment. In some cases it is not feasible for practical and foremost ethical consideration to perform a classical randomized controlled experiment (e.g. asking random subjects to start smoking in order to see whether they die of cancer). An alternative solution is to perform a retrospective analysis in which the *cases* (people with a known illness) are compared with a random population of *controls* (randomly selected patients with the same characteristics). A famous application of the methodology is the 1950 study by [49], where the authors showed the correlation between smoking habits and the presence or absence of cancer of the lungs by performing a case-control study with data on hospitalization. Since we can not ask users to stay vulnerable and get their bank accounts emptied, a case-control study looks the appropriate instrument to evaluate risk factors for vulnerability exploitation.

We will use this methodology to assess whether the CVSS score can be a good predictor for vulnerability exploitation, or whether it can be improved by additional information.

We start by giving the reader some terminology:

- *Cases.* The cases of a control study are the subjects that present the observed effect. For example, in the medical domain the cases could be the "patients" whose status

has been ascertain to be “sick”. In a computer security scenario, a “case” could be a vulnerability that has been exploited in the wild.

- *Explanatory variable or risk factor.* A risk factor is an effect that can explain the presence (or increase in likelihood) of the illness (or attack). For cancer it is smoking habits. We use the CVSS score of the vulnerability as well as other information such as existence of proof of concept exploits, or presence in kits traded in black-markets. Another possibility (not investigated here) could be to use some of the CVSS subscores.

- *Confounding variables* can be other variables that may be alternative explanations of the effect, or correlate with its observation. For example, patient age or sex may be confounding factors for some types of cancer. In our case the existence of an exploit in SYM may depend on factors such as the type of vulnerability, its time of disclosure and the affected software.

- *Control group.* A control group is a group of subjects chosen at random from a population with similar characteristics (e.g. age, social status, location) to the cases. In the original formulation of case-control study, the control group was composed of healthy people only. With that application of the case-control study we can only ascertain whether the observed effect (e.g. cancer of the lung) is related to a particular risk factor (e.g. smoking habits) by a greater or lower degree than to other confounding variables (e.g. living in polluted cities). We relax this condition and leave open the (random) chance that cases get included in the control group. This relaxation allows us to perform additional computations on our samples (namely CVSS sensitivity, specificity and risk reduction). This, however, introduces (random) noise in the generated data. To address this issue, we perform the analysis with bootstrapping.

- *Bootstrapping.* Bootstrapping is a technique that allows us to derive sound conclusions from the data by re-sampling multiple times, with repetition, from the same data. This mitigates the effects, in the final analysis, of a random observation showing up in an iteration.

In our case-control study the cases are the vulnerabilities in SYM (loosely corresponding to cases of lung cancer); we consider as risk factors (1) the CVSS level; (2) the existence of a Proof-of-Concept exploit (EDB); (3) the presence of an exploit in the black markets (EKITS).

Confounding variables Deciding which confounding factors to include in a case-control study is usually left to the intuition and experience of the researcher [49]. Because SYM is the “critical point” of our study (as it reports our cancer patients), we consulted with Symantec to decide which factors to consider as confounding. While this list can not be considered an exhaustive one, we believe the variables we identify in the following capture the most important aspects of the inclusion of the vulnerability in SYM. More details on this process are discussed in the Threats to Validity Section.

- *Software.* Symantec is a security market leader and provides a variety of security solutions but its largest market share is in the consumer market. In particular, the data in

SYM is referenced to the malware and attack signatures included in commercial products that are often installed on "consumer" machines. These are typically Microsoft Windows machines running commodity software like Microsoft Office and internet plugins like Adobe Flash or Oracle Java [50]. Unix software is also included in SYM. However we do not consider this sample to be representative of Unix exploited vulnerabilities. Because of this selection problem, SYM may represent only a subset of all the software reported in NVD or EDB or EKITS. We therefore consider "software" to be a confounding variable to the presence of the vulnerability in SYM.

- *Year.* Symantec's commitment in reporting exploited CVEs may be influenced by time as well. From a detailed conversation with Symantec, the inclusion of a CVE in an attack signature is an effort on Symantec's side aimed at enhancing the usefulness of their datasets. Specifically, Symantec recently opened a data sharing program called WINE whose aim is to share attack data with security researchers [47]. The data included in the WINE dataset spans from 2009 to the present date. Given the explicit sharing nature of their WINE program, we consider vulnerabilities disclosed after 2009 to be better represented in SYM. We therefore consider only those in our study.

- *Impact type.* Our analysis showed that some CIA types are more common in SYM than elsewhere (e.g. CIA="CCC"). An explanation for this may be that attackers contrasted by Symantec may prefer to attack vulnerabilities that allow them to execute arbitrary code rather than ones that enables them to get only a partial access on the file system. We therefore also control for the CVSS Confidentiality, Integrity and Availability assessments.

ANNEX2.4.1 Experiment run

We divide our experiment in two parts: sampling and execution. In the former we generate the samples from NVD, EDB and EKITS. In the latter we compute the relevant statistics on the samples. What follows is an English description of these processes. Our R script to replicate the data analysis is available.

Sampling The sampling procedure generate a sample with the same number of vulnerabilities that are present in SYM. Every sample vulnerability sv_i is representative of the v_i-th vulnerability in SYM. In other words, they must have the same control factors identified in this study: *CIA CVSS Impact, Year* and *Software*.

The only control that we can straightforwardly enforce is the CIA Impact, as CIA values are strictly defined by the CVSS framework. For year and software we can match the control variables in several ways.

- "Year" represents the year of disclosure of the vulnerability. To enforce this control we can *censor* all vulnerability data disclosed before 2009 and after 2012, and randomly pick a vulnerability within this time-frame. An alternative is *exact match* for the date of v_i and sv_i. We believe the former is preferable because the timing data reported in NVD is noisy due to how the vulnerability disclosure mechanism works [41, 40]. We used only coarse data granularity (year).

Table 15: Output format of our experiment.

	$v \in SYM$	$v \notin SYM$
Above Threshold	a	b
Below Threshold	c	d

Table 16: Sample thresholds

CVSS $\geq$ 6
CVSS $\geq$ 9
CVSS $\geq$ 9 & v $\in$ EDB
CVSS $\geq$ 9 & v $\in$ EKITS

- "Software" is the name of the software affected by the vulnerability. It is represented by a string reported in the NVD entry of the vulnerability. However, no standardized way to report vulnerability software names exists.

  For example CVE-2009-0559 (in SYM) is reported as a "*Stack-based buffer overflow in Excel*", but the main affected software reported is (Microsoft) Office. In contrast, CVE-2010-1248 (in SYM as well) is a "*Buffer overflow in Microsoft Office Excel*" and is reported as an Excel vulnerability. Thus, performing a perfect string match for the software variable would exclude from the selection relevant vulnerabilities affecting the same software but reporting a different software name. To consider *all* software entries reported in an NVD report is equally a bad idea: which value should one then consider as the valid one? Keeping the example of CVE-2009-0559, the correct value should be Office? or Excel? or Office Share point server?

  The problem with software names extends beyond this. Consider for example a vulnerability for *Webkit*, an HTML engine used in many browsers. This vulnerability may affect not only Webkit, but also the Safari, Chrome and Opera web browsers that use Webkit as a rendering engine. Because a Webkit vulnerability in Apple Safari might also be a Webkit vulnerability in Google Chrome, making a 1:1 selection of software names is prone to unknown noise.

  For these reasons we also checked that the software for sv_i is included in the list of software for $\forall v_i \in SYM$.

To create the samples, for each of NVD, EDB and EKITS we randomly select, with repetition, a sample vulnerability sv_i that satisfies the discussed conditions enforced by the control values for v_i. We then include sv_i in the list of selected vulnerabilities for that dataset sample. We repeat this procedure for all vulnerabilities in SYM. Eventually we obtain three samples of vulnerabilities in NVD, EDB and EKITS that are *identically distributed* to SYM with respect to our confounding variables. The EDB/EKITS samples will be used to test the presence in the corresponding database as an additional risk factor.

The sampling has been performed with the statistical tool R-CRAN [51].

Execution Once we collected our samples, we compute the frequency with which each risk factor identifies a vulnerability in SYM. Our output is represented in Table 15. Each risk factor is defined by a CVSS threshold level t in combination with any other identified risk factor. Examples of thresholds are reported in Table 16. We run our experiment for all CVSS thresholds t_i with $i \in [1..10]$. For each risk factor we evaluate the number of vulnerabilities in the sample that fall *above* and *below* the CVSS threshold, and that are included (or not included) in SYM: the obtained table reports the count of vulnerabilities that each risk factor

correctly and incorrectly identifies as "at high risk of exploit" ($\in$ SYM) or "at low risk of exploit" ($\notin$ SYM).

The computed values entirely depend on the random sampling process. In an extreme case we may therefore end up, just by chance, with a sample containing only vulnerabilities in SYM and below the current threshold (i.e. $[a = 0; b = 0; c = 1277; d = 0]$). Likely such an effect would be due only to chance rather than representing the reality. To mitigate this we repeat, for every instance of the risk factors, the whole experiment run 400 times and keep the median of the results. We chose 400 times because we observed that the distribution of results was already markedly Gaussian. Any statistic reported in this paper is to be intended as the median of the generated distribution of values.

ANNEX2.4.2 Parameters of the analysis

Sensitivity and specificity In the medical domain, the sensitivity of a test is the conditional probability of the test giving positive results when the illness is present. The specificity of the test is the conditional probability of the test giving negative result when there is no illness. Sensitivity and specificity are also known as True Positive Rate (TPR) and True Negatives Rate (TNR) respectively. In our context, we want to assess to what degree a positive result from our current test (the CVSS score) matches the illness (the vulnerability being actually exploited in the wild and tracked in SYM). The sensitivity and specificity measures are computed as:

$$Sensitivity = P(v\text{'s Risk factor above } t |\ v \in SYM) = a/(a+c) \tag{24}$$

$$Specificity = P(v\text{'s Risk factor below } t |\ v \notin SYM) = d/(b+d) \tag{25}$$

where t is the threshold. Sensitivity and specificity outline the performance of the test in identifying exploits, but say little about its effectiveness in terms of diminished risk.

Risk Reduction and Odds Ratio To understand the effectiveness of a policy we adopt an approach similar to that used by Evans in [52] to estimate the effectiveness of seat belts in preventing fatalities. In his case, the "effectiveness" was given by the difference in the probability of having a fatal car crash when wearing a seatbelt and when not doing it ($Pr(Death \,\&\, Seat\ belt\ on) - Pr(Death \,\&\, not\ Seat\ belt\ on)$.

In our case, we measure the ability of the CVSS score (combined with the existence of a proof-of-concept exploit or an exploit in the black markets) to predict the actual exploit in the wild (i.e. present in SYM). Formally, the risk reduction is calculated as

$$RR = P(v \in SYM | v\text{'s Risk factor above } t) - P(v \in SYM | v\text{'s Risk factor below } t) \tag{26}$$

and the odds ratio is given by

$$OR = \frac{|v \in SYM \wedge v\text{'s Risk factor above } t|}{|v \notin SYM \wedge v\text{'s Risk factor above } t|} \Big/ \frac{|v \in SYM \wedge v\text{'s Risk factor below } t|}{|v \notin SYM \wedge v\text{'s Risk factor below } t|} \tag{27}$$

Both measures give an evaluation of the relative distance of the two risk levels identified by the threshold. Note that the metrics are not always computable. For example, if no exploited vulnerability is below the threshold, the Odds Ratio loses its mathematical utility (because the ratio at the denominator goes to zero). We always report both RR and OR for the sake of completeness.

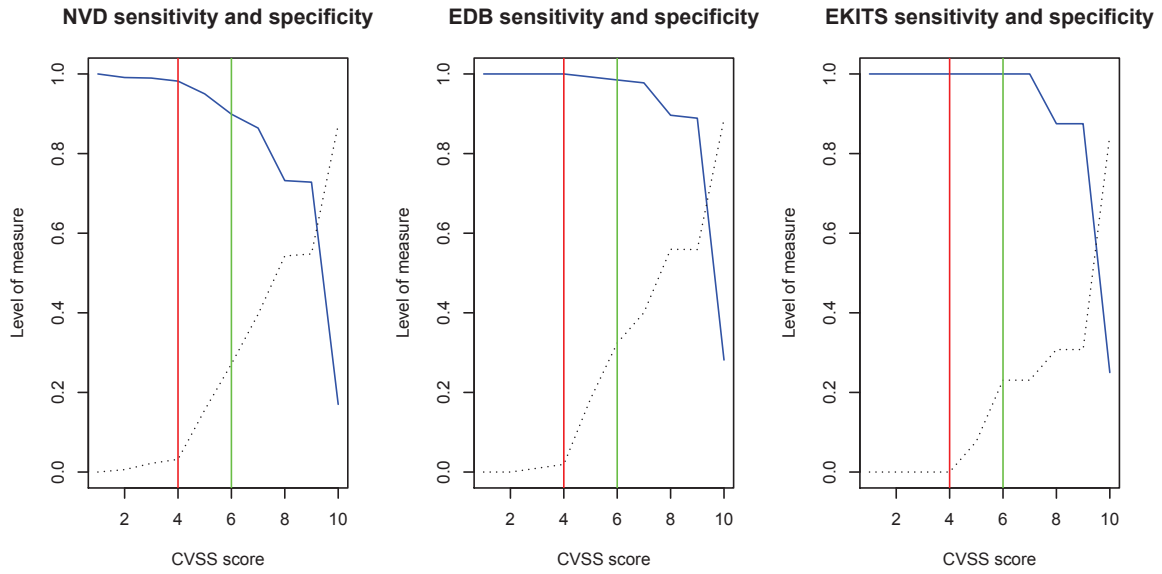


Figure 13: Sensitivity and specificity levels for different CVSS thresholds. The red line identifies the threshold for PCI DSS compliance (*cvss* = 4). The green line identifies the threshold between LOW and MEDIUM+HIGH vulnerabilities (*cvss* = 6, see histogram in Figure 9).

ANNEX2.4.3 Data Analysis

Sensitivity and specificity Figure 13 reports the sensitivity and specificity levels respective to different CVSS thresholds. Sensitivity is represented by the blue solid line; specificity is represented by the grey dotted line. The vertical red line outlines the CVSS threshold fixed by PCI DSS (*cvss* = 4). The green vertical line marks the threshold that separates LOW CVSS vulnerabilities from MEDIUM+HIGH CVSS vulnerabilities (*cvss* = 6).

Unsurprisingly, low CVSS scores show a very low specificity, as most non-exploited vulnerabilities result *above* the threshold.

With increasing CVSS thresholds, the specificity measure gets better without sensibly affecting sensitivity: NVD and EDB can achieve a specificity of 60% with a threshold equal to 8, but at the price of a high ratio of false negatives (30%). To further increase the threshold causes the sensitivity measure to collapse. In EKITS, because most vulnerabilities in the black markets are exploited and their CVSS scores are high, the specificity measure can not significantly grow without collapsing sensitivity.

Risk reduction and Odds ratio In Figure 14 we report our results for risk reduction (RR) and odds ratio (OR). To use the mere CVSS score, irrespectively of its threshold level, to define a patching policy always entails a very low risk reduction. Patching strategies considering the existence of a public proof-of-concept exploit as a risk factor seem to yield much higher returns in terms of risk reduction. Presence in the black markets seems to be the most important risk factor to consider, as it almost doubles the risk reduction entailed by Proof-of-Concept exploits. Due to the scarce variance in CVSS scores in the EKITS dataset, not all values can be computed (dividend is zero because there are no selected vulnerabilities

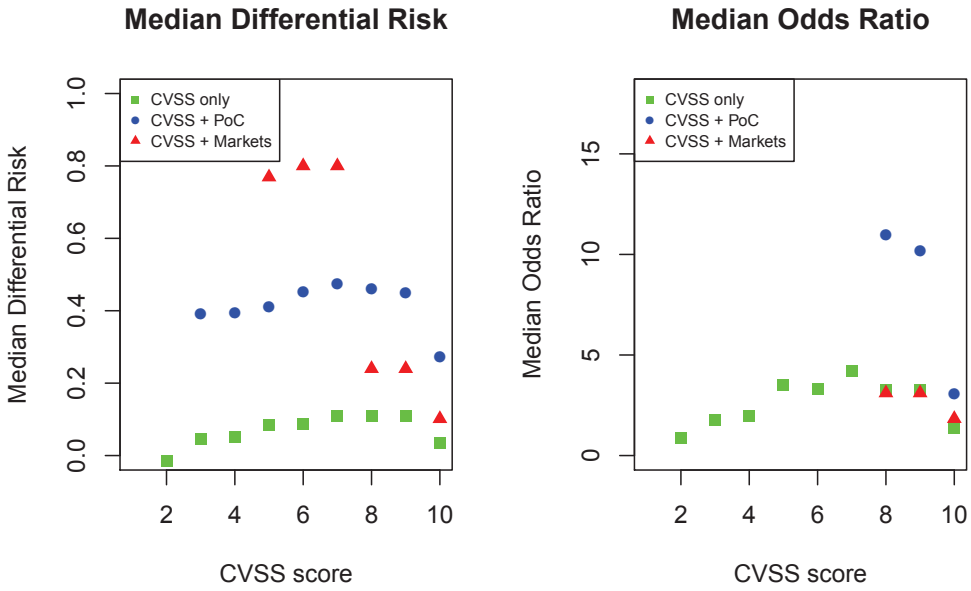


Figure 14: (left) risk reduction (RR) and (right) odds ratio (OR) entailed by different CVSS scores per dataset.

below the threshold).

Table 17 reports the numerical Risk Reduction for a sample of thresholds, and the analogous odds ratio.

The mere presence of the vulnerability (NVD) entails a reduction in risk that peaks at about 11% for rather high thresholds. In general, even considering the 95% confidence interval, we can conclude that CVSS-only based policies may be unsatisfactory from a risk-reduction point of view. Similar considerations can be outlined, for the "mere CVSS score", by looking at the Odds Ratio column: the gap between the odds of an exploit above the threshold and the odds of an exploit below the threshold is very low. The existence of a proof-of-concept exploit improves greatly the performance of the policy: with "CVSS $\geq$ 6 + PoC" a RR of 47% can be achieved. This result is comparable, and slightly preferable, to wearing a seat belt while driving: seat belts entail a reduction in risk of death of 43% [52]. A similar result, in terms of RR, is obtained by increasing the CVSS threshold to 9. The OR measure decreases sharply because the number of false negatives ($|SYM \ \& \ Below|$) rises. Finally, plugging into the methodology the existence of an exploit in the black markets rises the risk reduction up to 80%. The odds ratios are little informative here because the intersection between EKITS and SYM contains few vulnerabilities only, most of them with a high CVSS score (therefore rising the CVSS threshold doesn't make sense for black-marketed vulnerabilities).

ANNEX2.5 Discussion

We now summarize the main observations of our study. We focus on: (1) CVSS characteristics; (2) Risk reduction. Some conclusions are more absolute (exceptions counted on one's

Table 17: Risk Reduction for a sample of thresholds.

Threshold	RR	95% RR conf. int.	OR	95% OR conf. Int.	Significance
CVSS $\geq$6	9%	6% - 12%	3.32	1.94 - 6.53	***
CVSS $\geq$6 + PoC	45%	43% - 47%	31.36	28.37 - 34.96	****
CVSS $\geq$6 + Bmar	80%	$\pm$ 0	∞	$\pm$ 0	*
CVSS $\geq$9	11%	8% - 14%	3.24	2.26 - 4.68	****
CVSS $\geq$9 + PoC	45%	43% - 47%	10.16	9.12 - 11.5	****
CVSS $\geq$9 + Bmar	24%	$\pm$ 0	3.11	$\pm$ 0	

Note:Risk Reduction and Odds Ratio of vulnerability exploitation depending on policy and information at hand (CVSS, PoC, Markets). Significance is given by a Bonferroni-corrected Fisher Exact test (data is sparse) for three comparison (CVSS vs CVSS+PoC vs CVSS+BMar) per experiment [53]. **** indicates $p < 3.3E-5$; *** $p < 3.3E-4$; ** $p < 3.3E-3$; * $p < 1.6E-02$; nothing is reported for other values.

fingers), while others are only statistically significant.

1. *The CVSS Impact submetric assumes only a few of the possible values, as Confidentiality and Integrity losses usually go hand-in-hand*. The Availability CVSS assessment adds very little variability to the score, so of the 3 dimensions of the Impact subscore, only 2 are effectively relevant.

2. *The CVSS Exploitability metric shows little to none variability.* The only variability among the greatest majority of vulnerabilities in NVD is given for this metric by the Access Complexity variable. Authentication and Access Vector show very little (Access Vector) to almost none (Authentication) variability. The effect of this is that the Exploitability submetric results flattened around very high values. As a consequence, the Exploitability submetric is unsuitable as a characterization of "likelihood of exploit".

3. *The CVSS base score alone is a poor risk factor from a statistical perspective.* The de-facto usage in the industry of the CVSS base score as a metric for risk is therefore an unsatisfactory practice. Policies based on CVSS score, like the US Government NIST SCAP protocol or the world-wide used PCI DSS, may not make for effective strategies.

By considering risk factors other than the sole CVSS score it may be possible to obtain more effective strategies:

1. The existence of a proof-of-concept exploit is an interesting risk factor to consider. PoC-based policies can entail risk reductions up to 45% of the original risk. However, because of the nature of this data (see [40] and discussion in Section 1) special care must be taken to control for additional variables such as year and type of software. For example, a comparable policy ran *without* the control on software can obtain a risk reduction of only 10-11%.

2. The black markets are an even more important source of risk. Our results show that the inclusion of this risk factor can increase risk reduction up to 80%. Unsurprisingly, there is a distinction between the vulnerabilities identified by bad guys and good guys.

Most importantly, our methodology may be a useful tool for the practitioner that needs a scientific and precise way to assess the efficacy of a risk mitigation strategy. Our results can be reproduced by plugging in the methodology any risk factor the practitioner may find relevant to his/her use case. Different rationales than the mere risk reduction could be considered for the evaluation of the results: cost, time-to-deploy, or organizational effort are just example of an arbitrary set of options that can be factored in into a final evaluation.

ANNEX2.6 Threats to validity

Like any other empirical study based on field data, ours may be affected by a number of threats to validity [54] that we address in the following.

Construct validity Data collection is the main issue in an empirical study. SYM and EKITS may be particularly critical to the soundness of our conclusions. Because of the unstructured dataset of the original SYM dataset, to build SYM we needed to take some preliminary steps. We couldn't be sure about whether the collected CVEs were relevant to the threat. To address this issue, we proceeded in two steps. First, we manually analyzed a random selection of about 50 entries to check for the relevance of the CVE entries to the actual attack described in the signature. An informal communication with Symantec confirmed that the CVEs are indeed relevant to the attack.

Due to the shady nature of the tools, the list of exploited CVEs in EKITS may be not a representative of the population of CVEs in the black markets and/or affected by "false reporting" by the exploit kits authors (that, being criminals, may as well lie about what CVEs they infect). To mitigate the problem, we crossed-referenced EKITS entries with knowledge from the security research community and from our direct testing of tools traded in the the black markets [55].

External validity is concerned with the applicability of our results to real-world scenarios. Symantec is a world-wide company and a leader in the security industry. We are therefore confident in considering their data as representative of real-world attack scenarios. Yet, our conclusion can not be generalized to targeted attacks. These attacks in the wild usually target a specific platform or system and are less likely to generate an entry in a general purpose anti-virus product. We do not therefore extend our conclusions to Targeted attacks scenarios.

An important point to address is that our approach does not address *changing behavior* of the attacker. For example, if all vulnerabilities from the black markets with a certain characteristic get patched, the attacker may simply modify his own attack strategy in such a way to render the defender's strategy ineffective. This is a common problem in any game-theoretical approach: unfortunately the defender ought to move first, thus the attacker can always adapt to the defender's strategy (hence the definition of *equilibrium* as the state of the game into which neither attacker nor defender have a good reason to change their strategy). This problem is present in the application of any security technology or solution available. The game-theoretic nature of the problem is not addressed by our methodology either. We reserve the exploration of this issue for further work.

ANNEX2.7 Related work

Vulnerability studies Many studies before ours dealt with software vulnerabilities, software risk and risk mitigation. Among all, Frei et al. [37] were maybe the first to link the idea of life-cycle of a vulnerability to the patching process. Their dataset was a composition of NVD, OSVDB and 'FVDB' (Frei's Vulnerability DataBase, obtained from the examination of security advisories for patches). The life-cycle of a vulnerability includes discovery time, exploitation time and patching time. They showed that exploits are often quicker to arrive than patches are. They were the first to look, in particular, at the difference in time between time of first "exploit" and time of disclosure of the vulnerability. This work have recently been extended by Shahzad et al. [38], who presented a comprehensive vulnerability study on NVD and OSVDB datasets (and Frei's) that included vendors and software in the analysis. Many descriptive trends in timings of vulnerability patching and exploitation are presented. However, their use of EDB or OSVDB exploit data says little (if anything) about the actual exploitation of a vulnerability [56]. NVD timing data has also been reported to generate an unforeseeable amount of noise because of how the vulnerability disclosure process works [41, 56]. To avoid replicating these errors we make an effort in finding data on actual exploits, proof-of-concept exploits, and exploits in the black markets. We provide a descriptive analysis of this vulnerability data, and use our findings to provide sound advices to practitioners that desire to assess software vulnerability risk and efficacy of remediation strategies. For a thorough description of our datasets and a preliminary discussion on the data, see [57]; for additional details on Symantec's attack data we refer the reader to [47].

The idea of using vulnerability data to assess overall security is not new by itself. Attack surfaces [58] and attack graphs [59] are seminal approaches to the problem: the former uses vulnerability data to compute an "exposure metric" of the vulnerable systems to potential attacks; the latter aims at modeling consequent attacks on a system (or network of systems) the attacker might perpetrate to reach a (usually critical) component such as a data server. These approaches however lack of a characterization of risk or "likelihood of exploit". Our methodology integrates these approaches by providing sound risk estimations for vulnerabilities; our results can be plugged in both attack graphs and attack surface estimations to obtain more realistic assessments.

CVSS An analysis of the distribution of CVSS scores and subscores has been presented by Scarfone et al. [44] and Gallon [60]. However, while including CVSS subscore analysis, their results are limited to data from NVD and do not provide any insight on vulnerability exploitation. In this sense, Bozorgi et al. [42] were probably the first to look for this correlation. Unfortunately, they showed that the CVSS characterization of "likelihood to exploit" did not match with data on proof-of-concept exploits in EDB. We extended their first observation with a in-depth analysis of subscores and of actual exploitation data.

Vulnerability models Other studies focused on the modelling of the vulnerability discovery processes. As noted by [61], vulnerability models can help "*security engineers to prioritize security inspection and testing efforts*" by, for example, identifying software components that are most susceptible to attacks [62] or most likely to have unknown vulnerabilities hidden in the code [63]. Our contribution differs, in general, from work on vulnerability models in that

we do not aim at identifying “vulnerable components” or previously unknown vulnerabilities to point software engineers in the right direction. We instead propose a methodology to evaluate the risk of already known vulnerabilities that might (or might not) be exploited in the wild, and therefore may need immediate remediation or mitigation on the deployment side rather than on the development side. We find necessary to cover this part of the literature as “vulnerability discovery” necessarily lays the ground for the “vulnerability remediation” process that is the focus of our work.

Alhazmi et al.’s [64] and Ozment’s [65] work are both central in vulnerability discovery models research. Alhazmi et al. fit six vulnerability models to vulnerability data of four major operative systems, and show that Alhazmi’s ‘S shaped’ model is the one that performs the best. [61] suggest that vulnerability models might be substituted with fault prediction models, and showed that performances in terms of “recall” and “precision” do not differ sensibly between the two. However, as previously underlined by Ozment [65], vulnerability models may rely on unsound assumptions such as the independence of vulnerability discoveries. Current vulnerability discovery models are indeed not general enough to represent trends for all software [66]. Moreover, vulnerability disclosure and discovery are complex processes [67, 68], and can be influenced by {black/white}-hat community activities [68] and economics [40].

Markets for vulnerabilities Our analysis of vulnerabilities traded in the black markets is also interesting because it supports the hypothesis that the exploit markets are significantly different (and more stable) than the previous IRC markets frequented by cyber criminals were [69]. Previous work from the authors of this manuscript also experimentally showed that the goods traded in the black markets are very reliable in delivering attacks and are resilient to aging [55].

ANNEX2.8 Conclusion

In this paper we have proposed to use for security research the case-control study methodology.

In a case-control study the researcher looks backward at some the *cases* (for example vulnerabilities exploited in the wild) and compare them with *controls* (in our cases randomly selected vulnerabilities with similar characteristics such as year of discovery or software type). The purpose is to identify whether some *risk factor* (in our scenario a high CVSS score, or the existence of a proof of concept exploit) is good explanation of the cases and therefore represents a decision variable upon which system administrator must act upon.

The effectiveness of different risk factors can be compared by looking at the (statistical) *risk reduction*: the difference between the probability that a vulnerability with a high risk factor is exploited and the probability that a vulnerability with a low risk factor is exploited. Acting first on vulnerabilities whose risk factor has the highest risk reduction would then be the most effective strategy.

To illustrate the methodology we analyzed the CVSS score as international standards like PCI DSS and best practices like those identified by the NIST SCAP Protocol suggest to use it: as if it was a risk factor.

We first dug into the characteristics of the CVSS score to see whether it features an evaluation of the Impact of the undesired event and of its likelihood to happen. While the CVSS Impact assessment shows sufficient variability in its variables, its likelihood (i.e. Exploitability) metric does not. The lack of a characterization of likelihood-of-exploit in the CVSS metric, combined with a poor impact variability, make the CVSS score an unlikely risk metric.

To check this more formally we evaluated the CVSS score by performing a case-control study, in which we sample the data at hand to test how the CVSS score correlates with exploitation in the wild. Our results show that the CVSS base score never achieves high rates of identified true positives (sensitivity) simultaneously with a high rate of true negatives (specificity). Specificity is particularly unsatisfactory when using the CVSS thresholds indicated by standards and best practices alike.

Finally, we showed how to improve the analysis by considering additional risk factors such as existence of a proof-of-concept exploit, and existence of an exploit in the black markets. Our results show that markedly higher risk reductions can be obtained by considering risk factors other than the mere CVSS. For example, addressing vulnerabilities with a known proof-of-concept and a high CVSS score could yield a risk reduction of exploit comparable to the risk reduction in mortality obtained by wearing safety belts in cars.

Our methodology is fully reproducible and can be used as a tool for risk assessment by practitioners and researchers alike. Other datasets or any risk factors could be plugged into the model to fit company-specific scenarios.

For future work we plan to integrate our methodology with additional evaluation factors such as the cost of a strategy or the criticality of the assets. Another interesting venue would be to apply our methodology to other domains (e.g. critical infrastructures and targeted attacks).

There is a general issue behind the use of a case-control study. Is it as an appropriate scientific instrument to assess security? There are several trade- offs. A case control study is based on statistical evidence and therefore cannot offer a 100% security proof that a formal method offers. At the same time, being based on actual data, it is not prone to the security holes introduced by the gaps between a formal theory and its implementation. On the positive side, it avoids the ethical issues of randomized trials, as one cannot ask users to stay vulnerable[16]. On the negative side, it has less power to determine causality than controlled experiments because it looks backward. Can we prove that our conclusions will be applicable beyond 2013 (when our dataset stops)? We can't, in the same way Bradford Hill could not and cannot prove that their conclusion about smoking and lung cancer would apply beyond 1960 and a handful of English hospitals, a tiny speck against the world population. We can only argue that, due to the way the experiment is constructed, there is no apparent alternative explanation. Yet, if the methodology is accepted, it can be used by other scientists on other data sources and on other risk factors. Many of the risk factors we consider (such as CVSS, ExploitDB, etc) are the de- facto standards in industry, generating a multi-million business (a casual walk among the stands of BlackHat or RSA vendors would make it immediate). Many academics and industry experts grumble that these metrics are wrong, but offer no evidence. The usage of case-control studies can be a sound scientific method to evaluate those risk factors by using the very data that industry has available.

[16]Using honeynets for experiments would not give a controlled experiment either as they are artificial and not actually used.

Confirmed experiments could possibly lead to a community consensus that will replace the current industry witchcraft.

ANNEX3. An Experiment on Comparing Two Risk-Based Security Methods

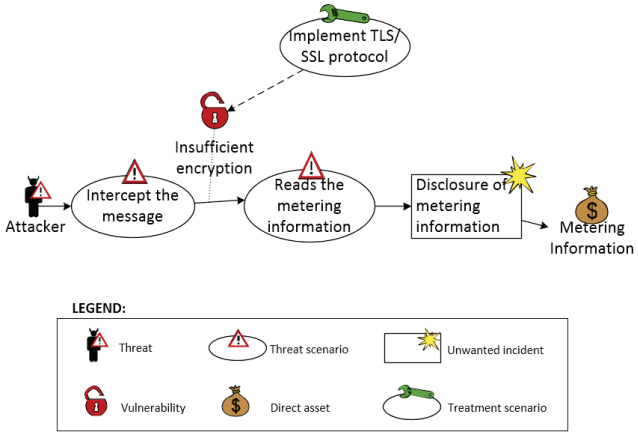
Many security risk assessment methods have been proposed both from academia and industry. However, little empirical evaluation has been done to investigate how these methods are effective in practice. In this paper we report a replication of a controlled experiment that we conducted to compare the *effectiveness* and participants' *perception* of visual versus textual methods for security risk assessment. As instances of the methods we have selected CORAS, an academic method by SINTEF used to provide security risk assessment consulting services, and SecRAM, a method by EUROCONTROL used to conduct security risk assessment within air traffic management domain. The experiment involved 29 MSc students who applied both methods to an application scenario from Smart Grid domain. The dependent variables were *effectiveness* of the methods measured as number of threats and security controls identified, and *perception* of the methods measured through post-task questionnaires based on the Technology Acceptance Model. The main findings are that the visual method has higher effectiveness and participants' perception. These results confirm the original experiment's findings on methods' perception but not the one on effectiveness.

ANNEX3.1 Introduction

Many security risk assessment methods, frameworks and standards exist - ISO 27005, NIST 800-30, STRIDE, CORAS, SREP - but they all face similar problems in practice. The security risk assessment process looks easy on paper - but it can turn into a complex and daunting task.

Despite the crucial role that security risk assessment plays in the identification of security controls, only few papers [70, 71, 72] have investigated which methods work better to identify threats and security measures and why. Evaluating a security risk assessment method is challenging because it includes a number of confounding variables: the type of training received (e.g. all papers on the ISACA journal reports methods applications by the method's expert), the previous expertise (e.g. student vs. practitioners is a key distinction here), the time allotted to the task, and the presence of three essential steps of the analysis (assets, threats and security measures identification depends on each other) so if one is badly performed the others may be poor as well.

In this paper we report on the replication of an experiment [70] we conducted to compare *effectiveness*, *perceived easy of use*, *perceived usefulness* and *intention to use* of visual versus textual methods for security risk assessment. In the original experiment, we selected CORAS [73] and SREP [74] as in instances of visual and textual methods respectively. CORAS is a *visual method* whose analysis is supported by a set of diagrams that represent assets, threats, risks and treatments. In contrast, SREP is a *textual method* whose artifacts are specified in tabular form. The original experiment has shown that there was no difference between the two methods in the number of threats, while the textual method was slightly



(a) CORAS - Threat Diagram

Threat Agent	Asset attacked	Attack Likelihood	Justification
Prosumer	Smart Meter	Occasional	He can send traffic the SM at willing since it connected to the HAN. The interest in compromised the SM is also quite high since he could alter energy consumption data and therefore pay less.
Prosumer	Data Communication Network	Remote	The interest in compromised the DCN is high since he could perform subsequent attacks in order to alter ABD and therefore pay less. However, in order to do that the prosumer must physically tap the power line.
Compromised Meter Point Operator	Billing Data	Remote	Even having access to the BD of the prosumer, the MPO could not make much out of it.
Hacker	Energy Management System, Smart Appliance	Probable	The asset is exposed on the Internet and could be targeted by random (i.e. random target) attacks. An hacker has interest in penetrating the system in order to install malware or "just for fun".
Hacker	Energy Supplier Server	Probable	Considering that the hacker could also have the role of prosumer, the IP address of the target could be obtained by capturing traffic carrying BDF between the EMS and the ESS. The data of many prosumer's on the ESS is at stake and a successful attack could give him fair media attention.
Hacker	Data Concentrator	Probable	The DC is exposed on the Internet and an hacker has fair interest in compromising it in order to gain fair media attention.

(b) SECRAM - Threat Agent Table

Figure 15: Examples of Visual (CORAS) and Textual (SecRAM) Methods' Artifacts Generated by Participants.

more effective for eliciting security controls. In addition, visual method overall perception and intention to use were higher than for the textual method. In this replication we replaced SREP with SecRAM, a method used by EUROCONTROL to conduct security risk assessment in the air traffic management domain. We involved 29 participants: 15 students of the MSc in Computer Science and 14 students of the EIT ICT LAB MSc in Security and Privacy of the University of Trento. Each participant applied both methods to identify threats and security controls for different facets (Network security and Database/Web application security) of a Smart Grid application scenario. The experiment was complemented with participants' interviews to explain possible differences in the effectiveness and the perception of the two methods.

The main findings on effectiveness are that visual method produces an higher number of threats and security controls than the textual one. While the original experiment has shown that the textual method leads to identify more security controls than the visual one. With respect to participants' perception we found that the visual method is preferred over the textual one with statistical significance. Thus only the results on perception from the original experiment are confirmed.

This paper is structured as follows. In the next section we discuss related works and then we present the design and execution of the experiment. The core of the paper reports on the analysis of the participants' reports, the post-task questionnaire, and the interviews. Then, the main findings are summarized and compared with the one of the original experiment and the threats to validity are discussed. Finally, we present conclusions and future work.

ANNEX3.2 Related Work

The few papers [71, 75, 76, 77, 78] that have attempted to evaluate if security risk assessment methods work in practice have adopted the Method Evaluation Model (MEM) [79] which

provides constructs to measure methods success. For example, Opdhal and Sindre [71] have carried out two controlled experiments (28 and 35 students) to compare two methods for threats identification, namely attack trees and misuse cases. In [78] Opdhal and colleagues have repeated the experiment with industrial practitioners. Both experiments show that attack trees help to identify more threats than misuse cases. Similar controlled experiments with students were reported by Stålhane et al. in [80, 76, 77, 75] where misuse cases are compared with other approaches for safety and security. In [80] Stålhane et al. report an experiment with 42 students where they compared misuse cases to Failure Mode and Effects Analysis (FMEA) to analyze use cases. They find that misuse cases are better than FMEA for analyzing failure modes related to user interactions. In a comparable setting [76], the authors compared misuse cases based on use-case diagrams to those based on textual use cases. The results of the experiment with 52 students show that textual use cases produces better results due to more detailed information. The e-RISE challenge organized by the University of Trento [81] also report an interesting protocol to perform empirical comparisons of different security and risk assessment methods by using both practitioners and students. More recently, Labunets et al. [70] have conducted a controlled experiment with 28 MSc students to compare two classes of risk-analysis driven methods, visual methods (CORAS) and textual methods (SREP). The experiment we report in this paper is a replication of the experiment of Labunets et al.

Most of these experiments have some limitations. Experiments with students such as [71, 75, 76, 77] usually have a short duration (less than two hours) and this may introduce bias in the evaluation of methods because subjects do not have enough time to understand the application scenario and to fully apply the methods under evaluation. Further, if the time for the execution of the experiment is short, it is impossible to use a realistically-sized application scenario. Hence, the methods under evaluation are applied to toy scenarios and the results might not generalize to real-world cases. The experiments in [81, 78] are a good compromise between experiment's cost on one side and scenario's complexity and participants' experience on the other (they last several days and includes practitioners). Yet, they have only focused on academic security methods so far.

The long running experiment of Labunets et al. [70] and its replication that we report in this paper addresses the issue of realism and the full application of the method, yet they miss practitioners.

ANNEX3.3 Research method

This section describes the design of the performed experiment, following the guidelines by Wohlin et al. [82].

ANNEX3.3.1 Research Questions

The *goal* of the experiment was to compare visual and textual methods for security risk assessment with respect to how successful they are in identifying threats and security controls. For this purpose we have adopted as dependent variables the success constructs defined in the Method Evaluation Model (MEM) proposed by Moody [79]: *effectiveness*, *perceived*

Variable	Scale	Means	Distribution
Gender	Sex		79% were male; 21%were female
Age	Years	25.72	48% were 21-24 years; 41% were 25-29; 10% were 30-40
Education Length		4.28	66% had <5 years; 17% had 5 years; 17% had >5 years
Work Experience		2.46	31% had no experience; 31% had < 2 years; 28% had 3-5 years; 10% had >6 years
Level of Expertise in Security Technology	1(Novice)-5(Expert)	2.31	28% novices; 28% beginners; 10% competent users; 31% proficient users; 3% experts
Level of Expertise in Security Regulation and Standards		1.86	45% novices; 17% beginners; 7% competent users; 31% proficient users
Level of Expertise in Privacy Technology		2.10	31% novices; 34% beginners; 28% competent users; 7% proficient users
Level of Expertise in Privacy Regulation		1.90	48% novices; 24% beginners; 7% competent users; 21% proficient users
Level of Expertise in RE		2.31	24% novices; 34% beginners; 14% competent users; 28% proficient users

Table 18: Demographic Statistics

easy of use, *perceived usefulness*, and *intention to use*. Therefore, we have specified the following research questions that match the constructs of the MEM:

RQ1 *Is the effectiveness of the methods significantly different between the two types of methods?*

RQ2 *Is the effectiveness of the methods significantly different between the two facets?*

RQ3 *Is the participants' overall perception of the method significantly different between the two type of methods?*

RQ4 *Is the participants' perceived usefulness of the method significantly different between the two type of methods?*

RQ5 *Is the participants' perceived ease of use of the method significantly different between the two type of methods?*

RQ6 *Is the participants' intention to use the method significantly different between the two type of methods?*

We have translated research questions $RQ1 - RQ6$ into a list of null hypotheses to be statistically tested. We do not list them here due to the lack of space. To answer $RQ1$ and $RQ2$ we have measured methods' *actual effectiveness* by counting the number of threats and security controls identified with each method application and we asked an external security expert to assess their quality. Research questions $RQ3$-$RQ6$ have been answered by administering to the participants a post-task questionnaire inspired to the Method Evaluation Model (MEM) [79] after they have completed each of the method applications. To gain a better understanding *why there is a difference in methods effectiveness and perception* we have conducted individual interview with participants.

ANNEX3.3.2 Methods Selection

As in our previous experiment [70], we have chosen as instance of the visual method CORAS [73] because is the only visual method for security risk assessment. CORAS is a method de-

signed at SINTEF, a research institution in Norway which is used to provide security risk assessment consulting services. It consists of three tightly integrated parts, namely, a method for risk analysis, a language for risk modeling, and a tool to support the risk analysis process. The risk analysis in CORAS is a structured and systematic process which uses diagrams (see Figure 15a) to document the result of the execution of each step. The steps are based on the international standard ISO 31000 [83] for risk management: context establishment, risk analysis (that identifies assets, unwanted incidents, threats and vulnerabilities), and risk treatments. Instead, we have replaced SREP, the instance of textual method used in the original experiment, with SecRAM [84], an industrial method by EUROCONTROL used to conduct security risk assessment in the air traffic management domain (ATM). SecRAM supports the security risk management process for a project initiated by an air navigation service provider, or ATM project, system or facility. SecRAM provides a systematic approach to conduct security risk assessment which consists of five main steps: defining the scope of the system, assessing the impact of a successful attack, estimating the likelihood of a successful attack, assessing the security risk to the organization or project, and defining and agreeing a set of management options. As shown in Figure 15b) tables are used to represent the results of each step's execution.

ANNEX3.3.3 Domain Selection

We selected the Smart Grid application scenario for our experiment as we had already used in the previous experiment so that we could compare the results from the two experiments. The Smart Grid is an electricity network that uses information and communication technologies to optimize the distribution and transmission of electricity from supply points to end-consumers. The application scenario focused on the gathering of metering information from the smart meters located in private households and its communication to the electricity supplier for billing purposes.

ANNEX3.3.4 Demographics

The participants of the experiment were recruited among MSc students enrolled in the Security Engineering course at the University of Trento. Table 18 presents descriptive statistics about the participants. Most of the participants (69%) reported that they had at least 2 years of working experience while the remaining said they had no working experiences. With respect to knowledge in privacy technologies and regulations, most of the participants had limited expertise. In contrast, they reported an extensive general knowledge of both security technologies and regulations and standards. Participants also reported good general knowledge in requirements engineering.

ANNEX3.3.5 Experimental design

We chose a within-subject design where all participants apply both methods to ensure a sufficient number of observations to produce significant conclusions. In order to avoid learning effects, the participants had to identify threats and security controls for different types of security facets of a Smart Grid application scenario. The security facets included Network

Table 19: Original experiment and replication settings

	Original	Replication
Subject Type	28 MSc students	29 MSc students
Subject Unit	16 groups of 1-2 students	29 Groups of 1 student
Subject Environment	Security Engineering course	Security Engineering course
Experiment Task	Identify threats & controls	Identify threats & controls
Time to complete the task	4 sessions of remote work	2 sessions of remote work
Experiment Design	Two factors (2 methods, 4 facets))	Two factors (2 methods, 2 facets))
Experiment Group	visual vs textual	visual vs textual
Variables	EFFECT, PEOU, PU, ITU	EFFECT, PEOU, PU, ITU

Facet/Method	Visual	Textual
Network Security	14	15
DB/Web App. Security	15	14

Table 20: Experimental design

Security (Network) and Database/Web Application Security (DB/WebApp). For example, for Network Security facet, participants had to identify network security threats like man-in-the-middle attack or DoS attack and proposed security controls to mitigate them.

The participants were randomly assigned to treatments: half of the participants applied first the visual method to network security facet while the second half applied the methods in the opposite order. Table 20 summarizes how the participants has been assigned to the methods.

ANNEX3.3.6 Experimental procedure

The experiment was performed during the Security Engineering course held at University of Trento from September 2013 to January 2014. The experiment was organized in three main phases:

Training. Participants were given a tutorial on the Smart Grid application scenario and a tutorial on visual and textual methods of the duration of two hours each. Then, participants were administered a questionnaire to collect information about their background and their previous knowledge of other methods and they were assigned to facets based on the experimental design.

Application. Once trained on the Smart Grid scenario and the methods, the participants had to repeat the application of the methods on two different facets: Network and DB/WebApp. For each facet, the participants:

- Attended a two hours lecture on the threats and possible security controls specific for the facet but not concretely applied to the scenario.
- Had 2,5 weeks to apply the assigned method to identify threats and security controls specific for the facet.
- Gave a short presentation about the preliminary results of the method application and received feedback.
- Had one week to deliver an intermediate report to get feedback.

At the end of the course in mid January 2014, each participants submitted a final report documenting the application of the methods on the two facets.
Evaluation. In this phase, the experimenters (the authors of this paper) assessed participants final reports while the participants evaluated the method through questionnaires and interviews. After each application phase the participants answered an on-line post-task questionnaire to provide their feedback on method application. In addition, after final report submission each participant was interviewed for half an hour by one of the experimenters to investigate which are the advantages and disadvantages of the methods. Then, at the end of January each participant gave a presentation summarizing their work in front of the experimenters and an expert in security for Smart Grid. The expert evaluated the quality of the threats and security controls delivered by the participants for the Smart Grid application scenario.

The interview guide contained open questions about the overall opinion of the methods, whether the methods help in identification of threats and security controls and about methods' possible advantages and disadvantages. The interview questions were the same for all the interviewees. The post-task questionnaires include the same questions of the one we administered for our previous experiment which was inspired to the Technology Acceptance Model (TAM) [85]. To avoid that the participants answered on "auto-pilot", 15 out of 31 questions were given with the most positive response on the left and the most negative on the right. The interview guide and the post-task questionnaire are reported in [86].

ANNEX3.3.7 Changes to the Original Experiment

The experiment reported in this paper differs from the original experiment in that the participants were asked to work individually than in pairs in order to correlate their performance with their perception of the two methods. In addition, we reduced the focus of security risk assessment only to Network security an Database/Web application security to increase the application time provided to the participants. In fact, in the original experiment, participants reported that the time for methods application was short. The main differences are reported in Table 19.

ANNEX3.4 Quantitative analysis

In this section we report the results from the analysis of the final reports delivered by the participants and of the participants' answers to the post-task questionnaires.

ANNEX3.4.1 Reports' Analysis

To assess the effectiveness of visual and textual methods, we reviewed the final reports delivered by the participants to count the number of identified threats and security controls and we assessed their quality. We followed the same coding process used in our previous experiment [70].

Quality of results Since a method is effective based not only on the quantity of results but also on the quality of the results that it produces, we asked a domain expert in Smart

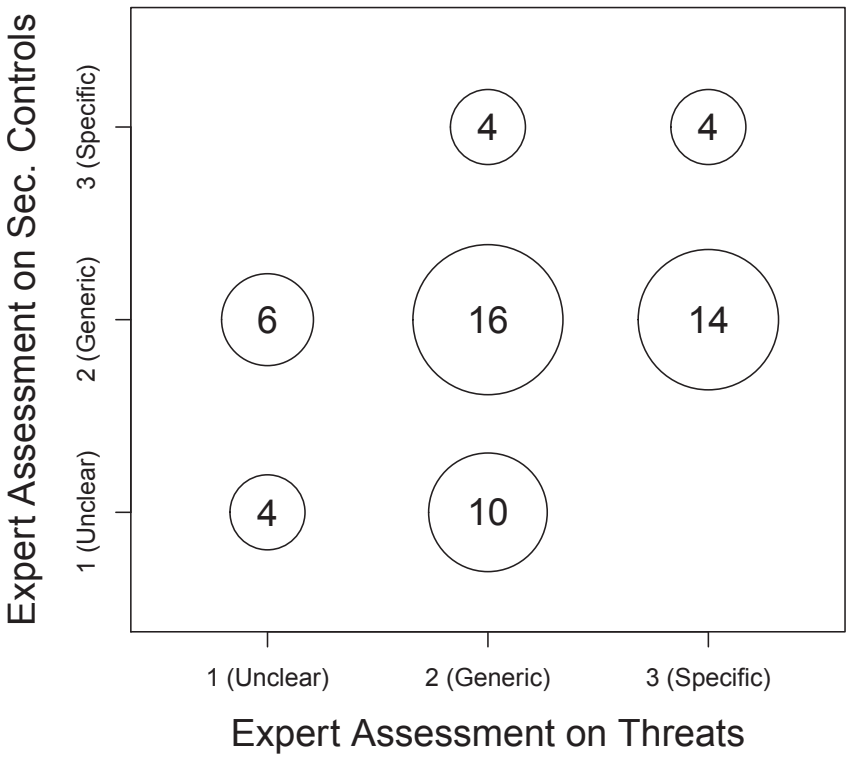


Figure 16: Expert assessment of threats and security controls for the two facets.

Grid security to evaluate the overall quality of the identified threats and security controls for Network security, and Database/Web application security of the smart metering application scenario. In fact if we consider only the number of results but not the quality, threats to conclusion validity may arise. To evaluate the quality of threats and security controls we used a four item scale: *Unclear* (1), *Generic* (2), *Specific* (3) and *Valuable* (4). Based on this scale the participants who achieved an assessment *Specific* or *Valuable* were classified as *good participants*.

Figure 16 illustrates the expert evaluation of all participants for the Network security facet and Database/Web application security facet. As each participant applied one of the methods on the two facets, there are 58 method applications in total. The number inside each bubble denotes the number of method applications which achieved a given expert's assessment for threats (reported on x-axis) and security controls (reported on y-axis). There were 14 method applications that generated some specific threats but generic security controls, while 4 method applications delivered both threats and security controls specific to the scenario. Additional 4 method applications delivered specific security controls but generic threats. The remaining method applications delivered unclear and/or generic threats and security controls. Overall, most of the method applications produced clear threats and security controls but they were generic.

Number of threats and security requirements As the design of our experiment is two factor (the method and the facet) block design, we could use two-way ANOVA test or Friedman test (non-parametric analog of ANOVA) to analyze the number of threats and security controls identified with each method and within each facet. To select a right test we checked whether our samples satisfy ANOVA's assumptions: 1. observations independence, 2. sphericity, and homogeneity of variance 3. normality of distribution of samples.
Observation Independence. We have *observation independence* by design because participants' worked individually. This gave us independence within sample and mutual independence within sample as the facets were different.
Sphericity and Homogeneity of Variance. Sphericity holds for our samples due to one degree of freedom for each factor [87]. We also checked the homogeneity of variance with Levene's test. This test returned p-value equal to 0.24 for threats and 0.46 for security controls. Therefore, we can assume homogeneity of variance for our samples.
Distribution Normality. To check this assumption we used Shapiro-Wilk normality test. This test returned p-value less than $1.75 \cdot 110^{-3}$ for threats and $9 \cdot 10^{-8}$ for security controls. So, normality assumption does not hold for our samples.

Because last assumption has been rejected we need to use non-parametric analog of ANOVA, Friedman test.

First, we analyzed the differences in the number of threats identified with each method. As shown in Figure 17 (left), if we consider all participants, there is no difference between the number of threats identified with the visual method and textual one. For the good participants visual method performed slightly better than textual method as shown in Figure 17 (right). This is also confirmed by the Friedman test which does not show any significant differences in the number of threats identified by all participants (*p-value* = 0.57), but reveals a statistically significant effect of methods on number of threats identified by good participants (*p-value* = $8 * 10^{-3}$).

Similarly, Figure 18 compares the means of the number of security security controls identified by all participants (left) and good participants (right). For all participants we can see that visual and textual methods produce the same number of security controls, while for good participants the visual method is better than the textual one in identifying security controls. This is attested also by the results of Friedman test which shows there is no statistically significant difference in the number of security controls identified by all participants (*p-value* = 0.57) while the difference is statistically significant for good participants (*p-value* = 0.012).

We have also investigated the differences in the number of threats and security controls identified with the visual and the textual method within each facet. The boxplots in Figure 19 shows that both for all and good participants there is no difference in the number of threats produced with the visual and textual method for DB/Web Application facet, while for Network security facets the number of threats identified with the visual method is higher. The results of Friedman test show that the difference in the number of threats between the two facets is not statistically significant for all participants (*p-value* = 0.85) while it is significant for good participants (*p-value* = $3.7 * 10^{-4}$).

Figure 20 reports the number of security controls identified with the visual and the textual method within each facet. If we consider all participants, there is no difference in the number of security controls identified with the two methods for the DB/Web Application facet, while the number of security controls identified with the visual method is higher for the Network

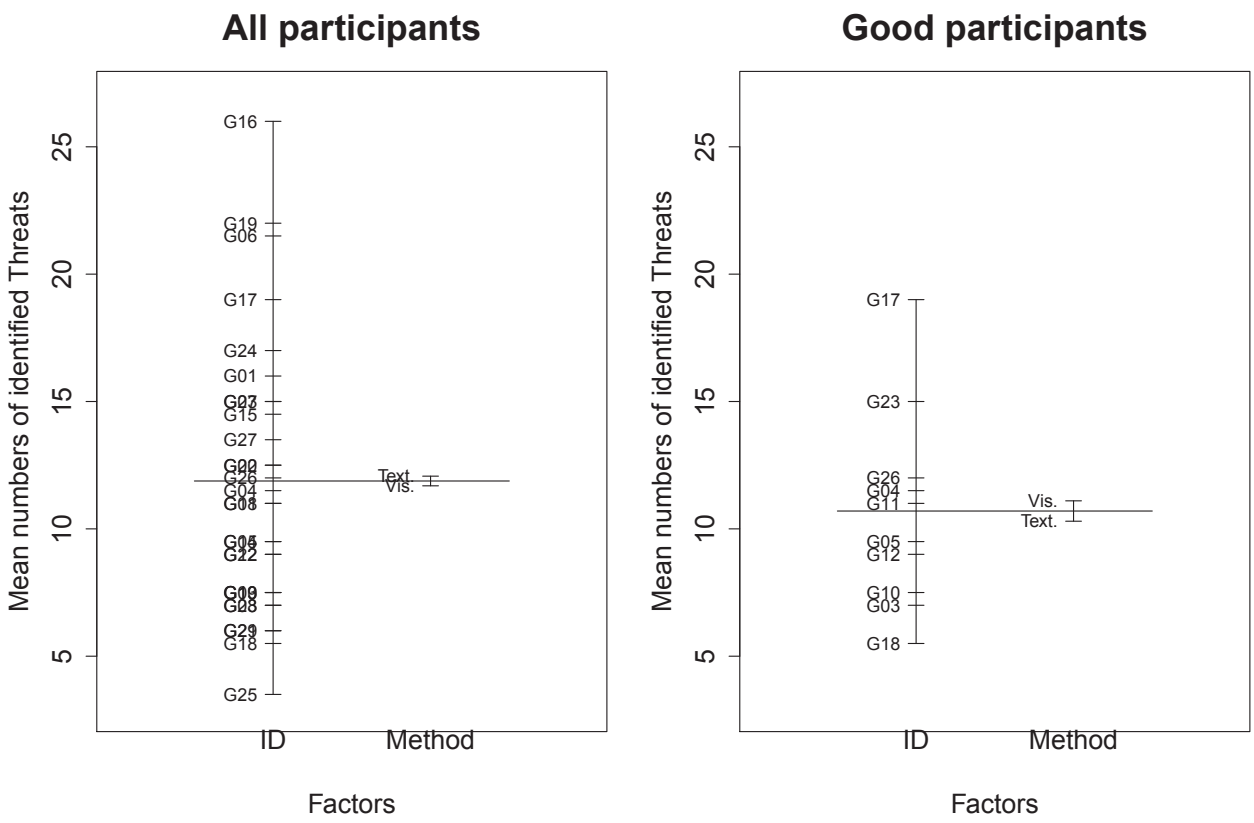


Figure 17: Means of identified threats by all participants (left) and good participants (right).

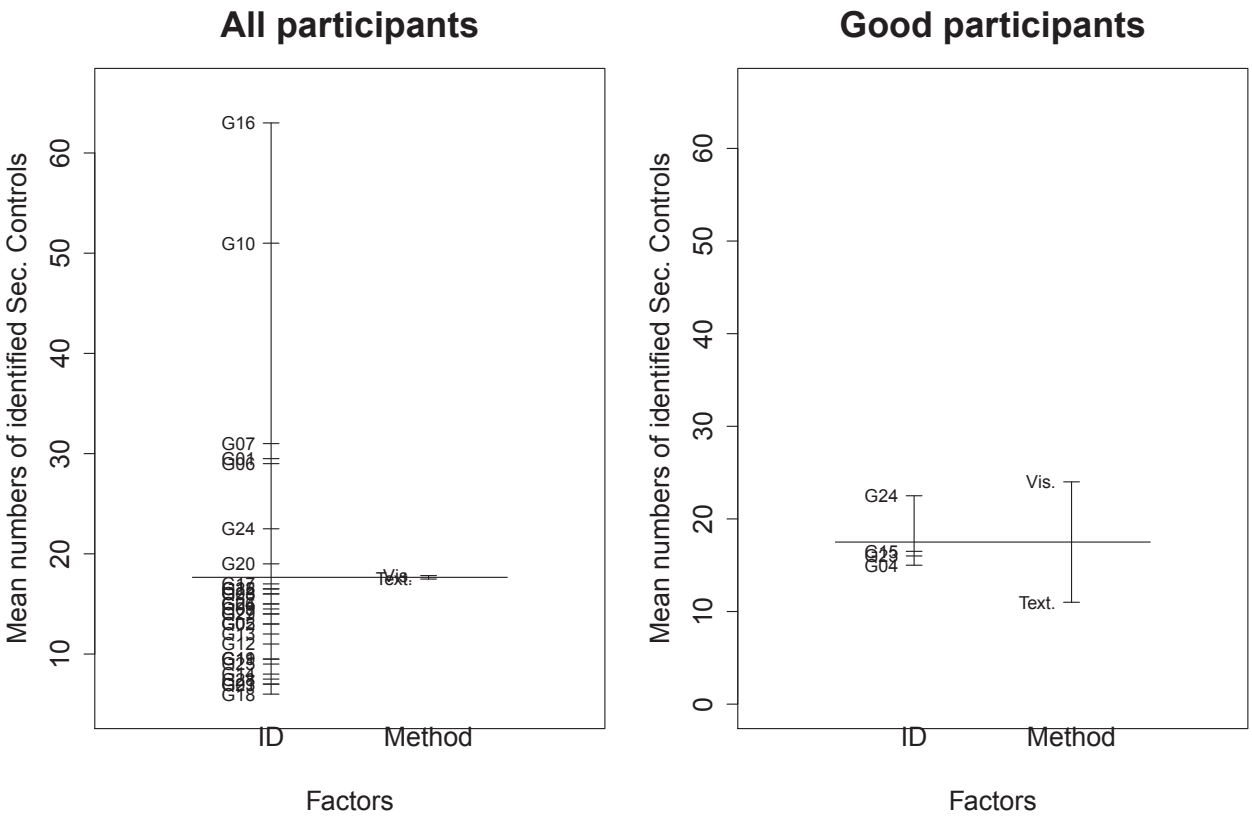


Figure 18: Means of identified security controls by all participants (left) and good participants (right).

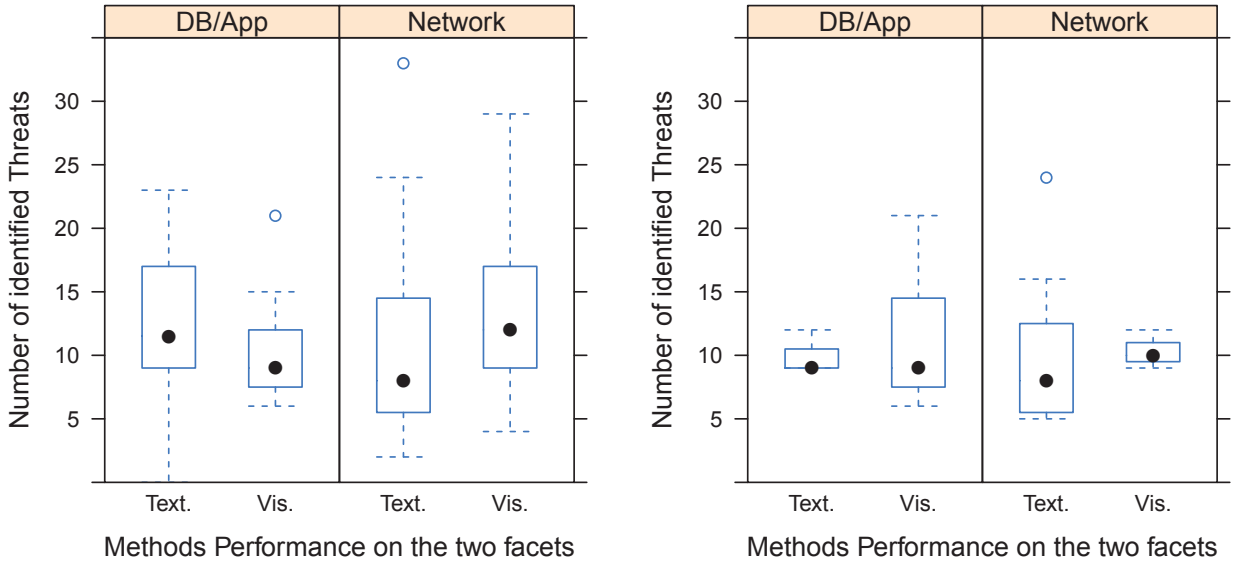
security facet. For good participants, instead, the visual methods leads to identify more security controls in both facets. Friedman test shows that the difference in the number of security controls between the two facets is not statistically significant for all participants (*p-value* = 0.08) while it is significant for good participants (*p-value* = $6.3 * 10^{-3}$).

To compare the results on methods' effectiveness with the one of the original experiment, we run Friedman test also on the number of threats and security controls produced in the original experiment. The main difference we found is that the original experiment showed there is no statistically significance difference in the number of threats identified with the two methods, while in this replication we found there is a statistically significance difference.

Questionnaire Analysis

The post-task questionnaires have been analyzed to identify the difference in participants perception of two methods. Before conducting analysis all responses have been reverted to 5 being the best. The questions were formulated in opposite statements format with answers on a 5-point Likert scale. As the data are ordinal, the responses are paired and had ties, we have used the exact Wilcoxon signed-ranks test with Wilcoxon method for handling ties. The significance level α is set to 0.05.

Table 21 presents the results of questionnaires' analysis and compare them with the



(a) All participants

(b) Good participants

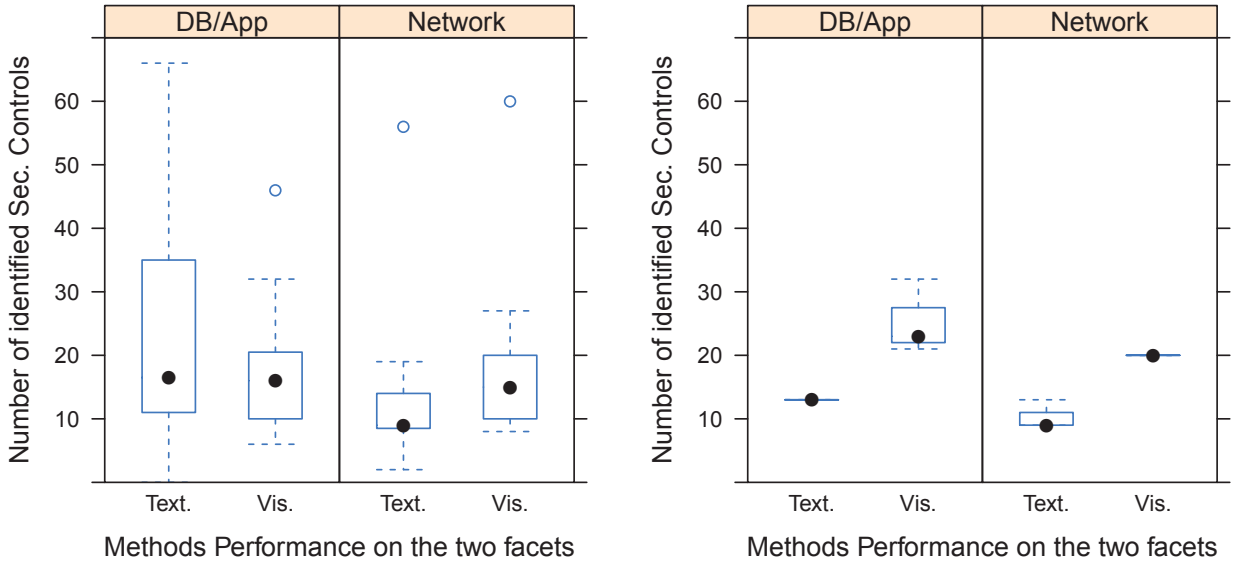
Figure 19: The distribution of identified threats for all participants (left) and good participants (right).

one from the original experiment. For each question, the table reports to which perception variable the question refers to (PEOU, PU, ITU), the mean of the answers by all and by good participants (the one who produced good quality threats and security controls based on expert's assessment), and Z statistics returned by the Wilcoxon test and the level of statistical significance based on the p-value returned by the test. The level of statistical significance is specified by • (p<0.1), or * (* p<0.05, ** p<0.01, *** p<0.001). For the questions that were asked in both experiments (top part of Table 21) we also report level of statistical significance.

Perceived Easy of Use. Visual method is better than the textual method with respect to overall PEOU across all participants and good participants and the difference is statistically significant. Wilcoxon test returned: Z = -5.4, p = $9.4 * 10^{-9}$, effect size (es) = 0.36 (for all subjects), and Z = -3.9, p = $2.6 * 10^{-5}$, es = 0.42 (for good subjects). The results of the original experiment also showed preference for visual method but there was no statistical significance for all participants and only 10% significance level for good participants.

Perceived Usefulness. Visual method is better than the textual one with respect to overall PU across all participants and good participants. Both all and good participants show a statistically significant preference. Wilcoxon test returned: Z = -4.8, p = $6.3 * 10^{-7}$, es = 0.26 (all), and Z = -4, p = $2.2 * 10^{-5}$, es = 0.35 (good). The results of the original experiment also showed preference for visual method but there was no statistical significance for all participants and only 10% significance level for good participants.

Intention to Use. Visual method is better than the textual one with respect to overall ITU across all participants with statistical significance. This is also true for good participants. Wilcoxon test returned: Z = -3.7, p = $2.1 * 10^{-4}$, es = 0.17 (all), and Z = -2.8, p = $4.7 * 10^{-3}$,



(a) All participants

(b) Good participants

Figure 20: The distribution of identified security controls for all participants (left) and good participants (right).

es = 0.21 (good). The results of the original experiment also showed preference for visual method but it was statistically significant only for good participants.

Overall Perception. The average of responses shows that participants preferred the visual method over the textual method with statistical significance for both all participants and good participants. Wilcoxon test returned: Z = -7.8, p = $8.7 * 10^{-16}$, es = 0.24 (all), and Z = -6.1, p = $2.5 * 10^{-10}$, es = 0.3 (good). The results of the original experiment showed preference for visual method but there was only 10% significance level for all participants. Instead, among good participants the preference for visual method was statistically significant.

ANNEX3.5 Qualitative analysis

In this section we report the results of the analysis of individual interview with participants. The interviews were transcribed and analyzed by two researchers independently using *coding* [88], a qualitative analysis method from grounded theory. The list of core codes was taken from analysis of previous experiments [70, 81].

Table 22 reports the positive and negative aspects of visual and textual method that may affect PEOU and PU while Table 23 illustrates *other* aspects that may influence methods' success. For each aspect we report the total number of statements made by participants as relative indicator of its importance. We report here only the aspects for which at least 10 statements were made by participants.

Perceived Ease of Use. The main aspect influencing PEOU of visual method is that it provides a *visual summary* of the results of the security analysis (29% of the positive state-

Q	Type	All subjects Mean Tex	Mean Vis	$Z_{repl.}$	$Z_{orig.}$	Good subjects Mean Tex	Mean Vis	$Z_{repl.}$	$Z_{orig.}$
Same questions asked in original experiment									
4	PU	3.1	3.8	-2.4 *		2.6	3.8	-2.2 *	
5	PU	3.2	3.6	-1.6		2.9	3.5	-1.2	
6	PEOU	2.9	3.9	-2.9 **		2.8	4	-2 ●	
7	PEOU	2.9	3.7	-2.6 **		2.5	3.5	-1.8 ●	-2.1 *
8	PU	3.6	4	-1.6		3	3.8	-1.6	-2.3 *
9	PEOU	2.8	3.8	-3.3 ***		2.7	3.5	-2 ●	
10	PU	3	3.8	-2.5 *	-2.1 *	2.7	3.7	-1.8 ●	
11	PU	2.9	3.5	-2.2 *	-2.1 *	2.7	3.5	-1.6	
15	ITU	3.1	3.7	-1.8 ●		2.7	3.3	-1.1	
16	ITU	3.3	3.4	-0.6		3.2	3.3	-0.3	
19	ITU	3.1	3.8	-2.2 *		2.8	3.5	-1.3	
20	ITU	3.2	3.6	-1.3		2.8	3.4	-1.4	
23	ITU	3.1	3.6	-1.7 ●		2.9	3.4	-1.1	
24	ITU	3.1	3.4	-1.1		2.8	3.4	-1.3	
26	PU	3.1	3.5	-1.5		2.6	3.3	-1.3	
28	ITU	3.1	3.5	-1.3		2.9	3.4	-1.1	
29	ITU	3.2	3.3	-0.2		3	3.3	-0.5	
31	PEOU	3.1	3.9	-2.1 *		2.8	3.9	-2.2 *	
2	Control	3.7	3.9	-0.4	3.1 ***	3.9	3.8	0.6	2.3 *
3	Control	3.8	4.1	-0.8		3.2	4.2	-1.8	
	PU_{repl}	3.1	3.7	-4.8 ***		2.8	3.6	-4 ***	-1.7 ●
	$PEOU_{repl}$	2.9	3.8	-5.4 ***		2.7	3.7	-3.9 ***	-2.0 ●
	ITU_{repl}	3.1	3.5	-3.7 ***		2.9	3.3	-2.8 **	-2.5 *
	$Total_{repl}$	3.1	3.6	-7.8 ***	-1.6 ●	2.8	3.5	-6.1 ***	-3.5 ***
New questions									
1	PU	3.7	4.1	-2.1 *		3.3	4	-1.9 ●	
12	PU	3.2	3.1	0.4		3	3.2	-0.5	
13	PU	3	3	0.2		2.8	2.8	0	
14	PU	3.2	3.4	-1.4		3	3.5	-1.3	
22	PU	3.1	3.4	-1		2.8	3.6	-1.7	
25	PU	3.3	3.7	-2.1 *		2.7	3.5	-2 ●	
27	PEOU	3	3.9	-2.8 **		2.8	3.9	-2.6 **	
30	PEOU	2.9	3.6	-2.8 **		2.7	3.7	-2 ●	
17	Control	2.9	3.7	-3.2 **		2.6	3.6	-1.9	
18	Control	3	3.6	-2.7 **		2.8	3.6	-1.8 ●	
21	Control	3.3	2.7	2 *		2.9	2.6	0.4	
	$PU_{repl+new}$	3.2	3.6	-5.2 ***		2.9	3.5	-5.2 ***	
	$PEOU_{repl+new}$	3	3.8	-6.6 ***		2.7	3.7	-5 ***	
	$ITU_{repl+new}$	3.1	3.5	-3.7 ***		2.9	3.3	-2.8 **	
	$Total_{repl+new}$	3.2	3.6	-8.9 ***		2.9	3.5	-7.8 ***	

● - p-value <0.1, * - p <0.05, ** - p <0.01, *** - p <0.001

Table 21: Wilcoxon Signed-Ranks Test of Responses

ments made by the participants on visual method's PEOU). Examples of these statements: " *there are many summary diagrams which are useful to summarize what has been done*" and " *the advantages is the visualization*". Other noteworthy positive aspects for visual method's PEOUS are that the visual method has *clear process* (19% of positive statements) and that it is *ease to use* (19% of positive statements). About these aspects participants made the following statements: "*I think it is easy to use*" and "*The advantages of CORAS is very clear structure*", respectively. Instead, the main aspects that can affect negatively the visual method's PEOU are that it is a *time consuming* method and it has a *primitive tool* (26% of negative statements). As participants indicated "*the diagrams are really time consuming*" and "*first I tried the CORAS tool. And somehow, it was confusing. So, I switched to the Visio*". Another negative aspect for visual method's PEOU is that the process has *redundant steps* (17% of negative statements): "*I think CORAS has some duplications.*".

PEOU Category	Vis.	Text.	Total
Positive Aspects			
Clear Process	28	18	46
Easy for Customer	13	2	15
Easy to Understand	18		18
Easy to Use	28	22	50
Time effective	7	16	23
Visual summary	43		43
Worked examples	12	4	16
Total Pos PEOU	149	62	211
Negative Aspects			
No Evolution Support	15	2	17
Not easy to Understand	3	11	14
Not easy to Use	6	18	24
Primitive Tool	30		30
Redundant Steps	19	4	23
Time consuming	36	7	43
Unclear Process	4	28	32
Poor worked examples	2	27	29
Total Neg PEOU	115	97	212
Total PEOU	264	159	423
PU Category	**Vis.**	**Text.**	**Total**
Positive Aspects			
Help in Identifying Security Controls	22	16	38
Help in Identifying Threats	39	18	57
Help to Model	10	2	12
Total Pos PU	71	36	107
Negative Aspects			
No Help in Identifying Security Controls	9	16	25
Visual Complexity	17		17
No Tool Support		21	21
Total Neg PU	26	37	63
Total PU	97	73	170

Table 22: Positive and Negative Aspects Influencing PEOU and PU

The participants of the original experiment reported the same positive and negative aspects for visual method with respect to PEOU. They appreciated that the visual method provides a *visual summary* of the results and that it is *easy to use*, but had a negative opinion about *primitive tool*.

The main positive aspect for the textual method's PEOU is *time effectiveness* (26% of positive statements). Typical statement about this aspect was "*I used very little time to do my work*". Instead, there is no consensus among participants about other two aspects: *clear process* and *ease of use*. In fact, participants made a similar number of statements that indicate these aspects as both positive and negative. For example: "*it's quite easy*" (positive statement) or "*it was sometimes a bit confusing how to apply the methodology*" (negative statement).

The main negative aspect (28% of negative statements) impacting textual method's PEOU is related to *poor worked examples* illustrating method application. As participants reported "*the main problem was about the example that it uses - instead of defining in more general way, and you are misguided by this example*".

In the original experiment participants reported as negative aspects for textual method's PEOU that process to follow was unclear and the use of tables to represent threats.

Other Category	Vis.	Text.	Total
Positive Aspects			
Catalog of Sec. Controls	23	31	54
Catalog of Threats	30	29	59
Total Pos Other	53	60	113

Table 23: Other Criteria Influencing Methods Success

Perceived Usefulness. There are two main aspects that could positively affect PU of visual method: *help in identifying threats* (55% of positive statements) and *security controls* (31% of positive statements). Typical statements about these aspects were: "*when you're doing a diagram you can actually see the flaw of the actions and it is easy to identify the threats, the attacks*" and "*I find it good for finding some security requirements and risk*". The negative aspect for visual method PU is that visual notation does not scale well for complex scenarios (65% of negative statements): "*these diagrams are getting very soon, very huge and very complex*".

Also participants of the original experiment complained about the visual notation that does not scale well.

Similarly, the main positive aspect for textual method PU is that it *helps in identifying threats* (50% of positive statements). They made such statements as "*it has detailed steps and helps to identify assets, threat agents and management options*". Instead, there is no consensus among participants about the textual method helping in the *identification of security controls*. In fact, they made equal number of positive and negative statements about this aspect. Here are examples of typical statements made by participants about it: "*After we already known that our system description, the vulnerabilities, the threat or agents is easy to identify the control.*" (positive statement) or " *I can't say that they allow you to find the threat, the security control, whatever you want. It's just a framework to help you.*" (negative statement).

The most significant negative aspect mentioned for textual method's PU is the fact there is no software supporting the execution of the steps of the textual method: "*It is needed because it will save half of the time if the table were generated automatically*" (57% positive statements).

According to the participants of the original experiment textual method *helps in identifying security controls*, but the tabular representation of threats makes it difficult to show the link among assets, threats and security controls, and thus to give a summary of the results of the security analysis.

Other Relevant Aspects. In participants' interview we have also identified other possible aspect influencing methods' success. Participants think that both methods would benefit from availability of catalogs of threats and security controls. Typical statements made by participants were: "*I think that SecRAM could just employ some catalog, I think, by default.*".

Discussion

In this section we present the main findings regarding each of the research questions and compare them with the findings from the original experiment (see Table 24).

Methods' effectiveness. As shown in the previous sections, visual method is more effective in identifying threats and security controls than textual method for good participants.

Table 24: Results of hypothesis testing

Id	Hypotheses	Original	Replication
$H1.1_A$	Difference in the number of threats found with visual and with textual method	NO(*)	YES
$H1.2_A$	Difference in the number of security controls found with visual and with textual method	YES(*)	YES
$H2.1_A$	Difference in the number of threats found with visual and with textual method within each facet	YES(*)	YES
$H2.2_A$	Difference in the number of security controls found with visual and with textual method within each facet	YES(*)	YES
$H3_A$	Difference in the participants preference for visual and textual method	YES	YES
$H4_A$	Difference in the participants perceived ease of use for visual and textual method	MAY BE	YES
$H5_A$	Difference in the participants perceived usefulness for visual and textual method	MAY BE	YES
$H6_A$	Difference in the participants intention to use for visual and textual method	YES	YES

* We re-done statistical analysis on data from original experiment with Friedman test used in this replication

This result is also confirmed if we consider the facets assigned to the *good participants*. Since the difference in the number of threats and security controls identified with the two methods is statistically significant, we can accept the alternative hypotheses $H1.1_A$, $H2.1_A$, $H1.2_A$ and $H2.2_A$. In contrast, in the original experiment $H1.1_A$ was rejected and even if $H1.2_A$ was accepted but textual method performed better in security controls identification rather than the visual one.

Methods' perception. Participants' *overall perception* is higher for visual than for textual method with statistical significance for all and good participants. Alternative hypothesis $H3_A$ of difference in the overall perception of the two methods is thus upheld. The same result holds for *perceived easy of use*, *perceived usefulness* and *intention to use*. Thus, the alternative hypotheses $H4_A$, $H5_A$ and $H6_A$ can be accepted. Similar results were found in the original experiment. The overall perception and intention to use were higher for the visual method, while for perceived usefulness and perceived easy of use there was no evidence to tell if there was a difference between the two methods.

Qualitative Explanation. The different number of threats and security controls identified with visual and textual methods can be likely explained by the differences between the two methods indicated by the participants during the interviews. Diagrams in visual method help participants in identifying threats and security controls because they give an overview of the possible threats (who initiate the threats), the threat scenarios (possible attacks) and the assets, while the identification of threats in textual method is not facilitated by the use of tables because it is difficult to keep the link between assets and threats and the process is unclear. Also, lower effectiveness and perception of textual method can be explained by a poor worked example illustrating method application, and the unavailability of the software that would help to generate a bulk of tables.

Threats to Validity

We discuss here the main types of threats to validity [82]. **Internal validity** is concerned with issues that may falsely indicate a causal relationship between the treatment and the outcome, although there is none. One possible threat to internal validity is related to *par-*

ticipants' background. The familiarity of the participants with the methods evaluated during the experiment is a threat to internal validity. At the beginning of the experiment, we have administered a questionnaire to check the background of the participants and their knowledge of security methods. The questionnaire has shown that all participants had a similar background and had no prior knowledge about visual and textual methods. Another threat is related to possible *bias in the tutorials*. Differences in the methods' performance may occur if a method is presented in a better way than the other. In our experiment we limit this threat by giving the same structure and the same duration to the tutorials on textual and visual methods. Finally, *bias in data analysis* was limited by having the participants' reports coded by the authors of the paper independently. In addition, the quality of the threats and security controls identified by each group was assessed by an expert external to the experiment.

Construct validity concerns generalizing the result of the experiment to the concept and theory behind the experiment. The main threat to construct validity in our experiment is the design of the research instruments: interviews and questionnaires. The questionnaire was designed following TAM with at least six questions for each of the independent variables we wanted to measure: *perceived usefulness*, *perceived easy of use*, *intention to use*. Three researchers independently have checked the questions included in the interview guide and in the questionnaire: therefore we are reasonably confident that our research instruments measured what we wanted to measure.

Conclusion validity is concerned with issues that affect the ability to draw the correct conclusion about the relations between the treatment and the outcome of the experiment. A main threat to conclusion validity is related to how to evaluate the effectiveness of the methods under evaluation. A method is effective based on the quality of the results that it produces. If we consider just the number of results (e.g., number of threats identified) but not the quality, threats to conclusion validity may arise. To mitigate this threat, we have asked an expert in security for Smart Grid to evaluate the results the subjects have produced.

External validity concerns the ability to generalize experiment results beyond the experiment settings. External validity is thus affected by the objects and the subjects chosen to conduct the experiment. The main threat is related to the *use of students instead of practitioners*. We mitigated this threat by using MSc students enrolled in a course on security engineering. This allowed us to rely on students with the required expertise in security and to ensure that they had the same level of knowledge on the subject. Another threat is the *realism of the experimental environment*. Our experiment had the duration of three months rather than two hours like most of the experiment. This allows us to use a realistically-sized application scenario and thus to generalize our results to real-world cases.

Conclusions

With this study we replicated a controlled experiment we conducted to compare the effectiveness and the perception of visual versus textual methods for security risk assessment. The main findings on effectiveness are that visual method produces an higher number of threats and security controls than the textual one. While the original experiment has shown that the textual method leads to identify more security controls than the visual one. With respect to participants' perception we found that the visual method is preferred over the textual one with statistical significance. Thus only the results on perception from the original experiment

are confirmed.

To sum up the intentions for future works, we plan to carry out a replication of this experiment with practitioners in order to increase the validity of our findings. In addition, we will conduct experiments to evaluate the effect that some of the aspects that we identified during interviews analysis have on the effectiveness and perception of the methods.