Cyber Threat Intelligence based Holistic Risk Quantification and Management

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften
(Dr.-Ing.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von
Florian Klaus Kaiser, M.Sc.
geb. in Kirchheim unter Teck

Tag der mündlichen Prüfung:	12. Dezember 2022
Hauptreferent:	Prof. Dr. Frank Schultmann
Korreferent:	Prof. Dr. Marcus Wiens

Abstract

Technology has deeply penetrated modern society and economies and is ubiquitous in daily live. It led to high levels of digitalization, digitization, informatization, automatization, technitization, and intelligentization which enable novel degrees of comfortability and productivity. However, the integration of technology to every part of modern life introduced novel vulnerabilities to society and economies. This is the benefits that are intended to be realized when digitalizing come at the cost of cyberrisks.

Nowadays, cyberrisks are one of the most pressing risks to economic prosperity and the stability of societies. Furthermore, the importance of managing cyberrisks is expected to increase with amplifying permeation of technology in every aspect of modern life. Although the great importance of cyberrisks for society and economies, currently there is a lack in ability to manage these and low capabilities of quantifying these with high reliability leading to significant misallocation of defensive resources.

The dissertation aims to propose *novel methods for quantifying cyberrisks* taking advantage of *cyber threat intelligence* and deliver *decision support for efficiently managing these risks*. Special emphasis is on quantifying and managing cyberrisks in a *holistic* manner taking into account different domains of cyberattacks which are in particular the technical as well as the human and societal domain.

Study A focuses on quantifying cyberrisks on the basis of artefacts from system monitoring and cyber threat intelligence. It proposes a data analytical approach and is directed towards highly precise and automated attack hypothesis generation. The approach contributes to enabling the quantification of the probability of an attack by suggesting the most probable attack given a set of current observables (network monitoring artefacts; real-time attack hypothesis generation). Study B proposes an alternative approach towards quantifying the probability of an attack which is more forward-looking. It takes a game theoretical perspective on cyberthreats using a weighted attack graph. The attack graph is based on cyber threat intelligence. Weighting is consistent to motivation based attacker modelling. The approach derives the most probable attack from the combination attack motivation, available attacks and vulnerabilities.

Study C is directed towards minimizing the limitations of cyberrisk analysis approaches that are based on cyber threat intelligence which is data quality and availability. The work proposes a data scientific approach for predicting and forecasting novel attacks that is able to augment the database (cyber threat intelligence). By applying the approaches presented for attack prediction and forecast the approaches for cyberrisk quantification can gain higher levels of accuracy and gain predictive capability.

Study D focuses on the human and societal domain and aims at delivering insights to understanding different susceptibility of human actors for cyberattacks. It uses a game theoretical model to investigate the influence of different factors on the susceptibility (i.e. stress and workload).

Study E and F propose the use of digital twinning for quantifying the impact of cyberattacks. An exemplary application of cyberrisk quantification for an attack on a hospital is presented in study E. It is motivated by past observed cyberattacks on hospitals and contributes to enabling the quantification of the impact of successful cyberattacks that lead to the unavailability of data (e.g. medical records) or unavailability of machinery that is used for providing medical treatment (e.g. medical imaging devices). Study F focuses on quantifying the effect of cyberattacks in automobile manufacturers. Special focus is set on understanding the effect of malfunctioning of machinery (e.g. industrial robots) which could lead to quality deviations in manufacturing. The study takes advantage of a customer model to estimate the monetary effect of attacks.

Study G takes advantage of the works on cyberrisk quantification and aims at delivering decision support. It proposes a means for automatically reacting to attacks and is hence thought to be applied in combination with real-time attack hypothesis generation. Yet, it can also be applied in combination with the other approaches presented to deliver forward looking advises. Within the study a baseline defensive investment is tested against an adaptive approach. For ensuring efficient cyberrisk management, the study proposes to take advantage of both sides by relying on a bio-inspired artificial immune system for systems.

Zusammenfassung

Technologie hat die moderne Gesellschaft und Wirtschaft tief durchdrungen und ist im täglichen Leben allgegenwärtig. Die damit einhergehende Digitalisierung, Informatisierung, Automatisierung, Technisierung und Intelligenzierung führte zu neuen Level an Komfort und Produktivität. Die Integration der Technologie bringt jedoch ebenso neue Risiken mit sich. Schon heute sind Cyberrisiken eine der größten Risiken für die wirtschaftliche Stabilität sowie den Wohlstand. Die Bedeutung dieser Risiken wird durch die weiter fortschreitende Digitalisierung noch zunehmen. Trotz dieser großen Bedeutung für die Gesellschaft und die Wirtschaft mangelt es derzeit an der Fähigkeit, diese zu managen. Ebenso ist bisher die zuverlässige Quantifizierung der Risiken schwierig, welches essentielle Hindernisse für die Entwicklung sicherer Systeme darstellt.

Die Dissertation zielt darauf ab, neue Methoden zur Quantifizierung von Cyberrisiken einzuführen, die auf der Nutzung von Cyber Threat Intelligence basieren. Zudem möchte die Arbeit Methoden liefern, welche unterstützend bei der Entscheidungsfindung hinsichtlich des effizienten Managements dieser Risiken eingesetzt werden können. Besonderes Augenmerk liegt auf der Quantifizierung und dem Management von Cyberrisiken in einer ganzheitlichen Weise. Dieser ganzheitliche Ansatz berücksichtigt die unterschiedlichen Domänen von Cyberangriffen, welche insbesondere den technischen sowie den menschlichen und gesellschaftlichen Bereich umfasst sowie deren Interaktion.

Im Zuge der Dissertation werden so unterschiedliche Methoden vorgestellt, welche der Risikoquantifizierung dienen. Hierbei wird ein Set von Methoden vorgestellt, welche genutzt werden können um basierend auf Ergebnissen eines Systemmonitorings die Wahrscheinlichkeit eines Angriffs zu quantifizieren (dynamische Risikobewertung). Zudem wird ein statischer Ansatz der Risikobewertung verfolgt, welcher basierend auf spieltheoretischer Methodik eine Quantifizierung der Wahrscheinlichkeit eines Angriffs ermöglicht. Beide Ansätze basieren auf der Analyse von Cyber Threat Intelligence. Eine zentrale Schwäche deren Analyse ist die Aktualität und Aussagekraft für zukünftige Angriffe innerhalb einer sich dynamisch verändernden Gefahrenlandschaft. Mit dem Ziel diese Limitationen zu minimieren wird eine Methodik vorgestellt, welche eine Vorhersage zukünftiger Angriffe ermöglicht. Die Quantifizierung der Auswirkungen eines Angriffs beruht auf der Analyse und Simulation eines erfolgreichen Angriffs innerhalb Digitaler Zwillinge. Basierend auf der Risikoquantifizierung wird abschließend eine Methodik zur Entscheidungsunterstützung vorgestellt.

Acknowledgements

This dissertation was written during my time at the Institute for Industrial Production (IIP) especially the Chair of Business Administration, Production and Operations Management and the Institute of Information Security and Dependability (KASTEL) at Karlsruhe Institute of Technology (KIT). In this time, I had the opportunity to receive a Networking Grant by the Karlsruhe House for Young Scientists (KHYS) as well as receive funding from the Helmholtz Information & Data Science Academy (HIDA) for an exchange with the Data Science Research Center (DSRC) at Ben-Gurion University (BGU) of the Negev, Beer Sheva, Israel, and was able to prepare parts of this thesis while being hosted by the BGU. I am very thankful to these institutions for providing me the opportunity and the environment to pursue my research interests.

Further, I experienced the support and trust of many people, for which I want to express my gratitude. First and most importantly, I want to thank my supervisor Prof. Dr. Frank Schultmann for opening this opportunity for me, his guidance, trust and support, for granting me the greatest possible freedom, and thus for providing me with the best possible environment for personal development and scientific research. I also want to thank Prof. Dr. Marcus Wiens for his supervision and wise academic advice during his time as my research group leader as well as within his role as Principal Investigator at KASTEL. Furthermore, I would like to acknowledge his role as co-reviewer. Special thanks to Prof. Dr Steffen Rebennack and Prof Dr. Hagen Lindstädt for serving as members of the committee. Moreover, I want to thank the senior researchers and professors I had the chance to collaborate with or who kindly agreed to discuss my work with them. Here, I would like to give special thanks to Prof Dr. Melanie Volkamer and Dr. Rami Puzis.

I also want to direct my deepest gratitude towards all my colleagues and friends at KIT for all the discussions, the support, the coffee breaks, and the remarkable conference and doctoral seminar experiences. In particular, I want to mention Dr. Florian Diehlmann, Katharina Eberhardt, Dr. Miriam Klein, Markus Lüttenberg, Amelie Schwärzel, and Rebecca Wehrle. I am sorry if anybody misses her or his name in the list and am sure she or he deserved to be mentioned. I would especially like to thank my friends outside the university world for being such important companions during the time of my doctorate - despite the considerable sacrifices in terms of time together - and of course for the always amusing and balancing meetings.

Above all, I would like to thank my family for their wholehearted support throughout my journey - Danke - Gracias - Thank you. Finally, I like to express my sincere gratitude to my caring and loving girlfriend, Sina. I consider myself very lucky to have her.

- I owe this to you -

Contents

List of Figures

List of Tables

Part I

Introduction and Background

1 Introduction and Motivation

> Everything is going digital. Our devices, our everyday life, and ourselves.
>
> (Neugebauer and Zanko, 2021)

1.1 Digital technologies in society and economy

Technology is an integral part of all our lives. When we wake up, technology may have already generated the first data points on some of us. *Smart watches, smart implantable devices* or *smart beds*, for example, enable the measurement of vital parameters and sleep times as well as record breathing patterns and heart rates. With the help of this information, these devices can quantify the health status as well as the quality of sleep and predict future sleep needs (Barnes, 2022; Neshenko et al., 2019). At this time, a *smart coffee machine* might already have brewed a fresh cup of coffee (Aitenbichler et al., 2007) enabling a convenient start in the day. After having taken a breakfast, *smart toothbrushes* can measure, if its users brush their teeth properly during their daily morning routine (Marcon et al., 2016). Furthermore, *smart toilets* can perform a daily health check (Bhatia et al., 2020). Within a *smart mirror*, people can get a first morning briefing on the weather or events that might have happened when they were asleep. Also, on our daily way to work, technology might guide our way when relying on *navigation systems* to find the shortest or quickest path to work, buy tickets for smart public transportation via *smart phones* (John et al., 2014), or when getting to work with a *smart* and *(semi-) autonomous vehicle* (Jovanov and Pajic, 2019; Yaqoob et al., 2019). When connecting with friends or searching for a partner, *social media applications* are omnipresent. Besides these applications, technology has become a companion in our daily lives such as *(online) banking, shopping*, *and education* (Yoon et al., 2017).

Yet, not only the private, but also the economic sphere is penetrated deeply by technology. Although this penetration strongly differs between various industry sectors, branches, and firms, (Friedrich et al., 2011), nearly every area of work confronts their employees with technology to a certain degree. Broad working areas are permeated with cyber-physical systems (Pasqualetti et al., 2015). In agriculture, digitalization and technization take an important role in ensuring food security (Abbasi et al., 2022). Digitalization is within the agricultural industry considered a key means of improving efficiency of food production for a rapidly growing population with increased demands for agri-foods in a world characterized by loss of arable land, water scarcity and environmental degradation. Agricultural industry, hence, is transforming to *smart agriculture* and *smart farms* including the

use of *autonomous robotic systems* like *self driving tractors* or *drones* and *big data analytics* to guide agri-food production (Abbasi et al., 2022). In manufacturing industries, digitalization provides possibilities for efficiency gains. Digitalization of manufacturing industries is fueled by customer demands for individualized products, shortening of product life cycles, and rapid technological change (Bozkurt et al., 2021). Modern production systems are thereby characterized as *cyber-physical production systems* in which *robotic systems, driverless transportation systems*, and products that are going to be manufactured communicate with each other (Herrmann, 2018). They thereby form a *smart factory* respectively a *smart production network* (Herrmann, 2018). Through mobile devices, people are integrated to the communication within the industrial production system (Herrmann, 2018). *Cloud computing technologies* integrate physical and cyber-layer along the whole value chain in manufacturing. Applications thereby range from product development to manufacturing system management (Mourtzis and Vlachou, 2016). Manufacturing analytic systems enable a continuous real time analysis of production systems (Lade et al., 2017).

Services vice versa are deeply penetrated by technology. In healthcare, digital technologies have proven to be a helpful means in many areas of medical service provisioning (Kaiser et al., 2021b; Topol, 2019; McCullough et al., 2010). *Robot assisted* or *robotic surgeries* (van Mulken et al., 2020; P. P. Rao, 2018), *technology supported medical image analysis* (Shen et al., 2017), medical health applications (Kaiser et al., 2020), and *smart nursing tools* like *smart beds* (Ajami and Khaleghi, 2015) bear great potential to improve the quality of healthcare. Furthermore, the introduction of novel communication channels may improve accessibility of medical care, for example, via *telehealth applications* (Kaiser et al., 2021b).

Last, there is growing consensus among experts that technology can have disruptive power and may act as a "game-changer" in warfare and military combat (Sechser et al., 2019). Consistently, nations spend high efforts in increasing offensive and defensive capabilities and "are actively militarizing cyberspace" (Pupillo et al., 2018; Gertz, 2016). Also, governments are developing and stockpiling innovative and highly potential *cyber-weapons* that can be used for military operations against nations. Cyberattacks are thereby considered the "Swiss Army knife of war" (Schulze, 2020). Military application of cyberattacks offer potential for enabling highly precise attacks at low cost and devastating consequences (Schulze, 2020). Furthermore, cyberattacks have a global reach (Schulze, 2020). Consistently, a weaponization of cyberattacks can be observed. Cyberattacks are thereby witnessed to be used increasingly for reaching political goals (Pupillo et al., 2018). "Cyberattacks can be more dangerous to the stability of democracies and economies than guns and tanks" Juncker (2018).

As these examples substantiate, digital technologies are integral to peoples lives and often, people even cannot imagine their personal and professional lives without digital technologies (Cijan et al., 2019). Today, permeation with technology is further increasing and all aspects of modern societies and economies are subject to a persistent digital transformation. However even now, technology represents the backbone of modern societies and economies as we know them (Limba et al., 2017) and the use of technology is ubiquitous (Makhdoom et al., 2018). The further accelerating digitalization can foster the importance of technology for economies and societies (Kazanin, 2020).

1.2 Cyberrisks

The broad permeation of technology and its ubiquitous use in many areas of life, however, also introduce substantial risks to economic stability (Ameli et al., 2018; El-Gayar and Fritz, 2010). This is, besides improving operational efficiency, enabling novel degrees of comfortability (e.g., remote access), informatization, digitization, intelligentization, technitization, and digitalization open novel attack channels making systems vulnerable to internal and external perpetrators of cyberattacks (Lv et al., 2020; Asghar et al., 2019; Moustafa et al., 2018). In this context, *cyberattacks* can be described as "an intentional misuse of personal computers on the technology dependent corporations, companies and systems or sites" (N. U. Qamar et al., 2017). For this purpose, "cyberattacks use harmful and destructive code to change coding of computer, reasoning or data, leading to disruptive effects or repercussions that can destroy the actual data and leads to (...) identity or personal information theft" (N. U. Qamar et al., 2017). Cyberattacks can thereby cause severe risks to economic stability and functioning of critical infrastructure of modern live (so called *cyberrisks*). They represent a novel type of risk that is distinct from conventional risk. With the amplifying permeation of every aspect of our lives, the management of these risks and the process of ensuring security becomes an increasingly compelling issue for businesses and societies around the globe (Radanliev, De Roure, Cannady, et al., 2018; Nurse et al., 2017).

However, in practice, there are still many challenges. These challenges need to be overcome when aiming at efficiently managing cyberrisks. One reason is that cyberrisk management methods are oftentimes not practically useful or there is a "lack of (applicable) supporting tools" (Leszczyna, 2021). This "lack of supporting tools" is a paticular burden to cyberrisk management as technology and the possibilities of its use are oftentimes not properly understood (Lenka et al., 2017).

In this way, the challenges in understanding of the implications of digitalization need to be considered an essential burden for cyberrisk management and for ensuring that technology serves people. The knowledge gap applies to both, the positive and the negative, and leads to a kind of manic relationship with a glorification of the use of technology and its benefits, as well as a cursing of the risks. The difficulties in understanding of cyberrisks are also reflected in the poor ability to quantify cyberrisks. "Models to quantify (cyberrisks) are not well developed; therefore, (cyberrisk) management in most businesses depends on qualitative assessments" (Keskin, 2021).

Furthermore, cyberrisk quantification is challenged by a lack of historic data. In particular this lack in historic data is observed for strategic information (e.g., losses caused by an attack; strategic cyber threat intelligence; Zeller and Scherer, 2021; Chismon and Ruks, 2015) while technical data is shared as cyber threat intelligence (Elitzur et al., 2019).

1.3 Research objectives

The aim of this work is to provide insights to the holistic analysis of cyberrisks from the perspective of companies and their value chain, to evaluate it quantitatively in the form of metrics, and to derive specific packages of measures and support for decision-making that may help in increasing

the resilience of companies and value chains. This work aims at contributing to solving the aforementioned challenges to *efficient cyberrisk management* that is in particular the insufficient ability of *quantifying cyberrisks* and the lack in *descision support tools*. A main contribution of the work represents its *holistic* nature. Hereby a set of methods is presented that includes the human and societal domain (exploitations are also referred to as social engineering attacks) as well as the information respectively technical domain of cyberrisks (attacks exploiting technical vulnerabilities). Furthermore, the methodological set considers monetary as well as non-monetary impacts of cyberattacks.

Another key differentiator between related works is the necessary degree of practical applicability of the set of methods presented. This is reached by the concise reliance of cyberrisk quantification on open access *cyber threat intelligence* as well as the reliance on *digital twins* for simulating the effects of cyberattacks on an underlying system. The methodology is hence designed and calibrated on the basis real world data on attacks.

Compared to most current cyberrisk management approaches (mainly originating from the field of information technology), within this work a set of methods is provided that strives for *efficient security* rather than absolute or maximal achievable security. This gives tribute to the economic rational of opportunity costs of security investments given limited budgets and the strategic nature of attackers (as intelligently adapting actors) assumed within this work. Hence, in the business context, the "price of security" is an additional limiting factor. The quantification of cyberrisks creates the prerequisite for their economic management.

The work comprises two major building blocks:

1. Risk analysis and quantification

   Here, the complex interaction of threat, consequences and security measures is traced and evaluated. A set of methods is developed that provides novel perspectives and approaches to the current discussion on the quantification of cyberthreats. Cyberrisk quantification can thereby either be executed as an approach for *dynamic risk assessment* on the basis of an analysis of network traffic (for this purpose, chapter 6 introduces to different methods for *threat hunting*, were the probability of being under attack is quantified) or as a forward looking analysis of cyberrisks based on a game theoretical model (chapter 7). Both approaches rely on the analysis of cyber threat intelligence and are hence limited with regard to their actuality and ability in reflecting future developments. The special importance of attack prediction and forecast thereby originates from the high dynamic nature of cyberattacks and the whole field of technological development. Tackling this, the set of methods also includes methods for predicting and forecasting future developments of cyberattacks based on which the databasis can be augmented (chapter 8). Special emphasis is given to the aspect of the human and societal domain of cyberrisks were the work provides a game theoretical model for fostering the understanding of different susceptibilities to cyberattacks taking the example of phishing attacks (chapter 9). Furthermore, the approach presented takes advantage of digital twins for cyberrisk quantification. On the basis of digital twinning, the presented approach for cyberrisk quantification is tested by simulating cyberattacks on the digital

twin of digitalized medical service provisioning within a hospital (chapter 10) and a smart automotive production facility (chapter 11). The digital twin also represents a generator of synthetic data on cyberattacks tackling the challenge of a lack of historic data.

2. Cybersecurity investment decision support

   On the basis of the approaches for cyberrisk quantification, a tool for investment decision support is proposed that offers the possibility of automating incident responses. Furthermore, an approach for improving cybersecurity is proposed combining static base defense and means for adaptive defense. The proposed approach for cyberrisk management, hence, represents a *bio-inspired artificial immune system* for the cyber-sphere.

1.4 Structure of the work

The dissertation is structured in three parts. The first part describes the introduction and background (chapter 2 to 5). The second part comprises the published articles that are integral to this dissertation (chapter 6 to 12). The third part concludes the work (chapter 13 to 15).

Chapter 2 includes an overview of relevant terms in the field of cyberrisks, theoretical foundations of research in cyberrisks, their quantification, and challenges for efficient cyberrisk management. Chapter 3 provides a literature review giving an overview on the current state of scientific research on quantitative methods for cyberrisk assessment and management, placing the work in the wider scientific context, and identifying research gaps. The current state is summarized, challenges are identified and research gaps are pointed out in chapter 4. Thereafter, chapter 5 highlights the research objectives to which this work aims at contributing. The following chapters are dedicated to the articles that companion this dissertation. Chapter 6 presents the article "Attack Hypotheses Generation Based on Threat Intelligence Knowledge Graph" that presents a first methodological approach for quantifying the probability of a cyberattack. In chapter 7, another methodology is given for this purpose. However, comparing both approaches the later is more focused on providing forward looking probabilities while the first approach is directed towards delivering real-time estimations on the probabilities of cyberattacks given artefacts from network monitoring. Chapter 7 comprises the article "Cyber Risk Quantification - Using Weighted Attack Graphs for Behavioral Cyber Game Theory". Both approaches presented for probability quantification rely on a cyberthreat knowledge graph. Chapter 8 presents the work "Attack Forecast and Prediction" which represents a means for improving the accuracy of assessments of the probability of cyberattacks by reducing the limitations that are posed by the highly dynamic nature of the threat landscape. The approach can be used to augment the knowledge graph used in chapter 6 and 7. Within chapter 9, a method for understanding the susceptibility for attacks on the human flank taking the example of phishing mail detection is proposed. The chapter essentially covers the article "Too stressful to look closely? The Information Value of Signal Detection under Cognitive Contraints – A Decision-Theoretic Model for the Case of Phishing Mail Detection". Chapter 10 and 11 present approaches for quantifying the impact of successful cyberattacks relying on digital twinning. The case for a hospital ("Cyberattacks on hospitals and their impact on medical service") is presented within

chapter 10. A digital twin able to simulate the effect of cyberattacks on the automotive branch is given in chapter 11. The chapter comprises the article "Digital twins and their use for cyber risk quantification - Analyzing the impact of cyberattacks on an automobile manufacturer". Chapter 12 presents a decision support system that is based on cyberrisk quantification. This approach takes advantage of the method presented in chapter 6 for demonstrating its ability in enabling automated real-time incidence response and presents a bio-inspired artificial immune system. The article that is referenced within this chapter is "Transitions from threat hunting and automated incident response". Managerial implications and limitations of the introduced methodological set for cyberrisk quantification and management are discussed in chapter 13. Chapter 14 gives a critical review on the set of methods and discusses its limitations. Finally, the dissertation is summarized and future research directions are given in chapter 15.

2 Theoretical Foundation and Elementary Background

> "Elementary" does not mean easy to understand. "Elementary" means that very little is required to know ahead of time in order to understand it, except to have an infinite amount of intelligence.
>
> Richard P. Feynman

2.1 Digitalization

2.1.1 Differentiation of key terms in digitalization

The term *digitalization* does not have a single open agreed definition. Furthermore, the terms digitalization and *digitization* are oftentimes used synonymously (Brennen and Kreiss, 2016) although they should be separated from a scientific point of view (Schallmo and Daniel, 2018). The etymology of these terms is from the Latin word "digitus" - finger and toe or "digitalis" - belonging to the digitus (Cochoy et al., 2019; Ong, 2018). It has a numerical sense that originates from times where numbers were counted on fingers (Bruderer, 2018). The term digitization directly refers to this and means to take "analog information and encode it into zeroes and ones so that computers can store, process, and transmit such information" (Bloomberg, 2018). It should hence be seen as a synonymous term for *datafication* (Mayer-Schönberger and Cukier, 2013). Digitalization contrarily describes "the pervasive (use and) synergy of digital innovations in the whole economy and society" (Valenduc and Vendramin, 2017). Both, digitization and digitalization are hence linked with computer systems running on data and information encoded in the form of binary numbers (bits). Yet, digitization is more focussed on the transfer of analog information to the digital sphere (encoding information in binaries or bits Brennen and Kreiss, 2016) while digitalization describes the use of technology (Schallmo and Daniel, 2018). Undoubtedly, digitalization has led to the formation of a novel phase of industrial and technological development as well as economic growth (P. Xu et al., 2017). The digitalization leads to increasing levels of process automation through the introduction of robotic systems, computers, and other smart

devices as well as their monitoring allowing to manage systems with great flexibility, efficiency, and safety (Asghar et al., 2019).

The term *robot* was ushered into existence within Karel Capek's drama "R.U.R.: Rossum's Universal Robots" from the year 1920 (Čapek et al., 1993). "Robots are presented in this play as small artificial anthropomorphic creatures which obey strictly the commands of their master (..). These creatures are called 'robotnik' in the Czech and Russian language from the word "robota", which means 'forced labor'." (Iavazzo et al., 2014). In this play, the robot was "invented (as) a worker with the minimum amount of requirements (... rejecting) everything that did not contribute directly to the progress of work. (... In this sense,) they are more perfect than (... humans) are, (... by having) an enormously developed intelligence." (Čapek and Kallinikov, 1940)

The term *automation* is a composed word originating from the Greek words "autós" - self and "mãta" - to rule or reign (Frisk, 1972). The combined version "autómatos" - of its own initiative, happening on its own initiative - is also known in Greek language (Frisk, 1972). Furthermore, automation is linked to the Greek and Egypt mythology where "automata - creatures similar to robots" (Iavazzo et al., 2014) were described (e.g., the statue of Memnon). While most mythological robots were made out of flesh or stone, Hephaestus created some robots that were made out of metal and came close to the mechanical nature of robots today (e.g., Talos; Iavazzo et al., 2014). Technology and wisdom were thereby combined by Hephaestus for the creation of the robots (Iavazzo et al., 2014). The use of robots was, hence, since the beginning (in Greek and Egypt mythology) linked to manufacturing, care for disabled people, transportation and decision support (Iavazzo et al., 2014).

The widespread introduction of computers to systems is also referred to as *computerization* (Yang et al., 2013). Computerization led to wide societal changes as well as in particular the introduction of computerized expert systems (Kaufmann, 2021; Hudomiet and Willis, 2021).

Digitalization has led to a "multidisciplinary revolution" (Molina Zarca et al., 2019) in economies, which is also referred to as *Industry 4.0*. The term Industry 4.0 was thereby introduced by the German government and aims at describing a fourth industrial revolution that is considered to be the digitalization of industry (Moustafa et al., 2018).

Intelligentization describes the associated process that comes through the introduction of technologies (e.g., microprocessors) that enable data analytics (Podgorski et al., 2017). In other words, intelligentization can be defined as the recently observable "trend oriented towards embedding (...) intelligent chips into any object that surrounds human beings on a daily basis, e.g., homes, cars, tools, machines, clothing and even into the human body" (Podgorski et al., 2017). This is, recent times can be described as a *big data* era with the *internet of things* being prevalent in today's society and economies producing various data points (Elmasry et al., 2020; Pawlick et al., 2019).

Big data is defined by high volumes (magnitude of data), velocity (rate of data generation) and variety (structural heterogeneity) of data ("three Vs definition of big data") and their generation (Gandomi and Haider, 2015; McAfee et al., 2012; Russom et al., 2011; Laney, 2001). Further extensions beyond the "three Vs definition" include as main characteristics value, veracity (with respect to unreliability inherent in some sources) and variability (Gandomi and Haider, 2015).

Intelligentization hence refers to the process of analyzing big data to gain value when insights are used to drive decision making (Gandomi and Haider, 2015).

The term *internet of things* was introduced in 1999 in reference to the application of radio frequency identification technologies (Podgorski et al., 2017; Ashton, 2009) and describes "a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual 'things' have identities, physical attributes, and virtual personalities" (Sundmaeker et al., 2010). The technologies thereby frequently enable autonomy of the devices or the processes the technology is introduced to.

2.1.2 Necessity of digitalization

There are high levels of digitalization in modern power (e.g., smart grids; Mishra et al., 2020; Q. Wang et al., 2019; B. Wang et al., 2019; Zeng et al., 2015), water management (Aivazidou et al., 2021), transportation (e.g., fleet management (Xidias et al., 2022; Sergeeva, 2021), port infrastructures and maritime navigation; Polatidis et al., 2018), administration (Dormann et al., 2019; Gustafsson, 2017), government (Borg et al., 2018; Gustafsson, 2017) as well as education (Mikryukov et al., 2020) and finance systems (Sengan et al., 2020; Limba et al., 2017). Hence, understanding the causes of the widespread digitalization can contribute to understanding the need for reaching a state of cybersecurity and highlights the importance of cyberrisk management.

A main driver of the increasing permeation of broad aspects of daily lives are the great potentials offered including efficiency gains, increases in productivity, flexibility, comfortability, and interoperability (Q. Zhang et al., 2015). Hence, in strive of economies for reducing costs, enhancing productivity and thus, increasing efficiency, high levels of digitalization, computization, intelligentization, and automation were introduced to modern economies across all branches that makes these economies reliant on the proper functioning of technology and the internet (Garg et al., 2003). This is, broad digitalization of society promises great potentials for economic and social prosperity (Kiesel et al., 2020). In this sense, digital transformation of every aspect of human life is envisioned to be able to elevate the quality of life as well as economic activity (Neshenko et al., 2019).

Furthermore, economic systems benefit from increasing digitalization by enabling "novel business models" (Timonen, 2022; Alkhammash et al., 2022; Nübel et al., 2021; Bosch and Olsson, 2021), the creation of novel jobs and simplification of business operations and with this increased productivity (Mourtzis and Vlachou, 2016). As a consequence, economic competitiveness is closely linked to firms' abilities in capitalizing onthe digital transformation (Radanliev, De Roure, Cannady, et al., 2018). Thereby, in particular globalization amplifies competitive forces which leads to the need to digitalize and a race to automation (Lv et al., 2020; Sheehan et al., 2019; Komljenovic et al., 2016). In this sense, competition and competitive forces, have transformed traditional physical systems to cyber-physical systems defined by high levels of penetration through technology (Lv et al., 2020). Also, the immense meaning of digitalization to the competitiveness

of organizations causes that digitalization led to a reshaping of the economic landscape within the last years (Jones et al., 2021; Schreckling and Steiger, 2017; Ruan, 2017; Nanterme, 2016).

Falling prices of technology and increasing computational power give momentum to this development (oftentimes also referred to as Moore's law Mishra et al., 2020; Schaller, 1997; Moore, 1975, 1965; Englebart, 1960) This leads to the situation where "everything is going digital" (Neugebauer and Zanko, 2021). This is,"'digitally upgrading' conventional object in this way enhances their physical function by adding the capabilities of digital objects. Thus generating substantial added value is omnipresent in today's economic systems as well as products for consumers. Forerunners of this development are already apparent today – more and more devices such as vacuum cleaners, sewing machines, exercise bikes, electric toothbrushes, washing machines, electricity meters and photocopiers are being 'computerized' and equipped with network interfaces" (Mattern and Floerkemeier, 2010).

2.2 Concept of risk

Risk is a quite abstract construct which unifies a diversity of different concepts and definitions. While the origin of the word is widely discussed, the etymology is inconclusive (Aven, 2012). However, following the etymology of the word is considered to provide instructive insights to foster its understanding.

According to a first etymological stream, the word risk is considered to originate from the ancient Latin word "resecare" - that which cuts - and with this is synonymous to rock, crag, and reef (Hamilton et al., 2007). This word's origin and its transferred meaning of describing a danger or a hazard is hereby linked to the maritime use and context (Aven, 2012). Sailors and helmsman were at risk that rocks might pose a threat to their ships. Tightly linked to this assumed origin is the etymological foundation in the Greek word "rhiza" describing hazards of sailing too near to the cliffs (Aven, 2012; Hamilton et al., 2007). The neo Latin word "risicum" has a transferred meaning loosening the ties to the maritime environment. It can be translated with, "that which opposes security" and "the act of courage or boasting of those who deal with something of which they don't know the limits or characteristics" (Maso, 2018). Risk and security can hence be viewed as complements.

A further stream assumes the Arabic word "rizq" to represent the etymological origin, which is more at the sense of fortune, luck, destiny, and chance (Aven, 2012). "Rizq" thereby oftentimes is referred to from a religious point of view where the term was used to describe "the daily provision allotted by God to each man" (Bouslama and Lahrichi, 2017). Closely linked to this religious meaning of "rizq" is that "in the phase of the Arab expansion in the Mediterranean (...) the term rizq (Arabic) and rizikòn (Greek-Byzantine) is used to indicate the silver pay of the mercenary soldier. He is in fact the 'luck soldier', (...) who owes his life to 'fate' and "destiny' (to riziko) and which constitutes the means often used by 'fate' and 'destiny' to decide the life of other men." (Maso, 2018). It is assumed that the Aabic "rizq" originates from the Persian respectively more precise Pahlavi word

"rôcik" which takes the meaning of "daily bread" (Bouslama and Lahrichi, 2017). This origin focuses on the meaning of risk as an uncertain future income.

The different origins also highlight key differences that still exist in the understanding of risk which are neutral or have negative connotations. While risk according to the Latin origin is clearly negatively connotated (focusing on losses), risk according to the Arabic origin is more neutral (focusing on uncertainty of payments). In this sense, the etymological history of the word might explain the ambiguity and inconsistency in the understanding of the word risk (Bouslama and Lahrichi, 2017).

Consistent to the broad etymological understanding of the term risk, many risk definitions are coexistent in scientific discussions (Aven, 2012). A first traditional understanding of risk considers risk as a tuple of consequences and the likelihood of events causing these consequences. Similar, the tuple of severity and probability is a frequent understanding of risk (Q. Zhang et al., 2015). Similarly, definitions focusing on the dualistic decomposition of risk in the possibility and the potential of an unfortunate, unfavourable event are known (Aven, 2012). These definitions can be classified to belong to the class of probabilistic risk assessment approaches which consider risk as a doublet of some kind of impact and probability (Ralston et al., 2007; Keller and Modarres, 2005). Let R denote the risk, I denote the Impact, P denote the probability, and i be a index for a specific cause of the impact (e.g., an attack). Then the risk as a doublet of probability and impact could be described as follows.

$$R =< I, P >= \sum_i I_i \cdot P_i \tag{Exp. 2.1}$$

Quantitative analysis of risks within this conceptualization of risk means to assign numeric values to the probability and the impact of an event as well as to the costs and benefits that are linked to security investments (Bojanc and Jerman-Blažič, 2008). The probabilities can thereby be frequency based (Kaplan and Garrick, 1981). Let assume ϕ denoting the frequency distribution then, equation 2.1 can be adapted as follows.

$$R =< I, P(\phi) >= \sum_i I_i \cdot P(\phi)_i \tag{Exp. 2.2}$$

Besides probability assessed in a frequentist manner (frequentist probability; Aven, 2012), the probability can also be extracted from knowledge and subjective judgements (subjective probability; Aven, 2012). In the later case, probabilities can be understood in the sense of degrees of belief (Beyerer and Geisler, 2015). In this case, the probability is dependent on the knowledge of the risk manager that aims at quantifying the risk (Aven, 2012). Risk is inevitably fuzzy. Hence, subjective risk assessment is subject to biases in measurements as well as deviations due to biased beliefs and bounded perspectives on risk.

The natural expansion of this duality of risk is the scenario based probabilistic risk definition (Radanliev, De Roure, Nicolescu, et al., 2018; Kaplan and Garrick, 1981). This definition represents a triplet. In essence, it is based on the fact that "a threat itself does not necessarily manifest as an incident. (... It) is only harmful if there is a corresponding vulnerability in the

target system (...). If a threat and an existing vulnerability lead to an incident, the impact refers to the consequences, where the impacted assets can be tangible (e.g. direct financial consequences from fines) or intangible (e.g. loss of reputation)" Zeller and Scherer (2021). Let s describe the probability of a scenario and j be a differentiator between the scenarios. The differentiation between i and j is based on the probability that a cause (e.g., attack) may hit an underlying system in different scenarios which vise versa would lead to different impacts of the cause (e.g., attack). Hereby the definition gives tribute to the probabilistic nature of attack scenarios that may occur.

$$R =< I, P(\phi), s >= \sum_{i,j} s_j \cdot I_{i,j} \cdot P(\phi)_{i,j} \tag{Exp. 2.3}$$

Besides approaches of probabilistic risk assessment, there are approaches that further differentiate the probability as the product of threat and vulnerability. This is, another definition of risks can be seen as the triplet of threat, vulnerability and consequences (Ganin et al., 2020; Beyerer and Geisler, 2015; Kaplan and Garrick, 1981) respectively impact (Zeller and Scherer, 2021). If this concept of risk is applied to the assessment of risks caused by attacks, threats can be defined as "a person or an organization that intends to cause harm" (Mateski et al., 2012). Let T denote the threat and V the vulnerability, then the risk can be defined according to equation 2.4.

$$R =< I, T, V >= \sum_{i} I_i \cdot T_i \cdot V_i \tag{Exp. 2.4}$$

Another definition quite similar to the triad of impact, threat and vulnerability is the triad of exposition vulnerability and consequence that is also frequently referred in scientific risk and security research, where E denotes the exposure.

$$R =< I, E, V >= \sum_{i} I_i \cdot E_i \cdot V_i \tag{Exp. 2.5}$$

A further triplet component definition of risk within the area of intelligent threats, so called perpetrators of attacks (i.e., attackers, offenders) is linked to the decomposition of probability to the probability of launching an attack and the attack success probability. This definition takes into account that for establishing an attack, the possibility of executing an attack successfully (attack success probability) must be given and the attacker needs to be willing to conduct the attack (probability launching an attack; Q. Zhang et al., 2015). Let $P_{l,i}$ describe the probability that an attack is launched and $P_{s,i}$ describe the attack success probability if an attack is launched. It can thereby be assumed that the attack success probability is predominantly determined by the safeguards implemented.

$$R =< P_{l,i}, P_{s,i}, I_i > \sum_{i} P_{l,i} \cdot P_{s,i} \cdot I_i \tag{Exp. 2.6}$$

These risk definitions describe risk as the expected value respectively disutility (loss). Risk thereby comes down to an arithmetic mean (Aven, 2012). By this means, risk can be applied to decision theoretic considerations assuming rational decision makers that aim at minimizing their disutility. Hereby, risk is computed in accordance with the expected utility theory.

Besides these definitions, however, there are more focused definitions regarding the risk as a solitary (solitary risk definition) of the probability of an undesirable event or some kind of uncertainty (Aven, 2012). The probability based perspective on risk thereby focuses on the chance of damage or loss (Aven, 2012; Campbell, 2005; Haynes, 1895). This might be a reasonable abstraction of the previously discussed abstraction for cases, where a specific predefined negative event might emerge. Statements like "the risk is at 50-90%" (A. A. Rao et al., 2022) refer to this risk definition. According to the probability based definition of risk, the risk can be formulated as follows.

$$R = P_i \tag{Exp. 2.7}$$

Equivalently, risk definitions solely focusing on the probability were decomposed to a doublet of risk. Thereby a definition taking into account hazard and safeguards shall be given as an example (Kaplan and Garrick, 1981). "This equation also brings out the thought that we may make risk as small as we like by increasing the safeguards but may never, as a matter of principle, bring it to zero." (Kaplan and Garrick, 1981) Let H denote hazards and S denote safeguards then risk can be formulated as follows.

$$R =< H, S >= \frac{H}{S} \tag{Exp. 2.8}$$

Like the singular definitions on risk focusing on probability, a risk equation with a focus on the consequence is discussed. Risk can thereby be defined as the undesirable consequence of an event (Aven, 2012).

$$R = I_i \tag{Exp. 2.9}$$

Risk can also be described as volatility risk (Arrow, 1964). While the definition is oftentimes utilized when describing risk on the basis of expected utility theory, the risk definition that satisfies the equivalence to vulnerability can also be referred to as a stand alone definition (Aven, 2012). This definition is based on the "risk equals vulnerability of return paradigm" (Thomson, 2003). "Volatility risk describes the possibility of negative and positive deviations from an expected outcome" (Wiens, 2021). Risk thereby refers to the variability and unsteadiness of outcomes (Wiens, 2021; Aven, 2012; Arrow, 1992, 1964). This risk concept is especially established in finance, and applies to established portfolio optimization approaches (mean-risk respectively mean-variance approaches; Markowitz, 1952), whereby risk is a statistical measure for the probability of deviation from a mean (a priori assumed) characteristic (Hardy, 1923). Risk according to the uncertainty (where σ is a measure for the uncertainty) based definition describes a deviation from an a priori known characteristic of a subject of interest.

$$R = \sigma_i \tag{Exp. 2.10}$$

Within all these risk definitions, known and unknown threats and even "unknown unknowns" (Sharkov, 2016) can contribute to the overall risk. While the management of known threats is possible with high precision and data is available on these, the management of unknown threats is much more challenging.

In general, risk is frequently referred to as a relative concept (relative risk; Kaplan and Garrick, 1981). It describes an approach for understanding risk and security as an equivalent in terms of a comparison with reference points (i.e., reference for measurement). Risk and security is thereby defined with reference to a clearly defined system (e.g., initial state of a system).

Furthermore, in general, risks are multi-dimensional. This is, a well-designed categorization of risks within a "risk vector" or "index tuple" might be more promising for describing risks of different nature (Cai et al., 2022; Chung et al., 2005). Hereby the aggregation of risks within a single metric causes loss of precision and granularity. From an economic point of view, however, these single metrics can be very useful, meaningful, and more applicable than vectorial risk metrics. This is, all resources are for competition of use and hence, spending in security causes opportunity costs (Beevers et al., 2021; Rampini and Viswanathan, 2010). Multi-dimensional risk vectors can thereby be a burden to risk management and might cause high computational costs when decisions should be based on vectors (multi-criteria decision making). Also, risks can originate from different sources which may necessitate to differentiate between exogenous and endogenous risks (Ghadge et al., 2019; Johnson, 2015). The former can be interpreted as a priori non-influenceable "shocks" and are also the standard assumption in many purely stochastic risk models (Beyerer and Geisler, 2015). In contrast, endogenous risks are the result of the interaction of offender and defender as well as the system properties and demand for other methods (e.g., game theory; Beyerer and Geisler, 2015).

2.3 Cybersecurity and cyberrisk

2.3.1 The concept of cyberrisk

Although the term cyberrisk gained considerable prominence in scientific literature as well as in public perception, there is no commonly agreed definition of cyberrisk nor terminology on the field of cyberrisk available (Strupczewski, 2021; Ganin et al., 2020; Eling and Wirfs, 2019; von Solms and von Solms, 2018). This is at least partially due to the existence of various related risk definitions (reference to this is given in chapter 2.2). Yet, there is also uncertainty and considerable discussions in the scientific community regarding the scope of cyberrisks (e.g., which domains are affected - technical versus human and societal domain) and the effects caused by cyberattacks (including non pecuniary as well as pecuniary effects but also physical harm; von Solms and von Solms, 2018). Also, cyberrisks are composed by direct and indirect costs of an attack and by direct (e.g., defense investment costs in the actual sense) and indirect costs of defense (e.g., reduction in productivity or security circumvention practices; Fielder et al., 2016).

The term cyberrisk can be decomposed to its components "cyber" and "risk". However, cyber lacks a single unanimous definition and causes many associations (Böhme et al., 2019). Hence, "cyber is (...) a perfect prefix" (New-York-magazine, 1996) that has no inherent meaning. It is assumed that the term cyber is a short form of cybernetics. The term cybernetics comes from the Greek "kubernētēs" respectively "kubernáō", which describes the activity of the helmsman, to control, to guide or to lead (Frisk, 1972). Cyber is frequently used as a discriminator between conventional space and technology guided, electronic spaces (Böhme et al., 2019). Cyber thus takes on a meaning transferred from sea navigation to the technological domain and can be used as an indicator for the application of digital systems within a digitized environment (e.g., digital work and communication environment; Haas, 2016). Hence, the prefix cyber describes the place where the risks arise and has been inextricably linked with information technology in our linguistic usage since the 1970s (Metzger, 2018).

Cyberrisks describe "the level of impact on organizational operations (including mission, functions, image, or reputation), organizational assets, or individuals resulting from the operation of an information system given the potential impact of a threat and the likelihood of that threat occurring." (Pub, 2005). Thus, the term cyberrisks can be understood in a broader sense as those risks that arise when navigating within a digital, data-driven and networked world. Cyberrisks represent a class of operational risks. However, cyberrisks differ from other operational risks (Eling and Wirfs, 2019). This is, the increasing dependency of economies and society on technology leads to considerable spillover effects within networks (Eling and Wirfs, 2019). Furthermore, complex interrelationships and coupling between the cyber, socio-cognitive, and physical layer as well as a great scope and dynamic of threats define cyberrisks (Ganin et al., 2020; C. Chen et al., 2020). Within this work, the term cyberrisks is limited to risks that are provoked by offenders (the perpetrators of cyberattacks). However, other definitions also include unintentional human failures of operators (Eling and Wirfs, 2019).

2.3.2 The need for cybersecurity

The competitive need (as outlined in chapter 2.1.2) for digitalization (e.g., for enabling novel business models and innovative products or procedures) clashes with cybersecurity (Kiesel et al., 2020; Sury, 2019; Rothrock et al., 2018). This is, the benefits aimed to leverage on when introducing technologies to society and economies come at the cost of cyberrisks (Alsaedi et al., 2020; Taormina and Galelli, 2018) and risks to data privacy (Steinbrück et al., 2021; Debatin et al., 2009). Significant cyberattacks can thereby pose a severe threat to an affected organization's solvency and a firms continuation of operation (Eling and Wirfs, 2019). The broad reliance on information and communication systems in modern society and economies demands for securing information flows (Ardito et al., 2018). Consistently, ensuring cybersecurity became a pervasive issue for modern economies (Garg et al., 2003). Further increasing connectivity and standardization of computer networks is considered to increase cyberrisks as potential attackers may have higher gains from intensive knowledge on one operating system. The increasing connectivity may thereby

increase the susceptibility to attacks (Knowles et al., 2015) and raise the surface perpetrators of cyberattacks can take advantage of.

Amplifying digitalization and dependence on technology in interconnected systems further leads to an extending attack surface, highly attractive environments for cyberattackers especially when proper investments in systems' security are neglected and consistently raising incident counts (Alshehri et al., 2018; T. Liu et al., 2015). The current situation renders a picture where systems are highly susceptible to cyberattacks (Housh and Ohar, 2018). Consequently, cyber-incidents are a frequent event and the emergence of incidents within large organizations is nearly certain (Evans et al., 2016). This is, economic prosperity is increasingly exposed to cyberthreats (Knowles et al., 2015). Reliable and secure operations of technology, however, is essential for the deployment of industrial and private internet of things and digitalized networks as a whole (Alshehri and Hussain, 2019). Cyberattacks might thereby inflict significant cascading effects in highly interconnected and interdependent systems leading to catastrophic impacts (Ganin et al., 2020).

These substantial security risks emerging from great dependency on technology highlight the necessity for organizations of selecting the right digitalization strategy and in this sense a strategy for the successful and sustainable implementation of technologies within their business (Stewart and Jürjens, 2018; Yang et al., 2013). Yet, today, technology and the possibilities of its use are oftentimes not properly understood (Lenka et al., 2017). Furthermore, the majority of people cannot be expected to have a strong technical expertise. Hence, for many users the technical background of attacks is a closed book. consequently, for many people, the effects and possibilities of technology are something that touches the boundaries of mystery (evidence can be found in phrases like "cybercriminals (...) (are) able to attack companies in mysterious ways"; Grech, 2021). The lack in understanding of the implications of digitalization, however, can be considered an essential burden for cyberrisk management and hence for ensuring that technology serves people. The knowledge gap applies to both, the positive and the negative, and leads to a kind of manic relationship with a glorification of the use of technology and its benefits, as well as a cursing of the risks.

Cybersecurity was neglected for a long time when discussing technology and digitalization strategies (Chinn et al., 2019). In the year 2000, it was assumed that the functionality of a highly technologized and digitized infrastructure was primarily threatened by physical influences and the resulting material damage to the hardware (Haas, 2016). Successful cyberattacks and their (partially) devastating impacts however fostered the awareness of businesses and society of the necessity for efficient management of cyberrisks (also considering damages to software). Among those, the *WannaCry* attack on the English national health service (Infobox 1) as well as the *Stuxnet* attack on a nuclear plant in Natanz (Infobox 2) made the risks that come with the increasing penetration of modern society and economies more tangible and sharpened the perception of hitherto abstractly perceived cyberrisks (Brown et al., 2017). The increasing realization that damage can also occur by purely digital means and the abuse of technology (in the course of a cyberattack) represents the beginning of the conceptualization of cyberrisk (Ani et al., 2017; Haas, 2016).

Infobox 1: WannaCry attack on the English national health service

WannaCry is a ransomware that had global reach. Yet, the attack is most known for impairing the English medical health provisioning system between 12 May to 19 May 2017 (Morse, 2018), representing the first large-scale cyberattack that affected healthcare service (Mahler et al., 2018). The national health service "responsible for providing life-saving care, (was) suddenly unable to access critical systems and patient data, delaying treatments and sending emergency rooms into chaos." (Pupillo et al., 2018) Within the English national health service, the attack was widespread and "affected at least 80 out of 236 trusts across England" (Morse, 2018), which comes down to a disruption of operations in at least 34% of trusts in the national health service. The systems were thereby either encrypted by the software (infected) or systems were shut down as a precautionary measure (affected; Morse, 2018). "The global ransomware event (...) WannaCry demonstrated how the performance of vulnerable medical devices may be compromised by an exploit" (Connolly et al., 2018). In this way, the functioning of digital systems were influenced by the attack (Adams et al., 2019). These medical devices included inter alia medical imaging devices such as devices for magnetic resonance imaging (Tully et al., 2020; Mahler et al., 2018). The nonoperational devices caused that physicians were not able to rely on evidence that would have be gained from those medical images and needed to craft diagnoses on a limited set of observations. Besides, the attack caused that physicians were not able to access medical records (e.g., test results), radiology as well as pathology results, and medication plans (Ghafur et al., 2019; Morse, 2018).

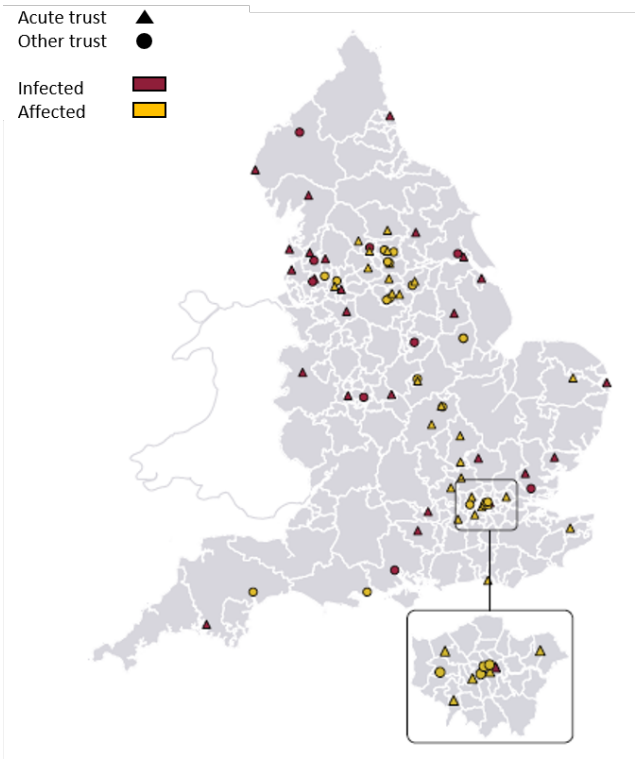

Figure 2.1: Affected trusts within the national health service through the WannaCry attack
Source: (Morse, 2018)

The attack caused that (outpatient) patient appointments as well as surgeries had to be cancelled (Tully et al., 2020; Ghafur et al., 2019; Morse, 2018; Martin et al., 2018). Furthermore, especially acute care was disrupted causing high deprivation costs. This is, patients needed to be diverted from

accident and emergency departments to non-infected trusts increasing the time to medical relief for these patients (Morse, 2018). Also, hospitals had to switch to manual workarounds. "The WannaCry attack (hence) had a significant negative impact on the delivery of care and cost to the health service in England." (Ghafur et al., 2019) Furthermore, from an economic point of view, the impact of the WannaCry attack on the NHS was considered to be "at least (...) $ 115 million" (Tully et al., 2020). The exploit WannaCry took advantage of is known as *EthernalBlue* (D. Rogers and Kanth, 2021; Mahler et al., 2018). Although a security patch was released before the exploit was made public, a significant share of devices was not patched. As medical devices are frequently in use for quite long times, legacy systems are predominant in the technical infrastructure of hospitals. This is, especially for systems of the national health service, WannaCry was a problem. Hereby, technical barriers as well as a lack of care (i.e., lack of proper patch management) caused the high susceptibility of the national health service (Mahler et al., 2018)

Today, the awareness of the importance of managing cyberrisks and the threats that cyberattacks pose to economic prosperity has greatly increased (Chinn et al., 2019). Meanwhile risk managers consider cyberrisk as a significant threat to businesses (Poppensieker and Riemenschnitter, 2018) and one of the greatest obstacles to digital transformation (Kiesel et al., 2020). However, the priorities that are set on ensuring cybersecurity differ between decision makers. It is observed, that the priority business leaders and senior management set on cyberrisk management oftentimes spikes in case of an incident for sharply decreasing afterwards (Choi et al., 2017).

With increasing awareness regarding the importance of cyberrisk management, also the confusion increased. This is, decision makers are confronted with bulky buzz words and novel terms related to cyberrisks and their management (Poppensieker and Riemenschnitter, 2018; Choi et al., 2017). Consistently, many myths in cyberrisk management exist and the understanding of cybersecurity (Poppensieker and Riemenschnitter, 2018).

Infobox 2: Stuxnet attack on a Nuclear plant at Natanz in Iran

The Stuxnet attack describes a *worm* that took advantage of four *zero-day vulnerabilities* and is designed to sabotage its victim's system (Baezner and Robin, 2017) and is considered the first cyber-weapon causing physical destructions (Langner, 2011). It is transmittable via USB-flash drives (Baezner and Robin, 2017). The attack impaired the proper functioning of a centrifuges for uranium enrichment within a nuclear plant in Natanz (Iran) between 2009 and 2011 and caused physical damage to the centrifuges (Baezner and Robin, 2017).
Stuxnet goes to action if all prerequisites are met and then causes a manipulation of centrifuges' rotor spinning speed and quick alternations of the speed which causes irreparable damage to the centrifuges and renders them inoperable (Langner, 2013; Farwell and Stuxnet, 2011). These changes were hidden from the operators (i.e., masked by the worm). To do so, Stuxnet replayed prerecorded input signals to the nuclear plants control systems during the attack (Langner, 2011). It took advantage of the (physical) weaknesses of the centrifuges used. "These centrifuges are (...) fragile and an abrupt change of speed could cause damage or even breakage." (Baezner and Robin, 2017). The fit of the exact structure of the centrifuges (groups of 164 centrifuges) that Stuxnet is designed for and the organisation of these in Natanz nuclear plant was clearly hinting at a targeted attack (Baezner and Robin, 2017; Albright et al., 2010).

The origin of the Stuxnet are not known today. Yet, there are hints such as "the presence of the word 'myrtus' in the code, which was the name of the file where the worm was stored when it was being developed. This word is believed to be a reference to Queen Esther who saved the Jews from a massacre from the Persians in the Bible and whose name in Hebrew refers to the word 'myrtle'. The involvement of Israel in the development of Stuxnet remains an uncertainty and the evidence pointing in that direction may also have been planted to mask the identity of the real perpetrator." (Baezner and Robin, 2017)

For conducting a successful attack, perpetrators need to conduct a series of multiple steps (so called cyber kill chain; Yadav and Rao, 2015; Q. Zhang et al., 2015; Hutchins et al., 2011). These are also referred to as atomic attacks. Each atomic attack thereby needed to be executed successfully. The course of atomic attacks differs between different attack families. The root causes of the vulnerabilities are oftentimes found in details such as faults within the source code or implementation failures (Ganin et al., 2020; Nicol et al., 2004).

To engineer absolutely secure systems is considered to be impossible (Nicol et al., 2004). Hence, to cyberattacks, there does not exist any system that is absolute secure (Nicol et al., 2004). One reason for this is the human nature (e.g., bounded rationality) and their current role in ensuring security when interacting with technology. This is, technology vendors respectively the creators of technology may unwittingly create vulnerable or even defect systems allowing their exploitation through perpetrators of cyberattacks (Sheehan et al., 2019). Furthermore, human actors do also have the main responsibility in engineering secure systems. They may thereby unwittingly introduce novel risks to the system.

Also, patching all vulnerabilities of a system is not possible due to insufficient budgets for engineering secure systems (Fielder et al., 2016). These limited budgets are due to strategic considerations of vendors in a highly competitive market, where customers do not directly value cybersecurity (Zander et al., 2020). Investments in the security of products are costly and the costs are unlikely to be compensated by customers as they (oftentimes) cannot differentiate between secure and insecure products and vendors can consequently not charge a premium for security (Bojanc and Jerman-Blažič, 2008). Hence, investing in security is frequently non lucrative and scarce resources are allocated to more salient product requirements (e.g., connectivity or infotainment systems in modern automobiles; Zander et al., 2020). From the vendors' point of view, security must hence be limited to the bare minimum.

Furthermore, reducing vulnerabilities and improvements of security operations should be higher up in scientific and practitioners' agendas for reaching a state of security allowing higher levels of digitalization. Hence, more efforts should be spent on reducing vulnerabilities to cyberattacks given the potentially devastating impact of successful cyberattacks (Hentea, 2008). It is thereby essential to provide a quantification of security that can be used as a signal for customers to differentiate between secure and insecure products. This could enable a situation where security may be a selling point for customers ("cybersecurity as a business model"; J. Rogers, 2016).

Besides engineering secure systems in the first place, cyberrisk management in the sense of supporting investment decisions in defensive means (e.g., safeguards respectively countermeasures

for hardening a systems properties) represents a main issue within digitalized economies and societies. Safeguards and their oftentimes synonymous used equivalent countermeasures describe "any process or technology developed to negate or offset offensive cyber-activities" (Kaloroumakis and Smith, 2021).

Furthermore, the dynamic technological development leads to systems quickly becoming obsolete (Falco et al., 2019). Obsolescence of technology (both soft- and hardware) are a further key burden to security (Ganin et al., 2020). However, investment costs for digitalization of industrial and service provisioning networks (e.g., digital health care solutions for medical imaging or industrial robots) are frequently high. Hence, systems are oftentimes deployed much longer as they are supported by their vendors and technical staff. Consequently, many of these, so-called legacy systems, are in operation (Infobox 1). These legacy systems introduce vulnerabilities to networks (Ralston et al., 2007). However, in some cases, legacy systems might also be more resilient to crises respectively more robust than their successors (e.g., if legacy systems allow a manual operation when necessary; Zeng et al., 2015).

The high dynamics of technological development also apply to the attackers' side manifesting in a dynamically changing attack landscape. Hereby, even today, attackers seem to have a *first mover advantages* and "the perpetrators of cyberattacks have been playing a dynamic cat and mouse game with cybersecurity analysts" (Elitzur et al., 2019). In essence, "computer and network security is (...) a battle of wits between a perpetrator who tries to find holes and the designer or administrator who tries to close them" (Almasizadeh and Azgomi, 2013).

2.4 Cyber threat intelligence

Although, the variety of cyberattacks is high and attackers follow different techniques to successfully execute an attack, cyberattacks in essence follow a common schema or similar life cycle which is known as the Cyber Kill Chain (Garba, 2019; Conti et al., 2018; Yadav and Rao, 2015). Cyber threat intelligence is evidence based collective knowledge along the Cyber Kill Chain and derived from past (recorded) attacks that can help in understanding attacks, improve security, and selecting appropriate defensive means (Sengupta et al., 2020; Conti et al., 2018; S. Qamar et al., 2017). Cyber threat intelligence describes information that is capable of describing attack behaviors, the proceeding of attacks as well as defenses against cyberattacks (H. Zhang et al., 2018). It comprises data on forensic artefacts of cyberattacks (e.g., virus signatures, attack files or hashes; Liao et al., 2016) and is frequently extracted from log data (Almasizadeh and Azgomi, 2013). "Such knowledge is essential for an organization to gain visibility into the fast-evolving threat landscape, timely identify early signs of an attack and the adversary's strategies, tactics and techniques, and effectively contain the attack with proper means" (Liao et al., 2016). Cyber threat intelligence describes "the set of data collected, assessed and applied regarding security threats, threat actors, exploits, (attacks ...), vulnerabilities and compromise indicators" (Shackleford, 2015). Furthermore, a key strength of cyber threat intelligence is its shareability among different organizations (Elitzur et al., 2019;

S. Qamar et al., 2017). This is essential, given that no organization can have a comprehensive picture on the threat landscape (Elitzur et al., 2019).

By sharing cyber threat intelligence, a more holistic view can be gained incorporating knowledge on attacks experienced as well as knowledge on those attacks that were experienced by peers (Elitzur et al., 2019; Stewart and Jürjens, 2018). Relying on cyber threat intelligence for security analyses is essential as each promising security means need to be founded on an in-depth understanding of both the attacked system as well as possible attacks targeting the system (Lv et al., 2020). For identification of possible attack paths, it is a feasible means to refer to previous successful attacks with the aim of understanding their proceeding and course of action (Almasizadeh and Azgomi, 2013). Cyber threat intelligence presents reliable data on cyberattacks and their (technical) impact on a real system whose specifics are oftentimes known.

2.5 Importance and necessity of holistic cyberrisk management

As already noted, in recent years, managing cyberrisks gained in importance within public discussions and the scientific literature (Strupczewski, 2021; Pasqualetti et al., 2015; J. Chen and Zhu, 2019). Rising incident counts, exploiting software, hardware and human operators (Hentea, 2008), as well as high needs to leverage on the benefits of digitalized systems (competitive need for digitalization - chapter 2.1.2) catalysed the increasing focus on cyberrisk management (Knowles et al., 2015; Bojanc and Jerman-Blažič, 2008). In the recent digitalized world of frequent cyberattacks, the management of cyberrisks is a crucial task for organizations for ensuring economic prosperity and competitiveness (Peters et al., 2015; Bojanc and Jerman-Blažič, 2008). This is as economies are highly dependent on the (proper) functioning of technology in digitalized systems (Knapp et al., 2009).

Cybersecurity, cyberrisks and reports on recent cyber-incidents, corporate security or "privacy disasters" (Culnan, 2019) are prevalent in today's news coverage (DiMase et al., 2015) exposing the public to information about the hazards introduced with digitalization (van Schaik et al., 2017). Thereby, cyberattacks have become so frequent, that today, only spectacular incidents with devastating impacts make headlines (insights on the development of media coverage regarding cyberrisks are provided in infobox 3; Rothrock et al., 2018). However, still reports on cyber-incidents are numerous. The high media attention also reflects the great interest that cybersecurity, cyberrisks and their management attract (von Solms and von Solms, 2018).

The frequent reports on current vulnerabilities and successful attacks have deteriorated trust in digitalized systems which represents a heavy burden to a sustainable digitalization of modern societies and economies as well as ensuring security of digitalized services as a major concern for their implementation and development of businesses based on these services (e.g., FinTechs; Stewart and Jürjens, 2018; Ani et al., 2017; Wong et al., 2012). Furthermore, ensuring secure systems is essential for their adoption (Stewart and Jürjens, 2018; Ani et al., 2017; Wong et al.,

2012). Thereby, in particular services relying on self disclosure or data sharing of their customers are highly dependent on the trust and hence, the security offered by the service provider for being adopted (Pawlick et al., 2019; Stewart and Jürjens, 2018). This is, trust takes a paramount role in every business relationship (Stewart and Jürjens, 2018; Wiens, 2013).

Trust is a multidimensional phenomenon that is defined by multiple elements (e.g., trust regarding the security of a system might be influenced by the triad of confidentiality, integrity, and availability or extended definitions of security including concepts like authentication and accountability as well as usability; Stewart and Jürjens, 2018; Wiens, 2013). It is in particular important in situations characterized by high levels of uncertainty or when deciding on the usage of products and services that expose its users to various risks (Stewart and Jürjens, 2018) and refers to positive beliefs about a specific characteristic (e.g., security, dependability, reliability or confidentiality of an entity; Pawlick et al., 2019) and is consequently closely linked to the concept of reliability. Decisions regarding investments in cybersecurity are such situations, where trust plays an important role. This is, a great need for enhancing trust in digitalized systems can be noted which is essential when aiming at being able to leverage on the benefits of digitalization (Lv et al., 2020). Efficient and holistic cyberrisk management considering different domains of cyberrisks is essential in enabling smart, digitalized, and modern societies and economy leveraging on the advantage that are promised by technological advances (Geer et al., 2003). Hence, cyberrisks and unresolved cybersecurity issues of technology are a key obstacle to digitalization, technization, and intelligentization (Chang et al., 2021).

Infobox 3: Media coverage analysis regarding cyberattacks

The analysis of media coverage of cyberattacks is based on the LexisNexis media database and research tool. LexisNexis represents an academic research tool that provides access to the most important international press sources. The search was conducted for five countries (United States of America, United Kingdom, Israel, Germany, and France). Details regarding the search term and resulting articles were presented below.

Country code	Language	Search term	Title of news paper	Number of articles
USA	English	cyber AND (risk* OR secur* OR attack*)	The New York Times	2,589
			Tampa Bay Times	732
			USA Today	867
			Star Tribune (Minneapolis MN)	363
			Wall Street Journal Abstracts	231
			Los Angeles Times	26
			Los Angeles Times Online	18
UK,	English	cyber AND (risk* OR secur* OR attack*)	Metro	757
			Daily Mail and Mail on Sunday	2,274
			The Times (London)	3,797
			Mirror (Daily and Sunday)	2,267
			The Guardian (London)	5,095
			The Independent	6,695
			The Daily Telegraph (London)	3,772
			The Daily Telegraph Online	143
			The Sunday Telegraph	624
ISR	English	cyber AND (risk* OR secur* OR attack*)	Jerusalem Post	1,850
			Haaretz	37
			The Times of Israel	911
GER	German	cyber UND (risik* ODER sicher* ODER attack* ODER angr*)	Der Spiegel	241
			Spiegel Online	914
			Die Zeit (incl. ZEIT Magazin)	137
			Zeit-online	225
			Der Tagesspiegel	504
			Die Welt	647
			Welt Online	1,537
			Die Welt am Sonntag	209
			taz, die Tageszeitung	313
FRA	French	cyber AND (risqu* OR sécur* OR attaqu*)	Le Figaro	608
			L'Express	71
			Sud-Ouest	480
			Les Echos	926
			La Tribune	825
			Ouest-France	1,121

Table 2.1: Literature search details

The systematic literature search resulted in 41,298 articles from the five countries. After removal of duplicates, the final set included 38,378 articles. The distribution over time of these articles for each country is presented in figure 2.2. Thereby, a normalization according to the number of newspapers per country is given. The figure gives evidence to the rising media interest in cyberattacks and shows increasing media coverage. The trend towards an enhancing media attention towards cyberattacks is observable for each country. The increasing number of reported cyber-incidents throughout the years is significant.

However, the observed trends may be caused by different reasons, whereby rising numbers of cyberattacks may be a potential contributor. With respect to the amplifying interest of cyberrisk related issues, the rising number of published reports on cyber-incidents can also be (at least partially) explained (Evans et al., 2016). Hence, the strengthening awareness and public interest in cyberattacks might also contribute to that trend.

Besides different importance of cyberrisks for each country (e.g., as a result of the economic structure when high shares of the gross domestic product are gained within a highly susceptible branch), the deviations between the countries can also be caused by the underlying data. As the number of different newspapers is low for each country, the reasons for high deviations between the considered articles may be hard to interpret. Besides different interest of people (e.g., as a cause of the importance of cyberrisks for businesses and economic prosperity like in economies highly dependent of highly technologized branches such as the banking branch), biases within the data structure could have a significant effect on the number of published articles (e.g., different distributions of daily/weekly news papers). Thus, cross country comparison should be done by a review which takes these aspects into consideration.

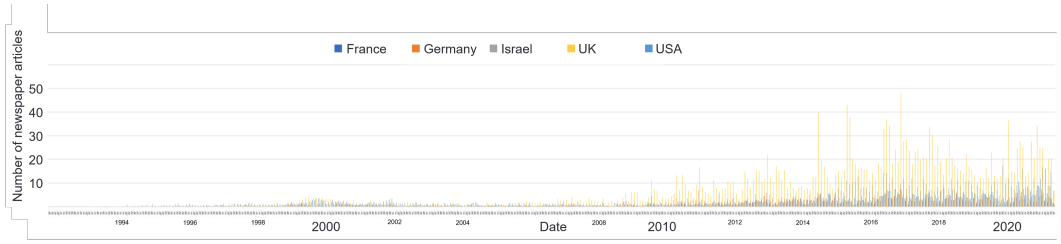

Figure 2.2: Time series of the number of media articles

Besides its effects on awareness, reports on cyberattacks led to an increase in public knowledge regarding cybersecurity and cyberrisk management (Evans et al., 2016). At least, the media coverage led to a situation where some buzz words from the field of cyberrisks came to the "forefront of people's minds" (Evans et al., 2016). With this growing smattering on cyberattacks, resulting risks, as well as preventive means for security management, broad awareness of internet users was generated, so that interest in management of these risks in the private as well as in the economic sphere increased.

Like public coverage within news papers, also academic research on cyberrisks spikes. However, academic research on cyberrisk is still limited compared to the high importance for economic prosperity, reported public interest, and media coverage (Eling and Wirfs, 2019). Yet, scientific research gained considerable shares of scientific investments as well as research efforts and thereby especially research on cybersecurity and cyberrisk assessment, attack and countermeasure modelling for cyberrisk quantification, measurement and assessment of the resiliency of a system (Q. Zhang et al., 2015; Pudar et al., 2009). With this, cyberrisk quantification and management became a research hotspot (H. Zhang et al., 2018). Consistently, there is a growing set of literature on cybersecurity (Eling and Wirfs, 2019).

Furthermore, politics pushed forward research and development in cyber-capabilities calling for greater levels of security and resilience (DiMase et al., 2015). Cyberrisk management is hence of great political importance and high upon the agenda of political parties (Ganin et al., 2020; Sechser et al., 2019; Pupillo et al., 2018; Trump, 2017; Obama, 2013). Increasing defensive capabilities demand for a holistic approach to cyberrisk management considering technical, human and societal factors of cyberattacks. Albeit this demand for increasing cyber-capabilities, there are only minor advances in improving cybersecurity (Eling and Wirfs, 2019). In the course of this, strong asymmetries in defense and attack costs increase the need for effective cyberrisk management. This is, launching cyberattacks is disproportionally inexpensive compared with the consequences to which the attack could lead and the costs for decreasing the attack surface in the highly interconnected society and economy (Ganin et al., 2020). While defenders need to comprehensively protect against the full scope of possible attacks, attackers can focus on a specific attack (Ganin et al., 2020; Almasizadeh and Azgomi, 2013). Furthermore, attacker adapt to the defensive measures employed within a targeted system and have access to a medley of efficient and simple attacks and tools for hiding from safeguards reflecting a significant *second mover advantage* of the attacker (Almasizadeh and Azgomi, 2013). Also, the level of automation is considerably higher on the attackers' side (Almasizadeh and Azgomi, 2013).

Although managers increasingly consider cyberrisk management to be an essential responsibility they need to take, the high importance of cyberrisk management is frequently not reflected in their actions. Management of cyberrisks should be considered an optimization problem of a cost-benefit-calculus where the prevention of losses induced by a cyberattack demands for continuous investments in security means (Bojanc and Jerman-Blažič, 2008). In this sense, cyberrisk management and ensuring security involves a trade-off including monetary as well as non-monetary components (e.g., convenience or productivity; Falco et al., 2019). Yet, in practice, the expenses in security are treated as spending rather than as investments (Bojanc and Jerman-Blažič, 2008) and hence, do not get the strategic importance they should be treated with. A common perception is that with "higher investment in information security comes more protection and resilience to malicious attacks" (Ryan et al., 2012). This is, investment sums in means for ensuring information security are spiking (Limba et al., 2017; Bojanc and Jerman-Blažič, 2008). Yet, security professionals still consider scarcity of security budgets a key obstacle to the security of systems and observe vulnerable systems (Limba et al., 2017; Bojanc and Jerman-Blažič, 2008). Hence, it is questionable whether the common perception comes close to reality or if misallocations in cyberrisk investments are dominant so that the perception does not hold in reality (Ryan et al., 2012).

Hereby "throwing resources at the problem" (Poppensieker and Riemenschnitter, 2018) is neither efficient nor effective in cyberrisk management and will not increase a firms capability of keeping pace with "the blistering pace of changes in cyberrisk" (Poppensieker and Riemenschnitter, 2018). Rather, "blunt implementation of controls across all assets is a key factor behind cybersecurity waste and productivity loss" (Poppensieker and Riemenschnitter, 2018).

As cyberrisks are not a limited technological issue, but do also comprise human and societal factors, as well as the multi-dimensionality of the causes of an attack, methods of their management need to accommodate to a holistic perspective on cyberrisks (El-Gayar and Fritz, 2010). However, although the multi-domain nature of cyberrisks is commonly agreed to, most methods that are currently discussed focus on a specific domain which in most cases is a solitary perspective on the technology (Ganin et al., 2020; Limba et al., 2017).

2.6 Safeguards, potential means of countering cyberrisks, and defensive capabilities

Efficient cyberrisk management comprises the optimality of security investment allocation to safeguards, and with this the selection of the most feasible safeguards from a growing number of options in the market of cybersecurity products. Investment decisions with regard to cybersecurity have the potential to diminish disruptions of operations and financial harm caused by cyberattacks (Poppensieker and Riemenschnitter, 2018). Risk management strategies can be differentiated along the risk management cycle as avoidance, reduction, transferring, and accepting (Masso et al., 2020; Teixeira et al., 2015; Bojanc and Jerman-Blažič, 2008).

These investments are integral to decisions regarding the digitalization strategy followed. With regard to these decisions, the marginal benefits of digitalization decrease while cyberrisks increase (Paté-Cornell et al., 2018). Thus, in some cases, the associated risk might outweigh the benefits linked to digitalization (Paté-Cornell et al., 2018). In this case, avoidance can be an efficient means of cyberrisk management (e.g., full internet connectivity; Bojanc and Jerman-Blažič, 2008). However, selecting efficient degrees of digitalization as a means for cyberrisk management is oftentimes out of the scope of scientific research. Furthermore, variations in the degree of digitalization are out of the merits of models for cyberrisk management.

In the focus of current cyberrisk management decisions are reduction strategies referring to the implementation of feasible safeguards for ensuring cybersecurity (Bojanc and Jerman-Blažič, 2008). Reduction is oftentimes in the focus of cyberrisk managers and a primary means for ensuring appropriate levels of cybersecurity (Bojanc and Jerman-Blažič, 2008). Hereby, technological upgrades and updates are a first means of ensuring cybersecurity. Linked to this security means, proper *patch management systems* can represent a significant safeguard and an efficient solution to security of systems and networks (Ganin et al., 2020; Sengupta et al., 2020). Operators thereby, however, frequently fear the degradation of quality and productivity losses when updating their systems (Sengupta et al., 2020). Losses to productivity need to be taken into account when calculating the efficiency of this safeguard. Yet, keeping up to date with technological development does not provide high levels of security as novel products are oftentimes insecure (i.e., and the development of secure systems is economically not feasible, possible efficiency constraints of technology and engineering secure systems is limited by bounded rationality of human developers; Almasizadeh and Azgomi, 2013). This is, under some circumstances, legacy systems might even be more robust than novel modern infrastructures (Zeng et al., 2015). It is thereby a widely unsolved problem of answering the question, whether a system is more secure after a patch than before, as even if the patch closes a vulnerability correctly, novel vulnerabilities might be introduced (e.g., zero-day vulnerabilities; Strom et al., 2017). Hence, for managing risks, a first means is vulnerability management and specifically the reduction of vulnerabilities (Hentea, 2008). Furthermore, updating technology within systems may also have impacts on the continuation of operations and may affect the performance and interoperability of systems (Wu et al., 2018).

Cryptographic methods and tools represent a further means for ensuring data integrity (P. Xu et al., 2017). Encryption can be seen as one of the most common and most successfully used approaches to ensuring security of systems (Kang et al., 2011). Recently, the use of *blockchain* technologies gained in public interest and its use has grown significantly in business applications (e.g., power systems; Dehghani et al., 2021; B. Wang et al., 2019). However, the continuous use of these methods can introduce significant costs, id est indirect costs of the implementation of the cryptographic methods in the form of computation and communication overheads that can affect performance, safety, and reliability (Jovanov and Pajic, 2019; Zheng et al., 2016). Furthermore, passwords as security means represent one of the most frequently used techniques for ensuring cybersecurity that is based on cryptography (Böhme et al., 2019).

Intrusion detection systems are considered to be a helpful means for protecting systems and networks in various domains (Alhakami et al., 2019). These are introduced in great richness to the scientific

literature offering great effectiveness (Buczak and Guven, 2015). Equivalently, *anti-virus programs* can help to protect against malicious programs. Anti-virus programs can thereby be classified either as being signature- or behavior-based (Sengupta et al., 2020). *Firewalls* are a further frequently employed safeguard. These have the potential to filter incoming traffic like anti-virus systems these can either be signature- or behavior-based (Allodi and Massacci, 2017).

Intrusion response systems provide detective capabilities and reactive capabilities (Chen et al., 2014). Yet, current systems suffer from practical limitations and high operation overheads due to imprecision (e.g., high false positive rates; Chen et al., 2014). Intrusion response systems can form the basis for generating self protective means in systems which are characterized by the ability to eliminate or mitigate cyberattacks without the need of human intervention (Chen et al., 2014). Given the high scarcity of human resources (scarce time and availability of security personnel), self protective means pose a great promise for ensuring cybersecurity.

Besides these safeguards focusing mainly on the technological domain, there are safeguards that are directed more towards the human and societal domain. These are inter alia *awareness and phishing detection training* (i.e., sending a mail with a manipulated link leading to manipulated websites with the aim of accessing the users credentials; Ganin et al., 2020). Equivalently, social engineering attacks are directed towards the human and societal domain. A social engineering attack can be defined as "persuading a person to allow the attacker increased access to the system that can later be exploited in a malicious way" (Ganin et al., 2020). This is, phishing detection capabilities and reactions to phishing mails are observed to be highly dependent on the humans' detection ability, environmental factors (e.g., working situation and usability of technology) as well as perception of possible consequences (Canfield et al., 2016). Phishing training is necessary and important as phishing attacks are not preventable by technological means entirely (Canfield et al., 2016) and attacks targeting the human domain are frequent (Ganin et al., 2020). Training of personnel, aims at decreasing the exploitability of human and societal vulnerabilities (Ganin et al., 2020). Further resiliency can be gained trough introducing redundancies both within the network structure (e.g., possibilities to switch back to manual operation) and within the implementations of safeguards (i.e., defense in-depth or multilayered defense). However, redundancies (e.g., excessive use of backups or the operation of a dual system of data storage) might be non-economic and hence, not possible to implement within an economic system (Mishra et al., 2020).

Within an attempt to generate "a precise, unambiguous, and information-dense knowledge graph of cybersecurity countermeasures" (Kaloroumakis and Smith, 2021), researchers generated a graph that builds a comprehensive basis and systematizes available countermeasures for cyberrisk management. The knowledge graph on countermeasures comprises information on "how those threats are addressed from an engineering perspective, and under what circumstances the solution would work" (Kaloroumakis and Smith, 2021). Knowledge for building the systematic overview on available countermeasures is extracted form literature (e.g., analysis of patents; Kaloroumakis and Smith, 2021). The potentials of countermeasures are thereby reported according to attack means offering vendors the possibility to "describe what specific adversary behaviors their products were able to detect, prevent, or monitor" (Kaloroumakis and Smith, 2021).

Risk transfer is a frequently discussed means of cyberrisk management that gained considerable importance within the last years (Bojanc and Jerman-Blažič, 2008). Transferring risks frequently is associated to the buying of appropriate insurances providing coverage to some risks (Bojanc and Jerman-Blažič, 2008). The market for such *cyberrisk insurances* is observed to grow steadily within the last years (Shetty et al., 2018). However, insurance coverage is limited due to multiple issues challenging the insurability of cyberrisks (e.g., correlated losses and interdependent security; Shetty et al., 2018; Böhme and Kataria, 2006).

Risk acceptance refers to the last technique for cyberrisk management. Accepting risks is a feasible strategy for risks where the investment costs for safeguards, cyberrisk insurances, and avoidance strategies are not feasible. This is if the costs for taking active risk management strategies are greater than the expected losses inflicted by a specific threat (Bojanc and Jerman-Blažič, 2008).

Although there are many defensive methods for cyberrisks, ensuring an efficient functioning of digitalized systems and with it, modern societies and economies, requires significant advances in defensive capabilities (Hahn and Govindarasu, 2011). This is, capabilities of defending against perpetrators of cyberattacks can be considered an important determinant of the performance of cyber-physical systems (Lv et al., 2020).

3 Literature Review on Cyberrisk Quantification and Management Methodologies

> The degree to which you can express something in numbers is the degree to which you really understand it.
>
> William Thompson, "Lord Kelvin" (1824–1907)

3.1 Methodological procedure and literature search

Research on cyberrisk analysis spikes and knowledge is developed with high pace. This is, many different methodologies are presented in scientific literature (Bojanc and Jerman-Blažič, 2008; Nicol et al., 2004). Yet, comparing qualitative approaches and quantitative approaches for cyberrisk analysis, there is an imbalance towards qualitative methods and research on quantitative methodologies is comparably scarce (Patel et al., 2008; Nicol et al., 2004). Regarding cyberrisks, there is a plethora of different quantification methods(Nurse et al., 2017) and combinations of different methods in the sense of multi method approaches or multi model based approaches (Q. Zhang et al., 2015).

This section reviews existing quantification methods and highlight their essential limitations. By doing so, scientific research gaps are identified and the need for alternative means for cyberrisk quantification is substantiated. In this section, a systematic literature review in the form of a content analysis is presented. The systematic procedure for the search and analysis of the literature ensures that - as far as possible - no relevant literature is omitted and that the review can be utilized for analyzing the current state of research. The content analysis is performed in a quantitative as well as in a qualitative way. Thereby, it places the work in the wider scientific context and identifies research gaps within the literature.

The literature review is oriented on the following research questions:

(i) Which methods for cyberrisk quantification are given in the scientific literature?
(ii) Which methods for supporting managerial decisions (cyberrisk management methods) are provided by scientific literature?

Content	*"cyber"* AND (*"security"* OR *"risk*"*) AND (*"quanti*"* OR *"manage*"*) AND *"method*"*	Title-Abstract-Keywords
Type	Proceeding OR Journal Article	
Language	English	

Table 3.1: Details on the inclusion criteria for the systematic literature search

For answering these research questions, a systematic literature search was conducted. The selection of keywords for the search term was extracted from the research questions and combined relying on Boolian logic as given in table 3.1. Keywords were truncated when feasible to reach comprehensiveness of the literature search. For the literature search, the Elsevier Scopus literature base of peer reviewed scientific literature was used. It offers possibilities for systematic literature searches providing a comprehensive coverage of high-quality scientific journals and research articles. Furthermore, features for automated literature extraction are given diminishing errors. The search term is presented in table 3.1.

The literature search results in 1,018 journal articles or proceedings. After scanning the literature, 25 articles were excluded for being duplicates. Furthermore, 81 articles were excluded for not providing insights on the research questions. The quantitative content analysis was conducted on the resulting 911 articles. For the qualitative analysis, the set of articles was reduced by the inclusion of a minimum citation criteria. The threshold was set to 25 citations. A further criterion was the check, whether each important keyword (identified through the quantitative content analysis esp. the clustering approach that was undertaken for the literature mapping) was included within those articles. The qualitative content analysis comprises 140 articles.

3.2 Literature mapping and quantitative content analysis

Figure 3.1 presents the count of publications and citations within each year. The figure shows the increasing scientific relevance of cyberrisk quantification methods and management practices.

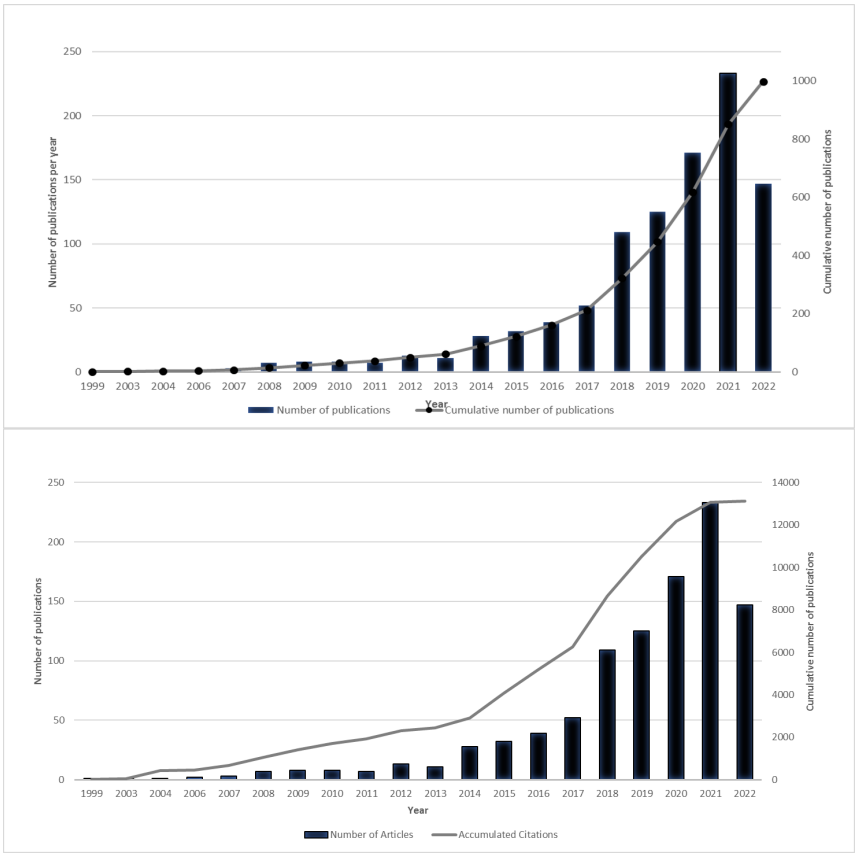

Figure 3.1: Time series of the count of publications (left) and citations (right)

Furthermore, figure 3.2 shows research hotspots and highlights differences in research interest within different countries. Each bubbles size thereby encodes for the number of publications from researchers of a country. The research hotspots highlight differences in research effort dedicated to the topic of cyberrisks in different countries.

Figure 3.2: Map on research hotspots for cyberrisk quantification and management

For the quantitative content analysis, a keyword co-occurrence map is presented in figure 3.3. The figure shows a bubble chart where links between each bubble are given. The strength of each link is determined by the number of co-occurrences within articles. Furthermore, the size of each bubble is defined by the number of occurrences within the scientific literature. The clustering approach identified three clusters of different scientific approaches for cyberrisk management and quantification. These are deductive approaches (red), inductive approaches (blue) as well as mixed methods (green; mainly the use of automation tools and algorithms for cyberrisk quantification respectively cyberrisk management). The co-occurrence map shows in particular the importance of game theoretical methodology for cyberrisk management and quantitative assessment. Furthermore, the use of Bayesian networks, artificial intelligence and deep learning is frequently discussed.

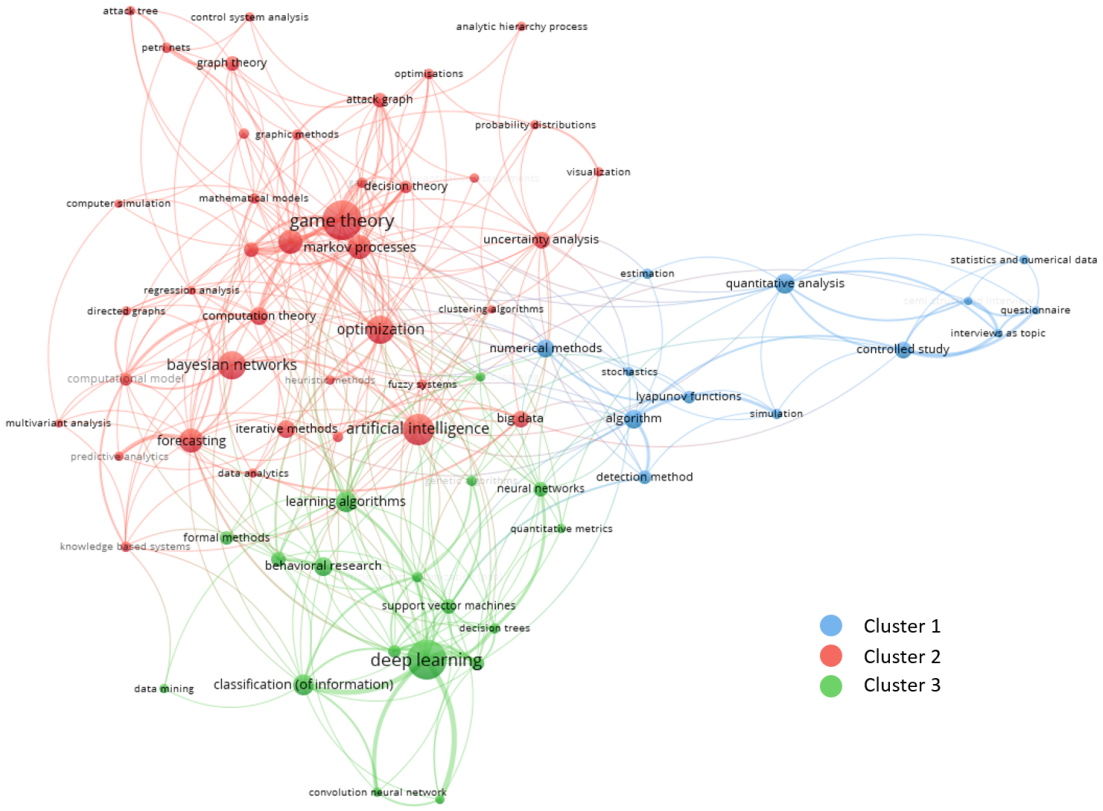

Figure 3.3: Co-occurence map of scientific methods for cyberrisk management and quantification

3.3 Qualitative content analysis

Some of the most widely used approaches in cyberrisk asssessment and management, are semi-quantitative or even qualitative (e.g., norm based cyberrisk assessment frequently relying on qualitative scales of risks ranging the risk on a equidistant risk scale from "low" to "high"; Polatidis et al., 2018). These are frequently based on risk rating approaches for coming up with numeric values for attack likelihoods and the impact of a specific attack. Such simplistic semi-quantitative approaches based on risk rating approaches and risk matrices are also used in norm based cyberrisk

respectively management (e.g., ISO/IEC27001; Allodi and Massacci, 2017; Nurse et al., 2017). Norm based cyberrisk management (ISO/IEC27001) represents today's standard in the field of cyberrisks and legal requirements build a key driver for security investments (Evans et al., 2016). Yet, given the dynamic development within the field of cyberrisks, legal requirements and norms need to be adapted to keep pace with this development and being more stringent with the current threat landscape to be able to provide efficient security (Evans et al., 2016). Risk metric calculations based on these simplistic semi-quantitative approaches can be criticized for high levels of subjectivity and the reliance on arbitrary risk factor scales (Ralston et al., 2007). Furthermore, risk matrices need to be questioned for their loss of resolution that is caused by the classification within categories provided within the matrices (Allodi and Massacci, 2017). From a management perspective, cyberrisk management and investment optimization based on return on investment (respectively adaptions such as the return on security investment) metrics, return on attack, net present value methods or methods based on the internal rate of return were studied (Pudar et al., 2009; El-Gayar and Fritz, 2010). However, the main obstacle for efficient cyberrisk management seems to be the challenges of cyberrisk quantification (chapter 4.2) and the need for adapting traditional cyberrisk management practices by means of presenting quantitative cyberrisk metrics that enable the management of cyberrisks by means of established methods in risk management (Knowles et al., 2015; Almasizadeh and Azgomi, 2013). The limited ability in cyberrisk quantification thereby constrain the application of risk management methods (El-Gayar and Fritz, 2010).

The predominant use of non-quantitative risk assessment approaches (Nicol et al., 2004) shows the low maturity levels of cyberrisk quantification. The adoption of these approaches can be considered as a three phased process where the first step is the adoption of best practices without relying on risk quantification (Allodi and Massacci, 2017). The second phase is the integration of quantitative security analyses (Allodi and Massacci, 2017). Within the last phase, quantitative security respectively risk analysis is trusted and decisions are based on risk and security metrics (Allodi and Massacci, 2017).

Furthermore, cyberrisk quantification approaches are oftentimes domain-specific to enable an appropriate cyberrisk quantification that takes into account the specifics of the domains (e.g., critical infrastructures). Also, current cyberrisk quantification methods are domain specific focusing on specific aspects or specific components of cybersecurity. For example, some approaches focus on dependability (Nicol et al., 2004) or the quantification of the consequences of an attack (Ganin et al., 2020; Ryan et al., 2012). A consequence of the domain specificity is the use of multi criteria decision analyses in cyberrisk quantification and management (Ganin et al., 2020; El-Gayar and Fritz, 2010).

Within this qualitative content analysis, an overview shall be given on the methods and the current status of research on these methods that are discussed in scientific research. The classification of proposed cyberrisk quantification and management practices seems to be challenging. In literature, classifications of methods are presented that address the differentiation based on which aspects or domains of cyberrisk were included (Ralston et al., 2007). A further approach for categorizing the methods discussed is based on the divers phases of the risk assessment process, which are risk identification, risk analysis, risk evaluation and ranking, as well as risk management

and treatment (Ralston et al., 2007). Also, a differentiation between deductive and inductive approaches is followed. In this context, deductive approaches outline knowledge creation that is based on mathematical analyses and formal logic, where inferences are made from theoretical considerations on the causes and effects as well as their interrelations (El-Gayar and Fritz, 2010). It is considered to be normative and rational. Contrary, inductive approaches describe knowledge creation through observations and their analysis, which is empirical and objective (El-Gayar and Fritz, 2010). The work follows this classification of methodological approaches to deductive and inductive approaches and classifies the content of the literature according to this schema. Furthermore, another category is introduced that describes the prerequisites of each approach, which are data extraction methods and methods for data representation as a first step of data preparation.

3.3.1 Data extraction methods and approaches for data representation

The lack of data on cyberattacks that were executed in real systems challenges cyberrisk quantification. Therefore, a first section of the literature on cyberrisk quantification and management methodologies shall be dedicated to this scarcity of historic data and scientific approaches to tackle this challenge. The section focuses on reviewing different methods discussed in scientific articles and providing a concise picture on the different approaches.

Data for cyberrisk quantification is oftentimes derived from expert estimations, questionnaires, surveys and the mining of their opinions or judgements given a lack of rich enough empirical data that has predictive as well as descriptive power and is available to cyberrisk quantification (Chang et al., 2021; Ganin et al., 2020; Cilliers, 2020; Q. Zhang et al., 2015; Ryan et al., 2012; N. Liu et al., 2010; Bojanc and Jerman-Blažič, 2008). Expert estimations represent a commonly used method, if not the most popular method for deriving insights on cyberrisks and vulnerabilities of a system (Ganin et al., 2020). Although challenging, expert based risk assessments can provide analysts with estimates on conditional probabilities of attacks and harm parameters (Q. Zhang et al., 2015). This is, expert opinions are currently oftentimes essential for calibration of risk quantification methods and their use is justified by the lack of quantitative data and its questionable or insufficient quality that oftentimes challenges the use of inductive, statistical cyberrisk quantification (Ryan et al., 2012). Cyberrisk assessment based on expert's beliefs is inter alia promoted by international standards in cyberrisk quantification and frequently referred handbooks for cyberrisk management (Allodi and Massacci, 2017). Furthermore, expert knowledge might be extracted relying on literature reviews and their analysis (Ganin et al., 2020; Atoum et al., 2014) as well as content analyses on the basis of social media postings from expert communities (Lee et al., 2017). Assisting literature reviews, practical experiences can be included (Atoum et al., 2014). However, cyberrisk assessments based on expert estimations and hence, the quantification of cyberrisks as a function of expert opinions can be criticized for being biased as well as for being time intensive and for the resulting low reproducibility (Allodi and Massacci, 2017). These biases are inherent to their estimations due to limited perspectives, high levels of uncertainty, and bounded rationality of

experts (Ganin et al., 2020; Allodi and Massacci, 2017). The feasibility of expert estimation based cyberrisk assessment is thereby questioned in terms of high complexity of systems, strategically acting adaptive attackers, and the dynamic threat landscape (Komljenovic et al., 2016).

Another frequently used technique in cyberrisk quantification is the use of experimental techniques and simulations for estimating properties of an underlying system (Ralston et al., 2007; Nicol et al., 2004) or the reliance on *honypots* that monitor, collect, and hence generate real data on attacks, attackers, attack methods, and attack techniques from online attacks on these systems (Ryan et al., 2012). In this sense, experiments can be used to derive information on parameters that can later on be used by deductive, model based approaches or inductive empirically based approaches to cyberrisk quantification. Experimental approaches hereby suffer from challenges when employed in systems that need to be in operation (e.g., due to criticality of operations; N. Liu et al., 2010). Yet, they provide the possibility of simulating various different attacks, vulnerabilities, and attack vulnerability combinations in different situations providing insights (e.g., loss frequency and severity) on different attack scenarios or case studies (Colicchia et al., 2018; Eling and Wirfs, 2019; Ruan, 2017; Q. Zhang et al., 2015; T. Liu et al., 2015; Genge et al., 2015; Hahn and Govindarasu, 2011). This can be to run an attack on a system to test the security properties and their functioning.

Simulations can be based on statistics on past attacks and may represent scenario analyses. Regarding the use of simulations for analytical purposes, especially the simulation of both the cyber and the physical layer is essential. In this way, the physical response of systems on cyberattacks and cascading effects of cyberrisks within interconnected systems can be assessed (Housh and Ohar, 2018; Wu et al., 2018; X. Liu et al., 2016). Methods for cyberattack simulation include Monte Carlo simulations (Eling and Wirfs, 2019; Ruan, 2017; X. Liu et al., 2016), discrete event simulations, historical simulations (Ruan, 2017), and Markov models (Q. Wang et al., 2019). Within these systems, network traffic analyses and data mining techniques can be applied providing technical data on real attacks (K. Xu et al., 2008). Thereby, the collection of network traffic and its analysis provides promising insights for cyberrisk quantification and management of these risks. Network traffic can also be analyzed without experiments on attacks and is, hence, not linked to simulations or experiments (Knowles et al., 2015). Rather network traffic can be extracted from systems that are in operation and perceive real attacks. In this case, there would, however, be uncertainty regarding the causes that may lead to network traffic (e.g., if anomalies would be detected, further analyses need to be conducted to differentiate them regarding their cause - benign or malicious). Furthermore, methods for vulnerability assessment exist to test a system and its properties that is beside network traffic analysis. These are experimental techniques for penetration testing, as well as the performance of attacks within controlled environments (Allodi and Massacci, 2017; Ralston et al., 2007). For (Monte-Carlo) simulations, discrete event simulations are commonly used and provide methodological support for reasonable estimates for cyberrisk quantification (Ralston et al., 2007). In particular, the use of simulations (e.g. of different attack scenarios) is predominant for assessing the impact of successful cyberattacks (Moness and Moustafa, 2015). A further option for gathering data for cyberrisk quantification on the technical layer is the use of process algebra to infer insights on attacks such as the costs of conducting a successful attack on a system by mathematical analyses

of cryptographic protocols (Almasizadeh and Azgomi, 2013) as well as formal reviews of software functionalities (Hahn and Govindarasu, 2011).

Data gained from different methods can also be combined, where applicable for risk management. Hereby, the weighting of statistical inferences and expert estimation based inference is critical as quantification of combined approaches is oftentimes considered to be the weighted average of both inferences (Ryan et al., 2012). This is, great research efforts were focused on suitable weighting schemes for such combined approaches (Ryan et al., 2012). Yet, most of the time, semi-quantitative assessments based on expert estimations can be reached (Ganin et al., 2020). However, these are rather simple risk scorings (Ganin et al., 2020).

Given the high complexity within the field of cyberrisks, data representation is essential for being able to understand the interactions in cyberrisk management and estimating risk exposure of a system. This is, data visualizations are not only important for risk communication but for exploration and analysis of patterns in data (Mohammadpourfard et al., 2017). Some of the most widely used methods for data representation and visualization include scatter and bar plots, heatmaps, matrices (e.g., frequently referred to in norm based cyberrisk management), and graphs (McKenna et al., 2016). Graph based methods can give relieve to high computational burdens and enable analyses that are beyond pure black box methods (C. Chen et al., 2009). Within these graphical methods for data representation, graph based approaches take a special role because of their frequent use in risk management (Sengupta et al., 2020; Pasqualetti et al., 2015; Knowles et al., 2015). Attack graphs illustrate attack descriptions such as paths and the course of an attack in different nodes. Different nodes can thereby represent different layers within an attack ontology, where attack graphs can have multiple root nodes (e.g., attack goals Sengupta et al., 2020). Within these approaches, a plethora of different graph based approaches is discussed in literature. These are inter alia attack graphs (Polatidis et al., 2018; Hahn and Govindarasu, 2011), attack-state graphs (Nicol et al., 2004), compromise graphs (Pudar et al., 2009; Ralston et al., 2007), vulnerability graphs (Almasizadeh and Azgomi, 2013; Hahn and Govindarasu, 2011), exposure graphs (Hahn and Govindarasu, 2011), interaction graphs (J. Wang and Paschalidis, 2016), and privilege graphs (Almasizadeh and Azgomi, 2013) and their analysis. The quality of attack graphs is tightly linked to the understanding of the system and its properties, attackers action taking as well as their goals (Hahn and Govindarasu, 2011). Also, graph based approaches suffer the challenge of scalability to large networks and systems of systems (Hahn and Govindarasu, 2011). Thereby especially high computational burden of analysis, error prone graph development as well as subgraph isomorphism represent challenges for the practical application of graph based approaches for cyberrisk quantification when large scale systems as well as systems of systems are considered (C. Chen et al., 2009). In case of such bulky graphs and if isomorphisms exist, graph summarization techniques should be followed for enabling efficient analyses (C. Chen et al., 2009). Furthermore, error prone development of attack graphs demand for automation of attack graph crafting (Abraham and Nair, 2014).

Tightly linked to graph based approaches are tree based approaches. Hereby, trees are a special form of more generic graphs allowing only one root (e.g., attack goal) or vice versa, attack graphs can be seen as an extension of attack trees (Sengupta et al., 2020; Knowles et al., 2015; Hahn and

Govindarasu, 2011). Within tree based approaches, there is a variety of use cases representing different foci. These are inter alia fault trees, event trees, vulnerability trees, attack trees, defense trees, and their analysis (Paridari et al., 2017; Knowles et al., 2015; Pudar et al., 2009; Fovino et al., 2009; Ralston et al., 2007; Nicol et al., 2004). Quite simple approaches are frequently considered in cyberrisk quantification that are based on counting approaches. These oftentimes include the analysis of attack graphs. For deriving quantitative measures of cyberrisk, attack path lengths and quantities are used (Hahn and Govindarasu, 2011). Analysts can also refer to Boolean logic, if it is supported by the graphical representation (e.g., attack graph or attack tree; Q. Zhang et al., 2015). The analysis of attack graphs enables the identification of a system's strengths and weaknesses regarding the robustness against intrusion attempts (Pudar et al., 2009).

Fault and event trees are acyclic graphs with nodes representing system components and edges representing information flows (Nicol et al., 2004). The edges connect different nodes and are therefore able to show operational dependencies within the underlying, described system and its components (Nicol et al., 2004). Hereby, an interoperability input-output model such as a Leontief-based model accounting for issues of intra- and interconnections in infrastructures may be needed (Ralston et al., 2007). A root node can then determine whether a system is operable (no failure within the system) or not (Nicol et al., 2004). Fault trees without shared nodes are thereby equivalent to reliability block diagrams that are series parallel. The major advantage of fault trees compared to reliability block diagrams are in network analysis where shared nodes exist (Nicol et al., 2004). The analysis of fault trees demands for algorithms (Nicol et al., 2004). These approaches can be based on statistical principles (inductive means for cyberrisk quantification) or formal mathematical, deductive methodology. The use of fault tree analysis is predominant in reliability and availability modelling (security component specific approaches) or the modelling of the robustness of a system (Nicol et al., 2004). Fault tree analyses can draw on the knowledge of an organization's employees for identifying risks, estimating the current level of security and determining the feasibility of strategies for ensuring and enhancing the security of a system (Ralston et al., 2007). The analysis of fault trees is highly dependent on the analytical procedure followed and can in this sense either be based on logical or empirical inferences. This is, fault tree analysis can either be classified as "a deductive, failure based approach" (Ralston et al., 2007) or an inductive approach to cyberrisk quantification. Yet, frequently, fault tree analyses that are currently presented fall short in their ability of providing quantitative estimations and hence only "few quantitative results are presented" (Ralston et al., 2007).

Attack trees are closely tied to fault trees and provide a feasible means for formal analyses on the security of systems (Ralston et al., 2007). Yet, instead of modelling system failures as the root node, security breaches as a special case of system failures and the goal of an attack are the root node of attack trees (Nicol et al., 2004). The different courses of action of different attacks, are represented within the tree structure (different leaf nodes) decomposing each attack to a set of events (atomic attacks or attack steps; Hahn and Govindarasu, 2011; Nicol et al., 2004). Attack tree analysis follows the structure of the attack tree assigning node values (e.g., probabilities) to each node which represents a function of it's childrens' nodes (Nicol et al., 2004). The utility for cyberrisk quantification by means of attack tree analyses was demonstrated for industrial control

systems within different industries and their specifics (Knowles et al., 2015). Fault trees and attack trees provide a formal methodological way of describing the security of a system and its weaknesses (Nicol et al., 2004). These methods can be combined with event tree analyses, failure mode and effect analyses, couple failure mode analysis, and cause consequence analyses for deriving quantitative metrics on cyberrisks and giving estimates necessary for probabilistic cyberrisk assessment (Q. Zhang et al., 2015; Knowles et al., 2015; Ralston et al., 2007).

Furthermore, the use of vulnerability trees and their combination with attack trees in the form of an augmented vulnerability tree for cyberrisk quantification is discussed in literature (Ralston et al., 2007). All tree based approaches are similar in the possibilities of analysis mainly differing in their root nodes (top events; Ralston et al., 2007). The graph based approaches are most often used as qualitative approaches for logical data representation (e.g., for describing and structuring the threat landscape) and only in the second place used for cyberrisk quantification (Ralston et al., 2007). Hence, they provide important insights on the threat landscape and possibilities for the analysis of data (e.g., allowing the use of algorithms for attack path discovery; Polatidis et al., 2018; Almasizadeh and Azgomi, 2013) which is a necessary prerequisite for cyberrisk management and risk quantification. Combining both sides (attacker side as well as the defender side) can enable a precise and comprehensive picture on the threat landscape (Fovino et al., 2009).

For analytical purposes, attack graphs can be weighted (Pudar et al., 2009; Fovino et al., 2009). This means assigning values to the nodes, which can either be Boolean (e.g., possible or impossible) or continuous (e.g., assigning probabilities, attack costs, or values representing attack tendencies to the attack paths; Nicol et al., 2004). The weighting of graphs is a crucial step for enabling accurate and precise assessments. Yet, the process of assigning weights is scarcely researched. Rather, weights are oftentimes simply estimated based on expert opinions (Pudar et al., 2009). Hence, estimations on the weights of different attack paths oftentimes suffer the lack of objectivity, transparency, and reliability (Pudar et al., 2009; Allodi and Massacci, 2017). As there is high uncertainty and unknown variables that influence weighting, research on weighting cyberrisk graphs is essential (Pudar et al., 2009). These challenges of weighting attack trees, attack graphs as well as setting weights in other graphical representations is a main obstacle to practical application and a main reason for distorted outcomes of analyses (Pudar et al., 2009).

3.3.2 Deductive approaches for cyberrisk quantification and management

Deductive approaches are model based, formal methods were cyberrisk quantification is an inference of the models properties. Approaches for the quantification of the impact of a successful attack frequently take advantage of means for process modelling (Knowles et al., 2015) and their simulation (Moness and Moustafa, 2015). Hereby, different scenarios can be simulated providing insights to scenario analytical purposes for cyberrisk quantification with a focus on the consequences of attacks (Ruan, 2017).

Petri net models were considered for cyberrisk quantification and metric based management as it represents a methodology that is able to model the interactions of the cyber and the physical layer within a combined comprehensive approach (X. Liu et al., 2016; Pudar et al., 2009). They can, hence, be used for attack modelling (Pudar et al., 2009). Within industrial control systems, petri net analyses show feasibility in providing reasonable estimates on the cyberrisk (Knowles et al., 2015). These formal analyses for security quantification are oftentimes limited to small parts of a system or network for sake of complexity reduction (Nicol et al., 2004). Taking advantage of stochastic distribution functions, Petri nets are also referred to as stochastic Petri net models (Almasizadeh and Azgomi, 2013). Also, applications of system dynamics can be used to provide risk estimates based on a deductive approach (Ameli et al., 2018; Genge et al., 2015). Furthermore, the use of Markov Reward Models (Nicol et al., 2004), beta mixture-hidden Markov models (Moustafa et al., 2018; Almasizadeh and Azgomi, 2013), and semi-Markov chains (Almasizadeh and Azgomi, 2013) as well as Markov and semi-Markov processes (Ralston et al., 2007) is discussed for deductive method based cyberrisk quantification calibrated with empirical data. Markov models and stochastic Petri nets thereby take advantage of stochastic distributions and models for analytically solving the problem of cyberrisk quantification (Almasizadeh and Azgomi, 2013). Hereby, setting the correct distribution function is essential and demands for sound statistical analyses form empirical data (Almasizadeh and Azgomi, 2013). Also fuzzy modelling is considered useful for cyberrisk quantification (Ralston et al., 2007).

Game theoretic methods, decision theory based models and agent-based models are further categories of quantification approaches in the field of deductive cyberrisk quantification (Knowles et al., 2015; Pudar et al., 2009; Bojanc and Jerman-Blažič, 2008; Ralston et al., 2007). Game theory can be considered a method for computing probabilities of attackers' behavior and provide insights to the quantitative analysis as well as input to stochastic models of security (Almasizadeh and Azgomi, 2013). Furthermore, game theoretical means provide insights on how to increase the effectiveness and efficiency of defensive means or strategies (Sengupta et al., 2020). The core assumption for applying inferring quantitative insights on cyberrisks is rationality or pseudo-rationality within the attackers' and defenders' decision making where decisions on conducting an attack are the outcome of a cost-benefit calculation (Katzir and Elovici, 2018). Attackers' rationality is supported by empirical findings including the identification of financial gain as a key driver for cyberattacks and the necessity of significant efforts in planning for reaching attackers' efficient resource allocation, coordination of atomic attacks and the need for strategic acting to capitalize on successful attacks (Katzir and Elovici, 2018). Extensions, relaxing the strong assumptions of rationality include the use of bounded rational decision making of human actors (e.g., attackers; J. Chen and Zhu, 2019).

For conducting the cost-benefit analysis and understanding attackers' action taking, game theoretical approaches take into account the consequences of an attack and their capitalization (with the possibility of including different motive structures), costs of conducting attacks as well as budgetary and resource limitations. A main advantage of game theory and its application to cyberrisk quantification is its ability to capture interactions between organizations, defenders, and attackers (C. Chen et al., 2020; Fielder et al., 2016). Within these game theoretic approaches, defender attacker games

are considered feasible for analyzing cyberrisks (Ameli et al., 2018) and have proven their practical usability (H. Zhang et al., 2018). Insights on attackers' behavior can by this means even be derived if appropriate information for statistical inferences is not available (H. Zhang et al., 2018). For this purpose, mathematical analyses of the game by means of determining the Nash-equilibrium are conducted (Pawlick et al., 2019; H. Zhang et al., 2018).

Game theoretic approaches can be devided in dynamic (e.g., sequential iterative games; Rana et al., 2017) and static games (Q. Wang et al., 2019) as well as within a categorization between deterministic and stochastic games (H. Zhang et al., 2018). Game theory can help in cyberrisk quantification as well as in their management. Insights derived from the analysis can be inter alia utilized for the optimization of resource allocation and defensive resource deployment as well as understanding attack and defense choices (Q. Wang et al., 2019). Game theoretic and decision theoretic methodology hereby oftentimes rely on information on the systems properties and the attackers preferences as well as motivations for estimating likely attack goals, describing adversary behavior (e.g., attack selection) and the results of such attacks (Almasizadeh and Azgomi, 2013). Works based on this methodology can be calibrated relying on different data including the use of cyber threat intelligence (H. Zhang et al., 2018). In recent years, automation of cyberrisk quantification is considered a key importance for cyberrisk management and increasing its efficiency. Approaches to automate deductive cyberrisk quantification thereby include approaches of adversarial learning taking advantage of game theoretical methodology and deep neural networks (Katzir and Elovici, 2018).

Combined, multi methods approaches are also used for cyberrisk quantification leveraging on the benefits that are provided by different methods. Furthermore, different viewpoints can be combined (e.g., vulnerabilities, faults and system failures, attacks; Hentea, 2008). Scientific works thereby propose the use of a combined approach relying on Markov chains, Petri nets and Bayesian networks (X. Liu et al., 2016) as well as combinations of clustering approaches and game theory (H. Zhang et al., 2018).

Although, there is much research and a plethora of different methods for a quantitative, model based evaluation, novel quantification methods are needed for quantifying cyberrisks in a methodologically sound and holistic way. This is inter alia as there is no standard methodology that is commonly accepted within cyberrisk management (Bojanc and Jerman-Blažič, 2008) and currently available methods mainly focus on specific domains of cybersecurity (e.g., technology).

3.3.3 Inductive approaches for cyberrisk quantification and management

Inductive approaches for cyberrisk quantification and their management comprise some of the most commonly used approaches within cyberrisk management. These inductive approaches take advantage of data science and methods for conducting inferences from data (McKenna et al., 2016). These are direct statistical analyses of historical data on past recorded attacks, regressions, time series analyses, and the analysis of evidence on incident counts or expert estimations (H. Zhang et

al., 2018; Q. Zhang et al., 2015; Almasizadeh and Azgomi, 2013; N. Liu et al., 2010). Inductive cyberrisk quantification, hence, is a data driven task relying on the analysis of empirics on cyberattacks (Paridari et al., 2017). This is, these statistical analyses are also referred to as empirical studies.

The quantification of cyberrisks by empirical studies was proposed for quantifying financial and safety impacts of cyberattacks (Paridari et al., 2017). Their application is quite common (Patel et al., 2008; Eling and Wirfs, 2019; M. Xu et al., 2018), although they suffer heavily from rare data. Depending on the available data, different statistical inferences for quantification of cyberrisks can be conducted. Within these statistical analyses, mean loss based cyberrisk extrapolations are frequent. However, the interpretation of mean values need to be reviewed critically (Eling and Wirfs, 2019). This is, especially loss distributions are highly skewed for cyberattacks (Eling and Wirfs, 2019). Also, such simple statistical analyses suffer from the fact that historic data might not reflect the future development which is even more important considering the highly dynamic nature of cyberrisks and the high velocity of technological development (Eling and Wirfs, 2019). However, the attackers behavior is observed to not follow any known distribution function that is used in statistical models (Pudar et al., 2009). With this, the use and reliability of such models is challenged and high precision of cyberrisk quantification is difficult to be reached (Pudar et al., 2009).

Given network traffic protocols, network traffic analysis can represent a helpful means for cybersecurity quantification. This is, in modern systems, network traffic analysis can be considered an imperative for cyberrisk quantification (K. Xu et al., 2008). Also. Network traffic canbe used for mimicking or twinning systems in operation. Furthermore, network traffic analysis is a feasible means for understanding the effects of dynamic alternations that take place in networks (e.g. through the permeation of a system with novel disruptive innovations) and may leave vulnerabilities within a system (K. Xu et al., 2008). In particular, semi-automated log data analysis gained considerable prominence in real time system analyses (Knowles et al., 2015). Log data analyses can thereby be used for estimating attack frequencies and probabilities (Patel et al., 2008). Furthermore, network traffic analyses can be used for real time threat analyses. Attacked systems show different behavior compared to systems in normal operation (Mohammadpourfard et al., 2017). This is, the detection of anomalies as a first step in real time threat analysis can be derived from network traffic analysis as the detection of deviations of current observations from normal empirical distributions of network traffic (Mohammadpourfard et al., 2017; J. Wang and Paschalidis, 2016). For doing so, anomaly detection follows statistical laws. Furthermore, the analysis of stock market prices and the analysis of market price reactions to attacks can be used as a means of cyberrisk quantification (Patel et al., 2008; Garg et al., 2003). When anomalies are detected, a next step is the classification of these to either benign or potentially harmful anomalies. This task is referred to as attack detection and is frequently based on statistical tests (Ameli et al., 2018). The duality of incident respectively anomaly and attack detection (which give estimates on the probability of an ongoing attack) can be combined with attack consequence prediction for means of cyberrisk assessment which is referred to as a dynamic risk assessment approach (Q. Zhang et al., 2015).

Given the high complexity and the big data analytics problem in cyberrisk quantification (especially referring to the analysis of technical cyber threat intelligence), the use of advanced approaches for data analytics is researched. Among these advanced approaches, the application of artificial intelligence and machine learning is central. Paridari et al. (2017) provide insights on the use of machine learning for cyberrisk quantification. Machine learning can be used in cyberrisk quantification and "allows to perform accurate quantitative assessments of security posture at a firm level" (Khalili et al., 2018). Hereby, cyberrisk detection can be based on statistical analyses and the use of rules derived from pattern recognition tasks (Paridari et al., 2017). Within this approach, the selection of appropriate rules and thresholds is essential. Further works focus on the use of deep learning, neural networks, and other algorithms in cyberrisk quantification (Taormina and Galelli, 2018).

Another approach is graph analytics leveraging on the benefits of graph based data representations 3.3.1. The analysis of graphs for identification of patterns and gaining understanding of the related concept in cyberrisk management is called graph pattern mining (C. Chen et al., 2009). Statistical models are thereby employed to derive insights (e.g., analysis of cyberattacks and their effects; C. Chen et al., 2009). Inter alia clustering approaches are used (C. Chen et al., 2009). The analysis of graph based data representations can follow statistical inferences such as Bayesian laws for information retrieval. An example is the analysis of Bayesian attack graphs (Wadhawan et al., 2018; Q. Zhang et al., 2017, 2015).

Within attack graph analysis, the most probable attack path is identified from a set of possible cyberattacks (Polatidis et al., 2018; H. Zhang et al., 2018). In this case, the graphical representations can serve as Bayesian networks (containing conditional probabilities on state transitions; Knowles et al., 2015; Q. Zhang et al., 2015) and cyberrisk quantifcation based on such Bayesian models (Alhakami et al., 2019; Sheehan et al., 2019; Huang et al., 2018; Q. Zhang et al., 2015). Bayesian networks are of particular importance for probabilistic risk assessment (Huang et al., 2018). The use of Bayesian inference for cyberrisk quantification is motivated by its benefit such as the robustness to over- and under-fitting other statistical approaches frequently lack of and its ability of formalization of uncertainty and knowledge (Alhakami et al., 2019). In particular, naive Bayesian approaches provide further benefits when compared with other methods, which is that large training sets for training purposes are not neccessary for reaching high accuracy and precision of estimations (Chen et al., 2014). Yet, the specification of priors (prior knowledge) represents a main disadvantage of relying on Bayesian methodology (Ryan et al., 2012).

Furthermore, Hoeffding's tree analysis can be used for deriving insights based on statistical means from graph based cyber threat representations as well as for deriving insights on how to effectively identify threats (attack detection) respectively on how to effectively defend against threats (Adhikari et al., 2017). Graph based approaches can also be a helpful means for analyzing the impact of attacks on a system (Q. Zhang et al., 2015). The use of graph based analysis is considered an effective means of cyberrisk assessment (Patel et al., 2008). The analysis of attack trees, fault trees as well as the reliance on graphs and their analysis can support the quantification of probabilities as well as attack impacts (Patel et al., 2008). However, changing threat landscapes, technological development, flexibility respectively variability of systems, and changes in the economic environment these are

employed (changes in businesses), challenge the use of graph based approaches (Q. Zhang et al., 2015).

4 Current status of and challenges in cyberrisk management and cyberrisk quantification

> In the era of digital information technology and connected devices, the most challenging issue is ensuring the security and privacy of the individuals' and organizations' data.
>
> (Conti et al., 2018)

Various quantitative approaches to cyberrisk management and approaches for cyberrisk quantification are introduced in scientific literature (chapter 3). This is, to date a plethora of different cyberrisk quantification methods is available (Shameli-Sendi et al., 2016). Yet, these have considerable shortcomings (refer to chapter 4.2). Sound quantitative cyberrisk management is hence not feasible in practice oftentimes. Consistently, today, quantitative, reliable security metrics with the ability to provide a holistic picture in the threat landscape are scarce (Hahn and Govindarasu, 2011). Today, a common understanding and a single upon agreed risk metric and methodology is not reached in cyberrisk management. Furthermore, the metrics that are currently used, are mainly comparative and do not reflect absolute, directly interpretive values (Ganin et al., 2020). Hence, currently, cyberrisk metrics are mainly unitless (due to their comparative nature) and do not allow the usage of conventional cyberrisk management practices established in cyberrisk management (e.g., the treatment of defensive investments with classical investment theory).

These conventional and established practices to cyberrisk management can foster the understanding regarding the importance of investments in cybersecurity and the benefits that come with increased cybersecurity (Falco et al., 2019). As those approaches are not feasible today, currently cyberrisk management is mainly based on checklist approaches such as those provided by international standards and norms for cyberrisk management (Böhme et al., 2019; Rees et al., 2011). However, compliance to these standards is only a weak guarantee for actual security (Böhme et al., 2019).

Furthermore, currently, cybersecurity, respectively cyberrisks, are poorly understood and sound cyberrisk management systems are scarce (DiMase et al., 2015). Essential for the scarcity of well grounded cyberrisk management approaches is the lack of appropriate, holistic risk quantification methods which frequently form the basis for risk management. This is, as the current understanding

of cyberrisks is beyond what is needed for accurate risk quantifications (M. Xu et al., 2018). Consequently, today, "many organizations are managing security in a somewhat inconsistent and superficial manner. Rather than taking a calculated or rational approach, they're emphasizing certain controls while leaving others, though no less important, poorly maintained" Baker and Wallace (2007), which is mainly due to limited abilities of consistent cyberrisk management. Hence, an existing "lack of comprehensive and rigorous theoretical fundamentals in the area of quantitative security evaluation has not been already filled" (Almasizadeh and Azgomi, 2013). This leads to a situation where today's cyberrisk quantification methodologies lack in their ability of providing efficient guidance to security professionals and risk managers (Ganin et al., 2020). For them, setting a precise metric for describing cyberrisks is hardly possible nowadays (M. Xu et al., 2018). In this sense, there are only limited means for cyberrisk management and risk quantification systems are largely fragmented (domain specific; DiMase et al., 2015).

Regarding the management of vulnerabilities, weakest link approaches are very common. Yet, to compose a holistic analysis of risks, methods limiting themselves on the weakest links are inappropriate or at least less promising than approaches that focus on all vulnerabilities (Holm et al., 2012). Attack paths describe feasible ways for attackers to get unauthorized access to a system and cause harm which is to violate the systems security (Almasizadeh and Azgomi, 2013). An attack path is followed successfully (i.e., an attack is executed feasibly) if all attack steps respectively all phases of the attack are executed with success (Almasizadeh and Azgomi, 2013). The steps an attacker need to take in order to successfully attack a system are described within the Cyber Kill Chain. The goals of cybersecurity are oftentimes defined as the triade of confidentiality, integrity, and availability (El-Gayar and Fritz, 2010).

4.1 Importance of cyberrisk quantification

Incident counts and reports on cyberrisks are perceived differently from decision makers. Hence, variations in risk perceptions may emerge that leads to different priorities professionals set in cyberrisk management (van Schaik et al., 2017). Considering these variations in risk perception and hence risk awareness, the need for concise upon-agreed cyberrisk quantification methods and derived quantitative metrics for describing cyberrisks gets obvious. Humans thereby use metrics as a means of understanding the world that surrounds them intuitively (Abraham and Nair, 2014). In this way, the use of metrics ensures that decisions within cyberrisk management are feasible reflecting a sound understanding of the cyberthreat landscape organizations are exposed to (Hahn and Govindarasu, 2011).

In this sense, the weak abilities in cyberrisk management and the dearth of feasible cybersecurity metrics represents a obstacle to the adoption of established risk management approaches to the domain of cyberrisks (Knowles et al., 2015). Decision makers in practice are thereby oftentimes confronted with dozens of security metrics (Boehm et al., 2017). Yet, they are usually inconsistent and of high granularity failing in providing a holistic picture of the cyberrisk exposure and convey

the impact of possible cyberattacks on business processes (Boehm et al., 2017). The risk status of a company, hence, oftentimes stays hidden to decision makers (Boehm et al., 2017).

Also, the use of metrics in cybersecurity can enable efficient communication of cyberrisks (Evans et al., 2016). This is, cyberrisk quantification is important to be able to give sound and meaningful statements on security (Nicol et al., 2004) and prioritize cyberdefense as well as scarce security budgets on specific vulnerabilities or against attacks the organization is most at risk (Kaminski et al., 2017; Fielder et al., 2016). Yet, determining the effectiveness of safeguards in enhancing cybersecurity today is very challenging (Hahn and Govindarasu, 2011). Cyberrisk quantification, however, is a necessary prerequisite for engineering secure systems (N. Liu et al., 2010). Due to the high practical importance of metrics on cybersecurity, enabling cyberrisk quantification is a priority of scientific efforts (Abraham and Nair, 2014).

Setting appropriate security budgets can only be based on a sound quantitative assessment of cyberrisks (Evans et al., 2016). This is, in practice, there is "no way of knowing what (...) (organizations) should be spending on cybersecurity" (Rauscher and Cox, 2013). Also, business decisions can only be based on quantitative security or risk metrics, in particular, if security investments should be a business case (e.g., when considering security as a selling point or competitive advantage; Ryan et al., 2012).

Insufficient budgets are frequently considered a top challenge by security professionals (Fielder et al., 2016). As stated, setting appropriate security budgets, however, demand for reliable quantification methods (Garg et al., 2003). Security metrics can be a helpful means for decision makers for objective quantification of the current security level, understanding the implications of cyberattacks, and their proceeding (Evans et al., 2016; Hahn and Govindarasu, 2011). Consequently, efficiently managing cybersecurity and risks of systems demands for quantification. For applying well founded risk management approaches to the field of cyberrisks, quantitative cybersecurity metrics are necessary.

4.2 Challenges in cyberrisk quantification

Besides the high importance of cyberrisk and its quantification for managing, there are many burdens to cyberrisk quantification of which the management of cyberrisks suffer. The difficulties of quantification result from several factors including the highly dynamic nature of attack behavior (dynamic development of attack strategies, multitude of flanks and dynamic choice of flanks), the inadequately determinable probability of attack in this context, the difficulty of estimating the reaction within the company (interaction of technological and behavioral or organizational factors), and the complexity of assessing the consequences of cyberrisks. Furthermore, a lack in historical data on past attacks. Hereby, patchy databases can be observed which is due to low willingness to share data. In particular data on mainly technical information (technical cyber threat intelligence) of forensic reports is available while information on impact, targets and frequency of attacks tends to be reluctant to share (Chismon and Ruks, 2015). The multi-dimensionality as well as the sheer volume of different attacks, potential attack surfaces, limited computational resources of human

actors as well as their bounded rational nature cause particular challenges in engineering secure systems (Zheng et al., 2016).

Consistently, there is a "lack of an established agreed-upon framework" (Zeller and Scherer, 2021) for risk quantification and based on it, decision support for professionals. Hence, today, compliance with security standards is the main driver of security investments rather than the aim for secure products, economic feasibility of investments, desire for protection or ethical reasons like moral obligations to protect customers' data (Zander et al., 2020; Evans et al., 2016). Yet, there is a wide variety of different risk quantification methods and the approaches that are used to measured cyberrisks are discussed controversially (Nurse et al., 2017).

4.2.1 Lack in historical data

Organizations can gain an overview on their network as well as measure and monitor the security of the system by penetration testing, vulnerability analysis, analysis of patching reports as well as internal and external incident statistics (Evans et al., 2016). Security event monitoring furthermore provides organization intern data on experienced attacks. Based on this data, organizations can accelerate the development of secure systems and their deployment (Hentea, 2008). Furthermore, companies can rely on the information gained for means of cyberrisk quantifications (Hentea, 2008).

Fears of liability, loss of reputation and competition among peers fuel the scarcity of historical data on cyberattacks representing an obstacle for data sharing on experienced security incidents (Limba et al., 2017; N. Liu et al., 2010) leading to a lack of sufficient (publicly available) data (Eling and Wirfs, 2019; Carfora et al., 2019; Nurse et al., 2017). Decisions on sharing threat intelligence represents a comparison between the costs and benefits for companies and individuals affected by an attack that oftentimes lead to a situation, where threat information is not shared (Menges, 2020). In this sense, affected companies suffering from cyberattacks mainly cover them up while security companies, forensic investigators, providers of safeguards, and consultants for cyberrisk management, either are not allowed to share data or/ and are not willing to share. Also, companies can leverage on this information, if they decide to keeping their data and experience private, for developing a "proprietary competitive advantage" (Ruan, 2017). Consequently, credible estimates on the (economic) consequences of a cyber-incident and estimates on the attack probability (e.g., via comprehensive cyberattack incident statistics) are limited (Zeller and Scherer, 2021; Garg et al., 2003) and pertinent data for cyberrisk quantification lacks in availability (Komljenovic et al., 2016). This is, currently predominant approaches in risk quantification lack in applicability within the field of cyberrisk management. Hereby, inter alia probabilistic cyberrisk quantification is challenged by the difficulties in assessing credible estimates on attack probabilities (Ralston et al., 2007).

Empirical data on cyberattacks are needed for cyberrisk quantification by enabling the parametrization and calibration of security models for measuring and analyzing the state of security (Almasizadeh and Azgomi, 2013). The lack of appropriate data on the level of the right abstraction level that can directly be used for cyberrisk quantification (i.e., directly applicable) significantly constrained

the capabilities of decision makers and risk management professionals to craft reliable estimates and make appropriate decisions (Almasizadeh and Azgomi, 2013; Garg et al., 2003). Hence, although attacks are frequent and many organizations are affected by attacks, cyberrisk quantification approaches oftentimes do not take advantage on this real world attack data (evidence on historic attacks; Pudar et al., 2009).

Last, for the evaluation of the effectiveness of cyberrisk quantification approaches, the lack of historic data that is directly applicable to the model limits security professional in their ability to evaluate the models and their quality in cyberrisk quantification (Ryan and Jefferson, 2003). This is, in the absence of reliable data, evaluations need to be based on assumptions (Almasizadeh and Azgomi, 2013). Also, simplistic case studies are frequently utilized for presenting the capabilities of developed methods respectively their evaluation by researchers that frequently lack in real world meaningfulness (Pudar et al., 2009). The analysis of methods based on real world systems and evidence from attacks experienced by organizations "in the wild" could thereby increase the expressiveness of evaluations and boost the ability of engineering secure system (Pudar et al., 2009). The lack of empirical data furthermore represents a challenge for efficient defense method development as methods developed were largely based on synthetic data (e.g., gained through simulations) or security testbeds rather than on real data (Knowles et al., 2015). While simulations can be a helpful means in situations where real world data is scarce, they are limited by their inability to perfectly mimic the underlying system (Knowles et al., 2015). However, for development of methods, their evaluation and the use of analytical procedures in cyberrisk management, data availability is critical and can hence be identified as a core challenge in cyberrisk management (Barker et al., 2017).

An approach that is frequently used today for tackling the lack of appropriate data is the reliance on expert estimations directly providing information on their subjective beliefs regarding the probability of attacks and their consequences (Ganin et al., 2020; Fovino et al., 2009). It is thereby questionable, whether this data enables to accurately estimate cyberrisks, as the complex task of estimating cyberrisk components might be beyond the capabilities of experts due to their limited perspectives, subjectivity, behavioural biases and bounded rationality. This is providing precise estimations on the risk components might be beyond the (computational) abilities of experts. Expert estimations therefore need to be reviewed critically. Furthermore, information from different sources might be combined to be able to derive artificial data on the right level of abstraction with high precision tackling the problem of scarce data on attacks. This is, today, cyberrisk assessment is performed mainly based on the basis of expert estimations and the reason why cyberrisk quantification today is pervaded by subjectivity of assessments and lacks in objectivity (Knowles et al., 2015; Fovino et al., 2009).

4.2.2 Adaptive behavior of attackers

A core difficulty of quantifying cyberrisks is that cyberrisks are so-called *adversarial risks*, where the threat and vulnerability are behavior-based and need to be determined simultaneously in the context of a strategic interaction. Consequently, behavioral patterns play a crucial role for

understanding cyberattackers, the incentivizing circumstances and resulting cyberrisks (Kaiser et al., 2021a; Lallie et al., 2021; Allodi and Massacci, 2017). This is, attackers need to be considered as human agents that are motivated by complex incentive structures including power (e.g., political, religious, or moral gains), achievement (e.g., curiosity), and affiliation motives (e.g., social acceptance; Kaiser et al., 2021a; Smith et al., 2010). The distributions of the motive structures within attackers are hard to assess which limits the practical applicability of motivation based attacker profiling (Pudar et al., 2009). Hence, attacker modelling can be considered a significant obstacle to cyberrisk management, efficiency of security investments, and cybersecurity (Nicol et al., 2004). Frequently, assumptions need to be made for assessing the distributions of motive structures, however, if attacker motive descriptions are comparable to attackers of other disciplines, methods can take advantage of the sophisticated knowledge on attack motivations from criminology.

There is broad evidence, that the incentivizing circumstances play a crucial role in attack execution. Cyberattacks are nowadays highly lucrative for low risk of being detected, low costs for launching an attack and high returns (not just limited on monetary profit or monetary equivalents) of successful attacks (Dayalan, 2017). Highly digitalized and valuable targets might thereby be especially attractive (e.g., the financial industry) although they might be especially protected (Eling and Wirfs, 2019). Hence, including attacker models in cyberrisk management is imperative for increasing the precision of risk metrics (Pudar et al., 2009). However, currently an essential burden for the precision of metrics is that cyberrisk quantification frequently considers attackers as illicitly malicious actors Vice versa representing a key obstacle in current cyberrisk management practices (Kaiser et al., 2021a). This demonization of perpetrators of cyberattacks and risks that come with the digitalization is also shown in the wording of the software used to abuse software functionality - malware (acronym for malicious software; Or-Meir et al., 2019).

Furthermore, intelligent attackers develop novel attack techniques circumventing current defensive techniques that are put in place by security personnel leading to a dynamic development of attack techniques (Zeller and Scherer, 2021). By doing so, attackers are honing their skills (Poppensieker and Riemenschnitter, 2018). Defense against such sophisticated attacks executed by intelligent attackers is challenging and defensive methods need to continuously adapt strategically to each attack (Sengupta et al., 2020; Shetty et al., 2018). With this, a highly dynamic arms race between attackers and defenders is sketched (Craig and Valeriano, 2016).

The fact that the adaptive behavior of attackers to evade defenses reduces the effectiveness of defenses over time (Sengupta et al., 2020), also means that preventing a particular type of cyberattack does not mean that its perpetrators give up, but merely that they change their tactics. Attackers are thus constantly on the lookout for new victims, new attack vectors, and new exploits (S. Xu, 2020; Elitzur et al., 2019; Winterrose and Carter, 2014). Also, the adaption of attack strategies on defense mechanisms employed also challenges cyberrisk quantification. In particular, the calculation of probabilities is impacted by the strategic reaction of an attacker to defense strategies. Thereby probabilities are considerable different if an attacker is aware of the implemented countermeasure compared to an unaware attacker (Ganin et al., 2020). Hence, analyzing the security of systems,demands for considering defender models and attacker models simultaneously (Almasizadeh and Azgomi, 2013). The adaption of attack behavior to the past and defensive means taken by the

defender, intelligently acting attackers and their ability to learn from their experience causes that attacks are not random and attack patterns may change with changing threat landscapes (Zeller and Scherer, 2021). Accordingly, it is not possible to take an overall statistical approach.

4.2.3 Interdisciplinarity

Cyberrisk represents a multi-dimensional and multi-disciplinary problem. This is, not only technological vulnerabilities (technological flank) are attacked but also its operators (oftentimes also referred to as human flank; Knowles et al., 2015; Abawajy, 2014). Vulnerabilities furthermore comprise different domains including a physical, information, cognitive and social domain (Ganin et al., 2020; Alberts and Hayes, 2006) opening a whole spectrum of different vulnerabilities (Ganin et al., 2020). Cyberrisk management, hence, needs to consider physical security as a component of cybersecurity. This physical security comprises physical barriers to or around a system or network for deterring unauthorized physical access to the system as well as protection that are linked to the location of the system or network such as cameras (Ganin et al., 2020). Information related vulnerabilities comprise vulnerabilities inherent to software such as software flaws (Ganin et al., 2020). Furthermore, the human, societal, and cognitive domain includes factors related to the personnel interacting with the systems and networks (Ganin et al., 2020). The human and societal domain includes cognition, physiology, psychology as well as the abilities and capabilities of the personnel in dealing with technology (Evans et al., 2016).

Cyberrisk quantification including different domains, hence, demands for the assessment of a high variety of factors (e.g., including human factors compared to a limited, singularly perspective on risk considering only technical factors; Evans et al., 2016). Risk metrics aiming at representing cyberrisks quantitatively thus, need to be able to combine risks emerging from different domains (different types of risk) to a single metric (Knowles et al., 2015). Although the interdisciplinarity of cyberrisks is commonly acknowledged and the value of taking a holistic approach is is well-known widely, most current works are domain specific, focusing only on some aspects of cyberrisk (Makhdoom et al., 2018).

The inclusion of human domain dependent cyberrisks is no common practice yet necessary to reach a state of acceptable security (Evans et al., 2016). The great importance of including the human domain is founded in attacks exploiting human behavior. These represent a main source of cyberrisks in today's infrastructure (Eling and Wirfs, 2019). Neglecting the human domain, hence, would lead to a non acceptable state of insecurity as human operators of a system are able to circumvent security practices and may therefore undermine all safeguards that were employed for securing the human operator as well as the system (Abawajy, 2014). Hereby, cyberrisk quantification taking into account the human domain need to consider aspects like the usability of systems and security awareness of operators. Hence, for security mechanisms to be able to contribute to enhancing the security of systems, usability, acceptance and adoption of mechanisms is elementary.

Furthermore, assessing the impact is oftentimes difficult as both cyber and physical layers as well as their interaction are exploited during an attack to cause substantial losses (Paté-Cornell et al., 2018; Genge et al., 2015; Pudar et al., 2009). In this sense, cyberattacks are observed to have repercussions on the real, physical world. These interactions between different domains demand for taking a holistic approach to cyberrisk management.

Specific characteristics of attacks, pose a further challenge demanding for interdisciplinary approaches. With the aim of considering these these special characteristics of an attack, attack specific cyberrisk quantification approaches are introduced (Pal et al., 2017). However, attack specificity of approaches can also be considered a significant obstacle to holistic cyberrisk quantification. Hereby, current cyberrisk management practices oftentimes consider only a strongly limited set of attacks (Eling and Wirfs, 2019). Also, limited awareness and asymmetric information on security leads to a situation where customers are not able to differentiate insecure systems from secure systems and technology vendors do not have an incentive to engineer secure systems within an increasingly competitive market with falling prices of technology.

Additionally, different natures (immaterial vs. material and direct vs. indirect) of impact challenge the quantification of damages caused by cyberattacks (Bojanc and Jerman-Blažič, 2008). This is, cyberrisks in critical infrastructures need to consider different characteristics of risks than those that come with cyberattacks within the private sphere. Hence, cyberrisk metrics need to be situation specific (e.g., targeted at the special issues relevant for a defined industy or industrial branch; Mishra et al., 2020). These interactions and growing interdependence of cyber and physical layers due to digitalization also leads to the need for branch or even firm specific cyberrisk quantification (Wadhawan et al., 2018) and demand for increased research effort as the interactions are not understood sufficiently to date (Pudar et al., 2009). Hence, general and reliable models that are readily available and applicable for organizations, lack in their ability to guide decision making regarding optimal, respectively appropriate or feasible, security investments (Bojanc and Jerman-Blažič, 2008). Yet, although risk quantification approaches need to be branch specific, they also need to be general in a sense, that they are able to simulate different types of attacks.

Furthermore, the impact caused by a cyberattack can be of multiple natures including tangible impact (e.g., financial loss, and disruptions of operations) as well as intangible impact (e.g., damage to an organizations reputation and disruption of trust; Sengupta et al., 2020; Sheehan et al., 2019; Garg et al., 2003). Singular methods of a specific scientific discipline thereby frequently (over-)weight dimensions that are important for their discipline and (under-)weight less important dimensions for their discipline limiting the broad adoption of a single quantification means. This outlines that ensuring cyberrisk is a multi-criteria decision problem where weighting approaches are prevalent (not only limited on different dimensions of the consequences but also different dimensions in other components of risk such as threats or vulnerabilities; Ganin et al., 2020). Hence, an interdisciplinary research approach is demanded considering multiple dimensions and different disciplinary approaches within a holistic approach combining different skill-sets and perspectives on cyberrisk (Nicol et al., 2004).

Recently, such holistic approaches that aim at achieving completeness in cyberrisk assessment are scarce in literature (Leszczyna, 2021). However, these approaches gained in importance. In the last decade, for example, economic perspectives on cyberrisks have gained in attention (J. Chen and Zhu, 2019). Economics thereby contribute to the discussion by providing insights on efficiency of security (Ruan, 2017). The inclusion of economics in cybersecurity research was motivated by research realizing that a technological point of view is not sufficient for reaching cybersecurity (Bojanc and Jerman-Blažič, 2008). By including an economic perspective, optimal security levels (economically efficient) can be determined which is a state, where the costs of security and the consequences of successful attacks are at a minimum (Ruan, 2017). At this state, enhancing the security of systems is inappropriate, as risk acceptance would be a superior risk management strategy (Ruan, 2017). Economics provide methodologies that enable to give founded estimates on the feasibility of security budgets and the optimality of spending (e.g., countermeasure selection). Almost in opposition to economic efficiency considerations, cybersecurity research from a information theoretic and technological point of view primarily follows the dictum of "maximum or absolute security" instead of "affordable or feasible security" respectively "efficient security". Yet, also other disciplines like juridical perspectives on cyberrisks and their management as well as interdisciplinary approaches combining different scientific fields of expertise were introduced to the scientific discussion (Winau et al., 2021; Steinbrück et al., 2021; Kaiser et al., 2020).

4.2.4 Dynamic evolution of threats and technology

The increasing digitalization and interconnection of different heterogeneous devices poses great complexity on cyberrisk assessment (Ge et al., 2017). Thereby different system characteristics and a high amount of distinct devices are combined within a system of systems to reach smart economic systems and smart societies. Oftentimes, these different devices are not even registered or known to the system administrator (e.g., in relation to bring-your-own-device policies). Also, legacy systems pose challenges, if different versions of operating systems are combined within networks.

For managing and quantifying security, the understanding of network traffic protocols has become essential, as these protocols are able to give a concise picture on the processes that take place within highly complex systems and systems of systems that are frequently in place in today's connected world (X. Liu et al., 2016; K. Xu et al., 2008). However, cyberattack practices and, hence, related cybersecurity issues evolve dynamically challenging analytical purposes (Fovino et al., 2009). This changes the threat landscape continuously with attackers that act increasingly sophisticated bringing novel types of cyberthreats with them (H. Zhang et al., 2018; Hentea, 2008). Furthermore, the attack surface is changing dynamically due to the highly dynamic development of technologies and their continuous introduction to systems. The pace of change and the evolution of systems might thereby be so high that professionals are forced to manage risks with limited knowledge on the underlying system that shall be defended as well as the attack landscape that shall be defended against (Nurse et al., 2017). With increasing penetration of technology, the attack surface is increasing perpetually (Alshehri et al., 2018).

Novel vulnerabilities are discovered each day by researchers as well as attackers. If attackers discover such vulnerabilities, the existence of the vulnerability is not known to the defender oftentimes . Those vulnerabilities respectively exploits on these vulnerabilities are referred to as zero-day-exploits (Q. Zhang et al., 2015). Attacks that are based on these vulnerabilities are referred to as zero-day-attacks (Sengupta et al., 2020). Equivalently, unseen anomalies and unknown attacks are detected perpetually giving emphasise on the dynamic development of attackers (Murugesan et al., 2020; Paridari et al., 2017). This is, relying on cyberrisk management models that depend strongly on known vulnerabilities and threats lack in accuracy for predicting the future and assessing the cyberrisk that come with attacks quantitatively (Hahn and Govindarasu, 2011). This motivates approaches for attack technique prediction and the deployment of defense in depth concepts.

4.2.5 Applicability

In the last years, research on cyberrisk quantification spiked and several methods were developed promising to be potent in cyberrisk quantification (Carfora et al., 2019). However, although there is a plethora of different approaches for cyberrisk quantification and management that are presented in scientific literature, they lack in applicability. This is, only few quantitative cyberrisk assessment methods have proven practical applicability, utility or even successful application in real interconnected, heterogeneous and complex systems rather applicability is oftentimes limited to simplistic models of real systems (Leszczyna, 2021; Nurse et al., 2017; J. Chen and Zhu, 2019). Furthermore, operational implementations of the developed methods are scarce (Leszczyna, 2021). Yet, "reliable, easy-in-use and economic tools are highly demanded" (Leszczyna, 2021). Current cyberrisk quantification methods lack in their ability for parametrization, which is the impossibility of including numerical data to the methods for obtaining quantitative metrics (Almasizadeh and Azgomi, 2013). Such parametrizations can be reached by enabling the integration of cyber threat intelligence to the risk assessment approaches (Leszczyna, 2021).

Furthermore, cyberrisk assessment methods are not standardized and in addition, no commonly accepted unit for measurement exists, which presents a heavy burden for their application and intuitive understanding (Ruan, 2017). Hence, there is an absence of readily available quantification methods practitioners can rely on. This is, professionals feel a lack in practical guidance for security analyses (Knowles et al., 2015). Also, approaches proposed by researchers fall short in their ability to be used for real time risk assessments (Knowles et al., 2015). A further obstacle for efficient cyberrisk quantification and their management (e.g., optimal investment) is its dependency on the underlying system demanding for great knowledge (Bojanc and Jerman-Blažič, 2008).

Cyberrisk management is also subject to a competence gap in managing these risks and understanding the consequences of broad and wide scale digitalization. Evidence is given to this research gap by the observation that most organizations are unprepared for tackling the needs of ensuring cybersecurity and fostering sustainable levels of digitalization (Rothrock et al., 2018). Businesses and society thereby seem to strive for leveraging on the benefits that come with digitalization neglecting that these come at the risk of cyberattacks. The competence gap within cyberrisk management is also reflected in lawsuits charging management for neglecting their fiduciary burdens in managing

customers data (Rothrock et al., 2018). It is therefore considered a key strategic necessity for companies to increase their capabilities in ensuring cybersecurity to stay ahead of attackers within the dynamic arms race with defenders (Craig and Valeriano, 2016).

4.3 Cyberthreats and offensive capabilities

The spectrum of cyberthreats and possible attacks is multifarious. Cyberattacks "come in wide range of variations like Virus, Worm, Trojan-horse, Rootkit, Backdoor, Botnet, Spyware, Adware etc. These classes of (cyberattacks) are not mutually exclusive meaning thereby that a particular (cyberattacks) may reveal the characteristics of multiple classes at the same time" (Gandotra et al., 2014). Hence, cyberattacks show great polymorphism (Abusitta et al., 2021).

Virus describe a form of cyberattacks that inject code to originary systems and have the ability to spread (Or-Meir et al., 2019). The category of cyberthreats is prominent, given the fact, that in media, computer virus are used for describing a wide variety of cyberattacks (Or-Meir et al., 2019). Worms are another class of cyberattacks that duplicate themselves and spread within networks (Or-Meir et al., 2019). Adware is an acronym of advertisement and software. The decomposition thereby describes the functionality convincingly. Adware is a form of cyberattacks that show advertisements to users (Or-Meir et al., 2019). Perpetrators of adware attacks capitalize on the attack by selling advertisement space to product vendors, companies or other parties searching for advertisements (Or-Meir et al., 2019). Further acronyms describing cyberattacks are scareware and ransomware. Scareware use malware detection messages mimicking anti-virus programs to trick users to purchasing products for the removal of the software (Or-Meir et al., 2019). Cyberattackers thereby directly draw reward from users. Ransomware encrypts files or other information, takes the information hostage, and demand for payment in form of a ransom for decryption of the files (Or-Meir et al., 2019). While the list is not considered to be comprehensive, this short overview introduces to the great variety of different cyberattack classes.

The cyberattacks rely on different attack techniques for reaching their goal. Some of the most common attack techniques are brute-force or bulk guessing attacks. These aim at getting access to an account by testing combinations of passwords and user identifiers. While brute force-attacks can oftentimes be detected quite easily and defense against such attacks can be realized quite easily (e.g., by relying on three strikes rules or unsuccessful to successful ratios), bulk guessing attacks are oftentimes much harder to detect. (Florêncio et al., 2007). These bulk guessing attacks rely on the same principle of testing user and password combinations, however, differently to brute force attacks, these vary the user instead of the password (Florêncio et al., 2007). While these attacks represent quite unsophisticated attacks, much more sophisticated and unconventional attacks exist. These are inter alia attacks that allow the data extraction form air gaped systems via signal encoding through heat emission manipulations (Guri et al., 2015) or encoding by radio frequencies (Guri et al., 2014). Equivalently, methods are known to exfiltrate data through power consumption manipulations (Guri et al., 2019).

Besides attacks focusing on technical vulnerabilities, social engineering attacks are targeted to exploit human vulnerabilities by tricking people to share data or install software that can be used by cyberattackers to reach their goals (Hong, 2012). Thereby, phishing attacks are a prominent example. Phishing attacks oftentimes circumvent technical safeguards that are implemented and come in the form of emails (Hong, 2012). These emails mimic legitimate mails for fooling human actors to reveal information (e.g., login details; Gupta et al., 2016; Hong, 2012).

To get an overview on the plethora of attacks and the threat landscape that characterize cyberrisk, an ontological knowledge graph is created that comprises information on adversarial tactics (attack goals), techniques (technical means to achieve the attack goals), and common knowledge (Oosthoek and Doerr, 2019; Strom et al., 2017). Within this knowledge base, information on cyberattacks is aligned to different phases of conducting cyberattacks (Strom et al., 2017). These are among others reconnaissance, initial access, defense evasion, and impact. The knowledge graph described covering cyberthreats and attacks is, hence, aligned to the cyber kill chain (Jasper, 2017; Lockheed-Martin, 2014). It provides a concise overview on different cyberattacks and the offensive techniques they rely on (Strom et al., 2017). Single instances of cyberattacks are thereby classified within different cyberattack families (Oosthoek and Doerr, 2019).

5 Research Methodology and Schema

5.1 Research gaps

The field of cyberrisk management suffers the scarcity of formal cyberrisk analysis methodology (Ryan et al., 2012). Especially due to the high dynamics relying on cyber threat intelligence is inevitably for efficient security management as it can enable the (semi-)automated calibration of management and quantification methods (Nurse et al., 2017). Automation of cyberrisk quantification and management is essential given the high velocity of changes in the field of cyberrisk with simultaneous scarcity of resources (e.g., time of security professionals; Chen et al., 2014). The development of automated approaches for continuous, dynamic real-time cyberrisk assessment in the highly versatile environment is an essential research gap and its current lack represents a main obstacle for ensuring security of systems and enable the development of secure systems as well as their engineering (Q. Wang et al., 2019; Nurse et al., 2017). Yet, suitable methodologies enabling the dynamic assessment of the state of a system during as well as after a cyber incident and quantify its risk exposure are scarce. Hence, profound methods can be considered a research gap (Mishra et al., 2020; Polatidis et al., 2018). These approaches furthermore need to take into account the complex interrelations of attackers and defenders as well as their interaction (Q. Wang et al., 2019). This is, giving estimations on the efficiency of countermeasures is challenging (Ganin et al., 2020; Ryan et al., 2012). Current cyberrisk quantification approaches are limited in the number, type of attacks, and defensive means that are considered and have been proven its feasibility (Zheng et al., 2016). Furthermore, sound decision support tools assisting security personnel and enable simulation of attacks are demanded and may increase predictive capabilities (Nurse et al., 2017).

In dynamic cyberrisk management, attack detection is essential. Thereby, current incident detection systems provide good detection capabilities for both, known as well as unknown attacks (i.e., zero-day attacks; Moustafa et al., 2018). However, current systems fall short in their ability to differentiate attack types and predict the further course of action (i.e., progression) of an ongoing attack (Moustafa et al., 2018; Almasizadeh and Azgomi, 2013).

Furthermore, although game theoretical models have been proven to be able to provide real world implications and valuable insights on cyberrisk (especially if the interactions of defenders and attackers are important), mathematical simplifications for describing attack and defensive mechanisms result in inaccuracies in modelling and imprecise estimations (quantifications) on cyberrisk (Q. Wang et al., 2019). Moreover, currently, there are no approaches for quantification of

cyberrisks induced by the human domain (e.g., human and social vulnerabilities) and consistent means for their management or associated metrics that are accepted in science or practice (Evans et al., 2016). Also, socioeconomic attacker models focusing on incentives need to be included in cyberrisk research for enhancing cyberrisk management capabilities (Allodi and Massacci, 2017). Yet, methods for assessing the reliability and quantification of risks related to human actors were implemented in various industries but cyberrisk and considering the human domain of cyberrisk is no common practice in today's society and economies (Evans et al., 2016). However, including the human and societal domain in cyberrisk quantification promises to increase the reliability and accuracy of cyberrisk assessments by providing a more holistic picture on the threat landscape systems are exposed to (Evans et al., 2016). It is, hence, imperative to develop assessment approaches of such holistic nature (Komljenovic et al., 2016). Game theoretical models might furthermore be helpful for weighting different attack paths. Currently, attack path weighting in the sense of giving priorities on the specific vulnerabilities or the hardening against specific attack techniques is inadequate and oftentimes based on simplistic criteria such as the length of attack paths and imprecise estimations on the capabilities of potential attackers (Polatidis et al., 2018). These mathematical simplifications are predominant in cyberrisk management. These also include today's attacker profiling approaches (Polatidis et al., 2018). This is, there is no profound model of attackers rather cybersecurity as it is performed today.

Calculating the consequences of a cyberattack suffers great insecurities and estimations, hence, are defined by great variability. This represents a great obstacle for implementing sound cyberrisk management practices. Estimating the true costs of a successful attack or a system fault represents a great research gap that needs to be solved when aiming at implementing efficient security in modern society and economies (Mishra et al., 2020). The methods for quantification of the consequences thereby need to consider the multi-domain characteristic of cyberattacks and their impacts (Ganin et al., 2020). Current cyberrisk quantification methods, however, lack in the ability to address the different components of cyberrisks (i.e., the triplet of threat vulnerability, and consequence) but rather focus on specific components (e.g., consequences or probabilities; Ganin et al., 2020). Furthermore, these methods fall short in their ability to be able to integrate across the different relevant domains of cyberrisk management (Ganin et al., 2020). This work aims at providing relief to the current limitations in cyberrisk quantification.

For both, probabilities of cyberattack as well as their consequences, this work tackles the lack of data (or at least the problem linked to the availability of appropriate data Komljenovic et al., 2016) by providing (1) a means for digital twin based simulation of the effects of cyberattacks and (2) methods for probability quantification of attacks. Here, the introduced set of methods offers two approaches. The first approach relies on network traffic recorded within a system relying on approaches from data analytics (i.e., threat hunting and prediction of the course of an attack). The second approach is an a priori quantification of attack probabilities given cyber threat intelligence by means of game theoretical modelling. The work provides insights that the data available (cyber threat intelligence) combined with sufficient knowledge on an organization's system enables cyberrisk quantification, if data preprocessing steps are conducted (e.g., cyber threat intelligence based attack simulation). The methods proposed within this thesis answer the research

question of which data is needed for performing risk analyses and whether data available suffices for quantitative risk assessments (Komljenovic et al., 2016). Also, by enabling parametrization with real world data, the methods surpass the challenge of applicability that is inherent to current cyberrisk management approaches and thereby, especially lacking possibilities of parametrization and calibration with real world data on past incidents (Almasizadeh and Azgomi, 2013).

Targeting the high uncertainties due to the highly dynamic development of attack techniques and offensive capabilities defenders need to keep pace with, attack prediction capabilities need to be increased (Hahn and Govindarasu, 2011). Basing the predictions on data of past experienced and recorded attacks as well as data on systems (e.g., known vulnerabilities) promise potentials for increasing the predictive capabilities and effective handling the dynamics in the field of cyberrisks as well as resulting uncertainties for defenders (Polatidis et al., 2018). In doing so, this work proposes a method for attack prediction and forecasting which is based on cyber threat intelligence.

Furthermore, when it comes to selecting defensive means for risk mitigation, methodological support is scarce and efficient decision support (i.e., considering the specific needs of organizations or industries) for security professionals in the current state of scientific research unavailable (Komljenovic et al., 2016). Traditional approaches from risk management such as reliance on returns on investment or net present value approaches have proven their feasibility in finance and real option management (Ruan, 2017). The adaption of these existing, traditional methods might hereby be necessary for increasing the efficiency of cyberrisk management. However, for enabling their application, sound quantitative cyberrisk metrics need to be provided (Komljenovic et al., 2016). As these quantifications are scarce, today, traditional approaches have not been successfully utilized in the field of cyberrisk management (Ruan, 2017).

Combining real time dynamic attack detection and risk quantification with attack mitigation decision support is scarcely researched. Research on the application of recommender systems for real time attack mitigation provides promises for enhancing cybersecurity (Polatidis et al., 2018). Hereby, automation of attack incident response (reactive means) as well as system hardening (proactive means) can be reached based on automated risk assessments. Hence, if a vulnerability would be exploitable within a system, the approach could automatically propose and deploy effective countermeasures (Polatidis et al., 2018) in the sense of a self-protective system (Chen et al., 2014). Hereby, the assessment of the effectiveness poses a significant burden to the utilization of traditional investment theory (e.g., return on investment or net present value; Ruan, 2017). The automation of attack path discovery is the main enabler for the transition of automated cyberrisk quantification to automation of reactive (e.g., incident response) and proactive cyberattack management (Polatidis et al., 2018). Novel approaches can provide superior defense, as current safeguards and cyberrisk management systems provide suboptimal defense and improper responses to attacks (Chen et al., 2014). This is, most systems suffer practical limitations such as weaknesses in the tackling of novel attacks, handling of insider attacks (Chen et al., 2014). Within this work, a method is presented combining effective countermeasure portfolio optimization capabilities with self protective functionalities in the form of a bio-inspired artificial immune system. The bio-inspired artificial immune system thereby represents a method enabling defense in depth aiming at

hardening a system against multiple attack techniques that are employed within an attack that is executed in a system.

Within this work, a contribution shall be given to the "lack of appropriate (cyberrisk) analysis and modeling methods, (and) scientific understanding of the complexity" (Komljenovic et al., 2016) related to these attacks. The continuous increase in the complexity of digitalized networks makes their understanding furthermore difficult (Komljenovic et al., 2016).

5.2 Overview of research

Effective cyberrisk management demands for the usage of well grounded risk management procedures. Such well grounded management practices are established within most firms. However, within the field of cyberrisk management, companies have trouble with applying the models. This is because conventional cyberrisk management practices demand for risk quantification. Cyberrisk quantification has been proven to be challenging and methodological support for professionals in cyberrisk management is scarce. The scarcity thereby emerges from low applicability of many methods and not merely of a shortage of methods introduced for cyberrisk quantification (rather there is a plethora of different methods - chapter 3; Kandasamy et al., 2020).

The aim of this work, hence, is to present a set of methods that allows automation in cyberrisk management and offers decision support for professionals (e.g., forensic analysts, security managers, and software engineers). Figure 5.1 provides an overview on the set of methods developed. A strong focus is thereby set on cyberrisk quantification (chapter 6 to chapter 11) presenting two methods for quantification of the probability of cyberattacks.

Beginning on the left side, the left upper box visualizes security monitoring contributing to the data problem. The left middle box shows a further data source which is information on past attacks (cyber threat intelligence). Also research conducted in this field contributes to the challenge of data scarcity. The methodology presented in chapter 6 takes advantage of these information sources. Within this chapter an approach for the dynamic real time quantification of the probability of attacks taking advantage of monitoring systems is introduced. Hereby a quantification of the probability of an attack is conducted based on a *multi-nominal naive Bayesian classifier*, a *multi-nominal multi-layer naive Bayesian classifier*, and an approach inspired by *term-frequency-inverse-document-frequency* for generating initial hypotheses. These initial hypotheses are refined in a second step. For refinement, the approach relies on a set of link prediction techniques (*Projected techniques, Link prediction on projected techniques, Link prediction on projected attack, Projected attack techniques*, and *Supervised link prediction*) and similarity metrics (*Jaccard's* and *Tanimoto coefficients*, *Adamic Adar, Friends* and *Katz measure* as well as *Preferential Attachment*). Furthermore, a heuristic is given for when refinement is feasible which is necessary when initial hypotheses are precise and refinements do not provide significant improvements. The method orders all known attacks (cyber threat intelligence) depending on their probability given a set of observables (network traffic). The method can hence be considered a means of real-time probability quantification within a risk monitoring system.

Chapter 7 introduces a methodology that enables the quantification of the probability of an attack that is based on *game theoretical modelling* that is expressive for longer horizons (in contrast to the approach presented in chapter 6). The game theoretical model gives tribute to the complexity of motivation structures, intelligently acting attackers and can be defined as a behavioural cyber security game (box in the middle an on the upper part of figure 5.1) Both approaches take advantage of Cyber Threat Intelligence for calibration of the underlying models.

Given the high dynamic nature of cyberattacks, high uncertainties challenge the models and the underlying data referring to the comprehensiveness. Chapter 8, therefore, introduces an approach that is designed to tackle this problem by forecasting the development of novel attacks and augment the calibration by synthetic attack data (left lower box in figure 5.1). The work contributes to understanding the dynamics in attack technique development and hence, contributes to solving the challenges posed by high dynamic nature of the cyberthreat landscape. Furthermore, it can be used as a means for augmenting the database with forward looking predictions on attacks. From a methodological point of view attack forecast and prediction is based on *simulations, genetic algorithms, time series approach* or an approach relying on *Generative adversial networks*.

A focus on cyberrisks originating from the human and societal domain is set in chapter 9. Hereby, a *game theoretical model* is employed for understanding and formalizing the cybersecurity problem that is introduced by human operators. Exemplary, the work focuses on phishing attacks and the quantification of the risks associated as well as on factors influencing the susceptibility to those attacks (e.g., time scarcity and work stress). The approach combines the game theoretical methodology and *Bayesian updating*. The model considers human operators as bounded rational agents aiming at contributing to the dynamic interaction of defenders and attackers given intelligent actors.

Furthermore, a method that is designed for quantifying the impact of cyberattacks (tested for cyberattacks on the healthcare branch - chapter 10- and within the automotive industry - chapter 11) is presented (lower, middle part of figure 5.1). The method is based on *digital twinning* is targeted towards contributing to solving the data availability challenge by generating synthetic data and the complexity in impact determination.

Last, chapter 12 leverages on the previously proposed methods for cyberrisk quantification and develops a *decision support system*. The decision support system (right part of figure 5.1) offers potentials for automation providing potential for the use as a automatic incidence response system when applied in combination with real-time risk assessment (e.g., one of the approaches introduced in chapter 6). The chapter also introduces to the use of a *bio-inspired artificial immune system* combining a defensive base investment and an adaptive defense directly targeted towards the defense against a currently observed attack.

Altogether, the work focuses on a holistic research approach combining technological, organizational, social, behavioral and economic research approaches. The work is thereby founded on game theoretical modelling, digital twinning, and data scientific methodology. The method for holistic cyberrisk quantification has a strong economic focus, in that the risk measure is monetary in nature, but can be broken down into a detailed risk vector comprising technical risks (e.g., risk

to quality or machine utilization) and other non-monetary risks (e.g., risks to intellectual property or reputation). The derivation of recommendations for defensive measures corresponding to the optimization of the security investment portfolio is based on the approach for holistic cyberrisk quantification and shows the possibility of using established methods that have proven themselves in risk management if the risk assessment approach is followed.

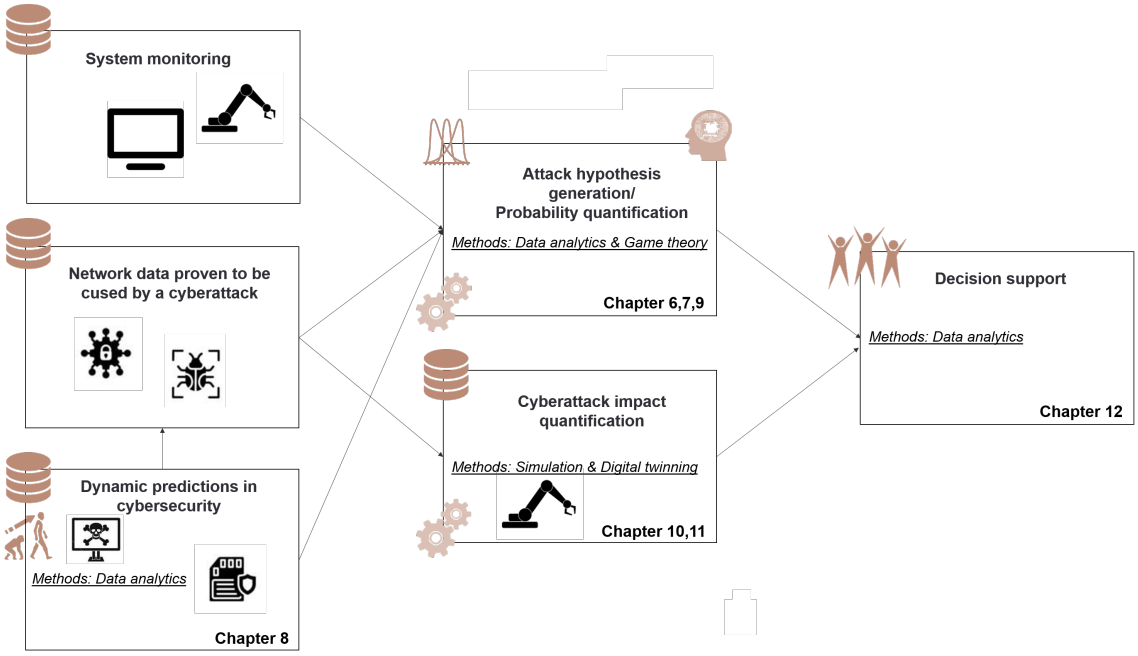

Figure 5.1: Research schema and overview

5.3 Contextualization

Digital systems are the biggest drivers of modern economic activity, because they enable business processes and influence business models. The resulting innovations and efficiency gains contribute to the prosperity of society. On the other hand, however, the increased digitization of broad areas of economic activity also harbors (cyber) risks. Cybersecurity quantification thereby proved to be challenging. These difficulties in quantification result primarily from the highly dynamic nature of attack behavior (dynamic development of attack strategies, multiplicity of flanks and dynamic choice of flanks), but also from the complexity and interconectedness of processes in digitalized environments and the interaction of people and technology. Despite the high (overall) economic costs caused by cyberrisks, quantification has not been satisfactorily possible so far. The work aims at contributing to the economic evaluation of cyberrisks, taking into account strategic aspects of the interaction of affected companies and their potential attackers. For this purpose, an industrial economic market model is developed that maps cybersecurity along the value chain. This model is intended to enable an economic assessment of cyberrisks and thus, serve as a basis for investment decisions. At the threat level, attackers are differentiated according to their relevant incentive structures (power, affiliation, and achievement). With the help of a defender-attacker game, relative (expected) attack frequencies can be derived approximately for each given attack strategy.

Dynamic cyberrisk assessment and management enabling the implementation of risk mitigation means in close timely relation are important steps in reaching a state of acceptable security. Thereby, the use of algorithms for automatic risk analysis is demanded. This is, the dynamic threat landscape and great computational efforts caused by the need for big data analyses in cybersecurity as well as the scarce time of security professionals demand for automation. Within literature, there are substantial progresses in this regard relying inter alia on statistical methods for information retrieval (e.g., Bayesian statistics or link prediction algorithms; Elitzur et al., 2019; Poolsappasit et al., 2011). Within chapter 6, a substantial expansion of the existing literature regarding this topic is provided. The work thereby focuses on the analysis of cyber threats where network traffic is observed and anomalies or indicators for intrusion are extracted. This data is correlated with known information on historic cyberattacks (cyber threat intelligence) for threat hunting and predicting the course of action of recently detected attacks. However, the approach can also be used for deriving estimates on the attack discovered (i.e., attack probability). Besides its practical use for forensic investigations, the introduced methodology can hence be used for dynamic cyberrisk quantification in the sense of a risk monitoring for systems that are in operation.

In order to be able to prepare for potential attacks as early as possible, taking the perspective and the mindset of an attacker may be helpful (Kaminski et al., 2017). It is thereby of utmost importance to understand the incentivizing environment, in which context an attack is executed as well as the motivations of attackers (Kaiser et al., 2021a). In this context, chapter 7 provides a scientific study on behavioral cyber game theory that aims at quantifying the probability of cyberattacks. Game theory is thereby used as a comprehensive set of methods for predictive, normative as well as descriptive, exploratory research on interactions between two parties (defender and attacker as a perpetrator of attacks). In essence the approach is founded on the assumption of rationality which, however, can be relaxed (bounded rational decision makers). Rationality in this sense means, that:

(i) Each player wants to maximize his utility (opportunism).
(ii) Each player understands the rules of the game and the consequences.
(iii) Statements (i), (ii) apply and each player has knowledge about them.
(iv) Statement (iii) is true and each player has knowledge about it.
(∞) Statement (n-i) is valid and each player has knowledge about it.

Within chapter 7, the attacker profiling and the distinction of different attackers within the classification schema of external and internal perpetrators is mainly with regard to their resources. Besides, attackers are assumed to be motivated by the incentives triad of power, affiliation, and achievement (Kaiser et al., 2021a). The spectrum of considered threats thereby ranges from hacktivism, industrial espionage, extremism, organized crime, and terrorism to state-controlled attacks. Yet, each attacker from these different groups can be described by a combination of these three main motives (e.g., a terrorist is mainly motivated by power motives where the power is oftentimes legitimized by moral or religious superiority). In this sense, the categorization of attackers follows the motive structures, where byzantine attackers and tullock attackers are differentiated. The motivational structures of system internal attackers thereby do not deviate substantially from external attackers. Yet, internal attackers have a number of advantages compared with external perpetrators, which are insider knowledge and contact to system gatekeepers. Furthermore, they

can take advantage of the fact that they are initially unsuspected as part of the organization. The approach is targeted to predict the attackers' action taking and the resources an attacker is likely to spend (monetary budget, technical know-how, insider knowledge, technical equipment). Attack anticipation is thereby considered key for efficient cybersecurity yet, "intelligence-heavy" and in need for precise input data (Brown et al., 2017). The introduced approach relies on the categorization of attack forms and attacker "types" (differentiation by motivation) based on historical cases and current literature (Kaiser et al., 2021a) and thus, uses attacker modelling within a game-theoretical analysis (serious games) of attack behavior as well as the identification of "attractive flanks". From the combination of attackers' goals and resources as well as the organization's vulnerabilities, the specific attack variant can be inferred as a possible attack in most cases. The attacker perspective is within the presented approach analyzed based on historical cases (based on recorded information on previous successful attacks - cyber threat intelligence) using a behavioral attacker-defender model. Cyber threat intelligence thereby is encoded within a multi-layer threat ontology which represents the game graph and is in essential a event database. Attackers' goals were extracted from relevant scientific studies on motivational patterns and goals of attackers (Kaiser et al., 2021a). The analysis (weighting of the cyber threat game graph and derivation of optimal attack strategies) is substantiated by means of game-theoretical modelling. The conceptual map of all conceivable attack sources can thereby be adapted. For each attack, a probability is derived (plausibility-based probability) and identified, which signs can be used as identifiers of such attacks. These findings are derived formally and mathematically to from the model and describe the probability of an attack quantitatively.

Chapter 8 is mainly targeted towards the handling of uncertainties emerging from the high dynamics in the field of cyberrisk management. Current means for cyberrisk management as well as cyberrisk quantification are mainly reactive limiting their normative power towards ensuring efficient cybersecurity (Husák et al., 2018). Taking cyberrisk quantification and management to the next level, the approach presented in chapter 8 represents a means to tackle the challenges proposed by the dynamic development of the cyber threat landscape. The high dynamics were considered a further key obstacle to cyberrisk quantification and data actuality provides a key limitation of applying the approaches presented within chapter 6 and chapter 7. To represent the dynamic evolutions in the cyberrisk environment, genetic algorithms are commonly used in academia to create mutations of known malware to simulate the dynamic evolutions in the cyberrisk environment (O'Reilly et al., 2020; Winterrose and Carter, 2014). The results derived from attack forecast and prediction can be used to augment the data bases on which the approaches presented in chapter 6 and 7 are based. By doing so, the risk quantification can gain normative power and be directed towards the management of future events. Also, decision support is backed by attack forecast and prediction for enabling more accurate support.

Human error is one of the most frequently underestimated sources of risk in risk management. Consistently, current cyber management methodology oftentimes lacks in considering the human and social domain adequately (Bailey et al., 2018). Risks that are driven by the human and social domain include aspects such as inadequate competence profiles and competence deficits, a lack of awareness of the importance of safety-relevant, proactive measures, and stress due to

excessive workload, fatigue, time pressure, or private problems. All those aspects might lead to high cognitive load ("the amount of stress placed on working memory"; Pfleeger and Caputo, 2012) which is considered a key behavioral aspect that need to be considered when leveraging behavioral science in cyberrisk quantification and management to reach a holistic view on the thereat landscape. Cognitive limitations (e.g., memory capacity and computational limitations for analytical tasks) and bounded rationality thereby interfere with the security behavior of operators and lead to rational inattentive behavior (Pfleeger and Caputo, 2012). Inattentional blindness supports significantly to human actor caused cyberrisks (Pfleeger and Caputo, 2012). Of particular relevance are risks specifically associated with the work of responsible engineers in control rooms and employees in the office. These include the problem of the drastic change from a job that is usually highly monotonous to extreme stress in exceptional situations. Another problem is the fear of those in charge of making wrong decisions in contexts very much regulated by rules. Chapter 9 aims at contributing to current discussions on cyberrisk originating from the human and social domain, by providing a (psychological) game theoretical model on phishing mail detection and contributing to people-centricity or some kind of "human-centered security" (Pfleeger and Caputo, 2012) in cyberrisk management. People-centric security thereby focuses on the behavior of human operators and their use of technology. It helps understanding the role human operators play in ensuring cybersecurity and identifying staff that is likely to be attacked or fall victim to a cyberattack (so called very attacked person; Proofpoint, 2019). The approach introduced for human centered cyberrisk analysis thereby focuses on the interaction of the way human actors behave when they perform a goal oriented task and the way they react when perceiving a disruption. In most cases, security operations (e.g., phishing mail detection) is the secondary, disrupting task while their normal, ordinary task gets disrupted (Pfleeger and Caputo, 2012). In this sense, security is intertwined with human goal oriented behavior (Pfleeger and Caputo, 2012). The work on people-centric security considering the bounded rational nature of human operators as a unchangeable fact, may furthermore help in developing feasible countermeasures (such countermeasures that are beyond a "fix the human" (Sasse, 2015) approach). Interventions aiming at fostering cybersecurity thus, need to take into account the cognitive characteristics of a human operator (Bada et al., 2019). This, however, is rarely done in today's cybersecurity market (Bada et al., 2019). However, when designing safeguards to meet the needs of human operators to support their decisions, humans can also be seen as "great assets in the effort to reduce risk related to information security" (Bulgurcu et al., 2010).

Efficient management of cyberrisks demands for not merely leaving cyberrisk management to technology experts (e.g., to the information technology department) but demands for addressing cyberrisks in the context of a business (Poppensieker and Riemenschnitter, 2018). This is, for quantifying the consequences of successful cyberattacks, understanding the influenced business processes and the mapping of digital assets along the value chain is essential (Kaminski et al., 2017). Cyberrisk quantification demands for approaches that take into account the whole value chain of a focal company and its specifics. In this context, chapter 10 and chapter 11 present methodological support. Within these works, two applications of digital twin based cyber incidents' consequence quantification are given. A digital twin can thereby be considered an "integrated multi-physics, multi-scale, probabilistic simulation of a (...) system that uses the best available physical models,

sensor updates,(...) etc., to mirror the (physical) twin. The digital twin is ultra-realistic and may consider one or more important and interdependent (...) systems (...). In addition to the backbone of high-fidelity physical models, the digital twin integrates sensor data (...) and all available historical (...) data obtained using data mining and text mining. The systems on board the digital twin are also capable of mitigating damage or degradation" (Shafto et al., 2012) by simulating possible consequences and supporting the implementation of feasible countermeasures. Chapter 10 thereby focuses on the health care branch while chapter 11 focuses on the industrial production of automobiles. The subject of these chapters is the assessment of the consequences of an attack in terms of direct and indirect damage. A distinction can be made here between a temporal, a monetary and a non-monetary component, whereby the non-monetary dimension can often also be expressed (with somewhat more additional information) in a monetary cost figure. For this purpose, an approach is used which, on the basis of different requirements of a customer for a product (compare Kano model), represents the essential market interactions (especially market equilibria) between customer and seller. Cyberattacks also result in intangible consequential damage, which on the one hand lies in the deprivation of services (e.g., healthcare), but on the other hand also in the loss of reputation of system operators and even a loss of confidence in the system as a whole. The potentials of the introduced approach for impact quantification are highlighted within these studies for service provisioning (chapter 10)and industrial production of physical products (chapter 11).

The increasing digitalization of health care is pronounced along the entire value chain (Kaiser et al., 2021b) to gain high-precision and personalization in medical care (Topol, 2019). Also, the digital transformation of health care service provisioning offers novel possibilities for providing medical relief to severe health conditions (van Schaik et al., 2017). Digitalization in health care provisioning has thereby already proven its potential in increasing the quality of medical care (McCullough et al., 2010). Assessing the cyberrisks that come with digitalization is, hence, of utmost importance for guaranteeing high quality health care. Risks that health care service provisioning are confronted with are of especial importance due to potential harm to human life and well being. Ensuring cybersecurity is, hence, of utmost importance for societies. Hereby, also the great importance of ensuring privacy of patients data is essential (Kaiser et al., 2020). Cyberrisk management in the context of the health care branch is thereby subject to special characteristics of health care provisioning. These are the participation of humans within medical service provisioning and long term effects low quality service provisioning in the case of quality deterioration through malfunctioning systems. Furthermore, complexity of medical service provisioning and ensuring human health lead to a challenging quantification of the effects of cyberattacks. The approach thereby takes advantage of empirical knowledge in the sense of evidence based medicine for deriving diagnoses and simulating the planning of medical treatment. Quality deterioration of medical service provisioning can thereby originate from bad data quality (e.g., if an attack prevent to derive medical images to support a diagnosis) or in availability of medical devices which may prevent medical service provisioners from relying on standard, established procedures of treatment. The digital twin is thereby able to derive quantitative estimates on the effects of cyberattacks within the health care branch.

Chapter 11 focuses on industrial production. Modern industrial production systems will be hallmarked by technology and a broad scale digitalization of the value chain. Thereby, the industrial internet of things contributes to a comprehensive interconnectivity of all components within the the factory. In this way, industrial robots, driverless transportation systems, and workpieces are communicating with each other in real time. The digitalization enables flexible production and permits fast reconfigurations of production lines. The changes with higher levels of digitalization introduce risks to security and safety of operation. The digital twin of the automotive production line is developed as a full-scale model factory with high levels of detail and permits the simulation of different degrees of digitalization and the production of a set of different products (different car body types and specifications) and the consequences an attack might have. The approach aims at enabling the evaluation of cyberrisks and demonstration of the application of the proposed methodology in an environment that comes as close as possible to a real factory. The digital twin thereby enables the quantification of the impact of cyberattacks, focusing on the specifics of industrial production. The approach thereby differs from the approach presented within chapter 10 by adaptions to the specifics of the automotive industry respectively industrial production respectively service provisioning within a hospital.

"Anticipating attacks, responding to them in real time, setting traps to contain them, and protecting assets according to their value" (Brown et al., 2017) is considered a key step to proactive and efficient defense against attacks that get increasingly sophisticated. This is, although, encryption of communication is essential for reaching cybersecurity, attackers strive to develop novel attacks for breaking into the most secure systems in a dynamic arms race with defenders. Hence, such static and non-specific defense mechanisms are not enough for reaching cybersecurity (Brown et al., 2017). Rather non-specific defense should be augmented with attack specific denfense. Chapter 12 presents an essay on the transition from automated threat hunting and real time attack detection to the automatic derivation of adequate safeguards for providing decision support in increasing the efficiency of cybersecurity investments. The work thereby introduces a bio-inspired artificial immune system composed by modules for unspecific defense against attacks and targeted defense against attacks that were detected. The approach descibed within this chapter takes advantage of the before introduced set of methods and aims at contributing to enabling decision support for increasing the efficiency of resource allocation. It thereby represents a first step in enabling cybersecurity investment portfolio management and the introduction of economic principles to cyberrisk management. It could thereby be expanded by the introduction of information on the costs of safeguards.

Bibliography

Abawajy, J. (2014). User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, *33*(3), 237–248.

Abbasi, R., Martinez, P., Ahmad, R. (2022). The digitization of agricultural industry–a systematic literature review on agriculture 4.0. *Smart Agricultural Technology*, 100042.

Abraham, S., Nair, S. (2014). Cyber security analytics: a stochastic model for security quantification using absorbing markov chains. *Journal of Communications*, *9*(12), 899–907.

Abusitta, A., Li, M. Q., Fung, B. C. (2021). Malware classification and composition analysis: A survey of recent developments. *Journal of Information Security and Applications*, *59*, 102828.

Adams, T., Connor, M., Whittaker, R. (2019). *Protecting our digital medicine infrastructure* (Vol. 2) (No. 1). Nature Publishing Group.

Adhikari, U., Morris, T. H., Pan, S. (2017). Applying hoeffding adaptive trees for real-time cyber-power event and intrusion classification. *IEEE Transactions on Smart Grid*, *9*(5), 4049–4060.

Aitenbichler, E., Lyardet, F., Austaller, G., Kangasharju, J., Mühlhäuser, M. (2007). Engineering intuitive and self-explanatory smart products. In *Proceedings of the 2007 acm symposium on applied computing* (pp. 1632–1637).

Aivazidou, E., Banias, G., Lampridi, M., Vasileiadis, G., Anagnostis, A., Papageorgiou, E., Bochtis, D. (2021). Smart technologies for sustainable water management: an urban analysis. *Sustainability*, *13*(24), 13940.

Ajami, S., Khaleghi, L. (2015). Hospital beds wireless sensor network and reducing decubitus ulcer. *Journal of Research in Medical Sciences*, *20*(6).

Alberts, D. S., Hayes, R. E. (2006). *Understanding command and control* (Tech. Rep.). Assistant secretary of defense (C3I/Command Control Research Program . . . .

Albright, D., Brannan, P., Walrond, C. (2010). *Did stuxnet take out 1,000 centrifuges at the natanz enrichment plant?* Institute for Science and International Security.

Alhakami, W., ALharbi, A., Bourouis, S., Alroobaea, R., Bouguila, N. (2019). Network anomaly intrusion detection using a nonparametric bayesian approach and feature selection. *IEEE Access*, *7*, 52181–52190.

Alkhammash, M., Alshahrani, M., Beloff, N., White, M. (2022). Revolutionising the approach to smart campus architecture through iot and blockchain technologies. In *Transformations through blockchain technology* (pp. 1–41). Springer.

Allodi, L., Massacci, F. (2017). Security events and vulnerability data for cybersecurity risk estimation. *Risk Analysis*, *37*(8), 1606–1627.

Almasizadeh, J., Azgomi, M. A. (2013). A stochastic model of attack process for the evaluation of security metrics. *Computer Networks*, *57*(10), 2159–2180.

Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., Anwar, A. (2020). Ton_iot telemetry dataset: A new generation dataset of iot and iiot for data-driven intrusion detection systems. *IEEE Access*, *8*, 165130–165150.

Alshehri, M. D., Hussain, F. K. (2019). A fuzzy security protocol for trust management in the internet of things (fuzzy-iot). *Computing*, *101*(7), 791–818.

Alshehri, M. D., Hussain, F. K., Hussain, O. K. (2018). Clustering-driven intelligent trust management methodology for the internet of things (citm-iot). *Mobile networks and applications*, *23*(3), 419–431.

Ameli, A., Hooshyar, A., El-Saadany, E. F., Youssef, A. M. (2018). Attack detection and identification for automatic generation control systems. *IEEE Transactions on Power Systems*, *33*(5), 4760–4774.

Ani, U. P. D., He, H., Tiwari, A. (2017). Review of cybersecurity issues in industrial critical infrastructure: manufacturing in perspective. *Journal of Cyber Security Technology*, *1*(1), 32–74.

Ardito, L., Petruzzelli, A. M., Panniello, U., Garavelli, A. C. (2018). Towards industry 4.0: Mapping digital technologies for supply chain management-marketing integration. *Business Process Management Journal*.

Arrow, K. J. (1964). The role of securities in the optimal allocation of risk-bearing. *The review of economic studies*, *31*(2), 91–96.

Arrow, K. J. (1992). Insurance, risk and resource allocation. In *Foundations of insurance economics* (pp. 220–229). Springer.

Asghar, M. R., Hu, Q., Zeadally, S. (2019). Cybersecurity in industrial control systems: Issues, technologies, and challenges. *Computer Networks*, *165*, 106946.

Ashton, K. (2009). That 'internet of things' thing. *RFID journal*, *22*(7), 97–114.

Atoum, I., Otoom, A., Ali, A. A. (2014). A holistic cyber security implementation framework. *Information Management & Computer Security*.

Aven, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety*, *99*, 33–44.

Bada, M., Sasse, A. M., Nurse, J. R. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672*.

Baezner, M., Robin, P. (2017). *Stuxnet* (Tech. Rep.). ETH Zurich.

Bailey, T., Kolo, B., Rajagopalan, K., Ware, D. (2018). Insider threat: The human element of cyberrisk. In *Perspectives on transforming cybersecurity* (pp. 33–40).

Baker, W. H., Wallace, L. (2007). Is information security under control?: Investigating quality in information security management. *IEEE Security & Privacy*, *5*(1), 36–44.

Barker, K., Lambert, J. H., Zobel, C. W., Tapia, A. H., Ramirez-Marquez, J. E., Albert, L., ... Caragea, C. (2017). Defining resilience analytics for interdependent cyber-physical-social networks. *Sustainable and Resilient Infrastructure*, *2*(2), 59–67.

Barnes, S. (2022). American dreams: Smart sleep, high-tech beds, and the national football league. *International Review for the Sociology of Sport*, 1012690221991778.

Beevers, N., Du Plessis, H., Martellini, L., Milhau, V. (2021). Measuring and managing the opportunity cost of downside risk protection. *The Journal of Portfolio Management*, *48*(1), 21–42.

Beyerer, J., Geisler, J. (2015). A quantitative risk model for a uniform description of safety and security. In *Proceedings of the 10th future security-security research conference" future security", berlin, 10, 2015* (p. 317).

Bhatia, M., Kaur, S., Sood, S. K. (2020). Iot-inspired smart toilet system for home-based urine infection prediction. *ACM Transactions on Computing for Healthcare*, *1*(3), 1–25.

Bloomberg, J. (2018). Digitization, digitalization, and digital transformation: confuse them at your peril. *Forbes. Retrieved on August*, *28*, 2019.

Boehm, J., Merrath, P., Poppensieker, T., Riemenschnitter, R., Stähle, T. (2017). Cyber risk measurement and the holistic cybersecurity approach. In *Perspectives on transforming cybersecurity* (pp. 61–74).

Böhme, R., Kataria, G. (2006). Models and measures for correlation in cyber-insurance. In *Weis* (Vol. 2, p. 3).

Böhme, R., Laube, S., Riek, M. (2019). A fundamental approach to cyber risk analysis. *Variance*, *12*(2), 161–185.

Bojanc, R., Jerman-Blažič, B. (2008). An economic modelling approach to information security risk management. *International Journal of Information Management*, *28*(5), 413-422. Retrieved from `https://www.sciencedirect.com/science/article/pii/S026840120800039X` doi: https://doi.org/10.1016/j.ijinfomgt.2008.02.002

Borg, M., Olsson, T., Franke, U., Assar, S. (2018). Digitalization of swedish government agencies-a perspective through the lens of a software development census. In *2018 ieee/acm 40th international conference on software engineering: Software engineering in society (icse-seis)* (pp. 37–46).

Bosch, J., Olsson, H. H. (2021). Digital for real: A multicase study on the digital transformation of companies in the embedded systems domain. *Journal of Software: Evolution and Process*, *33*(5), e2333.

Bouslama, G., Lahrichi, Y. (2017). Uncertainty and risk management from islamic perspective. *Research in International Business and Finance*, *39*, 718–726.

Bozkurt, A., Weiner, R., Rusch, I., Schulz, R. (2021). Exploring the requirements and challenges in production logistics for different sectors of the manufacturing industry. In *Towards sustainable customization: Bridging smart products and manufacturing systems* (pp. 475–482). Springer.

Brennen, J. S., Kreiss, D. (2016). Digitalization. *The international encyclopedia of communication theory and philosophy*, 1–11.

Brown, B., Ennis, D., Kaplan, J., Rosenthal, J. (2017). To survive in the age of advanced cyberthreats, use 'active defense'. In *Perspectives on transforming cybersecurity* (pp. 41–46).

Bruderer, H. (2018). Algorithms have been around for 4000 years. *Communications of the ACM*.

Buczak, A. L., Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, *18*(2), 1153–1176.

Bulgurcu, B., Cavusoglu, H., Benbasat, I. (2010). Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS quarterly*, 523–548.

Cai, J., Jia, H., Mao, T. (2022). A multivariate cvar risk measure from the perspective of portfolio risk management. *Scandinavian Actuarial Journal*, *2022*(3), 189–215.

Campbell, S. (2005). Determining overall risk. *Journal of risk research*, *8*(7-8), 569–581.

Canfield, C. I., Fischhoff, B., Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human factors*, *58*, 1158-1172.

Čapek, K., Comrada, N., Comrada, B. (1993). About the word robot.

Čapek, K., Kallinikov, I. (1940). *Rossum's universal robots*. Fr. Borovỳ.

Carfora, M., Martinelli, F., Mercaldo, F., Orlando, A. (2019). Cyber risk management: An actuarial point of view. *Journal of Operational Risk*, *14*(4).

Chang, C.-H., Kontovas, C., Yu, Q., Yang, Z. (2021). Risk assessment of the operations of maritime autonomous surface ships. *Reliability Engineering & System Safety*, *207*, 107324.

Chen, C., Cui, M., Fang, X., Ren, B., Chen, Y. (2020). Load altering attack-tolerant defense strategy for load frequency control system. *Applied Energy*, *280*, 116015.

Chen, C., Lin, C. X., Fredrikson, M., Christodorescu, M., Yan, X., Han, J. (2009). Mining graph patterns efficiently via randomized summaries. *Proceedings of the VLDB Endowment*, *2*(1), 742–753.

Chen, J., Zhu, Q. (2019). Interdependent strategic security risk management with bounded rationality in the internet of things. *IEEE Transactions on Information Forensics and Security*, *14*(11), 2958-2971. doi: 10.1109/TIFS.2019.2911112

Chen, Q., Abdelwahed, S., Erradi, A. (2014). A model-based validated autonomic approach to self-protect computing systems. *IEEE Internet of things Journal*, *1*(5), 446–460.

Chinn, D., Kaplan, J. M., Poppensieker, T. (Eds.). (2019). *Perspectives on transforming cybersecurity* (Vol. 32). McKinsey Global Institute.

Chismon, D., Ruks, M. (2015). Threat intelligence: Collecting, analysing, evaluating. *MWR InfoSecurity Ltd*, *3*(2), 36–42.

Choi, J., Kaplan, J., Krishnamurthy, C., Lung, H. (2017). Hit or myth? understanding the true costs and impact of cybersecurity programs. In *Perspectives on transforming cybersecurity* (pp. 8–17).

Chung, Y. J., Kim, I., Lee, N., Lee, T., In, H. P. (2005). Security risk vector for quantitative asset assessment. In *International conference on computational science and its applications* (pp. 274–283).

Cijan, A., Jenič, L., Lamovšek, A., Stemberger, J. (2019). How digitalization changes the workplace. *Dynamic relationships management journal*, *8*(1), 3–12.

Cilliers, L. (2020). Wearable devices in healthcare: Privacy and information security issues. *Health information management journal*, *49*(2-3), 150–156.

Cochoy, F., Hagberg, J., Kjellberg, H. (2019). The ethno-graphy of prices: on the fingers of the invisible hand (1922-1947). *Organization*, *26*(4), 492–516.

Colicchia, C., Creazza, A., Menachof, D. A. (2018). Managing cyber and information risks in supply chains: insights from an exploratory analysis. *Supply Chain Management: An International Journal*.

Connolly, J., Christey, S., Daldos, R., et al. (2018). *Medical device cybersecurity regional incident preparedness and response playbook*. Oct.

Conti, M., Dargahi, T., Dehghantanha, A. (2018). Cyber threat intelligence: challenges and opportunities. *Cyber Threat Intelligence*, 1–6.

Craig, A., Valeriano, B. (2016). Conceptualising cyber arms races. In *2016 8th international conference on cyber conflict (cycon)* (pp. 141–158).

Culnan, M. J. (2019). Policy to avoid a privacy disaster. *Journal of the Association for Information Systems*, *20*(6), 1.

Dayalan, M. (2017). Cyber risks, the growing threat. *IJNRD-International Journal of Novel Research and Development (IJNRD)*, *2*(9), 4–6.

Debatin, B., Lovejoy, J. P., Horn, A.-K., Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of computer-mediated communication*, *15*(1), 83–108.

Dehghani, M., Ghiasi, M., Niknam, T., Kavousi-Fard, A., Shasadeghi, M., Ghadimi, N., Taghizadeh-Hesary, F. (2021). Blockchain-based securing of data exchange in a power transmission system considering congestion management and social welfare. *Sustainability*, *13*(1). Retrieved from `https://www.mdpi.com/2071-1050/13/1/90`

DiMase, D., Collier, Z. A., Heffner, K., Linkov, I. (2015). Systems engineering framework for cyber physical security and resilience. *Environment Systems and Decisions*, *35*(2), 291–300.

Dormann, M., Hinz, S., Wittmann, E. (2019). Improving school administration through information technology? how digitalisation changes the bureaucratic features of public school administration. *Educational Management Administration & Leadership*, *47*(2), 275–290.

El-Gayar, O. F., Fritz, B. D. (2010). A web-based multi-perspective decision support system for information security planning. *Decision Support Systems*, *50*(1), 43–54.

Eling, M., Wirfs, J. (2019). What are the actual costs of cyber risk events? *European Journal of Operational Research*, *272*(3), 1109–1119.

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. In *2019 european intelligence and security informatics conference (eisic)* (pp. 40–47).

Elmasry, W., Akbulut, A., Zaim, A. H. (2020). Evolving deep learning architectures for network intrusion detection using a double pso metaheuristic. *Computer Networks*, *168*(107042).

Englebart, D. (1960). Microelectronics and the art of similitude. In *1960 ieee international solid-state circuits conference. digest of technical papers* (Vol. 3, pp. 76–77).

Evans, M., Maglaras, L. A., He, Y., Janicke, H. (2016). Human behaviour as an aspect of cybersecurity assurance. *Security and Communication Networks*, *9*(17), 4667–4679.

Falco, G., Eling, M., Jablanski, D., Weber, M., Miller, V., Gordon, L. A., . . . others (2019). Cyber risk research impeded by disciplinary barriers. *Science*, *366*(6469), 1066–1069.

Farwell, J. P., Stuxnet, R. R. (2011). the future of cyber war. *Survival. Global Politics and Strategy*, *53*(1), 25–28.

Fielder, A., Panaousis, E., Malacaria, P., Hankin, C., Smeraldi, F. (2016). Decision support approaches for cyber security investment. *Decision support systems*, *86*, 13–23.

Florêncio, D., Herley, C., Coskun, B. (2007). Do strong web passwords accomplish anything? *HotSec*, *7*(6), 159.

Fovino, I. N., Masera, M., De Cian, A. (2009). Integrating cyber attacks within fault trees. *Reliability Engineering & System Safety*, *94*(9), 1394–1402.

Friedrich, R., Le Merle, M., Grone, F., Koster, A. (2011). Measuring industry digitization: Leaders and laggards in the digital economy. *Booz & Co., London*.

Frisk, H. (1972). *Griechisches etymologisches wörterbuch/3 nachträge, wortregister, corrigenda, nachwort*. Winter.

Gandomi, A., Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137–144.

Gandotra, E., Bansal, D., Sofat, S. (2014). Malware analysis and classification: A survey. *Journal of Information Security*, *2014*.

Ganin, A. A., Quach, P., Panwar, M., Collier, Z. A., Keisler, J. M., Marchese, D., Linkov, I. (2020). Multicriteria decision framework for cybersecurity risk assessment and management. *Risk Analysis*, *40*(1), 183–199.

Garba, F. A. (2019). The anatomy of a cyber attack: dissecting the cyber kill chain (ckc). *Scientific and practical cyber security journal*, *3*(1).

Garg, A., Curtis, J., Halper, H. (2003). The financial impact of it security breaches: what do investors think? *Inf. Secur. J. A Glob. Perspect.*, *12*(1), 22–33.

Ge, M., Hong, J. B., Guttmann, W., Kim, D. S. (2017). A framework for automating security analysis of the internet of things. *Journal of Network and Computer Applications*, *83*, 12–27.

Geer, D., Hoo, K. S., Jaquith, A. (2003). Information security: Why the future belongs to the quants. *IEEE Security & Privacy*, *1*(4), 24–32.

Genge, B., Kiss, I., Haller, P. (2015). A system dynamics approach for assessing the impact of cyber attacks on critical infrastructures. *International Journal of Critical Infrastructure Protection*, *10*, 3–17.

Gertz, B. (2016). Pentagon developing pre-launch cyber attacks on missiles. *The Washington Free Beacon*.

Ghadge, A., Weiß, M., Caldwell, N. D., Wilding, R. (2019). Managing cyber risk in supply chains: A review and research agenda. *Supply Chain Management: An International Journal*.

Ghafur, S., Kristensen, S., Honeyford, K., Martin, G., Darzi, A., Aylin, P. (2019). A retrospective impact analysis of the wannacry cyberattack on the nhs. *NPJ digital medicine*, *2*(1), 1–7.

Grech, J. (2021). *A case study on cyber risk in financial institutions on the maltese islands and its reputation* (Unpublished master's thesis). University of Malta.

Gupta, S., Singhal, A., Kapoor, A. (2016). A literature survey on social engineering attacks: Phishing attack. In *2016 international conference on computing, communication and automation (iccca)* (pp. 537–540).

Guri, M., Kedma, G., Kachlon, A., Elovici, Y. (2014). Airhopper: Bridging the air-gap between isolated networks and mobile phones using radio frequencies. In *2014 9th international conference on malicious and unwanted software: The americas (malware)* (pp. 58–67).

Guri, M., Monitz, M., Mirski, Y., Elovici, Y. (2015). Bitwhisper: Covert signaling channel between air-gapped computers using thermal manipulations. In *2015 ieee 28th computer security foundations symposium* (pp. 276–289).

Guri, M., Zadov, B., Bykhovsky, D., Elovici, Y. (2019). Powerhammer: Exfiltrating data from air-gapped computers through power lines. *IEEE Transactions on Information Forensics and Security*, *15*, 1879–1890.

Gustafsson, M. S. (2017). *Reassembling local e-government: a study of actors' translations of digitalisation in public administration* (Unpublished doctoral dissertation). Linköping University Electronic Press.

Haas, A. (2016). Management von cyber-risiken und möglichkeiten des risikotransfers: eine ökonomische und versicherungstechnische analyse.

Hahn, A., Govindarasu, M. (2011). Cyber attack exposure evaluation framework for the smart grid. *IEEE Transactions on Smart Grid*, *2*(4), 835–843.

Hamilton, C., Adolphs, S., Nerlich, B. (2007). The meanings of 'risk': A view from corpus linguistics. *Discourse & Society*, *18*(2), 163–181.

Hardy, C. O. (1923). *Risk and risk-bearing*. University of Chicago Press.

Haynes, J. (1895). Risk as an economic factor. *The Quarterly Journal of Economics*, *9*(4), 409–449.

Hentea, M. (2008). Improving security for scada control systems. *Interdisciplinary Journal of Information, Knowledge, and Management*, *3*, 73.

Herrmann, F. (2018). The smart factory and its risks. *Systems*, *6*(4), 38.

Holm, H., Ekstedt, M., Andersson, D. (2012). Empirical analysis of system-level vulnerability metrics through actual attacks. *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 825-837. doi: 10.1109/TDSC.2012.66

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*(1), 74–81.

Housh, M., Ohar, Z. (2018). Model-based approach for cyber-physical attack detection in water distribution systems. *Water research*, *139*, 132–143.

Huang, K., Zhou, C., Tian, Y.-C., Yang, S., Qin, Y. (2018). Assessing the physical impact of cyber-attacks on industrial cyber-physical systems. *IEEE Transactions on Industrial Electronics*, *65*(10), 8153–8162.

Hudomiet, P., Willis, R. J. (2021). Computerization, obsolescence and the length of working life. *Labour Economics*, 102005.

Husák, M., Komárková, J., Bou-Harb, E., Čeleda, P. (2018). Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Communications Surveys & Tutorials*, *21*(1), 640–660.

Hutchins, E. M., Cloppert, M. J., Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, *1*(1), 80.

Iavazzo, C., GkegkE, X.-E. D., Iavazzo, P.-E., Gkegkes, I. D. (2014). Evolution of robots throughout history from hephaestus to da vinci robot. *AMHA-Acta medico-historica Adriatica*, *12*(2), 247–258.

Jasper, S. (2017). *Strategic cyber deterrence: The active cyber defense option*. Rowman & Littlefield.

John, R. M., Francis, F., Neelankavil, J., Antony, A., Devassy, A., Jinesh, K. (2014). Smart public transport system. In *2014 international conference on embedded systems (ices)* (pp. 166–170).

Johnson, K. N. (2015). Managing cyber risks. *Ga. L. Rev.*, *50*, 547.

Jones, M. D., Hutcheson, S., Camba, J. D. (2021). Past, present, and future barriers to digital transformation in manufacturing: A review. *Journal of Manufacturing Systems*, *60*, 936–948.

Jovanov, I., Pajic, M. (2019). Relaxing integrity requirements for attack-resilient cyber-physical systems. *IEEE Transactions on Automatic Control*, *64*(12), 4843-4858. doi: 10.1109/TAC.2019.2898510

Juncker, J.-C. (2018). European commission–speech: State of the union address 2017. *Journal of European Studies (JES)*, *34*(1), 128–145.

Kaiser, F., Wiens, M., Schultmann, F. (2020). Comparing the perception of privacy for medical devices and devices with medical functionality. *International Journal of Privacy and Health Information Management (IJPHIM)*, *8*(1), 52–69.

Kaiser, F., Wiens, M., Schultmann, F. (2021a). Motivation-based attacker modelling for cyber risk management: A quantitative content analysis and a natural experiment. *Journal of Information Security and Cybercrimes Research*, *4*(2), 132–147.

Kaiser, F., Wiens, M., Schultmann, F. (2021b). Use of digital healthcare solutions for care delivery during a pandemic-chances and (cyber) risks referring to the example of the covid-19 pandemic. *Health and Technology*, *11*(5), 1125–1137.

Kaloroumakis, P. E., Smith, M. J. (2021). Toward a knowledge graph of cybersecurity countermeasures. *Corporation, Editor*.

Kaminski, P., Rezek, C., Richter, W., Sorel, M. (2017). Protecting your critical digital assets: Not all systems and data are created equal. In *Perspectives on transforming cybersecurity* (pp. 27–32).

Kandasamy, K., Srinivas, S., Achuthan, K., Rangan, V. P. (2020). Iot cyber risk: A holistic analysis of cyber risk assessment frameworks, risk vectors, and risk ranking process. *EURASIP Journal on Information Security*, *2020*(1), 1–18.

Kang, D., Lee, J.-j., Kim, B. H., Hur, D. (2011). Proposal strategies of key management for data encryption in scada network of electric power systems. *International Journal of Electrical Power & Energy Systems*, *33*(9), 1521-1526. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0142061509000313` doi: https://doi.org/10.1016/j.ijepes.2009.03.004

Kaplan, S., Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, *1*(1), 11–27.

Katzir, Z., Elovici, Y. (2018). Quantifying the resilience of machine learning classifiers used for cyber security. *Expert Systems with Applications*, *92*, 419–429.

Kaufmann, E. (2021). Algorithm appreciation or aversion? comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence*, *2*, 100028.

Kazanin, A. (2020). Trends and prospects of development of the oil and gas sector in the context of digitalization. *Economics and Management*, *26*(1), 35–45.

Keller, W., Modarres, M. (2005). A historical overview of probabilistic risk assessment development and its use in the nuclear power industry: a tribute to the late professor norman carl rasmussen. *Reliability Engineering & System Safety*, *89*(3), 271–285.

Keskin, O. F. (2021). Quantifying cyber risk by integrating attack graph and impact graph.

Khalili, M. M., Naghizadeh, P., Liu, M. (2018). Designing cyber insurance policies: The role of pre-screening and security interdependence. *IEEE Transactions on Information Forensics and Security*, *13*(9), 2226–2239.

Kiesel, R., Heutmann, T., Dering, J., Kies, A., Vollmer, T., Schmitt, R. H. (2020). Cybersecurity in der vernetzten produktion. *Fraunhofer-Institut für Produktionstechnologie IPT, Aachen*.

Knapp, K. J., Morris Jr, R. F., Marshall, T. E., Byrd, T. A. (2009). Information security policy: An organizational-level process model. *Computers & security*, *28*(7), 493–508.

Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., Jones, K. (2015). A survey of cyber security management in industrial control systems. *International journal of critical infrastructure protection*, *9*, 52–80.

Komljenovic, D., Gaha, M., Abdul-Nour, G., Langheit, C., Bourgeois, M. (2016). Risks of extreme and rare events in asset management. *Safety science*, *88*, 129–145.

Lade, P., Ghosh, R., Srinivasan, S. (2017). Manufacturing analytics and industrial internet of things. *IEEE Intelligent Systems*, *32*(3), 74–79.

Lallie, H. S., Shepherd, L. A., Nurse, J. R., Erola, A., Epiphaniou, G., Maple, C., Bellekens, X. (2021). Cyber security in the age of covid-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *Computers & Security*, *105*, 102248.

Laney, D. (2001). 3-d data management: Controlling data volume, velocity and variety. application delivery strategies by meta group inc. *disponibile sul sito, available at: https://blogs. gartner. com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety. pdf (accessed 6 February 2001)*.

Langner, R. (2011). Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, *9*(3), 49–51.

Langner, R. (2013). To kill a centrifuge: A technical analysis of what stuxnet's creators tried to achieve. *The Langner Group*.

Lee, K.-C., Hsieh, C.-H., Wei, L.-J., Mao, C.-H., Dai, J.-H., Kuang, Y.-T. (2017). Sec-buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation. *Soft Computing*, *21*(11), 2883–2896.

Lenka, S., Parida, V., Wincent, J. (2017). Digitalization capabilities as enablers of value co-creation in servitizing firms. *Psychology & marketing*, *34*(1), 92–100.

Leszczyna, R. (2021). Review of cybersecurity assessment methods: Applicability perspective. *Computers & Security*, *108*, 102376.

Liao, X., Yuan, K., Wang, X., Li, Z., Xing, L., Beyah, R. (2016). Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In *Proceedings of the 2016 acm sigsac conference on computer and communications security* (pp. 755–766).

Limba, T., Plėta, T., Agafonov, K., Damkus, M., et al. (2017). Cyber security management model for critical infrastructure. *Entrepreneurship and Sustainability Issues*, *4*(4), 559–573.

Liu, N., Zhang, J., Zhang, H., Liu, W. (2010). Security assessment for communication networks of power control systems using attack graph and mcdm. *IEEE Transactions on Power Delivery*, *25*(3), 1492–1500.

Liu, T., Sun, Y., Liu, Y., Gui, Y., Zhao, Y., Wang, D., Shen, C. (2015). Abnormal traffic-indexed state estimation: A cyber–physical fusion approach for smart grid attack detection. *Future Generation Computer Systems*, *49*, 94–103.

Liu, X., Shahidehpour, M., Li, Z., Liu, X., Cao, Y., Li, Z. (2016). Power system risk assessment in cyber attacks considering the role of protection systems. *IEEE Transactions on Smart Grid*, *8*(2), 572–580.

Lockheed-Martin. (2014). Cyber kill chain. *URL: http://cyber. lockheedmartin. com/hubfs/Gaining the Advantage Cyber Kill Chain. pdf*.

Lv, Z., Han, Y., Singh, A. K., Manogaran, G., Lv, H. (2020). Trustworthiness in industrial iot systems based on artificial intelligence. *IEEE Transactions on Industrial Informatics*, *17*(2), 1496–1504.

Mahler, T., Nissim, N., Shalom, E., Goldenberg, I., Hassman, G., Makori, A., ... Shahar, Y. (2018). Know your enemy: Characteristics of cyber-attacks on medical imaging devices. *arXiv preprint arXiv:1801.05583*.

Makhdoom, I., Abolhasan, M., Lipman, J., Liu, R. P., Ni, W. (2018). Anatomy of threats to the internet of things. *IEEE communications surveys & tutorials*, *21*(2), 1636–1675.

Marcon, M., Sarti, A., Tubaro, S. (2016). Smart toothbrushes: inertial measurement sensors fusion with visual tracking. In *European conference on computer vision* (pp. 480–494).

Markowitz, H. (1952). *Portfolio selection* (Vol. 7) (No. 1).

Martin, G., Ghafur, S., Kinross, J., Hankin, C., Darzi, A. (2018). *Wannacry—a year on* (Vol. 361). British Medical Journal Publishing Group.

Maso, S. (2018). The philosophical category of" risk. *Philosophy & Epistemology International Journal*, *1*(1).

Masso, J., Pino, F. J., Pardo, C., García, F., Piattini, M. (2020). Risk management in the software life cycle: A systematic literature review. *Computer standards & interfaces*, *71*, 103431.

Mateski, M., Trevino, C. M., Veitch, C. K., Michalski, J., Harris, J. M., Maruoka, S., Frye, J. (2012). Cyber threat metrics. *Sandia National Laboratories*, 30.

Mattern, F., Floerkemeier, C. (2010). From the internet of computers to the internet of things. In *From active data management to event-based systems and more* (pp. 242–259). Springer.

Mayer-Schönberger, V., Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60–68.

McCullough, J. S., Casey, M., Moscovice, I., Prasad, S. (2010). The effect of health information technology on quality in us hospitals. *Health affairs*, *29*(4), 647–654.

McKenna, S., Staheli, D., Fulcher, C., Meyer, M. (2016). Bubblenet: A cyber security dashboard for visualizing patterns. In *Computer graphics forum* (Vol. 35, pp. 281–290).

Menges, F. (2020). *Cyber threat intelligence exchange* (Unpublished doctoral dissertation). Universität Regensburg.

Metzger, J. (2018). *Cyber risiken.* Gabler Springer Verlag.

Mikryukov, V. O., Melkov, S. A., Sushanskiy, A. S., Kholikov, I. V., Gruver, N. V. (2020). About the impact of the concept of e-government on modern education in russia. In *Shs web of conferences* (Vol. 79, p. 01007).

Mishra, S., Anderson, K., Miller, B., Boyer, K., Warren, A. (2020). Microgrid resilience: A holistic approach for assessing threats, identifying vulnerabilities, and designing corresponding mitigation strategies. *Applied Energy*, *264*, 114726.

Mohammadpourfard, M., Sami, A., Seifi, A. R. (2017). A statistical unsupervised method against false data injection attacks: A visualization-based approach. *Expert Systems with Applications*, *84*, 242–261.

Molina Zarca, A., Bernabe, J. B., Trapero, R., Rivera, D., Villalobos, J., Skarmeta, A., . . . Gouvas, P. (2019). Security management architecture for nfv/sdn-aware iot systems. *IEEE Internet of Things Journal*, *6*(5), 8005-8020. doi: 10.1109/JIOT.2019.2904123

Moness, M., Moustafa, A. M. (2015). A survey of cyber-physical advances and challenges of wind energy conversion systems: Prospects for internet of energy. *IEEE Internet of Things Journal*, *3*(2), 134–145.

Moore, G. E. (1965). *Cramming more components onto integrated circuits.* McGraw-Hill New York.

Moore, G. E. (1975). Progress in digital integrated electronics. In *Electron devices meeting* (Vol. 21, pp. 11–13).

Morse, A. (2018). Investigation: Wannacry cyber attack and the nhs. *Report by the National Audit Office. Accessed*, *1*.

Mourtzis, D., Vlachou, E. (2016). Cloud-based cyber-physical systems and quality of services. *The TQM Journal*.

Moustafa, N., Adi, E., Turnbull, B., Hu, J. (2018). A new threat intelligence scheme for safeguarding industry 4.0 systems. *IEEE Access*, *6*, 32910–32924.

Murugesan, M., Balamurugan, P., Santhosh, J., Arulkumaran, G. (2020). Threats and emerging developments in cyber security. *Webology*, *17*(2).

Nanterme, P. (2016). Digital disruption has only just begun. In *World economic forum.*

Neshenko, N., Bou-Harb, E., Crichigno, J., Kaddoum, G., Ghani, N. (2019). Demystifying iot security: an exhaustive survey on iot vulnerabilities and a first empirical look on internet-scale iot exploitations. *IEEE Communications Surveys & Tutorials*, *21*(3), 2702–2733.

Neugebauer, L. M., Zanko, I. (2021). How digitalization is changing the world? In *Lead community fundraising* (pp. 1–8). Springer.

New-York-magazine. (1996). Cyber extra! *New York magazine*.

Nicol, D. M., Sanders, W. H., Trivedi, K. S. (2004). Model-based evaluation: from dependability to security. *IEEE Transactions on dependable and secure computing*, *1*(1), 48–65.

Nübel, K., Bühler, M. M., Jelinek, T. (2021). Federated digital platforms: Value chain integration for sustainable infrastructure planning and delivery. *Sustainability*, *13*(16), 8996.

Nurse, J. R., Creese, S., De Roure, D. (2017). Security risk assessment in internet of things systems. *IT professional*, *19*(5), 20–26.

Obama, B. (2013). Presidential policy directive 21: Critical infrastructure security and resilience..

Ong, W. J. (2018). *Language as hermeneutic: A primer on the word and digitization.* Cornell University Press.

Oosthoek, K., Doerr, C. (2019). Sok: Att&ck techniques and trends in windows malware. In *International conference on security and privacy in communication systems* (pp. 406–425).

Or-Meir, O., Nissim, N., Elovici, Y., Rokach, L. (2019). Dynamic malware analysis in the modern era—a state of the art survey. *ACM Computing Surveys (CSUR)*, *52*(5), 1–48.

O'Reilly, U.-M., Toutouh, J., Pertierra, M., Sanchez, D. P., Garcia, D., Luogo, A. E., . . . Hemberg, E. (2020). Adversarial genetic programming for cyber security: A rising application domain where gp matters. *Genetic Programming and Evolvable Machines*, *21*(1), 219–250.

Pal, S., Sikdar, B., Chow, J. H. (2017). Classification and detection of pmu data manipulation attacks using transmission line parameters. *IEEE Transactions on Smart Grid*, *9*(5), 5057–5066.

Paridari, K., O'Mahony, N., Mady, A. E.-D., Chabukswar, R., Boubekeur, M., Sandberg, H. (2017). A framework for attack-resilient industrial control systems: Attack detection and controller reconfiguration. *Proceedings of the IEEE*, *106*(1), 113–128.

Pasqualetti, F., Dorfler, F., Bullo, F. (2015). Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems. *IEEE Control Systems Magazine*, *35*(1), 110–127.

Paté-Cornell, M.-E., Kuypers, M., Smith, M., Keller, P. (2018). Cyber risk management for critical infrastructure: a risk analysis model and three case studies. *Risk Analysis*, *38*(2), 226–241.

Patel, S. C., Graham, J. H., Ralston, P. A. (2008). Quantitatively assessing the vulnerability of critical information systems: A new method for evaluating security enhancements. *International Journal of Information Management*, *28*(6), 483-491. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0268401208000054` doi: https://doi.org/10.1016/j.ijinfomgt.2008.01.009

Pawlick, J., Chen, J., Zhu, Q. (2019, jun). Istrict: An interdependent strategic trust mechanism for the cloud-enabled internet of controlled things. *Trans. Info. For. Sec.*, *14*(6), 1654–1669. Retrieved from `https://doi.org/10.1109/TIFS.2018.2883272` doi: 10.1109/TIFS.2018.2883272

Peters, S., Chun, J.-H., Lanza, G. (2015). Digitalization of automotive industry–scenarios for future manufacturing.

Pfleeger, S. L., Caputo, D. D. (2012). Leveraging behavioral science to mitigate cyber security risk. *Computers & security*, *31*(4), 597–611.

Podgorski, D., Majchrzycka, K., Dąbrowska, A., Gralewicz, G., Okrasa, M. (2017). Towards a conceptual framework of osh risk management in smart working environments based on smart ppe, ambient intelligence and the internet of things technologies. *International Journal of Occupational Safety and Ergonomics*, *23*(1), 1–20.

Polatidis, N., Pavlidis, M., Mouratidis, H. (2018). Cyber-attack path discovery in a dynamic supply chain maritime risk management system. *Computer Standards & Interfaces*, *56*, 74–82.

Poolsappasit, N., Dewri, R., Ray, I. (2011). Dynamic security risk management using bayesian attack graphs. *IEEE Transactions on Dependable and Secure Computing*, *9*(1), 61–74.

Poppensieker, T., Riemenschnitter, R. (2018). A new posture for cybersecurity in a networked world. In *Perspectives on transforming cybersecurity* (pp. 18–26).

Proofpoint. (2019). *Proofpoint targeted attack protection transparenz und schutz vor hochentwickelten bedrohungen.*

Pub, F. (2005). Minimum security requirements for federal information and information systems.

Pudar, S., Manimaran, G., Liu, C.-C. (2009). Penet: A practical method and tool for integrated modeling of security attacks and countermeasures. *Computers & Security*, *28*(8), 754-771. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167404809000522` doi: https://doi.org/10.1016/j.cose.2009.05.007

Pupillo, L., Griffith, M., Blockmans, S., Renda, A. (2018). Strengthening the eu's cyber defence capabilities. *CEPS Task Force Report*.

Qamar, N. U., Razia, E. T., et al. (2017). An overview on cyber attacks and its types for enhancing data security in business world. *Lahore Garrison University Research Journal of Computer Science and Information Technology*, *1*(4), 37–50.

Qamar, S., Anwar, Z., Rahman, M. A., Al-Shaer, E., Chu, B.-T. (2017). Data-driven analytics for cyber-threat intelligence and information sharing. *Computers & Security*, *67*, 35–58.

Radanliev, P., De Roure, D., Cannady, S., Montalvo, R. M., Nicolescu, R., Huth, M. (2018). Economic impact of iot cyber risk-analysing past and present to predict the future developments in iot risk analysis and iot cyber insurance.

Radanliev, P., De Roure, D. C., Nicolescu, R., Huth, M., Montalvo, R. M., Cannady, S., Burnap, P. (2018). Future developments in cyber risk assessment for the internet of things. *Computers in industry*, *102*, 14–22.

Ralston, P. A., Graham, J. H., Hieb, J. L. (2007). Cyber security risk assessment for scada and dcs networks. *ISA transactions*, *46*(4), 583–594.

Rampini, A. A., Viswanathan, S. (2010). Collateral, risk management, and the distribution of debt capacity. *The Journal of Finance*, *65*(6), 2293–2322.

Rana, M. M., Li, L., Su, S. W. (2017). Cyber attack protection and control of microgrids. *IEEE/CAA Journal of Automatica Sinica*, *5*(2), 602–609.

Rao, A. A., Tan, S. Q. Y., Raghavi, R., Srivastava, A., Renumadhavi, C. (2022). Autism spectrum disorder therapy: Analysis of artificial intelligence integrated robotic approach. In *Journal of physics: Conference series* (Vol. 2161, p. 012038).

Rao, P. P. (2018). Robotic surgery: new robots and finally some real competition! *World journal of urology*, *36*(4), 537–541.

Rauscher, K. F., Cox, E. N. (2013). Measuring the cyber security problem..

Rees, L. P., Deane, J. K., Rakes, T. R., Baker, W. H. (2011). Decision support for cybersecurity risk planning. *Decision Support Systems*, *51*(3), 493–505.

Rogers, D., Kanth, V. (2021). Wannacry data collection system design and tutorial. In *Milcom 2021-2021 ieee military communications conference (milcom)* (pp. 1–6).

Rogers, J. (2016). *Public-private partnerships: A tool for enhancing cybersecurity* (Unpublished doctoral dissertation). Johns Hopkins University.

Rothrock, R. A., Kaplan, J., van der Oord, F. (2018). The board's role in managing cybersecurity risks. *MIT Sloan Management Review*, *59*(2), 12–15.

Ruan, K. (2017). Introducing cybernomics: A unifying economic framework for measuring cyber risk. *Computers & Security*, *65*, 77–89.

Russom, P., et al. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1–34.

Ryan, J. J. C. H., Jefferson, T. I. (2003). The use, misuse, and abuse of statistics in information security research. In *Proceedings of the 2003 asem national conference, st. louis, mo.*

Ryan, J. J. C. H., Mazzuchi, T. A., Ryan, D. J., De la Cruz, J. L., Cooke, R. (2012). Quantifying information security risks using expert judgment elicitation. *Computers & Operations Research*, *39*(4), 774–784.

Sasse, A. (2015). Scaring and bullying people into security won't work. *IEEE Security & Privacy*, *13*(3), 80–83.

Schaller, R. R. (1997). Moore's law: past, present and future. *IEEE spectrum*, *34*(6), 52–59.

Schallmo, A., Daniel, R. (2018). *Digital transformation now! guiding the successful digitalization of yourbusiness model*. Springer.

Schreckling, E., Steiger, C. (2017). Digitalize or drown. In *Shaping the digital enterprise* (pp. 3–27). Springer.

Schulze, M. (2020). Cyber in war: assessing the strategic, tactical, and operational utility of military cyber operations. In *2020 12th international conference on cyber conflict (cycon)* (Vol. 1300, pp. 183–197).

Sechser, T. S., Narang, N., Talmadge, C. (2019). Emerging technologies and strategic stability in peacetime, crisis, and war. *Journal of Strategic Studies*, *42*(6), 727-735. doi: 10.1080/01402390.2019.1626725

Sengan, S., Subramaniyaswamy, V., Nair, S. K., Indragandhi, V., Manikandan, J., Ravi, L. (2020). Enhancing cyber–physical systems with hybrid smart city cyber security architecture for secure public data-smart network. *Future generation computer systems*, *112*, 724–737.

Sengupta, S., Chowdhary, A., Sabur, A., Alshamrani, A., Huang, D., Kambhampati, S. (2020). A survey of moving target defenses for network security. *IEEE Communications Surveys & Tutorials*, *22*(3), 1909–1941.

Sergeeva, T. (2021). Private wagon fleet management in a digitised industry. In *International scientific siberian transport forum* (pp. 361–370).

Shackleford, D. (2015). Who's using cyberthreat intelligence and how. *SANS Institute*.

Shafto, M., Conroy, M., Doyle, R., Glaessgen, E., Kemp, C., LeMoigne, J., Wang, L. (2012). Modeling, simulation, information technology & processing roadmap. *National Aeronautics and Space Administration*, *32*(2012), 1–38.

Shameli-Sendi, A., Aghababaei-Barzegar, R., Cheriet, M. (2016). Taxonomy of information security risk assessment (isra). *Computers & security*, *57*, 14–30.

Sharkov, G. (2016). From cybersecurity to collaborative resiliency. In *Proceedings of the 2016 acm workshop on automated decision making for active cyber defense* (pp. 3–9).

Sheehan, B., Murphy, F., Mullins, M., Ryan, C. (2019). Connected and autonomous vehicles: A cyber-risk classification framework. *Transportation research part A: policy and practice*, *124*, 523–536.

Shen, D., Wu, G., Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, *19*, 221.

Shetty, S., McShane, M., Zhang, L., Kesan, J. P., Kamhoua, C. A., Kwiat, K., Njilla, L. L. (2018). Reducing informational disadvantages to improve cyber risk management. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *43*(2), 224–238.

Smith, S., Winchester, D., Bunker, D., Jamieson, R. (2010). Circuits of power: A study of mandated compliance to an information systems security" de jure" standard in a government organization. *MIS quarterly*, 463–486.

Steinbrück, A., Wiens, M., Birnstill, P., Kaiser, F., Zander, T., Raabe, O., . . . Volkamer, M. (2021). Ein neues" datenkartellrecht" zum schutz der informationellen selbstbestimmung im markt der sozialen plattformen? *Recht der Datenverarbeitung (RDV)*, *37*(1), 7–13.

Stewart, H., Jürjens, J. (2018). Data security and consumer trust in fintech innovation in germany. *Information & Computer Security*.

Strom, B. E., Battaglia, J. A., Kemmerer, M. S., Kupersanin, W., Miller, D. P., Wampler, C., . . . Wolf, R. D. (2017). Finding cyber threats with att&ck-based analytics. *The MITRE Corporation, Bedford, MA, Technical Report No. MTR170202*.

Strupczewski, G. (2021). Defining cyber risk. *Safety science*, *135*, 105143.

Sundmaeker, H., Guillemin, P., Friess, P., Woelfflé, S. (Eds.). (2010). *Vision and challenges for realising the internet of things*. Publications Office of the European Union. doi: 10.2759/26127

Sury, M. (2019). Digitization of tax administration in india. *VISION: Journal of Indian Taxation*, *6*(2), 79–91.

Taormina, R., Galelli, S. (2018). Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems. *Journal of Water Resources Planning and Management*, *144*(10), 04018065.

Teixeira, A., Sou, K. C., Sandberg, H., Johansson, K. H. (2015). Secure control systems: A quantitative risk management approach. *IEEE Control Systems Magazine*, *35*(1), 24–45.

Thomson, R. (2003). The use of utility functions for investment channel choice in defined contribution retirement funds. i: Defence. *British Actuarial Journal*, *9*(3), 653–709.

Timonen, K. (2022). The impact of fintech on the banking industry.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, *25*(1), 44–56.

Trump, D. (2017). Executive order 13800—presidential executive order on strengthening the cybersecurity of federal networks and critical infrastructure..

Tully, J., Selzer, J., Phillips, J. P., O'Connor, P., Dameff, C. (2020). Healthcare challenges in the era of cybersecurity. *Health security*, *18*(3), 228–231.

Valenduc, G., Vendramin, P. (2017). Digitalisation, between disruption and evolution. *Transfer: European Review of Labour and Research*, *23*(2), 121–134.

van Mulken, T. J. M., Schols, R. M., Scharmga, A. M. J., Winkens, B., Cau, R., Schoenmakers, F. B. F., ... van der Hulst, R. R. W. J. (2020). First-in-human robotic supermicrosurgery using a dedicated microsurgical robot for treating breast cancer-related lymphedema: a randomized pilot trial. *Nature communications*, *11*(1), 1–7.

van Schaik, P., Jeske, D., Onibokun, J., Coventry, L., Jansen, J., Kusev, P. (2017). Risk perceptions of cyber-security and precautionary behaviour. *Computers in Human Behavior*, *75*, 547–559.

von Solms, B., von Solms, R. (2018). Cybersecurity and information security–what goes where? *Information & Computer Security*.

Wadhawan, Y., AlMajali, A., Neuman, C. (2018). A comprehensive analysis of smart grid systems against cyber-physical attacks. *Electronics*, *7*(10), 249.

Wang, B., Dabbaghjamanesh, M., Kavousi-Fard, A., Mehraeen, S. (2019). Cybersecurity enhancement of power trading within the networked microgrids based on blockchain and directed acyclic graph approach. *IEEE Transactions on Industry Applications*, *55*(6), 7300–7309.

Wang, J., Paschalidis, I. C. (2016). Botnet detection based on anomaly and community detection. *IEEE Transactions on Control of Network Systems*, *4*(2), 392–404.

Wang, Q., Tai, W., Tang, Y., Ni, M., You, S. (2019). A two-layer game theoretical attack-defense model for a false data injection attack against power systems. *International Journal of Electrical Power & Energy Systems*, *104*, 169–177.

Wiens, M. (2013). *Vertrauen in der ökonomischen theorie: Eine mikrofundierte und verhaltensbezogene analyse* (Vol. 9). LIT Verlag Münster.

Wiens, M. (2021). *Resilient systems-an economic, operational, and behavioral perspective* (Unpublished doctoral dissertation). Karlsruher Institut für Technologie (KIT).

Winau, M., Kaiser, F., Wiens, M., Schultmann, F., Spiecker, I. (2021). Datenschutz durch technikgestaltung und unternehmerische strategie in der digitalwirtschaft. In *Lecture notes in informatics* (pp. 999–1018).

Winterrose, M. L., Carter, K. M. (2014). Strategic evolution of adversaries against temporal platform diversity active cyber defenses. *arXiv preprint arXiv:1408.0023*.

Wong, N., Ray, P. K., Stephens, G., Lewis, L. M. (2012). Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results. *Information Systems Journal*, *22*.

Wu, Y., Wei, Z., Weng, J., Li, X., Deng, R. H. (2018). Resonance attacks on load frequency control of smart grids. *IEEE Transactions on Smart Grid*, *9*(5), 4490-4502. doi: 10.1109/TSG.2017.2661307

Xidias, E., Zacharia, P., Nearchou, A. (2022). Intelligent fleet management of autonomous vehicles for city logistics. *Applied Intelligence*, 1–19.

Xu, K., Zhang, Z.-L., Bhattacharyya, S. (2008). Internet traffic behavior profiling for network security monitoring. *IEEE/ACM Transactions On Networking*, *16*(6), 1241–1252.

Xu, M., Schweitzer, K. M., Bateman, R. M., Xu, S. (2018). Modeling and predicting cyber hacking breaches. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2856-2871. doi: 10.1109/TIFS.2018.2834227

Xu, P., He, S., Wang, W., Susilo, W., Jin, H. (2017). Lightweight searchable public-key encryption for cloud-assisted wireless sensor networks. *IEEE Transactions on Industrial Informatics*, *14*(8), 3712–3723.

Xu, S. (2020). The cybersecurity dynamics way of thinking and landscape. In *Proceedings of the 7th acm workshop on moving target defense* (pp. 69–80).

Yadav, T., Rao, A. M. (2015). Technical aspects of cyber kill chain. In *International symposium on security in computing and communication* (pp. 438–452).

Yang, Y.-P. O., Shieh, H.-M., Tzeng, G.-H. (2013). A vikor technique based on dematel and anp for information security risk control assessment. *Information sciences*, *232*, 482–500.

Yaqoob, I., Khan, L. U., Kazmi, S. A., Imran, M., Guizani, N., Hong, C. S. (2019). Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. *IEEE Network*, *34*(1), 174–181.

Yoon, S., Ha, T., Kim, S., Lim, H. (2017). Scalable traffic sampling using centrality measure on software-defined networks. *IEEE Communications Magazine*, *55*(7), 43–49.

Zander, T., Birnstill, P., Kaiser, F., Wiens, M., Beyerer, J., Schultmann, F. (2020, Apr.). It security and competition in the automotive industry: A diffusion model. *TATuP - Journal for Technology Assessment in Theory and Practice*, *29*(1), 16–22. Retrieved from `https://www.tatup.de/index.php/tatup/article/view/6790` doi: 10.14512/tatup.29.1.16

Zeller, G., Scherer, M. (2021). A comprehensive model for cyber risk based on marked point processes and its application to insurance. *European Actuarial Journal*, 1–53.

Zeng, W., Zhang, Y., Chow, M.-Y. (2015). Resilient distributed energy management subject to unexpected misbehaving generation units. *IEEE Transactions on Industrial Informatics*, *13*(1), 208–216.

Zhang, H., Yi, Y., Wang, J., Cao, N., Duan, Q. (2018). Network security situation awareness framework based on threat intelligence. *Computers, Materials and Continua*, *56*(3), 381–399.

Zhang, Q., Zhou, C., Tian, Y.-C., Xiong, N., Qin, Y., Hu, B. (2017). A fuzzy probability bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems. *IEEE Transactions on Industrial Informatics*, *14*(6), 2497–2506.

Zhang, Q., Zhou, C., Xiong, N., Qin, Y., Li, X., Huang, S. (2015). Multimodel-based incident prediction and risk assessment in dynamic cybersecurity protection for industrial control systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *46*(10), 1429–1444.

Zheng, B., Deng, P., Anguluri, R., Zhu, Q., Pasqualetti, F. (2016). Cross-layer codesign for secure cyber-physical systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *35*(5), 699–711.

Part II

Articles

Overview of Companion Articles

Study A

Kaiser, F., Dardik, U., Elitzur, A., Zilberman, P., Wiens, M., Schultmann, F., . . . Puzis, R. (2022). Attack hypotheses generation based on threat intelligence knowledge graph. *Submitted to a scientific journal*.

Study B

Kaiser, F., Wiens, M., Schultmann, F. (2022). Cyber risk quantification - using weighted attack graphs for behavioral cyber game theory. *Accepted for publication within the book "Advances in Cyber Security and Intelligent Analytics" CRC Press, Taylor & Francis Group.*

Study C

Kaiser, F., Budig, T., Goebel, E., Fischer, T., Muff, J., Wiens, M., Schultmann, F. (2021). Attack forecast and prediction. In *Proceedings of the 28th c&esar: Computer electronics security application rendezvous.*

Study D

Wiens, M., Kaiser, F., Schultmann, F. (n.d.). Too stressful to look closely? the information value of signal detection under cognitive contraints – a decision-theoretic model for the case of phishing mail detection. *Submitted to a scientific journal*.

Study E

Kaiser, F., Wiens, M., Schultmann, F. (n.d.). Cyberattacks on hospitals and their impact on medical service. *Submitted to a scientific journal*.

Study F

Kaiser, F., Wiens, M., Schultmann, F. (n.d). Digital twins and their use for cyber risk quantification - analyzing the impact of cyberattacks on an automobile manufacturer. *Submitted to a scientific journal*.

Study G

Kaiser, F., Andris, L., Tennig, T., Iser, J., Wiens, M., Schultmann, F. (n.d.). Transitions from threat hunting and automated incident response. *Submitted to a scientific conference*.

6 Attack Hypotheses Generation Based on Threat Intelligence Knowledge Graph[1]

Abstract

Cyber threat intelligence on past attacks may help with attack reconstruction and the prediction of the course of an ongoing attack by providing deeper understanding of the tools and attack patterns used by attackers. Therefore, cyber security analysts employ threat intelligence, alert correlations, machine learning, and advanced visualizations in order to produce sound attack hypotheses. In this paper, we present AttackDB, a multi-level threat knowledge base that combines data from multiple threat intelligence sources to associate high-level ATT&CK techniques with low-level telemetry found in behavioral malware reports. We also present the Attack Hypothesis Generator which relies on knowledge graph traversal algorithms and a variety of link prediction methods to automatically infer ATT&CK techniques from a set of observable artefacts. Results of experiments performed with 53K VirusTotal reports indicate that the proposed algorithms employed by the Attack Hypothesis Generator are able to produce accurate adversarial technique hypotheses with a mean average precision greater than 0.5 and area under the receiver operating characteristic curve of over 0.8 when it is implemented on the basis of AttackDB. The presented toolkit will help analysts to improve the accuracy of attack hypotheses and to automate the attack hypothesis generation process.

Keywords: Cyber Threat Intelligence, Data Fusion, Attack Hypotheses, Link Prediction

6.1 Introduction

In the last years, the perpetrators of cyber attacks have been playing a dynamic cat and mouse game with those trying to stop them. In order to stay ahead of their opponents, cyber security analysts search for techniques that can assist them in threat hunting and intrusion detection, as well as in the forensic investigation of attacks, as they try to infer the attackers' objectives, trace back the

[1] This chapter includes the preprint of the article "Attack Hypotheses Generation Based on Threat Intelligence Knowledge Graph" by Uriel Dardik Aviad Elitzur Polina Zilberman, Marcus Wiens, Frank Schultmann, Yuval Elovici, Rami Puzis, and myself.

attack vector used for initial penetration, and reconstruct the intermediate attack steps. The general workflow of these investigations starts with sensors that send monitored data to an organization's security information and event management system (SIEM). The SIEM aggregates and correlates the data from the sensors and generates alerts when a suspicious event is detected. On the basis of these alerts, security analysts derive hypotheses on the state of the system and draw conclusions about the attacker's goals, the methods he/she uses to achieve these goals, and further plans which is critical to providing a quick response to an attack and may thus result in reduced damage from an attack. Despite the great importance of these tasks, security analysts have little time to devote to them due to a shortage of experienced security analysts. Currently, even the best security operation centers are not fully automated and human analysts have primary responsibility for understanding, prioritizing, investigating, and responding to the alerts raised by the SIEM. Accumulated cyber threat intelligence (CTI) can help analysts to understand the goals and methods of relevant attack actors (Bromiley, 2016; Mavroeidis and Bromander, 2017).

Besides attack hypothesis generation, the automated use of CTI, such as indicators of compromise (IoCs), is prevalent throughout endpoint detection and response (EDR), extends beyond EDR as well as security orchestration automation and response solutions. Although there are many tools that utilize CTI (e.g., EDR), most of them concentrate on low-level constructs such as IoCs. Some automation employs cyber analytics, e.g., the MITRE Cyber Analytic Repository[2], to detect specific techniques. These tools operate in a top-down fashion as they hunt for the techniques supposedly employed by the adversary. The methods presented in this work differ from the state of the art by (1) focusing on the high-level indicators of attack (IoAs), i.e., collections of techniques, and (2) operating in a bottom-up manner, inferring collections of techniques from observable artifacts (not necessarily IoCs). This is an important step towards improving the cyber security of systems, as deriving high-level CTI from low-level CTI may help analysts with their tasks and contribute to increased efficiency among security analysts. However, deriving these high-level insights is challenging. Currently, except for cyber analytics, there is a lack of analysis tools and algorithms that utilize both high- and low-level CTI.

In this paper, we build and extend AttackDB (Elitzur et al., 2019), a multi-level threat knowledge base that fuses data from the MITRE ATT&CK Enterprise knowledge base[3], the AlienVault Open Threat Exchange (OTX)[4], the IBM X-Force Exchange (X-Force)[5], and VirusTotal[6]. We also introduce the Attack Hypothesis Generator (AHG), a toolkit that (1) infers adversarial techniques from low-level telemetry data using multiple techniques for information retrieval; and (2) refines a given hypothesis using various link prediction techniques.

The main contributions of this paper are as follows:

(i) We contribute a comprehensive multi-level threat knowledge base that fuses multiple open-source threat intelligence sources (Dekel et al., 2021).

[2] https://car.mitre.org/
[3] https://attack.mitre.org/
[4] https://otx.alienvault.com/
[5] https://exchange.xforce.ibmcloud.com/
[6] https://www.virustotal.com/

(ii) We utilize the proposed multi-level threat intelligence knowledge base to generate and refine attack technique hypotheses using graph analytics. We also provide a heuristic, based on the expected number of techniques, which suggests when refinement is beneficial.

(iii) We evaluate the feasibility of AHG to generate attack hypotheses when attack indicators are not known (e.g. a novel malware family utilizing a zero-day exploit).

AHG's ability to infer and refine the set of adversarial techniques used by an attacker is a critical step toward increasing the automation level of threat hunting and forensic investigations, both of which will contribute to organizations' cyber defense and enable them to gain insight regarding the state of their system (system monitoring). AHG can improve an analyst's perception of an attack under investigation and result in actionable insights pertinent to an attack. Furthermore, AHG addresses the "lack of published or accessible methodologies" (Daszczyszak et al., 2019) for threat hunting based on high-level attack patterns.

The rest of the paper is structured as follows. Relevant background and related work are described in section 6.2. Section 6.3 presents the schema, insight into the construction process, and statistics regarding AttackDB. In section 6.4, we present the proposed algorithms, both for hypothesis inference and hypothesis refinement; the various algorithms are evaluated in section 6.5. section 6.6 contains a summary, our conclusions, and plans for future research.

6.2 Literature review

6.2.1 Cyber threat intelligence

CTI is structured, actionable information for identifying adversaries and their motives, goals, capabilities, resources, and tactics. It includes evidence-based knowledge in the form of measurable events and the context for the events' interpretation. This information can be clustered into four categories: (i) technical, (ii) tactical, (iii) operational, and (iv) strategic CTI (Chismon and Ruks, 2015). Information extracted from CTI improves an analyst's ability to recognize relevant threats and respond to them in a timely manner (Jasper, 2017; Tyler Technologies, 2018). This is, methods for CTI analysis can provide analysts with a list of related information, supporting their decision-making as they handle cyber incidents (Settanni et al., 2017). Hence, CTI is a powerful means of increasing the efficiency of various security solutions, such as intrusion detection, incident response, real-time analytics, forensic investigation, and threat hunting. The practical use is validated within a survey of various cyber security and information technology management professionals presented by Shackleford (Shackleford, 2015). According to the study, 48% of the respondents said their use of CTI has reduced incidents through early prevention, and 51% said they are able to respond more quickly to incidents.

CTI can be acquired by a victim organization that records attack investigation artifacts (such as IoCs), e.g., through anomaly detection systems respectively through intrusion detection systems.

However, differentiating a benign anomaly (e.g., caused by irregular user behavior or the implementation of a new device) from an attack is oftentimes challenging leading to high false positive rates (fpr) and false negative rates (fnr) (Landauer et al., 2019). Therefore, assessing the relevance of observables (i.e., IoCs) is crucial for effective attack hypothesis generation. Capturing CTI is also critical and has been the subject of much research. Wheelus et al. (2016) proposed a tiered big data architecture for the automated capturing and handling of network traffic; this enables the generation of features and artifacts for machine learning algorithms and anomaly detectors. Samtani et al. (2017) suggested collecting CTI proactively from large international underground hacker communities without waiting for attacks to happen. They developed a framework for storing and analyzing malicious assets, such as crypters, keyloggers, and web and database exploits collected from the dark web. Landauer et al. (2019) presented a methodology for automatically or semi-automatically transforming raw log data to actionable CTI. By doing so, they identified relevant information for threat hunting based on a continuous flow of raw log data using a parser tree and anomaly detection algorithms.

Furthermore, these low-level attack artefacts (technical CTI) are quickly actionable; however, very likely to become obsolete in a short time. Therefore, the value of low-level CTI (e.g., IoCs) for crafting attack hypotheses is debatable (Landauer et al., 2019). Hence, it is necessary to combine the benefits of more abstract and thus more robust (in terms of obsolescence) tactics and techniques (high-level CTI) with actionable IoCs (low-level CTI).

Since no organization possesses complete understanding of the threat landscape from recording attack artefacts, the importance of CTI lies in its ability be to shared among partners in a machine-to-machine manner. By sharing the who, what, where, how, and when of malicious activities, organizations obtain a holistic view of the threat landscape thus increasing their cyber security readiness (Jasper, 2017). However sharing CTI introduces novel risks for the trustworthiness of CTI. Thus, assessing the correctness and reliability of CTI is essential, as there may be untrusted data sources. Furthermore, a major challenge in sharing CTI is that it is shared in various different formats from a variety of sources. Therefore, information sharing needs to be streamlined and structured (Bromander et al., 2020). In an effort to formalize a standard language for sharing CTI, the US Department of Homeland Security's Office of Cybersecurity and Communications provided funding to MITRE to develop the Structured Threat Information eXpression (STIX) language.[7] STIX covers the entire range of cyber security concepts, including observables, IoCs, attack patterns, tools, malware, threat actors, courses of action, and more. A STIX element is denoted as a STIX Domain Object (SDO). An IoC is an artifact or pattern which, if found, indicates that malicious activity is being performed. SDOs, such as observables and IoCs, are considered low-level CTI, while SDOs, such as attack patterns, tools, and threat actors, are considered high-level CTI. In the literature the term IoA is defined as the entirety of CTI available on an attack, including high-level descriptions of tactics, techniques, and procedures (TTP) (DeCianno, 2014).

[7] `https://oasis-open.github.io/cti-documentation/stix/intro`

Additional CTI languages include OpenIOC,[8] Trusted Automated eXchange of Indicator Information,[9] and the Incident Object Description Exchange Format,[10] as well as proprietary languages and ontologies developed, e.g., Global Threat Intelligence by McAfee (McAfee, retrieved 2019b) or IntelGraph by Accenture (Plona et al., 2017).

Yet, according to a study provided by Sauerwein et al. (2017), STIX is the de facto standard language for sharing CTI. Zhao et al. (2017) presented an unified cyber threat ontology integrating heterogeneous CTI languages. An ontology can thereby be considered a meta-model of the knowledge graph which presents general domain concepts (Iqbal and Anwar, 2016). There are many different sharing platforms with practical relevance, including the Malware Information Sharing Platform,[11] OpenCTI,[12] the Collective Intelligence Framework,[13] Anomali STAXX,[14] and the OTX platform [15]. de Melo e Silva et al. (2020) compared and investigated a large variety of CTI languages and platforms and evaluated their strengths and weaknesses. For a review on the various languages we refer to their work.

While existing CTI ontologies, languages, and repositories are integral to the approach presented in this study, they are not sufficient for the effective generation of attack hypotheses. Mavroeidis and Bromander (2017) reviewed existing ontologies and concluded that there is "not any (...) ontology readily available for use" (Mavroeidis and Bromander, 2017). According to their study, existing ontologies need to be criticized for their "lack of expressiveness" and missing holistic view. Motivated by their study, we partially bridge this gap using the proposed AttackDB in section 6.3. AttackDB facilitates a holistic view of the different techniques used by a variety of malware, effectively connecting them with low-level CTI.

6.2.2 Threat hunting

In this subsection, we provide background on threat hunting. Cyber security experts are divided regarding the exact stages of the threat hunting cycle and its reactive or proactive nature. On the one hand, some experts define threat hunting as proactively looking for early indications of presumably ongoing attacks without waiting for alerts to indicate suspicious activity (Alonso, 2016). On the other hand, threat hunting may refer to an investigative process initiated in response to an alert. This process may include advanced analytics, forensic investigations, targeted data collection, or policy updating (Rasheed et al., 2017; Sqrrl Data, 2016). The main difference between proactive and reactive threat hunting is the trigger for the investigation. Proactive threat hunting relies on CTI to actively search for potentially malicious behavior. Reactive threat hunting involves forensic investigation and attack hypothesis testing in response to alerts indicating such behavior.

[8] https://github.com/mandiant/OpenIOC_1.1
[9] https://oasis-open.github.io/cti-documentation/taxii/intro
[10] https://datatracker.ietf.org/doc/html/rfc5070
[11] https://www.misp-project.org/
[12] https://www.opencti.io/en/
[13] https://csirtgadgets.com/collective-intelligence-framework/
[14] https://www.anomali.com/resources/staxx
[15] https://otx.alienvault.com/

A significant amount of effort has been invested in the seamless integration of machines and human analysts within the threat hunting cycle (McAfee, retrieved 2019a; R. M. Lee and Lee, 2018; Sqrrl Data, 2016; Tyler Technologies, 2018). A noteworthy product that provides human-machine teaming capabilities is the McAfee Investigator (McAfee, retrieved 2019a). This product can be considered a reactive threat hunting product, because it starts with choosing an incident for detailed investigation. Advanced machine learning algorithms choose the most relevant insights for the human analysts who can then determine the risk and urgency of the incident. After the analyst has chosen an incident for detailed investigation, the machinery uses human input to gather relevant information and provides a summary to the analyst.

Mavroeidis and Jøsang (2018) presented an ontological approach for automating threat hunting using system monitoring logs. The authors discussed the potential benefits of CTI in investigating attack events and anticipating the next attack steps. However in contrast to our approach, they did not present hypothesis generation at the level of adversarial techniques.

Homayoun et al. (2018) used sequential pattern mining techniques to identify features (i.e., activity logs) that are used for classification relying on J48, random forest, bagging, and multi-layer perceptron to detect and hunt ransomware. The authors showed that different types of ransomware can be identified by their frequent patterns and that this differentiation can be used to build CTI from log data. However, their work mainly focused on the classification task and does not consider the possibility of crafting high-level CTI.

Ranveer and Hiray (2015) presented an overview of methods that can be used in different stages of malware detection. They focused on the phase of feature extraction and performed a comparative analysis of feature extraction methods for malware detection.

While most prior research on threat hunting has focused on automated detection and response, the current work focuses on the hypothesis generation phase of the threat hunting process.

6.2.3 Hypothesis generation

6.2.3.1 Reasoning with operational and strategic CTI

Attack reconstruction, which is often the output of a successful threat hunting procedure, refers to describing a threat by presenting the different steps the attacker successfully executed. The security analyst should be able to explain how each step was achieved by pointing to the relevant events based on the evidence collected and its analysis (Strom et al., 2017).

6.2.3.1.1 Causal attack graphs Polatidis et al. (2018) proposed an approach for cyber attack (path) prediction using visualized attack graphs and recommender systems, They used naive Bayesian and random forest classifiers for attack prediction and reasoning with CTI. The victim organization structure and situational awareness are out of the scope of this paper. However, we employ Bayesian inference for the analysis of the CTI knowledge graph.

Milajerdi, Gjomemo, et al. (2019) employed causal provenance graphs to model the organization structure and processes alongside the attacker activities. Their objective was to infer the high-level TTPs from system logs. The inference of the causal relationships for the construction of the provenance graph from logs requires extensive threat emulation within the target environment. Such information is usually not provided by major CTI sources. In contrast, the TTP inference proposed in this article relies on readily available general-purpose CTI published by a variety of sources.

6.2.3.1.2 Attack detection AlEroud and Karabatis (2017) used domain knowledge to improve initial predictions and create an accurate attack profile. Attacks were represented as nodes on semantic link networks. First, the authors ranked predictions according to the networks, and then, with domain knowledge and a taxonomy, they adjusted their predictions according to the predictions' correlation to the taxonomy. Fard et al. (2020) presented a multi-view ensemble threat hunting model. For threat hunting they relied on weighted majority voting using sparse representation-based classifiers. Thereby, each classifier detects malware, using one feature set as a single modality of available CTI. In an experimental setting using three different data sets (e.g., extracted from the VirusTotal Threat Intelligence platform), they showed that the proposed approach delivers high accuracy with little computational burden. Bhatt et al. (2014) also suggested a threat detection model. The proposed model is used to improve hypothesizing on ongoing attacks given correlated events, and knowledge on the Cyber Kill Chain[16] (CKC; Hutchins et al., 2011). In contrast to works in this category, AHG is not intended for attack detection. Rather, we aim to infer the most probable attack techniques given a set of suspicious artifacts under the assumption that an organization is already under attack.

6.2.3.1.3 Inferring TTPs/CKC phases from artifacts Wafula and Wang (2019) suggested a threat hunting hypothesis development methodology for identifying the threat actor, target assets, relevant vulnerabilities, and artifacts. They suggested using exploratory data analysis of CTI for both the generation of an initial hypothesis and validation of the hypothesis. Giura and Wang (2012) proposed an attack pyramid that aims to capture the movements of an attacker through CKC phases (represented by the levels of the pyramid) and the organization's environments (represented by the planes of the pyramid), e.g., physical, network, user, and application. The attacker's goal is located at the top of the pyramid, and it is reachable by stepping from one event to another. The pyramid enables the detection of the attack path and the attacker's goal. Iqbal and Anwar (2016) built a unified graph representation for CKC and Pyramid of Pain models by extracting entities from textual reports. They compared two variants of the same attack and showed that they could map TTPs for each stage of the attack using the graph which would allow analysts relying on the methods presented by the authors to derive and predict missing TTPs from the graph. However, the approach presented does not provide automation and demonstrated the predictive power for only two malwares. Rubinshtein and Puzis (2016) built an attack ontology from logs and reconstructed

[16] CKC was developed by Martin Lockheed to model the diverse threat landscape. The CKC describes seven steps that an attacker must perform in order to accomplish the goal of an attack.

attack steps based on this ontology. In addition to generating actionable CTI from logs, Landauer et al. (2019), processed the CTI generated using advanced unsupervised machine learning methods to transform the anomalies detected into hypotheses about abstract attack patterns.

Taken together, these pioneering works form the foundation upon which we build our hypothesis generation approach. Most prior state-of-the-art methods demonstrate inference of TTPs using expert based rules on just a few attacks that exhibit a similar attack flow. We build AttackDB – a comprehensive knowledge graph constructed top down, relying on major CTI sources (MITRE ATT&CK, AlienVault, X-Force Exchange, and VirusTotal). AttackDB enables to streamline the process of TTP inference by relying on network based inference instead of manually defined sets of rules. In addition, as proposed by Rubinshtein and Puzis (2016), we add similarity scores and hypothesis generation algorithms to increase the attack reconstruction and threat hunting capabilities. In this paper we investigate a wider set of approaches for crafting robust hypotheses.

6.2.3.2 Reasoning using CTI knowledge graphs

In the field of cyber security, the use of knowledge graphs, which are systematics representing CTI with the help of directed labeled graphs, is in its early stages and is mainly used for visualization and less for prediction. One of the first studies on the use of knowledge graphs for analysis was performed by S. Lee et al. (2018) who built a knowledge graph from open-source intelligence. Based on established graph algorithms, they improved the identification of malicious nodes and attack infrastructures, as well as the relationship among attack groups and their similarity. In particular, they used page rank and betweenness metrics to detect relevant information in the graph. Gao et al. (2018) presented a trust evaluation mechanism to assess which information is relevant (malicious) and which is not. They trained supervised classification algorithms based on a random forest classifier to distinguish between trusted and untrusted information. They found that graph-based features increase the accuracy of the trained models. Link prediction techniques have shown to provide valuable insights from CTI-based knowledge graphs (Xiao et al., 2019; Garrido et al., 2021). Inspired by these works, we employ supervised machine learning and link prediction methods for the inference of TTPs from artifacts through knowledge graph analysis.

Najafi et al. (2019) proposed a novel graph-based inference algorithm to evaluate the maliciousness of IoCs. The proposed algorithm outperformed established prior state-of-the-art algorithms, such as belief propagation and SimRank, and showed that the proposed algorithm was particularly effective in identifying previously unknown IoCs. Milajerdi, Eshete, et al. (2019) modeled threat hunting as an inexact graph pattern matching problem based on kernel audits, with relationships between CTI used as reliable artifacts. However, the query graphs are manually constructed by the researchers for each instance of the attack.

S. Qamar et al. (2017) and Riesco and Villagrá (2019) proposed data-driven analytics based on the STIX ontology. They used logic-based deductive inference rules (defined in Semantic Web Rule Language (Horrocks et al., 2004)) and defined queries for evaluating threat likelihood and managing cyber threat response activities. Although logic rules may detect complex patterns in CTI, a domain expert is required to define the rules, which is the main drawback of this method.

Ulicny et al. (2014) highlighted the need for inferences to be automated to cope with the high dynamic developments in the field of cyber security and motivate the automation of threat hunting relying on CTI. The authors showed the possibility of automated threat detection based on Web Ontology Language (Antoniou and Van Harmelen, 2004). They thereby introduced an approach which mimics the work of a human analyst. We aim at contributing to both the automated inference of IoAs and the refinement of IoAs inferred by a human analyst. In contrast, to graph alignment and logical rules mentioned above, the methods proposed in this paper allow full automation of TTP inference. We rely on CTI sources that include many attack variants and are readily available to security teams. The threat intelligence knowledge base, AttackDB, that we provide (Dekel et al., 2021) in this paper is unique, since it is (1) comprehensive – comprising all malware families from MITRE ATT&CK and (2) multi-level – connecting high-level TTPs with low-level observable artifacts.

6.3 Multi-level threat knowledge base

6.3.1 Schema

In this section we describe AttackDB - a multi-level threat knowledge base. AttackDB contains SDOs at all levels of the Pyramid of Pain (Bianco, 2013), from abstract concepts, such as tactics and top-level techniques, down to IoCs and specific observables, such as hashes, Internet protocol (IP) addresses, and domain names.

Figure 6.1 depicts AttackDB's schematic structure. We utilize the definition of SDOs and relationships between them with a few exceptions; for example, we do not take advantage of all SDOs defined in STIX (e.g., location). We also use a single concept of attack instead of the intrusion set, campaign, and malware defined in STIX.

The top-level SDOs in AttackDB are attack patterns (a.k.a. tactics and techniques). Tactics are the most abstract representations of attacks in AttackDB and represent tactical goals of attackers. Techniques denote the actions attackers take to achieve the tactical goal. SDOs include the malicious activities exhibited by malware, campaigns, or intrusion sets. Malware is software that exhibits a set of malicious activities. Malware can be a part of multiple campaigns. A campaign is a set of malicious activities performed for a specific period of time against specific targets. Campaigns that are believed to be orchestrated by the same threat actor may be grouped into intrusion sets. Despite the semantic differences between them, malware, campaign, and intrusion set SDOs can be used to represent an abstract attack that is being hunted. A tool is software which can be used by an adversary.

AttackDB also contains observed data SDOs associated with an attack on the one hand and with a CTI report on the other. Report SDOs are included in AttackDB to trace back the CTI to its source (see section 6.3.3) but are not used for hypothesis generation. Observed data SDOs may aggregate hashes, IP addresses, domains, network, and host artifacts (i.e., telemetry), such as process names, services, registry keys, and other artifacts. Artifacts may be grouped together in a pattern and

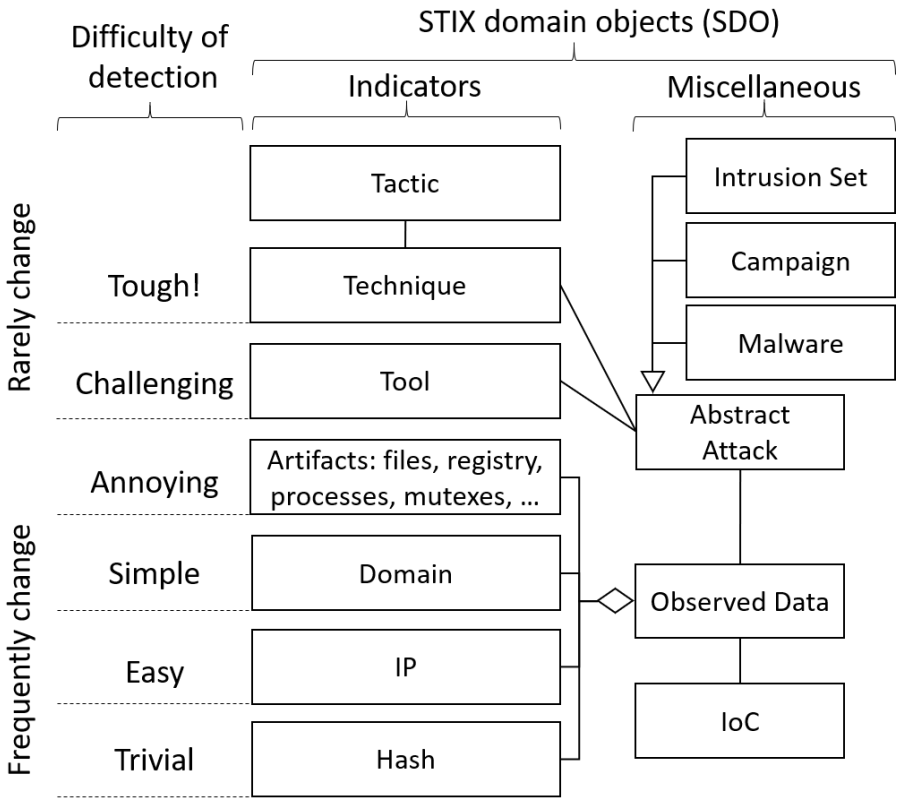

Figure 6.1: AttackDB schema with detection difficulties according to the Pyramid of Pain.

tagged as an IoC. IoCs can be used to identify attacks observed in the past but are usually easily modified by the attacker. All of the observed data stored in AttackDB is processed and used to create attack hypotheses, rank the hypotheses according to their probability, and generate workflows for proper response to (and in this sense defending against) the hypothesized attack.

6.3.2 Data fusion

In the current implementation, the AttackDB knowledge graph is stored in a Neo4j[17] database, however any graph database may be used for this purpose. With AttackDB, we constructed a rich knowledge base that consists of CTI from the MITRE ATT&CK Enterprise knowledge base, the OTX, the X-Force, and VirusTotal. Relationships between different objects are included to build the knowledge graph, as they are described in malware analysis reports extracted from the different CTI sources. Details regarding the data extraction process and data fusion are provided below. The process of constructing AttackDB is presented in figure 6.2.

MITRE's ATT&CK is an open CTI knowledge base that contains information on adversarial techniques and tactics, threat actors, mitigation, malware, and tools (Strom et al., 2017). First, we populate AttackDB with malware and techniques and the relationships between them, extracted from MITRE ATT&CK.

[17] `https://neo4j.com/`

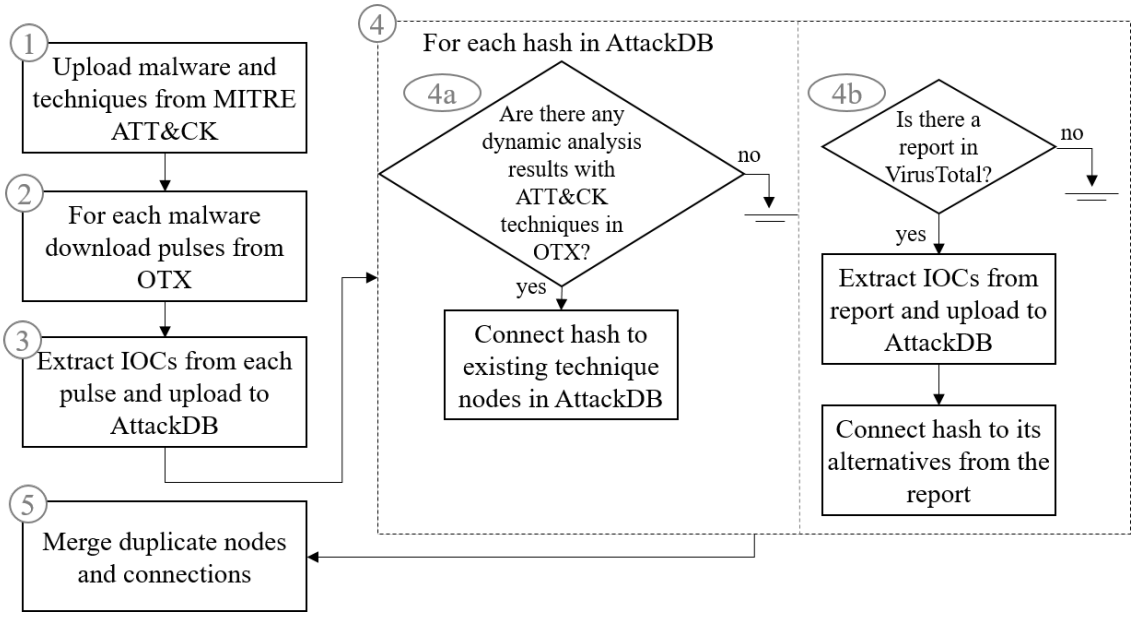

Figure 6.2: Flow chart of AttackDB's construction.

OTX provides CTI in the form of pulses, which contain one or more IoCs, such as file hashes, Uniform Resource Locators (URLs), and IPs. Pulses can be tagged with malware names, threat actors, and additional information. In the second step, we search for pulses, using malware names from MITRE ATT&CK, via the OTX Application Programming Interface and link malware nodes in AttackDB to IoCs from the respective pulses. However, AlienVault allows anyone to post a set of IoCs as a pulse, which may lead to unreliable data. Therefore, we only use pulses posted by the top 20 publishers with the most subscriptions who posted pulses related to the malware searched for: AlienVault, MalwarePatrol, jnazario, niddel, Metadefender, cyberprotect, popularmalware, Malwaremustdie, Cyber_Hat, burberry, bartblaze, ESET-Spain, julsec, zer0daydan, rpsanch, erik, milind, BLUELIV, techhelplist, BotnetExposer, and nightingale.

IBM X-Force provides malware reports, which contain IoCs of various types, such as URLs, domain names, filenames, and processes. As with AlienVault, we search for reports using malware names and link malware nodes in AttackDB to IoCs from the respective reports. We use file hashes retrieved from AlienVault and X-Force to fuse the pulses with VirusTotal reports.

In the third step, we enrich AttackDB with malware telemetry, such as network and host artifacts. For this purpose, we retrieve behavioral analysis data from *VirusTotal* for all hashes obtained from OTX and X-Force. The behavioral data retrieved from VirusTotal includes file names (opened, created, searched, etc.), URLs, domains, IPs, process names, registry keys, mutual exclusions (mutexes), emails, and more. Note that during the population of AttackDB, $IoCs$ with identical patterns should be represented by the same node, as well as observables with identical values.

Figure 6.3 provides an illustration of AttackDB's structure containing two attack nodes (A1 and A2) and all of the relevant connections. At the top, we see malware and associated techniques extracted from MITRE ATT&CK. The three respective IoCs (two hash values and one URL) are extracted from OTX and X-Force. Following the STIX format, the relevant observed data nodes

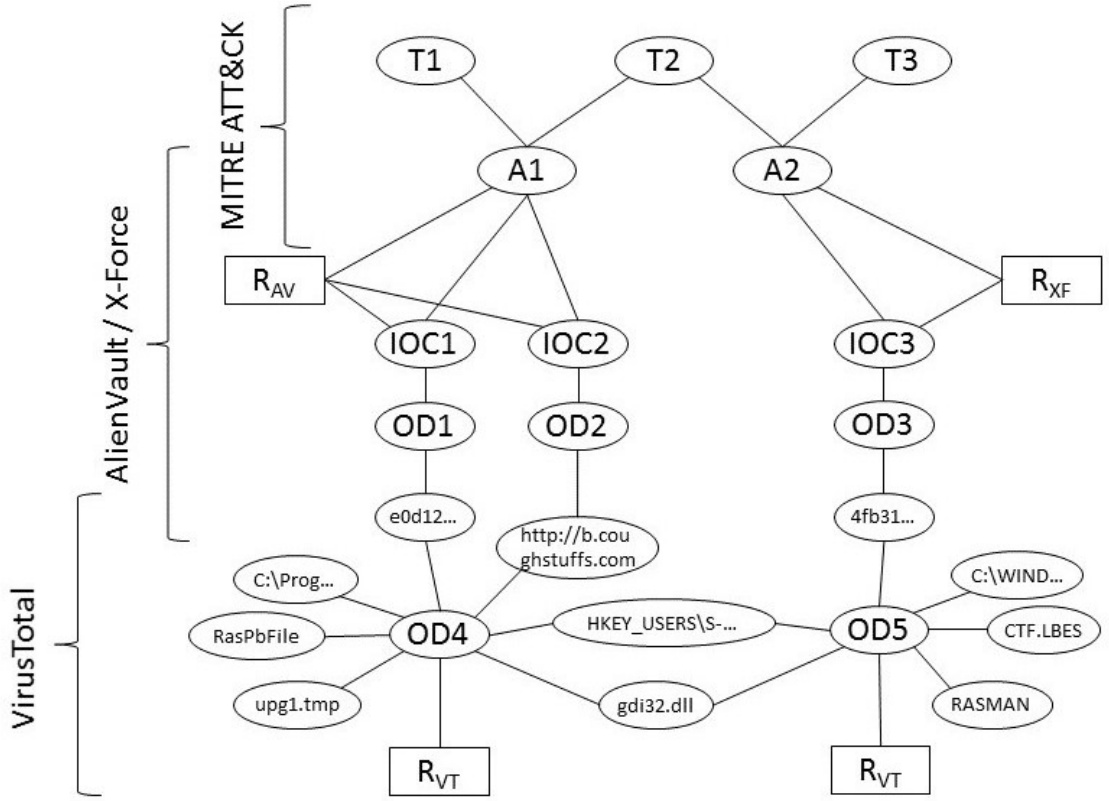

Figure 6.3: An illustration of AttackDB's structure.

are connected with IoCs connected with the relevant attack (i.e. malware). Finally, the behavioral data extracted from VirusTotal is displayed at the bottom of the figure. This observed data is not connected with IoC nodes, because it is not necessarily a strong indication of the attack but is merely a collection of artifacts generated by the malware during dynamic analysis.

Note that a behavioral artifact may be connected to malware through multiple paths. This happens when there are multiple instances of the same malware analyzed by VirusTotal. Also note that observables are indirectly connected to techniques through the respective malware. We use this connection to build attack hypotheses, as described in section 6.4.3.

6.3.3 Duplicate data and malware aliases

In the resulting AttackDB there are hash nodes that are connected to two or more malware nodes. Since hash nodes represent specific malware instances, ambiguous connections to malware nodes require additional clarification. The possible reasons for this phenomenon are described below.

We analyze all occurrences of hash nodes shared by attack nodes. For each shared hash we analyze the relevant malware analysis reports in order to categorize the relationships between the attack nodes sharing the same hashes. The nature of these relationships is diverse: ambiguous naming (aliases), shared actor, shared campaign, belonging to the same malware family, shared infection mechanism, errors in the reports' parsing process (i.e., extraction error), etc..

Roughly, we divide reports into two types: (1) reports that describe some factor, such as actors, campaigns, families, infection mechanisms, or IoCs overlaps, common to multiple malware instances; and (2) reports that describe a single specific malware. Reports of the first type often contain references to multiple malware binaries. In such cases, an attack node is connected not only to its representative hash but to all the hashes provided in malware analysis reports. Reports of the second type describe a single malware containing a single representative hash, however there may be different reports describing the same malware with different names. In such cases a single hash is connected to all the malware aliases that appeared in the reports.

Ambiguous connections whose origin is in both types of reports provide meaningful information, and therefore, we include them in AttackDB. Connections caused by errors in the process of extracting data from the malware analysis reports have been removed.

We analyze the reports connected to hash nodes with ambiguous connections, searching for connections erroneously omitted and connections caused by extraction errors. This manual analysis resulted in the connection of 74 reports to 93 malwares and disconnecting 54 reports from 35 malwares. Connecting a report to a malware means connecting the malware to all observables, including IoCs, that appear in the report, while disconnecting a report from a malware means disconnecting the malware from all IoCs that appear in the report.

Example 1 (Adding connections) *The XTunnel malware appears in a pulse in OTX and therefore is connected to the corresponding report in the graph. There are also IoCs from the CORESHELL and USBStealer malwares in the same report, because XTunnel, CORESHELL, and USBStealer belong to the APT28 group. However, during the data fusion process, when searching for CORESHELL and USBStealer, the report did not come up, resulting in missed connections.*

Example 2 (Removing connection) *We removed connections between the RTM malware and reports that contain the word "depa**rtm**ent.".*

6.3.4 Knowledge base summary

The resulting AttackDB fuses data from 1,675 AlienVault pulses, 281 IBM X-Force reports, and 53,005 VirusTotal reports. It contains 253 malware nodes associated with 144,216 IoCs. Around 60,000 of the IoCs are file hashes, and the rest are domain names, IPs, etc. In total there are over half a million observables in AttackDB. All of the techniques in the graph (190) are connected to some malware. The average number of techniques per malware is 10.3. We only include techniques that are associated with a malware by MITRE within this study.

6.4 Attack hypothesis generation

6.4.1 High-level overview

Assume that suspicious events are taking place in an organization. The goal of AHG is to propose an hypothesis on the course of the possible attack. Figure 6.4 provides an overview of the hypothesis generation process. The resulting hypothesis consists of a set of MITRE ATT&CK techniques that are closely related to: (1) the observed data, and (2) each other.

A hypothesis that consists of ATT&CK techniques that are closely related to the observed data can be obtained by an analyst who investigates the suspicious events and formulates an initial hypothesis based on alerts and various artifacts recorded by the organization's SIEM (see figure 6.4 (1)). In addition, an initial hypothesis can be inferred by ranking adversarial techniques based on currently observed data (COD) stored in the system (see figure 6.4 (2)). In section 6.4.3, we describe two approaches for inferring an initial hypothesis: inferring techniques related to IoCs in AttackDB and inferring techniques related to telemetries in AttackDB. These two approaches are implemented by using three methods (initial hypothesis generation (ih) algorithms) that produce a relationship that maps the observed data to techniques: A method inspired by term frequency inverse-document-frequency ($TFIDF$) (see section 6.4.3.1) a multi-nomial naïve Bayesian classifiers (NB-C) (see section 6.4.3.2), and a multi-nomial multi-layer naïve Bayesian classifiers ($MLNB$-C) (see section 6.4.3.3).

A hypothesis that consists of ATT&CK techniques that are closely related to each other can be crafted by refining the set of techniques (relying on hypothesis refinement (rh) algorithms) comprising the initial hypothesis using a recommender system-based technique (see figure 6.4 (3)). Section 6.4.4 describes five recommender system techniques for refining the initial hypothesis: the projected technique (ProjT, section 6.4.4.1), link prediction on projected technique (LPProjT, section 6.4.4.2), link prediction on projected attack (LPProjA, section 6.4.4.3), projected hypothesis (ProjAT, section 6.4.4.4), and supervised link prediction (SupLP, section 6.4.4.5).

To make it easier for the analyst, all methods used to produce the initial and refined hypotheses aim to accurately rank the ATT&CK techniques so that the most probable techniques appear first. In section 6.4.5, we propose a method for estimating the number of related techniques. The refinement mechanism selected is applied based on the results of the method for estimating the number of related techniques.

6.4.2 Problem definition

Assume a knowledge graph as described in section 6.3.2. Malware SDOs that are used to represent an attack are grouped in a super-set denoted as $\mathbb{A}$. An *attack descriptive* SDO contains information about an element that has been used or targeted by the attack. SDOs that can be used to describe an attack, specifically techniques (a.k.a. attack patterns), IoCs, observed data and specific observables, are grouped in a super-set denoted as $\mathbb{D}$.

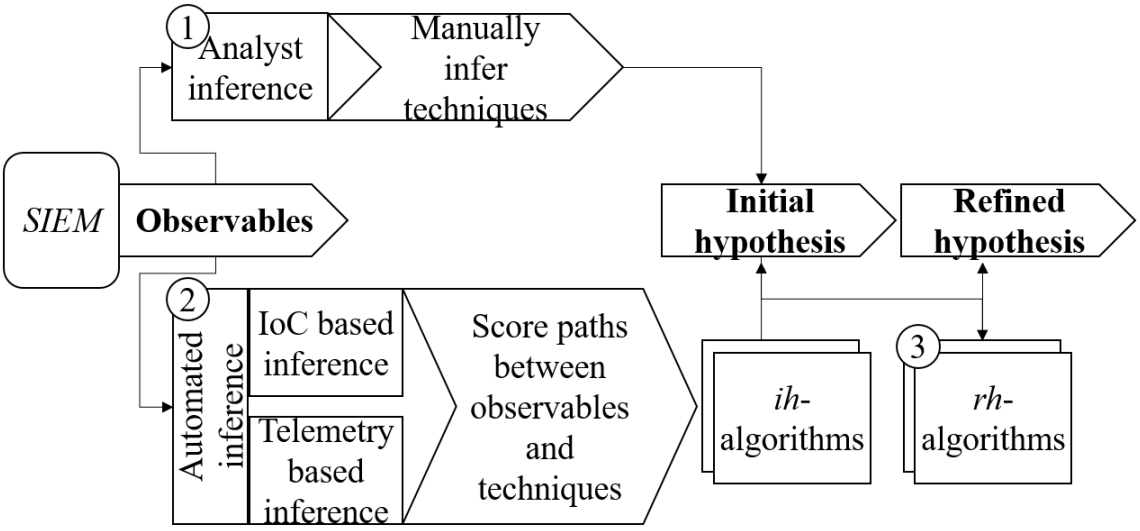

Figure 6.4: Overview of the hypothesis generation process.

Definition 1 (Cyber security knowledge graph) *A cyber security knowledge graph* $KG = \langle \mathbb{A}, \mathbb{D}, R \rangle$ *is a graph where* $\mathbb{A}$ *contains the nodes representing past attacks, and* $\mathbb{D}$ *contains the description nodes (specifically, techniques,* $IoCs$*, observed data, and observables).* R *is the set of directed links connecting related SDOs according the schema depicted in figure 6.1.*

Definition 2 (Attack descriptions) *Given an attack representation* $a \in \mathbb{A}$*, we refer to the set of* $v \in \mathbb{D}$ *that are at most five hops away from* a *as the* attack description $AD_a = \{v | dist(a, v) \leq 5\} \subseteq \mathbb{D}$*. We refer to all* $v \in AD_a$*, where* v *is a technique, as the* attack techniques $AT_a = \{v | v \in AD_a \wedge type(v) = Technique\}$.

Assume an unknown ongoing attack a_{new} currently being investigated by the analyst. If an analyst constructs the initial hypothesis, then he/she adds a_{new} to the KG and begins a preliminary investigation by concentrating on the recent alerts and related telemetry.

Problem 1 (Initial hypothesis generation problem) *Given a knowledge graph* KG *and* COD *in the SIEM, generate an initial hypothesis of an attack, denoted as* $AT^{init}_{a_{new}}$*, consisting of techniques closely related to* COD*.*

According to a preliminary investigation, the analyst connects a_{new} to technique SDOs, denoted as $AT^{init}_{a_{new}}$. $AT^{init}_{a_{new}}$ may also be referred to as an attack hypothesis consisting of a set of techniques.

The initial hypothesis can also be inferred automatically by ranking techniques based on COD, as will be elaborated on in section 6.4.3.

AHG consists of a module that refines $AT^{init}_{a_{new}}$ by recommending more relevant techniques and ignoring or omitting techniques that are not relevant.

Problem 2 (Hypothesis refinement problem) *Given an initial hypothesis $AT^{init}_{a_{new}}$ and a knowledge graph KG, generate a new hypothesis, denoted as $AT^{ref}_{a_{new}}$, that is more accurate than $AT^{init}_{a_{new}}$ with respect to the correct description of the real attack $AT^{*}_{a_{new}}$.*

Table 6.1 lists the notations and abbreviations used in this paper.

6.4.3 Initial hypothesis generation

$AT^{init}_{a_{new}}$ is constructed by selecting the techniques most relevant to COD. For this purpose we employ different methods that produce relationships mapping COD to techniques. The initial hypothesis generation is either based on a simulated human analyst (a strategy denoted as H) or automated, relying on one of the ih algorithms presented below. Note that COD may be either IoCs or telemetries, depending on the inferring technique approach employed.

6.4.3.1 Term frequency-inverse document frequency

We propose the use of a technique scoring mechanism based on $TFIDF$ common in information retrieval. In this case, techniques are analogous to documents, and observables are analogous to search terms for the purpose of $TFIDF$ computation. A technique t is relevant to an observable obs if obs appears in a report on attack a that uses t.

An observable may appear in several reports and be connected to a technique through multiple attacks.

Definition 3 (Connected observables to a technique) *$TF(obs, t)$ is the number of paths from obs to t in the knowledge base.*

In the discussions that follow, we use $x : Y$ notation to indicate $type(x) = X$, and we use dot $x : X - y : Y$ to indicate that x and y are connected in KG. When a technique's scoring is based on IoCs (obs is an IoC), then the set of paths between obs and t is defined as follows:

$$TF(obs, t) = |\{obs - (od : OD) - (i : IoC) - (a : A) - t\}|$$

When a technique's scoring is based on telemetries (obs is a telemetry, $type(obs) = OD \wedge type(obs) \neq IoC$), then the set of paths between obs and t is defined as follows:

$$TF(obs, t) = |\{obs - (od : OD) - (h : hash) - (od : OD) - (i : IoC) - (a : A) - t)\}|$$

Note that the hash is a subset of IoC ($hash \subseteq IoC \subseteq OD \subseteq \mathbb{D}$). For example, in figure 6.3 there are three 6-hop paths connecting the *gdi32.dll* observable at the bottom of the figure (in the middle) with the technique T2.

Algorithm 1 CountPaths

Input: $observables, AttackDB, ih_{strategy}$
Output: $num\_techs\_per\_obs, related\_techs$

$num\_techs\_per\_obs \leftarrow \{\}$
$related\_techs \leftarrow Set([])$
for $obs \in observables$ **do**
  $num\_techs\_per\_obs[obs] \leftarrow \{\}$
  if $ih_{strategy} == IoC$ **then**
    $techs_{obs} \leftarrow GetTechsConnectedToIoC(obs)$
  else
    $techs_{obs} \leftarrow GetTechsConnectedToTel(obs)$
  for $tech \in techs_{obs}$ **do**
    add $tech$ to $related\_techs$
    $num\_techs\_per\_obs[obs][tech] \leftarrow CountPaths(obs, tech)$
return $num\_techs\_per\_obs, related\_techs$

There are several different versions of inverse document fequency (IDF) assessment available in the literature. We use the simplest one which is the logarithm of the total number of techniques divided by the number of relevant techniques in AttackDB.

Definition 4 (Term frequency-inverse document frequency of techniques) *Let $\mathbb{T} = \{t\}$ be the set of techniques in AttackDB. Let be T a subset of $\mathbb{T}$ denoting the relevant t given obs; then IDF can be defined as the following*

$$IDF(obs, t) = Log_e\left(\frac{|\mathbb{T}|}{|T|}\right) \tag{Exp. 6.1}$$

The score of t is the sum of the $TFIDF(obs, t)$ values for all COD.

$$TFIDF(obs, t) = TF(obs, t) \cdot IDF(obs, t) \tag{Exp. 6.2}$$

$$TFIDF(t) = \sum_{obs \in COD} TFIDF(obs, t) \tag{Exp. 6.3}$$

6.4.3.2 Naïve Bayesian inference

We implement a multi-nomial NB-C. Multi-nomial NB-Cs can be based on word vector counts as well as on $TFIDF$ (Rennie et al., 2003). Here, we present a NB-C based on vector counts. Therefore, the prior probabilities are extracted from the KG referring to the number of relevant observables. Hence, a priori probabilities represent prior knowledge that can be extracted from the KG.

Definition 5 (Prior probabilities) *We assume the prior probability of the utilization of a technique $P(t)$ to be a priori equal for each t. Consequently, $P(t)$ can be defined as follows:*

$$P(t) = \frac{1}{|\mathbb{T}|} \tag{Exp. 6.4}$$

We furthermore assume, $P(a)$ the prior probability of the utilization of an attack to be equal for each a.

$$P(a) = \frac{1}{|\mathbb{A}|} \tag{Exp. 6.5}$$

We assume the occurrence of an obs to be deterministic in the occurrence of the related attack. Let A denote all relevant (related to the specific obs) attacks. Then, the prior probability of the observation of a specific observable $P(obs)$ can be defined as follows:

$$P(obs) = \sum_{a \in A} P(a) \tag{Exp. 6.6}$$

Based on these prior probabilities, the posterior probability that attack technique t is relevant given COD is defined by equation 6.7:

$$P(t|COD) = \frac{P(t) \cdot \prod_{obs \in COD} P(obs|t)}{P(COD)} \tag{Exp. 6.7}$$

$$P(obs|t) = \frac{TF(obs,t) + \alpha}{\sum_{t \in T} TF(obs,t) + \alpha} \tag{Exp. 6.8}$$

$$P(COD) = \prod_{obs \in COD} P(obs) \tag{Exp. 6.9}$$

where α is a smoothing prior. This smoothing prior can be used for Laplace smoothing ($\alpha = 1$) or Lidstone smoothing ($\alpha < 1$). With the smoothing prior it is possible to account for obs and connections that are not present in AttackDB (e.g., unknown). Furthermore, it prevents probabilities of zero. Therefore, the use of this additive smoothing parameter may hence increase the accuracy of classification.

6.4.3.3 Multi-layer naïve Bayesian inference

Leveraging on the ontology of the KG, we use $MLNB\text{-}C$. This inference algorithm allows to use causal relationships presented in AttackDB and therefore step-wise reduces the assumption of conditional independence of $NB\text{-}C$. The proposed method consists of two tiers (see figure 6.5): (I) $NB\text{-}C$ for technique inference from malware; (II) $NB\text{-}C$ for malware inference from COD.

According to the law of total probability, the probability of a technique t given COD can be defined as described in equation 6.10.

Definition 6 (Connected attacks to a technique) *$TF(a,t)$ is the number of paths from a to t in KG. Equivalently, $TF(obs,a)$ describes the number of paths from obs to a in KG, where $TF(a,t)$ and $TF(obs,a)$ are calculated symmetrically to $TF(obs,t)$*

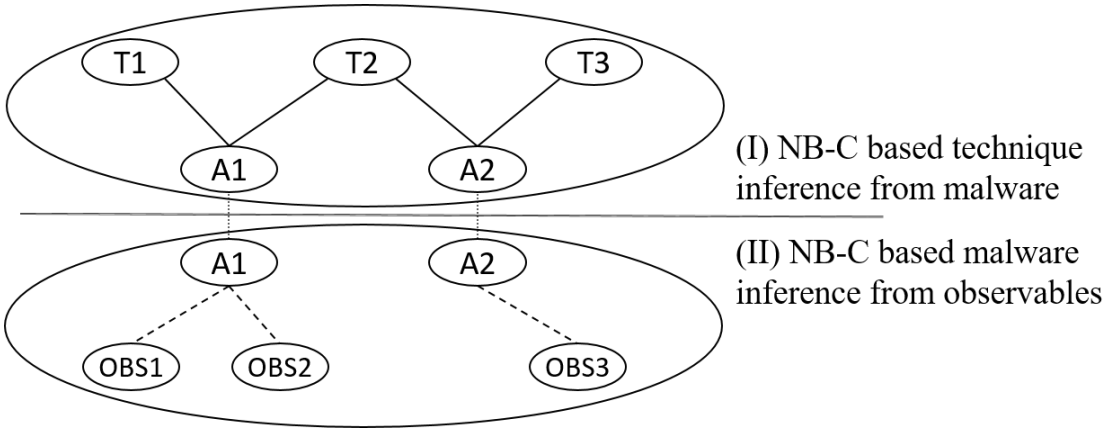

Figure 6.5: Framework of multi-layer naïve Bayesian inference of initial hypotheses within AttackDB.

$$P\left(t|COD\right) = \sum_{a \in \mathbb{A}} \frac{P\left(t\right) \cdot P\left(a|t\right) \cdot P\left(a|COD\right)}{P\left(a\right)} \tag{Exp. 6.10}$$

$$P\left(a|t\right) = \frac{TF(a,t) + \alpha}{\sum_{t \in T} TF(a,t) + \alpha} \tag{Exp. 6.11}$$

$$P\left(a|COD\right) = \frac{P\left(a\right) \cdot \prod_{obs \in COD} P\left(obs|a\right)}{P\left(COD\right)} \tag{Exp. 6.12}$$

$$P\left(obs|a\right) = \frac{TF(obs,a) + \alpha}{\sum_{a \in \mathbb{A}} TF(obs,a) + \alpha} \tag{Exp. 6.13}$$

6.4.4 Hypothesis refinement

Next, we refine $AT^{init}_{a_{new}}$, relying on rh algorithms. This stage accounts for interdependence between techniques by increasing the score of techniques which are often used together in the same attacks.

For hypothesis refinement, we rely on a set of link prediction (LP) techniques and similarity metrics, namely Jaccard's/Tanimoto coefficient (Tanimoto, 1958; Jaccard, 1912), Adamic Adar(Adamic and Adar, 2003), Friends measure/ Katz measure (Katz, 1953), and Preferential Attachment (Barabási and Albert, 1999). Here we define these algorithms, applying them to a, although the algorithms can analogously be used for AT and t.

Definition 7 (Jaccard's/Tanimoto coefficient) *The Jaccard's/Tanimoto coefficient $J_{\mathbb{A}}$ is a similarity metric describing the probability that a_i and a_j share the same AT for an AT that either a_i or a_j has. It is defined as follows.*

$$J_{\mathbb{A}}(a_i, a_j) = \frac{\left|AT_{a_i} \cap AT_{a_j}\right|}{\left|AT_{a_i} \cup AT_{a_j}\right|} \tag{Exp. 6.14}$$

Definition 8 (Adamic Adar) *Adamic Adar $AA_{\mathbb{A}}$ describes a frequency-weighted metric formalizing the notion that rarely observed techniques of a are more informative than frequently observed techniques. $AA_{\mathbb{A}}$ is defined as follows.*

$$AA_{\mathbb{A}}(a_i, a_j) = \sum_{t \in AT_{a_i} \cap AT_{a_j}} \frac{1}{log\,|AT(t)|} \tag{Exp. 6.15}$$

Definition 9 (Katz measure) *Katz measure $K_{\mathbb{A}}$ is a heuristic defined by the sum of the path length of connections. It is assumed that the shorter the paths connecting a_i and a_j are and the more paths that exist, the more similar the attacks. Therefore, it can be described as a variant of the shortest-path measure*

$$K_{\mathbb{A}}(a_i, a_j) = \sum_{l} \beta^l \cdot TF^l(a_i, a_j) \tag{Exp. 6.16}$$

where β is a scaling parameter, and $TF^l(a_i, a_j)$ is the number of paths from a_i to a_j of length l.

For the undirected KG, the Katz measure is equal to the Friends measure, where $\beta = 1$ and $l_{max} = 2$.

Definition 10 (Preferential Attachment) *Preferential Attachment $PA_{\mathbb{A}}$ calculates the similarity without relying on relational information. $PA_{\mathbb{A}}$ can be computed as follows.*

$$PA_{\mathbb{A}}(a_i, a_j) = |AT_{a_i}| \cdot \left|AT_{a_j}\right| \tag{Exp. 6.17}$$

Given a snapshot of KG, LP techniques identify missing relationships within the knowledge graph that are likely to occur. LP techniques can also be used to identify relationships that are likely to be incorrect within the KG. We apply five techniques to predict the relationships between the SDO representing the new attack a_{new} and the relevant technique SDOs. We rank each technique t and select the top N to serve as $AT^{ref}_{a_{new}}$.

As described in previous sections, KG is a graph linking attacks to relevant techniques. Following the common practice in recommender systems (Sarwar et al., 2001), in some of the following approaches we rely on the one-mode projection of the KG on either the set of attacks ($\mathbb{A}$) or the set of techniques ($\mathbb{T} \subseteq \mathbb{D}$).

Definition 11 (Attack similarity graph $KG_{\mathbb{A}}$) *$KG_{\mathbb{A}} = \langle \mathbb{A}, E_{\mathbb{A}}, J_{\mathbb{A}} \rangle$ is a unipartite weighted graph, where $(a_i, a_j) \in E_{\mathbb{A}}$ if $AT_{a_i} \cap AT_{a_j} \neq \emptyset$, and the edge weight $J_{\mathbb{A}}(a_i, a_j)$ is the Jaccard coefficient of the respective attack descriptions.*

We define the weighted techniques similarity graph ($KG_{\mathbb{T}} = \langle \mathbb{T}, E_{\mathbb{T}}, J_{\mathbb{T}} \rangle$) symmetrically as the one-mode projection of KG on $\mathbb{T}$.

6.4.4.1 Projected techniques

We begin with the simplest approach that aggregates the link weights between the hypothesis provided by the analyst or the initial hypothesis generation and each of the techniques $t \in \mathbb{T}$. The relevance score $ProjT$ of the techniques is calculated as follows:

$$ProjT(a_{new}, t) = \sum_{t' \in AT^{init}_{anew}} J_{\mathbb{T}}(t', t) \quad \text{(Exp. 6.18)}$$

6.4.4.2 Link prediction on projected techniques

Although $ProjT$ takes into account the similarity between techniques in terms of the attacks associated with them, the topology of KG_T is not considered. We utilize several common link prediction measures to improve $ProjT$ by taking the neighborhoods of the techniques suggested by the analyst into account. Given $KG_{\mathbb{T}}$, we apply known LP measures such as $J_{\mathbb{A}}, AA_{\mathbb{A}}, K_{\mathbb{A}}$, and $PA_{\mathbb{A}}$(Fire et al., 2014) on $KG_{\mathbb{T}}$.

The likelihood of each technique $t \in T$ is calculated as follows:

$$LPProjT(a_{new}, t) = \sum_{t' \in AT^{init}_{anew}} lp^{KG_T}(t', t) \quad \text{(Exp. 6.19)}$$

6.4.4.3 Link prediction on projected attack (LPProjA)

Equation 6.20 is used to rank each attack $a \in \mathbb{A}$ and select the top k attacks, denoted as $\mathbb{A}_{top}$.

$$score_a = lp^{KG_{\mathbb{A}}}(a_i, a_{new}) \quad \text{(Exp. 6.20)}$$

Similar to Equation 6.19, we use a range of LP measures to find the attack's most similar a_{new}. The score of each technique $t \in \mathbb{T}$ is calculated as:

$$LPProjA(a_{new}, t) = \sum_{a \in \mathbb{A}_{top}: a \neq a_{new} \wedge t \in AT_a} score_a \quad \text{(Exp. 6.21)}$$

In section 6.5.2, we present the results obtained with the $J_{\mathbb{A}}$ as the LP measure, since in preliminary experiments this method obtained the best results. Let $\Gamma_a = \{a' : (a, a') \in E_{\mathbb{A}}\}$ denote the neighbors of an attack a in the attack similarity graph $KG_{\mathbb{A}}$. The hypothesis based on $J_{\mathbb{A}}$ is:

$$LPProjA(a_{new}, t) = \sum_{a \in \mathbb{A}_{top}: a \neq a_{new} \wedge t \in AT_a} \frac{|\Gamma_{a_{new}} \cap \Gamma_a|}{|\Gamma_{a_{new}} \cup \Gamma_a|} \quad \text{(Exp. 6.22)}$$

6.4.4.4 Projected attack techniques (ProjAT)

Intuitively, the more an arbitrary attack a is similar to a_{new}, the higher the chance that a_{new} uses the same AT as a. In this approach, we rank the techniques $t \in \mathbb{T}$ according to $J_{\mathbb{A}}$ between $AT^{init}_{a_{new}}$ and all AT_a that include t ($\{AT_a | a \in \mathbb{A} \wedge t \in AT_a\}$). Equation 6.23 is used to calculate the likelihood score for each technique $t \in \mathbb{T}$:

$$ProjAT(a_{new}, t) = \sum_{a \in \mathbb{A}_{top}: a \neq a_{new} \wedge t \in AT_a} J_{\mathbb{A}}(AT_a, AT_{a_{new}}) \qquad \text{(Exp. 6.23)}$$

where $a \in \mathbb{A}_{top}$ are the top k attacks that obtain the highest score when applying $J_{\mathbb{A}}(a, a_{new})$.

6.4.4.5 Supervised link prediction (SupLP)

Here, we describe how to formulate LP as a supervised learning problem and apply a random forest classifier to predict the probability of a link in KG. First, we build a training set that consists of pairs of nodes $\{(u, v) | u \in \mathbb{A} \wedge v \in \mathbb{T}\}$. Then, we extract a variety of features for each pair following the methodology presented by Fire et al. (Fire et al., 2014):

(i) the nodes' topological attributes (e.g., degree, page rank, hubs, and authorities),

(ii) neighborhood-based metrics (e.g., $J_{\mathbb{A}}$, $PA_{\mathbb{A}}$), and

(iii) distance-based measures (e.g., shortest path length, number of shortest paths).

Next, given the initial hypothesis $AT^{init}_{a_{new}}$ we want to estimate the likelihood of a link existing between a_{new} and $t_j \in \mathbb{T}$ ($1 \leq j \leq |\mathbb{T}|$) in KG. As in the model's training phase, we (1) extract the topological attributes and distance-based metrics of a_{new} and t_j from KG, and (2) use the projections $KG_{\mathbb{A}}$ and $KG_{\mathbb{T}}$ to obtain neighborhood-based metrics for a_{new} and t_j, respectively. Finally, we feed the extracted features for all (a_{new}, t_j) tuples ($1 \leq j \leq |\mathbb{T}|$) to the trained model. The probability of the positive class (link exists) is used to rank the techniques.

On the one hand, building the SupLP model requires a large number of examples in the learning phase, while on the other hand, SupLP makes it possible to combine several features and be aware of wider contexts than the other four algorithms, which only take into account one feature.

6.4.5 The expected number of techniques and adaptive hypothesis refinement

Preliminary experiments on the initial inference of attack hypotheses and their refinement show that in cases where the real attack uses a large number of techniques the refinement may reduce the quality of AT^{init} (figures 6.8 and 6.10). We therefore introduce an adaptive refinement procedure. A dynamic threshold which states when to refine and when to rely on the AT^{init} is established based on the expected number of techniques related to the investigated attack.

Similar to the ih algorithms, we estimate the probability of each attack being the investigated attack. For $NB\text{-}C$ and $MLNB\text{-}C$ the probability of an attack is given by $P(a|COD)$, as specified in equation 6.12. Given $P(a|COD)$ and the number of techniques related to a specific attack $AT(a)$, the number of expected techniques related to the investigated attack (denoted as $|AT'|$) can be calculated for the naive Bayesian methods as follows:

$$|AT'_{NB\text{-}C}| = |AT'_{MLNB\text{-}C}| = \sum_{a \in \mathbb{A}} P(a|COD) \cdot |AT(a)| \quad \text{(Exp. 6.24)}$$

The decision on the refinement of $TFIDF$ based AT^{init} is based on $TFIDF(a)$, which is similar to $TFIDF(t)$ and is defined as follows:

Definition 12 (Term frequency inverse document frequency of attacks) *Let $\mathbb{A}$ be the set of attacks in AttackDB; let $A \subseteq \mathbb{A}$ be the subset of relevant attacks having a path to a given obs; and let $a \in A$ be some relevant attack. Then IDF can be defined as follows:*

$$IDF(obs, a) = Log_e \left(\frac{|\mathbb{A}|}{|A|} \right) \quad \text{(Exp. 6.25)}$$

The score of a is the sum of the $TFIDF(obs, a)$ values for all COD.

$$TFIDF(obs, a) = TF(obs, a) \cdot IDF(obs, a) \quad \text{(Exp. 6.26)}$$

$$TFIDF(a) = \sum_{obs \in COD} TFIDF(obs, a) \quad \text{(Exp. 6.27)}$$

where $TF(obs, a)$ is as in Definition 6.

$|AT'|$ can then be calculated for the $TFIDF$ method as the product of the probability for a and the number of related techniques $AT(a)$:

$$|AT'_{TFIDF}| = \sum_{a \in \mathbb{A}} \frac{TFIDF(a)}{\sum_{a \in \mathbb{A}} TFIDF(a)} \cdot |AT(a)| \quad \text{(Exp. 6.28)}$$

The decision whether to refine an initial hypothesis or not depends on the expected number of techniques $|AT'|$ and a configurable adaptive refinement threshold ($arth$). The refinement is only performed when $AT' \geq arth$.

6.5 Evaluation

6.5.1 Experimental setup

For every attack $a \in \mathbb{A}$, we use AT^{init}, AT^{ref}, and AT^{*} to respectively denote the initial hypothesis, the refined hypothesis, and the ground truth of the utilized techniques.

We consider three strategies for constructing an initial hypothesis: (1) by a simulated human analyst H, (2) by automatic inference from IoCs, and (3) by automatic inference from telemetry data Tel. The initial hypotheses constructed according to these strategies are denoted as AT_{H}^{init}, AT_{IoC}^{init}, and AT_{Tel}^{init} respectively.

For H based AT^{init}, we assume an analyst attempting to make the right decisions regarding the investigated attack. However, there may be errors in the analyst's hypothesis regarding the attack due to a lack of knowledge or insufficient forensic evidence. To challenge the robustness of the hypotheses refinement algorithms we simulate two types of errors: false positives – selection of unrelated techniques ($AT_{H}^{init} \setminus AT^{*}$) and false negatives – omission of related techniques ($AT^{*} \setminus AT_{H}^{init}$). The errors are captured by two configurable parameters false positive rate (fpr_H) and false negative rate (fnr_H).

$$fpr_H = \frac{\left|AT_{H}^{init} \setminus AT^{*}\right|}{|\mathbb{T} \setminus AT^{*}|} \tag{Exp. 6.29}$$

$$fnr_H = \frac{\left|AT^{*} \setminus AT_{H}^{init}\right|}{|AT^{*}|} \tag{Exp. 6.30}$$

For IoC and Tel based inferences of AT^{init}, we rely on the ih algorithms presented in section 6.4.3. Similar to fpr_H and fnr_H used to challenge the hypotheses refinement algorithms we challenge the whole hypothesis generation pipeline based on IoCs and Tel by introducing errors into the currently observed log data (COD).

We generate various input data samples using various false positive (fpr_{COD}) and false negative rates (fnr_{COD}): Let COD^{*} denote the optimal set of observations. Let $\mathbb{OBS}$ denote all obs in AttackDB. fpr is the fraction of erroneous SDOs associated with the attack out of the total number of irrelevant SDOs:

$$fpr_{COD} = \frac{|COD \setminus COD^{*}|}{|\mathbb{OBS} \setminus COD^{*}|} \tag{Exp. 6.31}$$

fnr_{COD} is the fraction of SDOs used by the attacker, which were not observed.

$$fnr_{COD} = \frac{|COD \setminus COD^{*}|}{|COD^{*}|} \tag{Exp. 6.32}$$

We run the evaluation for a wide range of error rates:

$$\begin{aligned} fpr_H, fpr_{COD} &\in \{0.0, 0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.5\} \\ fnr_H, fnr_{COD} &\in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\} \end{aligned}$$

In the rest of the paper, we will omit the subscripts of fpr, fnr when their use is apparent from the context.

6.5.1.1 Evaluation metrics

In the evaluation, for each $a \in \mathbb{A}$, we measure the gap between the actual techniques in AT^* and the initial hypothesis AT^{init}, examine whether the hypothesis refinement algorithms were able to decrease the gap and specify the extent to which they were able to do so, and improve AT^{init}.

Note that the proposed approaches described in section 6.4.4 rank all $d \in \mathbb{D}$ according to their likelihood to be associated with the attack, similar to search results in a typical recommender system. Therefore, we use average precision (AP) to evaluate the TTP inference approaches.

Given the actual techniques in AT^*, the initial hypothesis AT^{init}, and the refined hypothesis AT^{ref}, we compute the AP score for AT^{ref} and AT^{init} by comparing each of the hypotheses to AT^*. Next, we evaluate the improvement of the refined hypothesis in comparison to the analyst's hypothesis by calculating the difference between the AP score of AT^{init} and the AP score of AT^{ref}, i.e., we compute $AP(AT^*, AT^{ref}) - AP(AT^*, AT^{init})$.

6.5.1.2 Outline of the experiment

In order to evaluate the hypothesis generation methods, we perform a leave-one-out cross-validation procedure (LOOCV). In every iteration of LOOCV we choose one attack from AttackDB as the test set and rely on the remaining AttackDB content to reconstruct the attack's technique set.

For each attack $a \in \mathbb{A}$, we evaluate each combination of initial hypothesis inference strategy, hypothesis refinement algorithm, and fpr and fnr values. The hypothesis inference strategy is determined by two parameters: $ih\_approach \in \{H, IoC, Tel\}$ and $ih\_impl \in \{TFIDF, NB\text{-}C, MLNB\text{-}C\}$. The hypothesis refinement algorithm procedure is likewise controlled by two parameters: $rh\_approach \in \{arh, rh\}$ and $ih\_impl \in \{ProjT, ProjAT, LPProjT, LPProjA, SupLP\}$.

For each attack, we also investigate whether to apply the refinement or not (ahr). In doing so, we determine a threshold ($arth$) for each a_{new}. $arth$ thereby maximizes the AP for AHG for $a \in \mathbb{A} | a \neq a_{new}$.

Algorithm 2 delineates one iteration of the LOOCV and its evaluation. Each iteration of the LOOCV begins, in line 1, with the preparation of the subsets of the AttackDB, which are the relevant sets for the LOOCV ($AT*$ and the sample with errors (COD) according to the combination of fpr and fnr). The type of relevant SDOs is determined by $ih\_approach$.

Next, in line 2, the investigated attack a is removed from AttackDB. We remove the attack family, all IoCs connected to it, and all related behavioral reports. This step allows the evaluation procedure to assess the performance of AHG as if was facing zero-day exploits. In line 3, the GenInitHyp 3 procedure is called to generate AT^{init} according to the given sample of SDOs and

the initial hypothesis generation approach. For approach H (i.e., simulated analyst), the input is a sample of techniques that simulates AT_H^{init}, thus it is returned by GenInitHyp as is (see lines 1-2 in Algorithm 3). For IoC and Tel based approaches, the input is either a set of IoCs or Tel, respectively.

In line 4, one of the rh algorithms from section 6.4.4 is used to improve AT^{init} and return AT^{ref}. If the threshold is not reached, the adaptive refinement mechanism prevents the use of refinement algorithms so that $AT^{ref} = AT^{init}$. Finally, we restore a to AttackDB (line 5) and evaluate the performance of the ih and rh algorithms (lines 6-8).

The experiment is performed for each strategy (H, IoC, and Tel), combination of ih and rh algorithm, and for each combination of fpr and fnr.

Algorithm 2 Evaluation Procedure

Input: $fpr, fnr, ih\_approach, ih\_impl, rh\_approach, rh\_impl, a$
Output: ΔAP

1: $input \leftarrow sample(COD^*, 1 - fnr) \cup sample(\mathbb{OBS} \setminus COD^*, fpr)$
2: $RemoveAttack(a)$
3: $AT^{init} \leftarrow$ GenInitHyp$(input, ih_approach, ih_impl)$
4: $AT^{ref} \leftarrow$ GenRefHyp$(AT^{init}, rh_approach, rh_impl)$
5: $RestoreAttack(a)$
6: $AP^{ref} \leftarrow AP(AT^*, AT^{ref})$
7: $AP^{init} \leftarrow AP(AT^*, AT^{init})$
8: $\Delta AP \leftarrow AP^{ref} - AP^{init}$
9: **return** ΔAP

Algorithm 3 GenInitHyp

Input: $input, ih\_approach, ih\_impl$
Output: $ih\_impl(paths\_count, techs)$

1: **if** $ih\_approach == H$ **then**
2: **return** $input$
3: **else**
4: $paths\_count, techs \leftarrow CountPaths(input,$
5: $AttackDB, ih\_approach)$
6: **return** $ih\_impl(paths\_count, techs)$

Algorithm 4 GenRefHyp

Input: $AT^{init}, rh\_approach, rh\_impl$
Output: $rh\_impl(paths\_count, techs)$

1: **if** $rh\_approach == arh$ **then**
2: **if** $|AT'| \leq arth$ **then**
3: $AT^{ref} \leftarrow rh(AttackDB, AT^{init})$
4: **else**
5: $AT^{ref} = AT^{init}$
6: **return** $arh(paths\_count, techs)$
7: **else**
8: $AT^{ref} \leftarrow rh(AttackDB, AT^{init})$
9: **return** $rh\_impl(paths\_count, techs)$

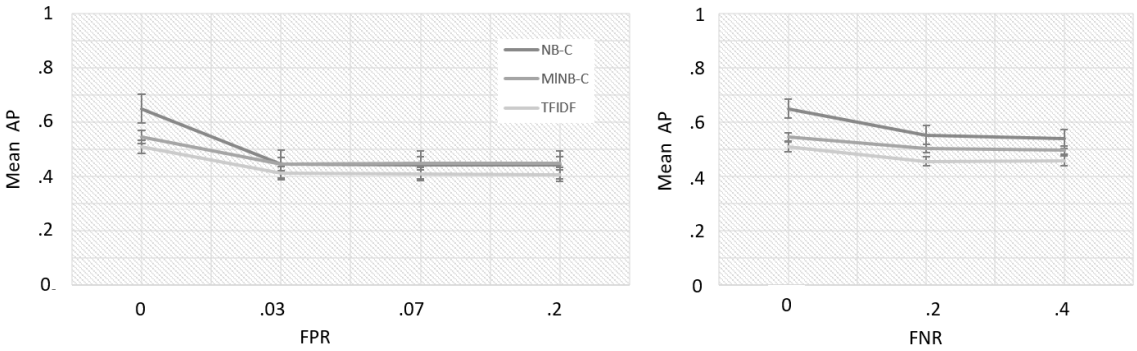

Figure 6.6: Initial hypotheses from IoCs. (left) AP as a function of fpr_{COD} when $fnr_{COD} = 0$. (right) AP as a function of fnr_{COD} when $fpr_{COD} = 0$.

6.5.2 Results

6.5.2.1 Initial hypothesis generation problem results

Figure 6.6 shows the performance of the initial hypotheses generation from IoCs as a function of the noise in the observed data. As expected the performance is the highest in absence of additional noise $fpr_{COD} = fnr_{COD} = 0$. Yet ih algorithms retain reasonable performance even for relatively high error rates in the observed data. In the following results we compare the performance of the algorithms averaged over the various error regimes.

Figure 6.7 presents the mean AP for each ih algorithm (from left to right: simulated human analyst (H) based inference, automated inference of ih based on IoCs, and automated inference of initial attack hypotheses AT^{init} based on telemetry). For automated inference of hypotheses, the results for each algorithm described in section 6.4.3 are provided. We also present a box-whisker plot to provide an overview of the behavior of the algorithms, highlighting the performance of the ih algorithms. As figure 6.7 presents the mean AP of refined attack hypotheses AT^{ref}, we use the identity function to represent the plain vanilla initial hypotheses (AT^{init}) and include their mean AP in the figure to show the effects of refinement.

The black line indicates the mean AP of random attack hypotheses. The expected AP for a baseline extracting 10 random techniques out of 190 is 0.077 (calculated according to Bestgen (2015)). A mean AP of 0.08 was measured empirically. The $MLNB$-C outperforms the other ih algorithms (NB-C and $TFIDF$), providing AT^{init} with the highest AP of the ih algorithms.

Figure 6.8 shows the dependence of the precision of the AT^{init} inference on the number of related techniques. In the upper part a histogram on the number of attacks is given; in the middle part of the figure, the mean AP is presented; while the bottom part presents the receiver operating characteristics area under the curve (ROC-AUC). As can be seen, the precision of attack hypotheses increases with an increasing number of related techniques, and the variance decreases with the number of related techniques. These effects can be seen for both AP and ROC-AUC. $MLNB$-C shows superiority for attacks that are linked to a large number of techniques, although it is not as effective for generation of AT^{init} on attacks that are linked to a small number of techniques (compared to $TFIDF$ and NB-C). The precision of AT^{init} does neither increase

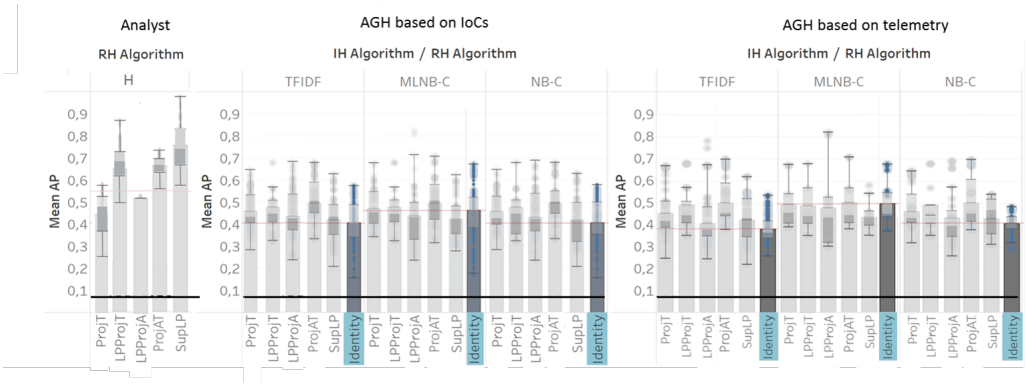

Figure 6.7: Mean AP of analyst based (left) and automatic inference of attack hypotheses based on IoCs (middle) and Tel (right). Black line represents the AP of a random baseline.

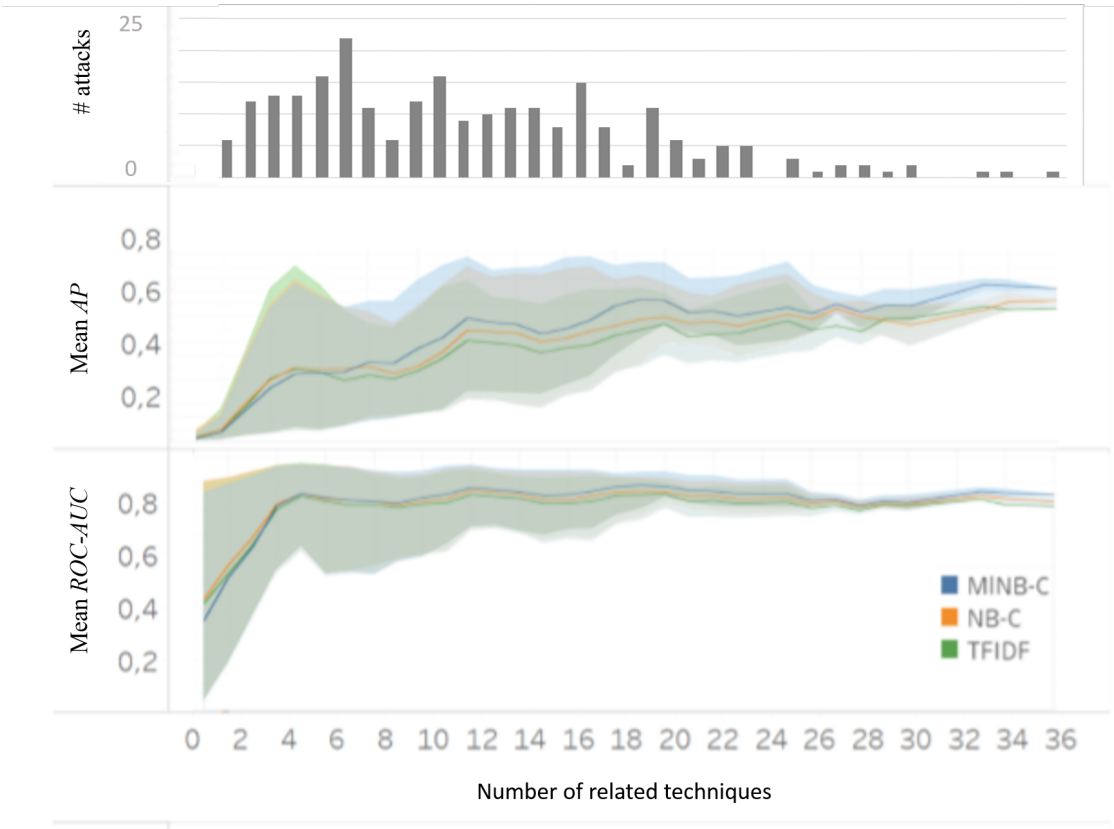

Figure 6.8: The number of attacks (top), mean AP (middle) and $ROC\text{-}AUC$ (bottom) of ih algs. as a function of the number of related techniques.

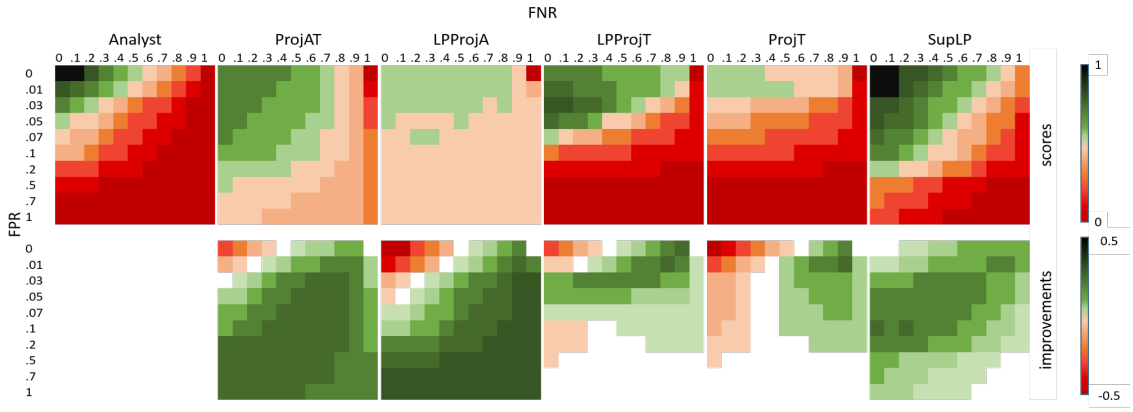

Figure 6.9: Mean AP for initial hypothesis generation and refinement for analyst based initial inferences.

with an increasing number of related telemetries nor related IoCs. These results show that the number of IoCs and telemetries does not significantly influence the performance of AHG. Rather the specificity of COD is more important.

6.5.2.2 Hypothesis refinement problem results

Figure 6.9 presents the mean AP for H based AT^{init} (left) and AT^{ref} for each rh algorithm (top) and the improvements that can be reached through improving the analysts' hypotheses when employing AHG based refinement (bottom). Improvements of AT^{init} are highlighted in green, while deterioration is indicated in red. As can be seen, each rh algorithm has the potential for improving the AT^{init} for specific combinations of fnr and fpr. However, especially for combinations of both, low fnr and low fpr, where the simulated human analyst (H) is able to generate very precise AT^{init}, there may be a deterioration of the precision of attack hypotheses. The extent of the deterioration of AT^{init}, precision is highly dependent on the rh algorithm employed. $SupLP$ decreases the precision of AT^{init} the lowest for combinations of low fnr and low fpr (AT^{init} that is close to AT^{*}), while $LPProjT$ performs the poorest under these circumstances. Furthermore, $SupLP$ achieves the best results for refining H based inferences of AT^{init} for reasonable combinations of fnr and fpr. However, it is shown to work worse for combinations of high fnr and high fpr than $LPProjA$ and $ProjAT$.

Figure 6.7 presents the results for automatic hypothesis generation for each rh algorithm (section 6.4.4) applied to AT^{init} crafted relying on ih algorithms, based on IoCs (left) and Tel (right). We highlight the identity function (as a plain vanilla initial hypothesis (AT^{init}) without refinement) to disentangle the evaluation of the of hypothesis refinement.

When the attack hypotheses are refined (if AT^{init} gets refined), $TFIDF$ and $NB\text{-}C$ can lead to comparable precise hypotheses than $MNB\text{-}C$. What stands out is that when applying Tel based inference of attack hypotheses, the precision of $MLNB\text{-}C$ increases, while a slight deterioration (compared to IoC based inference of attack hypotheses) can be seen for both $TFIDF$ and $NB\text{-}C$. Furthermore, $ProjAT$ is shown to be the most efficient for refining automatically generated AT^{init}.

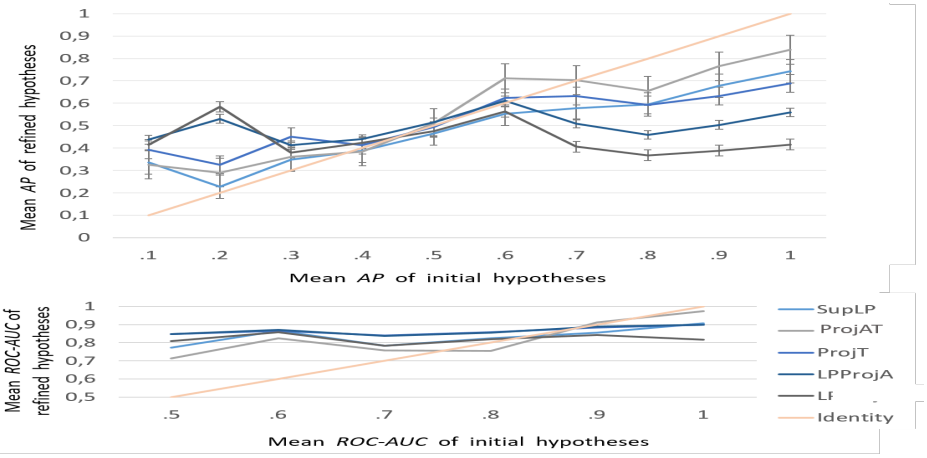

Figure 6.10: Mean AP and $ROC\text{-}AUC$ of refined hypotheses depending on the precision of the initial hypotheses.

Regarding the behavior of rh algorithms, it can be seen that $SupLP$ is ineffective for automatically generated AT^{init} (relying on ih algorithms) and obtains the worst precision of AT^{ref}.

Figure 6.10 shows the efficiency of rh algorithms depending on the quality (precision) of AT^{init}. This figure presents the behavior for each rh algorithm. While it can be seen that the rh algorithms react differently to the precision of AT^{init}, rh algorithms prove to be most useful when the accuracy of AT^{init} is low. However, for an AP of AT^{init} of around 0.5 and higher, rh algorithms are more likely to deteriorate the accuracy of AT^{init}. This is consistent with the results obtained for analyst crafted hypotheses (see figure 6.9). It is also shown that, the behavior of $LPProjA$ and $LPProjT$ is comparable.

6.5.2.3 Adaptive refinement

Finally, we address the limitations of link prediction in cases of accurate AT^{init} using the adaptive refinement method (arh algorithm) described in section 6.4.5. Figure 6.11 shows that the arh algorithm improves the precision of AT^{init} in the vast majority of cases. Moreover, arh based refinement does not lead to a deterioration of the mean AP for either combination of ih and rh algorithms).

For each ih algorithm, the optimal adaptive refinement threshold ($arth$) is calculated on the basis of the remaining data (according to the LOOCV procedure), leading to a refinement in 40% of the cases for $MLNB\text{-}C$ and approximately 84% for $TFIDF$. In the case of the $NB\text{-}C$, the precision of AT^{ref} is always better than the AT^{init}; thus, the arh procedure activates the rh algorithm in every case.

6.5.3 Discussion

We evaluated AHG based on a LOOCV procedure. Therefore, we consider AHG to be applicable for the analysis of novel (yet unobserved) attacks (e.g., zero-day exploits).

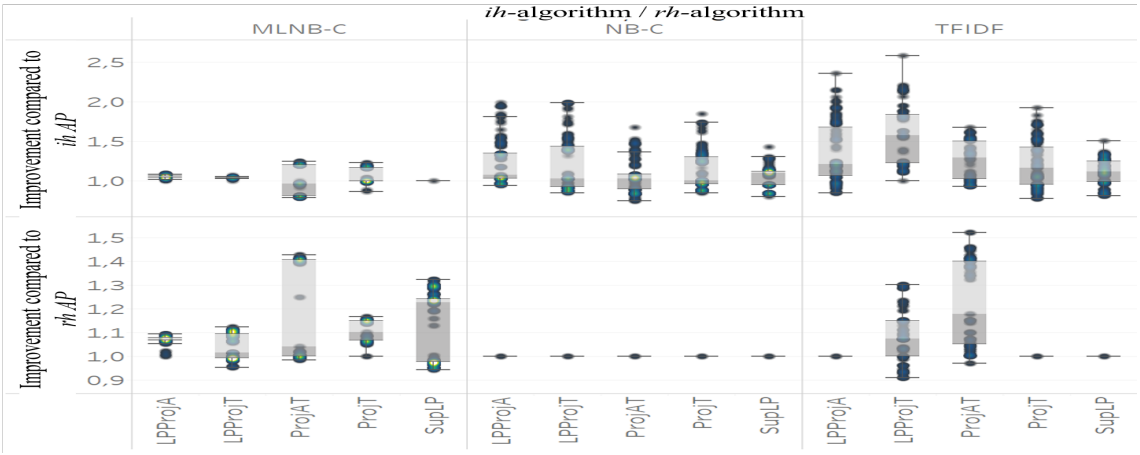

Figure 6.11: Improvements of mean AP when applying the arh-procedure.

For analyst (H) based initial hypotheses ($AT^{init}_{a_{new}}$), hypothesis refinement is able to increase the precision of attack hypotheses for a wide range of combinations of fnr and fpr. Experimental results show the superiority of $ProjAT$ and $SupLP$ for refining H based AT^{init}. The supremacy of $ProjAT$ for refining H based AT^{init} over the other rh algorithms indicates that (1) relying on similar attacks is a good approach for hypothesizing about new attacks, and (2) the Jaccard's/ Tanimoto coefficient between attack descriptions is a good measure for identifying similar attacks. Furthermore, $SupLP$ based hypothesis refinement can be integrated to the process of attack hypothesis generation within the threat hunting cycle for H generated AT^{init} without the risk of a significant deterioration of the precision of attack hypotheses. Evaluations of the results obtained for H based attack hypotheses need to consider that the real performance of H can differ from analyst to analyst. Furthermore, we simulated the accuracy of AT^{init} crafted by H rather than assessing the accuracy of H based attack hypotheses empirically. Also, we considered an analyst with infinite computational resources who produces a reasonable AT^{init} given a set of observables. If there are multiple reasonable AT^{init}, we assume that the analyst randomly selects one of those. In this study we did not account for the possibility that the structure of the database (AttackDB) could cause difficulties for a human analyst (e.g., due to the large volume of obs). Consequently, it is likely that we systematically overestimated the precision of H based AT^{init} when relying on the simulation approach.

For hypothesis refinement, our results show that $LPProjA$ provides a stable mean AP of around 0.5, even for high fpr and high fnr. Since some of the attack patterns are very common, $LPProjA$'s performance (measured in terms of AP and $ROC - AUC$) is reasonably high. Thus, recommending popular techniques can have significant value, especially when there is a lot of uncertainty, i.e., when there is no context to the attack or observations (COD) are weak. Furthermore, $ProjAT$ and $SupLP$ show the best results (high precision of AT^{ref})for reasonable combinations of fnr and fpr, highlighting their potential use for refining hypotheses.

Additionally, the results raise some interesting insights for establishing fully automated hypothesis generation that combines the use of ih and rh algorithms. Since the ih algorithms seem to be robust, even for combinations of high fnr and high fpr, the effectiveness of using some rh algorithms seems to be questionable. For most observed cases, the use of $LPProjT$ and $LPProjA$ resulted

in a deterioration of the precision of AT^{init}. This indicates that there is a need for an effective combination of rh and ih algorithms to craft attack hypotheses with high precision.

6.6 Conclusions and future work

In this work we proposed a comprehensive multi-level threat knowledge base derived from multiple open-source threat intelligence sources called AttackDB. AttackDB can be used to generate attack hypotheses which include high-level descriptions of the investigated attack. We focus on the automated inference of adversarial techniques from observable artifacts found within the attacked systems relying on multiple initial attack hypothesis generation algorithms. The inference is demonstrated using a large collection of VirusTotal behavioral reports. Further, we employ a variety of techniques inspired by recommender systems to refine initial attack hypotheses suggested by one of the algorithms or an analyst. We show that such refinement works best when the number of techniques used in the attack is sufficiently large. Based on this insight we developed an heuristic to decide whether to refine an initial hypothesis or not. Automated attack hypothesis generation relying on this adaptive hypothesis refinement procedure based on the expected number of techniques works the best, achieving a mean AP greater than 0.5 and a $ROC\text{-}AUC$ above 0.8.

Future research on evaluating the performance of the presented algorithms empirically against professional security analysts or against rule-based TTP inference such as Milajerdi, Gjomemo, et al. (2019) would highlight the pros and cons of automated CTI driven TTP inference. The former will assess the usefulness of the AHG algorithms in real security operations center settings with human in the loop. The latter will be possible once rule base for TTP inference will be expanded to cover a large fraction of ATT&CK techniques and allow inference based on VirusTotal behavioral reports.

MITRE ATT&CK does not include information about the order of adversarial techniques employed by malware. The same is true for VirusTotal behavioral reports. As a result, AttackDB does not include information about the sequences of attack steps. Once this limitation is removed by the major CTI sources, AHG can be augmented to consider the order of techniques in an IoA, similar to what has taken place in the area of expert based TTP inference.

While the high-level attack hypotheses provided in form of a collection of adversarial techniques are useful for analysis and reporting, future development is required to close the loop and automate data collection based on these hypotheses. Such automation is possible using analytics such as those provided in the MITRE Cyber Analytic Repository.[18] Future research may also utilize AttackDB in order to generate additional analytics, possibly by relying on machine learning.

[18] `https://car.mitre.org/`

Table 6.1: Summary of notations and abbreviations

α A smoothing parameter	fpr False positive rate	N, k, j Various counts	$SupLP$ Supervised link prediction
Γ Neighbors of an attack in $KG[A]$	h A hash value	NB-C Naive Bayesian Classifier	$\mathbb{T}$ Set of all techniques
a An attack	H Simulated human analyst (an ih strategy)	$\mathbb{OBS}$ Set of all observables	T A set of techniques
$\mathbb{A}$ Set of all attacks	i Specific Indicator of an Attack	obs An observable	t A technique
AA Adamic Adar	ih Initial hypothesis generation	OD Observed data	Tel Telemetry
AD Attack description	IoA Indicator of Attack	P Probability	TF Number of paths connecting two elements in KG
AP Average precision	IoC Indicator of Compromise	PA Preferential attachment	**TTP** Tactics, Techniques, and Procedures
arh Adaptive hypothesis refinement	**IP** Internet Protocol	$ProjA$ Projected attack	$TFIDF$ Term Frequency-Inverse Document Frequency (an ih strategy)
AT A set of attack techniques	J Jaccard's/Tanimoto coefficient	$ProjAT$ Projected attack techniques	**URL** Uniform Resource Locator
CKC Cyber Kill Chain	K Katz measure	$ProjT$ Projected technique	v A component of an attack description
COD Currently observed data	KG The AttackDB knowledge graph	R Set of directed links connecting related SDOs	w Ordering function on the basis of similarity metrics
CTI Cyber Threat Intelligence	$LOOCV$ Leave-one-out cross-validation	rh Hypothesis refinement	
$\mathbb{D}$ Set of all attack descriptions	LP Link prediction	ROC-AUC Area under the receiver operating characteristic curve	
E Expectation	$LPProjA$ Link prediction on projected attacks	**SDO** STIX Domain Object	
EDR Endpoint Detection and Response	$LPProjT$ Link prediction on projected techniques	**SIEM** Security Information and Event Management	
fnr False negative rate	$MLNB$-C Multi-layer naive Bayesian classifier	**STIX** Structured Threat Information eXpression	

References

Adamic, L. A., Adar, E. (2003). Friends and neighbors on the web. *Social networks*, *25*(3), 211–230.

AlEroud, A., Karabatis, G. (2017). Methods and techniques to identify security incidents using domain knowledge and contextual information. In *Integrated network and service management (im), 2017 ifip/ieee symposium on* (pp. 1040–1045).

Alonso, S. (2016, January). *Cyber threat hunting (1): Intro.* Retrieved from `https://cyber-ir.com/2016/01/21/cyber-threat-hunting-1-intro/`

Antoniou, G., Van Harmelen, F. (2004). Web ontology language: Owl. In *Handbook on ontologies* (pp. 67–92). Springer.

Barabási, A.-L., Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Bestgen, Y. (2015). Exact expected average precision of the random baseline for system evaluation. *The Prague Bulletin of Mathematical Linguistics*, *103*(1), 131.

Bhatt, P., Yano, E. T., Gustavsson, P. (2014). Towards a framework to detect multi-stage advanced persistent threats attacks. In *2014 ieee 8th international symposium on service oriented system engineering (sose)* (pp. 390–395).

Bianco, D. (2013, March). *The pyramid of pain (2014)*. Retrieved from `http://detect-respond.blogspot.nl/2013/03/the-pyramid-of-pain.html`

Bromander, S., Swimmer, M., Eian, M., Skjotskift, G., Borg, F. (2020). Modeling cyber threat intelligence. In *Proceedings of the 6th international conference on information systems security and privacy (icissp 2020)* (pp. 273–280).

Bromiley, M. (2016). Threat intelligence: What it is, and how to use it effectively. *SANS Institute InfoSec Reading Room*, *15*.

Chismon, D., Ruks, M. (2015). *Threat intelligence: Collecting, analysing, evaluating* (Tech. Rep.). MWR InfoSecurity.

Daszczyszak, R., Ellis, D., Luke, S., Whitley, S. (2019). *Ttp-based hunting* (Tech. Rep.). MITRE CORP MCLEAN VA.

DeCianno, J. (2014). Indicators of attack vs. indicators of compromise. *CrowdStrike*.

Dekel, L., Zilberman, P., Puzis, R., Dardik, U., Elitzur, A. (2021). *Attack db otx-xforce-vt*. IEEE Dataport. Retrieved from `https://dx.doi.org/10.21227/f74t-gh08` doi: 10.21227/f74t-gh08

de Melo e Silva, A., Costa Gondim, J. J., de Oliveira Albuquerque, R., García Villalba, L. J. (2020). A methodology to evaluate standards and platforms within cyber threat intelligence. *Future Internet*, *12*(108), 1–23.

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. In *2019 european intelligence and security informatics conference (eisic)* (pp. 40–47).

Fard, S. M. H., Karimimpour, H., Dehghantanha, A., Jahromi, A. N., Srivastava, G. (2020). Ensemble sparse representation-based cyber threat hunting for security of smart cities. *Computers and Electronical Engineering*, *88*(106825), 1–13.

Fire, M., Tenenboim-Chekina, L., Puzis, R., Lesser, O., Rokach, L., Elovici, Y. (2014). Computationally efficient link prediction in a variety of social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 1–25.

Gao, Y., Li, X., Li, J., Gao, Y., Guo, N. (2018). Graph mining-based trust evaluation mechanism with multidimensional features for large-scale heterogeneous threat intelligence. In *2018 ieee international conference on big data (big data)* (pp. 1272–1277).

Garrido, J. S., Dold, D., Frank, J. (2021). Machine learning on knowledge graphs for context-aware security monitoring. In *2021 ieee international conference on cyber security and resilience (csr)* (pp. 55–60).

Giura, P., Wang, W. (2012). A context-based detection framework for advanced persistent threats. In *Cyber security (cybersecurity), 2012 international conference on* (pp. 69–74).

Homayoun, S., Dehghantanha, A., Ahmadzadeh, M., Hashemi, S., Khayami, R. (2018). Know abnormal, find evil: Frequent pattern mining for ransomware threat hunting and intelligence. *IEEE Transactions on Emerging Topics in Computing*, 1–1. doi: 10.1109/TETC.2017.2756908

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al. (2004). Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, *21*(79), 1–31.

Hutchins, E. M., Cloppert, M. J., Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, *1*(1), 80.

Iqbal, Z., Anwar, Z. (2016). Ontology generation of advanced persistent threats and their automated analysis. *NUST Journal of Engineering Sciences*, *9*(2), 68–75.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37–50.

Jasper, S. E. (2017). U.s. cyber threat intelligence sharing frameworks. *International Journal of Intelligence and CounterIntelligence*, *30*(1), 53–65. Retrieved from `https://doi.org/10.1080/08850607.2016.1230701` doi: 10.1080/08850607.2016.1230701

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.

Landauer, M., Skopnik, F., Wurzenberger, M., Hotwagner, W., Rauber, A. (2019). A framework for cyber threat intelligence extraction from raw log data. In *2019 ieee international conference on big data (big data)* (pp. 3200–3209).

Lee, R. M., Lee, R. T. (2018, September). Sans 2018 threat hunting survey results. *SANS Institute*.

Lee, S., Cho, H., Kim, N., Kim, B., Park, J. (2018). Managing cyber threat intelligence in a graph database: Methods of analyzing intrusion sets, threat actors, and campaigns. In *2018 international conference on platform technology and service (platcon)* (pp. 1–6).

Mavroeidis, V., Bromander, S. (2017). Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In *Intelligence and security informatics conference (eisic), 2017 european* (pp. 91–98).

Mavroeidis, V., Jøsang, A. (2018). Data-driven threat hunting using sysmon.

McAfee. (retrieved 2019a). *Mcafee investigator: Transform analysts into expert investigators*. Retrieved from `https://www.mcafee.com/enterprise/en-us/assets/data-sheets/ds-investigator.pdf`

McAfee. (retrieved 2019b). *Mcafee threat intelligence exchange (datasheet)*. Retrieved from `https://www.mcafee.com/enterprise/en-us/assets/data-sheets/ds-threat-intelligence-exchange.pdf`

Milajerdi, S. M., Eshete, B., Gjomemo, R., Venkatakrishnan, V. N. (2019). Poirot: Aligning attack behavior with kernel audit records for cyber threat hunting. In *Proceedings of the 2019 acm sigsac conference on computer and communications security* (pp. 1795–1812).

Milajerdi, S. M., Gjomemo, R., Eshete, B., Sekar, R., Venkatakrishnan, V. (2019). Holmes: real-time apt detection through correlation of suspicious information flows. In *2019 ieee symposium on security and privacy (sp)* (pp. 1137–1152).

Najafi, P., Mühle, A., Pünter, W., Cheng, F., Meinel, C. (2019). Malrank: A measure of maliciousness in siem-based knowledge graphs. In *Proceedings of the 35th annual computer security applications conference* (pp. 417–429).

Plona, T., Maxwell, K., Varni, B. (2017). *Threat intelligence in splunk*. Retrieved from https://conf.splunk.com/files/2017/slides/do-you-really-know-my-adversaries-prove-it.pdf

Polatidis, N., Pimenidis, E., Pavlidis, M., Papastergiou, S., Mouratidis, H. (2018). From product recommendation to cyber-attack prediction: Generating attack graphs and predicting future attacks. *Evolving Systems*, 1–12.

Qamar, S., Anwar, Z., Rahman, M. A., Al-Shaer, E., Chu, B.-T. (2017). Data-driven analytics for cyber-threat intelligence and information sharing. *Computers & Security*, *67*, 35–58.

Ranveer, S., Hiray, S. (2015). Comperative analysis of feature extraction methods of malware detection. *International Journal of Computer Applications*, *120*(5), 1–7.

Rasheed, H., Hadi, A., Khader, M. (2017). Threat hunting using grr rapid response. In *New trends in computing sciences (ictcs), 2017 international conference on* (pp. 155–160).

Rennie, J. D. M., Shih, L., Teevan, J., Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (icml-03)* (pp. 616–623).

Riesco, R., Villagrá, V. (2019). Leveraging cyber threat intelligence for a dynamic risk framework. *International Journal of Information Security*, 1–25.

Rubinshtein, S., Puzis, R. (2016). Modeling and reconstruction of multi-stage attacks. In *2016 ieee international conference on software science, technology and engineering (swste)* (pp. 135–137).

Samtani, S., Chinn, R., Chen, H., Jr., J. F. N. (2017). Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *Journal of Management Information Systems*, *34*(4), 1023–1053. Retrieved from https://doi.org/10.1080/07421222.2017.1394049 doi: 10.1080/07421222.2017.1394049

Sarwar, B. M., Karypis, G., Konstan, J. A., Riedl, J., et al. (2001). Item-based collaborative filtering recommendation algorithms. *Www*, *1*, 285–295.

Sauerwein, C., Sillaber, C., Mussmann, A., Breu, R. (2017). Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives. In *Proceedings der 13. internationalen tagung wirtschaftsinformatik (wi 2017)* (pp. 837–851).

Settanni, G., Shovgenya, Y., Skopik, F., Graf, R., Wurzenberger, M., Fiedler, R. (2017, June). Acquiring cyber threat intelligence through security information correlation. In *2017 3rd ieee international conference on cybernetics (cybconf)* (pp. 1–7). doi: 10.1109/CYBConf.2017.7985754

Shackleford, D. (2015). Who's using cyberthreat intelligence and how. *SANS Institute*.

Sqrrl Data, I. (2016). *A framework for cyber threat hunting.* Retrieved from https://sqrrl.com/media/Framework-for-Threat-Hunting-Whitepaper.pdf,https://www.threathunting.net/files/framework-for-threat-hunting-whitepaper.pdf

Strom, B. E., Battaglia, J. A., Kemmerer, M. S., Kupersanin, W., Miller, D. P., Wampler, C., ... Wolf, R. D. (2017). *Finding cyber threats with att&ck™-based analytics* (Tech. Rep.). MITRE Technical Report MTR170202. The MITRE Corporation.

Tanimoto, T. T. (1958). Elementary mathematical theory of classification and prediction.

Tyler Technologies, I. (2018). *A guide to cyber threat hunting.* Retrieved from https://www.tylertech.com/services/ndiscovery/nDiscovery-Threat-Hunting.pdf

Ulicny, B., Moskal, J., Kokar, M. M., Abe, K., Smith, J. K. (2014). Inference and ontologies. In A. Kott, C. Wang, R. F. Erbacher (Eds.), *Cyber defense and situational awareness* (Vol. 62, pp. 167–199). Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-319-11391-3_9 doi: 10.1007/978-3-319-11391-3_9

Wafula, K., Wang, Y. (2019). Carve: A scientific method-based threat hunting hypothesis development model. In *2019 ieee international conference on electro information technology (eit)* (pp. 1–6).

Wheelus, C., Bou-Harb, E., Zhu, X. (2016, Nov). Towards a big data architecture for facilitating cyber threat intelligence. In *2016 8th ifip international conference on new technologies, mobility and security (ntms)* (pp. 1–5). doi: 10.1109/NTMS.2016.7792484

Xiao, H., Xing, Z., Li, X., Guo, H. (2019). Embedding and predicting software security entity relationships: A knowledge graph based approach. In *International conference on neural information processing* (pp. 50–63).

Zhao, Y., Lang, B., Liu, M. (2017). Ontology-based unified model for heterogeneous threat intelligence integration and sharing. In *2017 11th ieee international conference on anti-counterfeiting, security, and identification (asid)* (pp. 11–15).

7 Cyber Risk Quantification - Using Weighted Attack Graphs for Behavioral Cyber Game Theory[1]

Abstract

Operating and engineering secure systems is challenging yet a necessary prerequisite for modern life as we know it, flourishing economic systems and society as a whole. This is as digitalisation penetrated broad aspects of every facet of life realising many opportunities. However, digital transformation also leads to increasing vulnerability to cyber threats. Cyber risk quantification thereby has a crucial role as anything that "is not measured cannot be improved. [And] what is not improved will always degrade" (Thomas Kelvin). However, quantifying cyber risks respectively as an inverse quantifying cyber security is still in its infancy and a largely unsolved problem.

We propose a novel methodology for cyber risk quantification based on weighted attack graphs. By doing so, we introduce a multi-layered attack ontology which is the basis of the attack graph. The attack graph is developed relying on cyber threat intelligence. We weight each attack path using computational models of motivation. The attack graph is the basis of a defender attacker game. We analyse and solve the game for deriving quantitative measures describing the risk of getting attacked.

7.1 Introduction

Society is undergoing a comprehensive digital transformation. This can be seen in the increasing penetration with information and communication technology which is also known as digitalisation or digital transformation. Discussions on digitalisation investments frequently are strongly focused on the benefits neglecting any risks that may come with digitalisation. However, the increasing penetration of broad aspects of modern life with information and communication technology does

[1] This chapter includes the preprint of the article "Cyber Risk Quantification - Using Weighted Attack Graphs for Behavioral Cyber Game Theory" by Marcus Wiens, Frank Schultmann, and myself accepted for publication within the book "Advances in Cyber Security and Intelligent Analytics" CRC Press, Taylor & Francis Group.

not only introduce benefits bringing high levels of comfort, efficiency, and productivity but also novel risks making modern life more vulnerable to attacks. This is, cyber risks pose an increasing threat to modern life. Consistently, cyber attacks are observed to have tremendous impacts on worldwide economic performance leading to a necessity of managing these risks (World Economic Forum, 2022). Hence, ensuring cyber security while leveraging on the benefits of digitalisation is a key challenge of modern life and a necessary prerequisite for the successful and sustainable digitalisation.

However, managing cyber risks is challenging for professionals as currently there is no profound and commonly accepted methodology for quantitative cyber risk assessment. This is because established methods for quantitative risk assessment are oftentimes not suitable for the field of cyber risk. Hence, security experts rely on heuristics or qualitative risk assessment (e.g. expert judgements or guesswork). Besides neglecting strategic interactions between defenders and attackers, current approaches to quantitative cyber risk management oftentimes suffer from oversimplified attacker models considering attackers to be inherently and illicitly malicious (Kaiser et al., 2021a). Yet, defending against illicitly malicious attackers can easily lead to a distorted and undifferentiated picture of the attack landscape. If we consider that ”everything out there is hackable with enough time, with enough tools, with enough expertise” (Stephanie Domas), there would not be a state of security given this attacker model. However, if attackers are not considered to be illicitly malicious but motivated by a more realistic and differentiated mix of motives like affiliation, achievement or power, it is possible to use defense strategies in a more targeted and thus resource-saving way. This is, attackers' motivations in conducting an attack should have a more prominent role in discussions on cyber security than they have currently. Thus, cognitive science and motivational foundations of attacker models can be considered as an important step ahead in cyber risk analysis (Veksler et al., 2020).

Yet, till now, the quantification of cyber risks seems to be an unsatisfactorily solved scientific problem. Zeller and Scherer (2021) state that there is no scientifically recognised methodology for quantifying cyber risks with high accuracy. Consistently security professionals oftentimes need to trust on their gut feeling in risk management (Romanosky et al., 2019). Within this work, we contribute to close this research gap by combining technical analyses of a system relying on cyber threat intelligence and cognitional, behavioural factors. In doing so, our work provides the first actionable approach for attacker quantification based on behavioural cyber game theory. Our contributions are the following:

- We introduce a behavioural attacker-defender game;
- operationalise the game based on cyber threat intelligence;
- encode the attacker defender game in an attack graph;
- weight the attack graph relying on computational models of motivation taking into account different incentivising factors;
- solve the defender attacker game enabling an a priori quantification of attack probabilities in

- a single player game and
- a two player game.

The rest of the work is organised as follows. First, an introduction to the theoretical foundations and related work in cyber risk quantification is given. In section 7.3, we introduce the proposed methodology for risk quantification. Afterwards, we present the results of methodology, discuss the main insights and come to a brief conclusion.

7.2 Literature overview and theoretical foundations

7.2.1 Cyber risk quantification

Although academic research increasingly acknowledges the impact of cyber risks on economic performance, there is a lack of a systematic formal methodology for their quantification (Zeller and Scherer, 2021). Setting up models for understanding and quantifying cyber risks is challenging as cyber risks are in many aspects far from being understood comprehensively (Woods and Böhme, 2021). Common risk management practices, which are mainly based on statistical approaches, thereby seem not suitable or at least insufficient as defenders and attackers are considered to play "a dynamic cat and mouse game" (Elitzur et al., 2019).This reflects the adaptive behaviour of perpetrators to the (defensive) actions taken by the defenders (e.g. security analysts or business managers) and highlights the importance of risk quantification approaches that allow to consider this processes of strategic adaption (e.g. game theory). Beyerer and Geisler (2016) highlight the importance of focusing on the strategic interactions for quantifying security and line out that these must not be ignored. Since the counterparties on both sides (attacker and defender) make an adjustment to the strategy chosen by the other (strategic/ intelligent adversaries), risk quantification must take into account this adaptive behaviour and can learn from their experience. Consistently, Rass et al. (2017) state that game theoretical models can be applied to the field of cyber risk quantification. Zarreh et al. (2019) demonstrate the possibility to predict attackers' behaviour (probability of choosing a specific attack strategy) in the field of cyber risk relying on game theoretical approaches and use this information to quantitatively assess cyber security. Durkota et al. (2019) propose a defender-attacker game focusing on security investments given limited security budgets. The space of possible actions that may be taken by attackers is represented by attack graphs.

According to the quantification method taken, there are different measures for cyber risk. Taking game theoretical methodology, practical security against rational attackers can be defined as the state where every attack is non-profitable for attackers (Buldas et al., 2006). Yet, other approaches (e.g. technical analyses) propose different measures (e.g. k-zero-day safety, Wang et al. (2013). Taken together, cyber risk quantification, and cyber risk management, which needs to be based on these quantitative measure for cyber risk seems to be "more art than science" (Woods and Böhme, 2021) at the moment.

7.2.2 Behavioural cyber game theory

Recent research in cyber security mainly focuses on an operational level, presenting attack techniques for exploiting specific vulnerabilities or discussing the feasibility of specific defensive means against a specific attack (Attiah et al., 2018). Research on the dynamic interaction of defenders and attackers and the interdependence of their actions are on the other hand rare (Attiah et al., 2018). Game theoretic models aim at providing insights to those dynamics between attackers and defenders (Veksler et al., 2020). Furthermore, game theory has proven to be a promising technique for solving many real world problems not only in the field of cyber risk.

In general, game theory is a mathematical framework for analysing interdependent decisions taken by strategically acting rational agents (Durkota et al., 2019; Do et al., 2018). The players' payoffs are not restricted to monetary units (which is often the standard case of objective functions in game theory) but can include further components such as monetary equivalents, psychological payoffs or social motivations (often in the context of behavioural game theory). Actionable computational models for the motivational differentiation are inter alia presented by Merrick (2016). Game theoretical models can be descriptive but may also have normative power. Furthermore, each game can be depicted as a game graph.

A special case of these game graphs may be attack graphs. Attack graphs present a graphical representation of attack scenarios consisting of different layers representing an ontological model of an attack. Furthermore, attack trees are a special form of attack graphs where only one root is allowed (Ingoldsby, 2010). "Attack tree analysis (ATA) is a prominent graphical model technique used for modelling attack scenarios (...) Quantitative attack tree analysis utilises data analytics to predict attack rates and probability of success" (Verma et al., 2019). These concepts are used for decomposing the complex interrelations in cyber security making them accessible for analysis (Nguyen et al., 2017). It is a helpful means for understanding the interaction of vulnerabilities, exploits, attack techniques and patterns. For this purpose, attack graphs frequently consist of multiple ontologies that can help to describe the state of a system. In the same sense, attack-graphs can encode every action an attacker can undertake and hence can be considered a tool for modelling network security with game theoretical means (Anwar and Kamhoua, 2020). Kamdem et al. (2017) employ attack graphs and focus their analysis on the impact of vulnerabilities on the security on energy systems. Therefore, they craft a vulnerability multi-graph. Kamdem et al. (2017) analyse the vulnerability multi-graph based on a two-player defender-attacker zero-sum Markov game. Furthermore, Wang et al. (2008) employ a vulnerability-focused attack-graph for giving a probabilistic security metric. Anwar et al. (2020) presents a network-graph analysed in terms of network security. They thereby focus on the effects of network diversity on security. The graph is analysed in terms of a two-player non-zero-sum game. Anwar et al. (2020) focus within their work on the trade-off between costs for deploying a diversified network and reachable security. Gratifications or payoffs can be included in the game graph as instant or terminal gratifications. If gratifications are included in the game graph or probabilities for breaking a specific node (e.g. successfully taking a specific action) the graph frequently is called a weighted game graph.

Current game theoretical models in cyber risk management frequently consider attackers to be malicious utility maximisers, where attackers aim at maximising the impact inflicted on the defenders' side. This is, attacker modelling nowadays oftentimes reflects oversimplified cognitive models together with extreme motivational assumptions (Kaiser et al., 2021a). The inclusion of more differentiated cognitive models in attacker modelling including individual attacker preferences is considered to be able to provide "fairly high improvements over normative GT approaches in reducing the potential for successful attacks" (Veksler et al., 2020). Such behavioural game theoretical models have the potential to give a more precise picture of the threat landscape. According to criminology and psychology, three incentive-based theories of motivation (achievement (ach), affiliation (aff), and power (pow)) form the basis of differentiating attackers. "Achievement motivation drives humans to strive for excellence by improving on personal and societal standards of performance." (Merrick, 2016). Achievement motivation includes the strive for success, individual development, creativity, curiosity, and cognitive challenge. Power motivation describes the aim to reach a superior position among individuals, social groups and nations respectively to gain influence over others. "Power-motivated individuals select high-incentive goals, as achieving these goals gives them significant control of the resources and reinforces of others." (Merrick, 2016) Power can for instance be legitimised by moral superiority. The motive of affiliation represents social components of motivation - both, the hope of social acceptance as well as the fear of social exclusion. "Affiliation refers to a class of social interactions that seek contact with formerly unknown or little known individuals and maintain contact with those individuals in a manner that both parties experience as satisfying, stimulating, and enriching." (Merrick, 2016) Thereby subtypes of each motivation can be identified (Merrick, 2016). Prestige, social status and social security can be seen as submotives of the motive for affiliation.

"The key idea (of employing game theoretical approaches) is to model attackers/defenders to have multiple levels of attack/ defence strategies that are different in terms of effectiveness, strategy costs, and attack gains/damages. Each player adjusts his strategy based on the strategy's cost, potential attack gain/damage, and effectiveness in anticipating of the opponent's strategy." (Attiah et al., 2018) Different game theoretical approaches can thereby be classified according to the involved players. In cyber risk management, single player games (also called games against nature) can be used to evaluate the security of a specific state of a system (de Gusmão et al., 2016) or deciding on the optimal spending on cyber security measures (Kissoon, 2020; Panaousis et al., 2014). However, if the interdependency of the players and their strategic interaction is the focus of analysis, classic or "actual" game theory analysing n-player games ($n > 1$) is the tool of choice.

7.2.3 Cyber threat intelligence

Cyber threat intelligence is structured, actionable, evidence-based information on past attacks (Elitzur et al., 2019). It enables to gain a deeper understanding of the threat landscape and may help to gain a holistic view on the threat landscape. Cyber threat intelligence can be acquired by victims of a cyber attack recording the attack and extracting knowledge over this attack. Since cyber threat intelligence is information on experienced attacks, each defender can only acquire a

limited set of knowledge which represents a limited perspective. This is, an important feature of cyber threat intelligence is its shareability (Elitzur et al., 2019). By sharing cyber threat intelligence between different parties, each party can gain a holistic view on the current threat landscape. A major obstacle of cyber threat intelligence is that the information is oftentimes provided in various different formats depending on the preferences of the respective source. To encounter this obstacle, the US Department of Homeland Security's Office of Cybersecurity and Communications and MITRE developed and established a standardized and language for sharing cyber threat intelligence (Structured Threat InformationeXpression). However, other cyber threat intelligence languages exist as well. These include OpenIOC, Incident Object Description Exchange Format and proprietary languages. Yet, Sauerwein et al. (2017) provides evidence, that Structured Threat Information eXpression can be seen as de-facto standard language of cyber threat intelligence sharing. Cyber threat intelligence furthermore is shared by many platforms (e.g. Malware Information Sharing Platform, OpenCTI, Collective Intelligence Framework, Anomali STAXX, Open Threat Exchange platform, and many more). An understanding of the specific benefits as well as the weaknesses of these different formats of cyber threat intelligence sharing respectively the different languages and platforms (de Melo e Silva et al., 2020) can be helpful. In the research paper, they provide a comprehensive evaluation of a set of those different cyber threat intelligence languages and platforms. Additionally, information regarding vulnerabilities is shared. MITRE provides for this purpose the common vulnerability enumeration. Furthermore, for understanding cyber threat intelligence, a profound understanding of the procedures used during cyber attacks need to be established. Therefore, the Cyber Kill Chain was developed by Lockheed Martin Corporation (Muckin and Fitch, 2014). It describes on a generic level the steps an offender needs to take to achieve a specific attack goal. Hence, Cyber Kill Chain helps in understanding the threat landscape and establishes a means of understanding attack vectors as an attack strategy that is followed by an attacker. The Cyber Kill Chain includes seven steps of attacker action taking. It can be closely aligned to the tactics presented in various cyber threat intelligence languages. Cyber threat intelligence has proven to be a helpful means for forensic operations, attack hypothesis generation (Elitzur et al., 2019), attack forecasting (F. K. Kaiser et al., 2021) and for increasing the efficiency of intrusion detection systems and many other approaches in cyber risk mitigation, e.g. threat hunting. However, using cyber threat intelligence within game theoretical models for quantifying cyber risks and specifying game models is only rarely researched.

7.3 Game theoretical model on weighted attack graphs considering behavioural factors

Cyber risks can be represented and analyzed in the form of game-theoretical models. Here, attackers and defenders compete against each other in a game. The benefits of the game for the attacker can be of different nature. The motivating incentives can be understood according to different computational models of motivation (see e.g. Merrick (2016)) and describe the transition from a game theory model based primarily on monetary values to behavioral game theory. Thus, in the

game between defender and attacker, different attack strategies lead to different benefits (payoffs) depending on the type (different motivations) of the attacker.

In the simplest case, the game involves a defender i and an attacker j playing against each other. Let the defender's utility and the attacker's utility be given as follows. Where Ψ is the probability of success of an attack a, I is its impact, and V is its value to the attacker (here, the specification of utility can be done according to the motivating factors which means that V will be different for different Attacker types).

$$U_i = -\Psi_a \cdot I_a \tag{Exp. 7.1}$$

$$U_j = \Psi_a \cdot V_a \tag{Exp. 7.2}$$

For a first analysis, strategic interactions of the attacker and the defender can be abstracted from this point. In the sense of decision theory or a single player game, the analysis is limited to the execution of the attack (the choice of an attack strategy by the attacker). In this way, a risk analysis can be performed for a given system. The probability of success of an attack is determined by the existence of exploitable vulnerabilities for this attack. It is assumed that the attacker has complete information (the attacker's subjective beliefs about the network structure and its backups correspond to the real state of the system). For the analysis of real systems, an attack graph is created based on cyber threat intelligence, which includes information about the attack strategies (e.g. use of a special malware), the strategic actions (used attack techniques of the malware which represent the nodes), vulnerabilities that exploit the strategic actions, the products in which the vulnerabilities occur and the processes in which the products are used. A schema of the attack graph is presented in figure 7.1. It is valid that if no attack technique is prevented or hindered and there is a vulnerability for the attack technique, the attack will be successful. However, if a technique is prevented (by a suitable countermeasure being taken by the defender), the attack will be unsuccessful. In the structure of the attack graph, the different targets of the attackers are specified at the top level. Different types of attackers can be guided by different goals. For example, we could distinguish between a Byzantine player (the one who is solely interested in destruction possibly as a punishing power according to a power motive) and a tullock player (sees hacking as a game and is interested in challenging himself and being the first to hack a system that is considered secure according to achievement motives). In the attack tree, the nodes now represent those critical points that the attacker wants to reach (successfully mastering a subgoal of his attack, such as gaining initial access to the system). For an insecure system (a system with known vulnerabilities), the attacker can achieve this goal with negligible effort.

Given complete information, the attacker will now choose the attack that (I) will be successful (i.e., possible with a given infrastructure) and (II) promises the greatest benefit. The game can be solved as a simple single player game with a large size of the attack graph using backward induction. Under incomplete information, the attacker can only form a subjective belief about the actual effects of an attack and its probability of success. In this case, the attacker will not

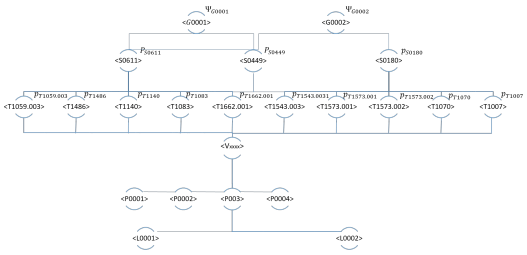

Figure 7.1: Exemplary attack graph with two targets and three possible attack paths

be able to form an exact picture of the benefits of the attack strategies and will choose the attack strategy that promises the highest benefit based on this subjective perception. This case can be represented as a Bayesian game, where the attacker does not know which defensive measures have been implemented and consequently which parts of the system are vulnerable. To understand the game, this makes the strategic interactions between attacker and defender important. Thus, if the system is secure in the sense that there are no known vulnerabilities (either because the attacker has no information about the defender's chosen security measures or because the system is zero-day secure), effort must be spent by the attacker (to find existing or new vulnerabilities). We describe this effort with the variable x. Specifically, x_{21} represents the attack intensity at node 2 expended by attacker 1. Attack intensities or attack effort have a cost associated with them (constant variable cost κ). A defender may also expend effort to keep the system secure (search for and close any vulnerabilities itself). The defender expends intensity/deployment d_k defense resources at node k. Deployment with defense resources also has a cost (constant variable cost c_d). This creates a competition between the defender and attacker to see who can find a vulnerability first and close or exploit it. This can lead to an arms race between attacker and defender. At the nodes, the respective attack intensities and defense intensities now determine the probabilities that the nodes can be overcome (an attack technique is successful and thus a partial goal of an attack can be achieved). The probability that node k can be overcome by attacker (type) j using the attack strategy a (the technique encoded in the node can be successfully applied) is given as p_{akj}.

$$p_{akj} = \frac{x_{akj}}{x_{akj} + d_k} \text{ (node probability)} \tag{Exp. 7.3}$$

$$P_{akj} = \prod_k p_{akj} \text{ (strategy success probability)} \tag{Exp. 7.4}$$

From these node probabilities (represented with a small p), the success probabilities of different strategies can be derived (the latter are represented with a large P). The node probabilities can thus be conceptualized as beliefs. Thus, P_1 is the probability that attack strategy 1 (of attacker type 1) will succeed. Ψ_1 is now the probability that an attack target will be reached. This can be determined now again from the success probabilities of the strategies. Here it must be noted that targets can be reached under use of different strategies (OR condition). For a simple example with two different targets and three strategies, where strategy 1 can be used to reach target 1 as well as target 2, it follows:

$$\Psi_1 = P_{11} + P_{21} - P_{11} \cdot P_{21} \text{ and } \Psi_2 = P_{12} + P_{32} - P_{12} \cdot P_{32} \text{ (target probability)} \quad \text{(Exp. 7.5)}$$

We now turn to the defender's objective function. In a more complicated but interesting version, the defender's objective can take into account the impact of digitization investments. A simpler version ignores digitalization investments and just focusses on the outcomes of the pure attack-defense-interaction. This results in two variants of the game. Whereby, taking into account the digitization investments, trade-offs are made between increased productivity due to the digitization investments on the one hand and increased vulnerability due to the rising attack surface and increased attractiveness for many attacker types (due to the greater impact of an attack) on the other.

Without consideration of digitalization investments and their effect on the utility of each player, the defenders objective function could be given as follows.

$$\begin{aligned} \Pi_D(d_k) &= z \cdot q - c_q \cdot q - c_d \cdot \sum_k d_k - \sum_a \Psi_a(d_{ak}) \cdot L_a \\ &= z \cdot q - c_q \cdot q - c_d \cdot \sum_{k=1}^{3} d_k - \sum_{a=1}^{2} \Psi_a(d_{ak}) \cdot L_a \end{aligned} \quad \text{(Exp. 7.6)}$$

When digitization investments are taken into account, the target function expands as follows:

$$\begin{aligned} \Pi_D(d_k) &= z \cdot q \cdot (1 + \gamma y) - c_q \cdot q \cdot (1 - \eta y) - c_d \cdot \sum_k d_k - c_y \cdot y - \sum_a \Psi_a(d_{ak}) \cdot y \cdot L_a \\ &= z \cdot q \cdot (1 + \gamma y) - c_q \cdot q \cdot (1 - \eta y) - c_d \cdot \sum_{k=1}^{3} d_k - c_y \cdot y - \sum_{a=1}^{2} \Psi_a(d_{ak}) \cdot y \cdot L_a \end{aligned} \quad \text{(Exp. 7.7)}$$

Where z s the product price, q is the quantity of product sold, y is the amount of digitization investment, c_q is the cost parameter for producing quantity q, c_d is the cost parameter for defense units d, c_y is the cost parameter for digitization investment, Ψ_1 is the probability that attacker type 1 reaches its target, L_1 is the defender's loss if attacker reaches target 1, Ψ_2 is the probability that attacker type 2 reaches its target (target 2), and L_2 is the defender's loss if attacker reaches target 2.

In general, the attacker follows the following logic.

$$\max_{a \in \{n\}} \Psi_{aj}(x_{akj}, d_k) \cdot v_j - c(x_j) \quad \text{(Exp. 7.8)}$$

In the example used above, the objective function of the tullock player T as well as the byzantine player B can thus be represented as follows.

$$u_T = \Psi_{aT}(x_{akT}, d_k) \cdot v_T - c(x_t) = \Psi_1(x_{11}, x_{21}, d_1, d_2) \cdot M_1 - \kappa_1(x_{11} + x_{21}) \quad \text{(Exp. 7.9)}$$

M represents the value of the game for the tullock player (non-monetary value of the achievement). The attacker's gain does not correspond to the defender's loss (in this case, reputation among his customers).

$$u_B = \Psi_{aB}(x_{akB}, d_B) \cdot v_B - c(x_B) = \Psi_2(x_{12}, x_{32}, d_1, d_3) \cdot \beta \cdot L_2 - \kappa_2(x_{12} + x_{32}) \quad \text{(Exp. 7.10)}$$

In the following, we determine the optimal decision of the attackers and the optimal reaction of the defenders. For this purpose, we determine the reaction functions of the attackers and insert them into the target probabilities. From the target function of the defenders, this results in the optimal defense strategy (amount of defense expenditure) shown in figure 7.2 as a function of the amount of damage.

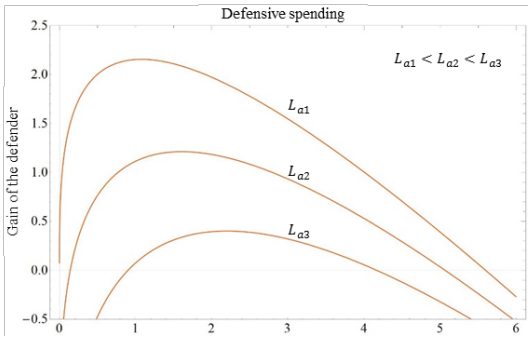

Figure 7.2: Defenders gain in dependence on defensive spending

Figure 7.2 shows that the defender's profit decreases with increasing expected damage of the defender, since the defense expenditure is higher (parametric variation among three values for the resulting damage). Correspondingly, a higher expected damage also results in a higher defense level. The model shows here that information technology monocultures (the attacker will be able to cause greater damage by attacking such a system than in a heterogeneous system) are particularly vulnerable to attacks, require higher defensive spending and are thus suboptimal for the defender. This is because these systems appear particularly attractive to an attacker. The attacker will therefore make a greater effort to crack these systems and their defense will require correspondingly more resources. In the present case with three nodes to be defended (three attack techniques to be prevented) as well as the two attack targets, the optimal resource allocation for the defender as well as the optimal attack intensity for the attacker (Nash equilibria) can be determined and on the basis of these the risk of an attack can be determined. The formal mathematical solution of the game is

given in the appendix. At this point, only the solution of the game presented at the beginning shall be presented (see figure 7.3).

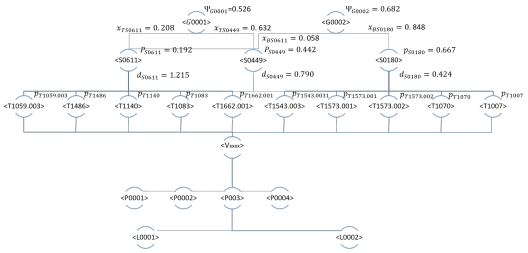

Figure 7.3: Numerical solution for the presented game within an exemplary firm

7.4 Conclusion

The presented quantitative approach on cyber risk management presents a systematic approach for analysing the state of security (single-player game) as well as investments to alter a specific state (two-player game). A significant contribution of our work is that the game can be calibrated relying on the available information on past attacks. This is, the game can be updated automatically to enable an up to date risk estimation. For this purpose, the approach can easily be implemented as an algorithm enabling automated risk assessments for a priori attack risk estimations. A further benefit of using the presented methodology is the familiarity of many decision makers with the established concept of game theory. Furthermore, to the best of our knowledge, this approach is the first of its kind demonstrating the possibility of extending quantitative risk estimations with behavioural aspects in decision making of attackers giving relief to the high burdens of oversimplified attacker modelling in risk quantification (Kaiser et al., 2021a). This is, the presented approach can support security professionals in their daily work providing quantitative security measures. The model thereby considers interdependencies between attackers and defenders yet, the game does not consider interdependencies between different defenders as potential targets of an attack nor between different attackers (e.g. assume an attacker that is motivated by breaching a system as the first one). Extending the model to include these effects between defensive spending of different defenders on target selection of attackers would provide substantial improvements to the model and should therefore be considered as a promising extension.

References

Anwar, A. H., Kamhoua, C. (2020). Game theory on attack graph for cyber deception. , 445–456.

Anwar, A. H., Leslie, N. O., Kamhoua, C., Kiekintveld, C. (2020). A game theoretic framework for software diversity for network security. *International Conference on Decision and Game Theory for Security*, 297–311.

Attiah, A., Chatterjee, M., Zou, Z. (2018). A game theoretic approach to model cyber attack and defense strategies. In *2018 ieee international conference on communications (icc)* (pp. 1–7).

Beyerer, J., Geisler, J. (2016). A framework for a uniform quantitative description of risk with respect to safety and security. *European Journal for Security Research*, *1*(2), 135–150.

Buldas, A., Laud, P., Priisalu, J., Saarepera, M., Willemson, J. (2006). Rational choice of security measures via multi-parameter attack trees. In *Critical information infrastructures security* (pp. 235–248). Berlin, Heidelberg: Springer Berlin Heidelberg.

de Gusmão, A. P. H., e Silva, L. C., Silva, M. M., Poleto, T., Costa, Ana Paula Cabral Seixas. (2016). Information security risk analysis model using fuzzy decision theory. *International Journal of Information Management*, *36*(1), 25–34.

de Melo e Silva, A., Costa Gondim, J. J., de Oliveira Albuquerque, R., García Villalba, L. J. (2020). A methodology to evaluate standards and platforms within cyber threat intelligence. *Future Internet*, *12*(6), 108.

Do, C. T., Tran, N. H., Hong, C., Kamhoua, C. A., Kwiat, K. A., Blasch, E., ... Iyengar, S. S. (2018). Game theory for cyber security and privacy. *ACM Computing Surveys*, *50*(2), 1–37.

Durkota, K., Lisý, V., Bošanský, B., Kiekintveld, C., Pěchouček, M. (2019). Hardening networks against strategic attackers using attack graph games. *Computers & Security*, *87*, 101578.

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. *2019 European Intelligence and Security Informatics Conference (EISIC)*, 40–47.

Ingoldsby, T. R. (2010). Attack tree-based threat risk analysis. *Amenaza Technologies Limited*.

Kaiser, F., Wiens, M., Schultmann, F. (2021). Motivation-based attacker modelling for cyber risk management: A quantitative content analysis and a natural experiment. *Journal of Information Security and Cybercrimes Research*, *4*(2), 132–147.

Kaiser, F. K., Budig, T., Goebel, E., Fischer, T., Muff, J., Wiens, M., Schultmann, F. (2021). Attack forecast and prediction. *Proceedings of the 28th C&ESAR*, 77.

Kamdem, G., Kamhoua, C., Lu, Y., Shetty, S., Njilla, L. (2017). A markov game theoritic approach for power grid security. In A. Musaev (Ed.), *Ieee 37th international conference on distributed computing systems workshops - icdcsw 2017* (pp. 139–144). Piscataway, NJ: IEEE.

Kissoon, T. (2020). Optimum spending on cybersecurity measures. *Transforming Government: People, Process and Policy*, *14*(3), 417–431.

Merrick, K. E. (2016). *Computational models of motivation for game-playing agents*. Cham: Springer International Publishing AG.

Muckin, M., Fitch, S. C. (2014). A threat-driven approach to cyber security. *Lockheed Martin Corporation*.

Nguyen, T. H., Wright, M., Wellman, M. P., Singh, S. (2017). Multistage attack graph security games: Heuristic strategies, with empirical game-theoretic analysis. *Proceedings of the 2017 Workshop on Moving Target Defense*, 87–97.

Panaousis, E., Fielder, A., Malacaria, P., Hankin, C., Smeraldi, F. (2014). Cybersecurity games and investments: A decision support approach. In (pp. 266–286). Springer, Cham.

Rass, S., König, S., Schauer, S. (2017). Defending against advanced persistent threats using game-theory. *PloS one*, *12*(1), e0168675.

Romanosky, S., Ablon, L., Kuehn, A., Jones, T. (2019). Content analysis of cyber insurance policies: How do carriers price cyber risk? *Journal of Cybersecurity*, *5*(1), tyz002.

Sauerwein, C., Sillaber, C., Mussmann, A., Breu, R. (2017). Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*.

Veksler, V. D., Buchler, N., LaFleur, C. G., Yu, M. S., Lebiere, C., Gonzalez, C. (2020). Cognitive models in cybersecurity: Learning from expert analysts and predicting attacker behavior. *Frontiers in Psychology*, *11*, 1049.

Verma, S., Gruber, T., Schmittner, C., Puschner, P. (2019). Combined approach for safety and security. In (pp. 87–101). Springer, Cham.

Wang, L., Islam, T., Long, T., Singhal, A., Jajodia, S. (2008). An attack graph-based probabilistic security metric. *IFIP Annual Conference on Data and Applications Security and Privacy*, 283–296.

Wang, L., Jajodia, S., Singhal, A., Cheng, P., Noel, S. (2013). k-zero day safety: A network security metric for measuring the risk of unknown vulnerabilities. *IEEE Transactions on Dependable and Secure Computing*, *11*(1), 30–44.

Woods, D. W., Böhme, R. (2021). Systematization of knowledge: Quantifying cyber risk. *IEEE Symposium on Security & Privacy*.

World Economic Forum. (2022). The global risk report 2022. *World Economic Forum*.

Zarreh, A., Wan, H., Lee, Y., Saygin, C., Janahi, R. A. (2019). Risk assessment for cyber security of manufacturing systems: A game theory approach. *Procedia Manufacturing*, *38*, 605–612.

Zeller, G., Scherer, M. (2021). A comprehensive model for cyber risk based on marked point processes and its application to insurance. *European Actuarial Journal*.

8 Attack Forecast and Prediction[1]

Abstract

Cyber-security has emerged as one of the most pressing issues for society with actors trying to use offensive capabilities and those who try to leverage on defensive capabilities to secure their assets or knowledge. However, in cyber-space attackers oftentimes have a significant first mover advantage leading to a dynamic cat and mouse game with defenders. Cyber Threat Intelligence (CTI) on past attacks bears potentials that can be used by means of predictive analytics to minimize the attackers first mover advantage. Yet, attack prediction is not an established means and automation levels are low.

Within this work, we present Attack Forecast and Prediction (AFP) which is based on MITRE Adversarial Tactics, Techniques and Common Knowledge (ATT&CK). AFP consists of three modules representing different analytical procedures which are clustering, time series analysis, and genetic algorithms. AFP identifies trends in the usage of attack techniques and crafts forecasts and predictions on future malware and the attack techniques used. We rely on time sorting to generate subgraphs of MITRE ATT&CK and evaluate the accuracy of predictions generated by AFP based on these. Results of an experiment performed on the basis of 493 different malware, validate the utility of using AFP for attack prediction. AFP reaches for each module an F-score which is higher than an extrapolation of observed probabilities (baseline) with an F-score of up to 0.83 for a single module. It can hence be considered an effective means for predicting future attack patterns and help security professionals with preparing for future attacks.

Keywords: Attack Prediction, Cyber Threat Intelligence, Genetic Algorithms

8.1 Introduction

8.1.1 Motivation

Cyber-security has emerged as one of the most pressing issues confronting our globally connected world. The World Economic Forum estimated that the damage related to worldwide cyber-crime

[1] This chapter includes the proceeding article "Attack Forecast and Prediction" by Tobias Budig, Elisabeth Goebel, Tessa Fischer, Jurek Muff, Marcus Wiens, Frank Schultmann, and myself published in the Proceedings of the 28th C&ESAR (2021): Computer Electronics Security Application Rendezvous, November 16-17, 2021, Rennes, France.

was $3 trillion in 2015. This number is expected to increase by 15% every year, reaching $10.5 trillion annually by 2025 (Boden, 2016; Freeze, 2021). Consistently, individuals, businesses and governments are becoming increasingly concerned about the costs and threats presented by cyber-crime, espionage, and cyber-warfare (Oberzaucher, 2019). It is hence expected that the worldwide information security market will reach $170.4 billion in 2022 (Contu, 2018).

Attackers' strategies quickly develop and are subject to dynamic innovations. Consequently, the cyber-criminal world is evolving to exploit vulnerabilities in a faster and more profitable way. To counter this threat, new approaches and investments in the field of cyber-security are essential. "The capabilities, persistence, and complexity of adversarial attacks in the present threat landscape result in an arms race between security professionals, and threat actors." (Mavroeidis and Bromander, 2017) Thereby the attacker seems to have a *first mover advantage*. Thus companies are often vulnerable even to relatively basic assaults on their computer networks.

According to the Global Information Security Workforce Study, the global cyber-security workforce will be short by 1.8 million people by 2022, a 20% increase since 2015 (Frost and Sullivan, 2017). 66% of respondents reported not having enough capacity to address current threats appropriately. The resulting consequences emphasise the importance of gaining knowledge about cyber-attacks to understand adversarial behaviour and increase the efficiency of dealing with threats (Frost and Sullivan, 2017). Understanding and analyzing cyber-attacks that happened in the past and predicting patterns of attacks for the future means an improvement of cyber-security in its ability to enhance one's position in the arms race between adversaries and defenders.

Therefore, predictive analysis could lead to an advantage for organisations to properly allocate their scarce defence resources. Although predicting attacks is not a new procedure, automating attack forecasting and predictions were options hardly used in the past. Rather, attack predictions were largely based on subjective perceptions of experienced experts from the cyber-threat landscape. Yet, experienced experts are rare and their time is even scarcer.

Automation of attack forecasting and prediction would substantially decrease attackers first mover advantage by decreasing biases in predictions and minimise experts' time spending on generating forecasts.

8.1.2 Problem statement

As described, predicting future malware and their functioning is of especial interest. We thereby consider a malware as a unique combination of attack techniques (software vector). Technically predicting future malware is the prediction of (co-)occurrences of techniques in future malware respectively predicting new malware nodes within the cyber-attack knowledge graph. These predictions can increase cyber-security maturity. Currently, predictions mainly focus on short term. However, medium respectively long term predictions are of high importance for security professionals. Consistently, research and development of defensive measures take a lot of time, whereas better prediction can reduce the time lead over attackers. Note that security spending is an investment in the future security of a company and should hence follow the dynamics of attacks guaranteeing the

feasibility of defensive means. Yet, as medium and long term predictions are scarce, research and development of defensive measures are subject to high risk of obsolescence.

CTI (e.g. as it is provided by MITRE ATT&CK)[2] contains information with the potential of making predictions more accurate representing a great opportunity for ensuring cyber-security. However, currently, predictions on future developments of attacks are rare and oftentimes are largely based on experiences of some analysts rather than CTI. This low level of automation regarding the prediction of new attacks causes many problems for cyber-security. First, experts were distracted from their operations and generating predictions means extra workload for them. Second, even the best and experienced experts perceive cyber-attacks only from a limited and in this way subjective perspective. This is, human crafted predictions are prone to biases.

8.1.3 Research question and course of the study

To support cyber-security staff on a strategic level and make predictions on the development of attacks more accurate, we investigate two questions:

1. Are there any patterns or trends in the MITRE ATT&CK that can be used for crafting medium to long term predictions on the threat landscape?
2. How do different algorithms perform to predict future malware in a medium to long term?

To address these questions, we collect data about observed ("historic") malware respectively CTI by scraping information about software and their used techniques from the MITRE ATT&CK database.

In a next step, we conduct a simple statistical analysis to identify the probability of techniques used by a software and the distribution of the number of techniques per software. Building on these results, there is a possibility to provide insights into the most important techniques and how often they have been used in the past. Furthermore, this will provide decision-makers with data instead of intuition to guide their decision-making process. We then predict new malware as a future combination of techniques. Thereafter, we compare approaches to predict impending attacks by fine-tuning the model by propagating trends.

Here we propose, on the one hand, to use genetic algorithms generating malware predictions (Ijaz et al., 2018). On the other hand, we make a dimensional reduction, followed by clustering approaches i.e. hierarchical clustering. A time-series and regression analysis is conducted to identify trends inside these clusters and craft predictions on new malware (Yang et al., 2014; Husák et al., 2018).

To evaluate our algorithms we craft subgraphs from knowledge graphs using time sorting and run the analysis on real data (knowledge at a specific point in time). We evaluate our predictions with walk-forward validation.

[2] For an in depth discussion on CTI and the relation to MITRE ATT&CK refer to section 8.2

In contrast to past work (please refer to section 8.2), we use predictive analytics not on an operational but on a strategic level.

8.2 State of research and related work

8.2.1 Cyber-situational awareness

Situational awareness describes the "perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in near future." (Endsley, 1988) In the context of cyber-security, situational awareness can be divided into three levels (Endsley, 2017). These are (1) monitoring of cyber-systems and intrusion detection, (2) understanding of the current situation and its significance for cyber-security, and (3) projection. The last aspect includes predictive capabilities and is hence the joint link to this work.

8.2.2 Cyber threat intelligence

CTI is structured information extracted from monitored systems or intrusion detection systems (Elitzur et al., 2019). It includes actionable information on past attacks (evidence based knowledge). CTI is often divided into four subcategories: (1) technical, (2) operational, (3) tactical, and (4) strategic threat intelligence (Chismon and Ruks, 2015). It includes tactics, techniques, and attack patterns (TTPs), indicators of compromise (IOCs), tools, threat actors, date of discovery, and other information. Information extracted from various sources of CTI is leveraged by many analysts to increase the efficiency of defensive measures such as anomaly detection systems, intrusion detection systems, and threat hunting (Elitzur et al., 2019).

Since a single individual, security analyst, security researcher or any other expert cannot acquire all information on all threats, there is a high importance of sharing CTI among different stakeholders to enable a holistic perspective (Elitzur et al., 2019). Hence, there were great efforts to formalise and standardise threat sharing and to develop a common language. One of those languages is the Structured Threat Information eXpression (STIX) language [3], which is also utilised by MITRE ATT&CK[4].

MITRE ATT&CK is a globally accessible curated knowledge base and a model about adversarial behaviour in cyber-attacks based on real-world observations. The MITRE Corporation created ATT&CK out of the need to categorize and structure data of adversarial behaviour due to the increasing number and relevance of cyber-attacks.

Through MITRE ATT&CK, a common taxonomy has been created to help understand adversarial behaviour and improving defensive actions. MITRE ATT&CK is used as the foundation for

[3] https://www.mitre.org/capabilities/cybersecurity/overview/cybersecurity-blog/stix-20-finish-line

[4] https://attack.mitre.org/

developing specific threat models by researchers, analysts and developers. The first model was created in 2013, primarily focusing on the Windows enterprise environment. Since then, the database has been extended to other platforms such as Linux, macOS and Android.

The foundation of MITRE ATT&CK is based on various techniques an adversary can use, representing how an opponent will carry out an attack tactic. Each technique is associated with one or more tactics. Tactics can be understood as phases of an attack and are therefore consistent to the cyber kill chain (Hutchins et al., 2011). These tactics answer the question why an adversary uses a particular technique. The sequence of several techniques used during an attack is defined as software. The software itself can be divided into malicious software (Malware) and legitimate software (Tools).

The most recent version of MITRE ATT&CK represents 552 techniques across 13 tactics and 585 different softwares. MITRE ATT&CK consists of three parts: The Enterprise version focuses on adversarial behaviour against enterprises, MITRE ICS version focuses consists of attacks within industrial control systems and the mobile version focuses on attacks against mobile devices.

Furthermore, beside CTI for attacker modelling, there is also information on defender modelling including information on system vulnerabilities. This is for example the National Vulnerability Database (NVD)[5] or the Common Vulnerabilities and Exposures (CVE) database[6]. This information can be used to understand trends, patterns and developments in software vulnerabilities that would affect the threat landscape.

8.2.3 CTI based predictions for cyber-security

Researchers showed how discrete or continuous models and machine learning methods could be applied to the cyber-security sector in recent years. Husák et al. (2018) made a detailed comparison of predictive methods applicable for both long term investigations and forecasts as well as short term predictions, e.g. used for efficient threat hunting in cyber-security and divided them into classes.

8.2.3.1 Short term predictions

Short-term attack projection assists security analysts in identifying the next step of an adversary. One example is the Attack Hypothesis Generator (AHG) by Elitzur et al. (2019) Within their work, they used a knowledge graph of historical malware based on the MITRE ATT&CK, AlienVault Open Threat Exchange (OTX), and VirusTotal. Based on the knowledge graph, they predict subsequent and linked attack techniques given some currently observed data. AHG performed significantly better than an analyst in estimating the next step of an ongoing cyber-attack (Elitzur

[5] https://nvd.nist.gov/
[6] https://cve.mitre.org/

et al., 2019). Within their work, Elitzur et al. (2019) proved that short term predictions can be beneficial for improving cyber-security and setting efficient defensive measures at place.

Furthermore, Zhan et al. (2013) demonstrated the usage of short term attack predictions relying on honeypots. They enabled attack predictions up to five hours ahead based on data gained from their honeypot. Furthermore, Fava et al. (2008) presents a methodology for projecting attacks based on information gathered from Intrusion Detection Systems (IDS).

Each of the techniques mentioned above craft predictions to support security analysts at an operational level in their day-to-day work. Husák et al. (2018) showed that a wide range of prediction methods achieve up to 90% accuracy in recognising adversarial network behaviour.

Further works are inter alia given by Qin and Lee (2004).

8.2.3.2 Medium to long term predictions

Zhang et al. (Zhang et al., 2015) provide a study applying data-mining and machine learning for predicting the "time to next vulnerability for a given software application". However, they concluded that the NVD has only low predictive power. This might be as there is a close link between the occurrence of new attacks and the exploitation of vulnerabilities. Furthermore, Ozment (Ozment, 2007) highlighted that there is not enough information in freely accessible vulnerability databases including NVD. This is, CTI about attacks might have a higher predictive power than data on vulnerabilities. Further contributions investigating the possibility of predicting vulnerabilities include the works from Alhazmi and Malaiya (Alhazmi and Malaiya, 2006), Abraham and Nahir (Abraham and Nair, 2015), and Nguyen and Tran (Nguyen and Tran, 2010).

8.3 Methodology

8.3.1 Hierarchical clustering

Al-Shaer and Spring (2019) proposed agglomerative hierarchical clustering as the most effective means of investigating associations within MITRE ATT&CK. Inspired by their approach we propose the use of hierarchical clustering for finding these associations and predicting yet not existing respectively unobserved links in the usage of attack techniques for crafting predictions on new attacks (new malwares). We thereby calculate for each software the distance to every other software and cluster the most similar ones. The distance between softwares is calculated based on the phi-coefficient. We define the phi-coefficient between software i and software j ($r_\phi(S_i, S_j)$) as follows. For doing so, we formalize each software as a binary vector each bit denotes the utilization or non-utilization of a specific technique.

$$r_\phi(S_i, S_j) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 0}n_{\cdot 1}}} \quad \text{(Exp. 8.1)}$$

where n_{11} describes the number of techniques that are used by both softwares, n_{00} the number of techniques not used by both softwares and n_{10} respectively n_{01} the number of techniques used by either software i or software j. Furthermore, $n_{1\cdot}$ respectively $n_{\cdot 1}$ represent the total number of techniques used by each software and $n_{\cdot 0}$ respectively $n_{0\cdot}$ the number of unused techniques by software i or j. For clustering we utilized Ward's linkage (Ward Jr, 1963) as it showed the best results.

8.3.2 Principal component analysis

To identify hidden factors inside the data, principal component analysis (PCA) is a feasible means (Shlens, 2014). It transforms the original data points to a new orthogonal basis. These basis vectors are sorted by the ratio of variance they cover and then interpreted as underlying factors.

8.3.3 Time series analysis

The time-dependent structure of the attack data implies a time series analysis and forecasting. We propose to use vector autoregression (VAR) (Lütkepohl, 2013). One software is represented by a binary vector of techniques with a date. Each technique can be interpreted as a time series in itself. VAR then regresses a technique to itself and all other previous techniques. Therefore, a one-step-ahead VAR(1) approach needs to fit k^2 parameters, where k is the dimension of techniques.

Given that the distribution of timestamps is highly uneven, we cannot perform a VAR model fitting directly mainly because multiple software have the same time-stamp. Instead, we introduced two different procedures. First, we did ordinary VAR with one software per point in time. Second, we aggregated all software per year and analyzed it to identify trends based on a yearly basis.

Within the first procedure, we extracted all data points where the timestamp is unique. The new data set X_{short} consists of 42 software ranging from *14-02-2019* to *10-06-2020* with non-uniform distributed dates. This is the training set for a VAR(n) time series analysis with a lag of n. We performed six regressions with lags $n \in \{1, \ldots, 5, 15\}$.

Predictions for future software can be made by the forecast method. It generates vectors of float values. We transformed the prediction to software vectors by setting values above 0.5 to one and the rest to zero. We compare the first prediction for evaluation as further steps of prediction generate the same software structure.

The second procedure involved aggregating all software per year. Next, all values are normalised by dividing them by the number of software in the respective year. This table of relative frequencies of techniques per year gets the subject of analysis. A time-series analysis like in the first procedure

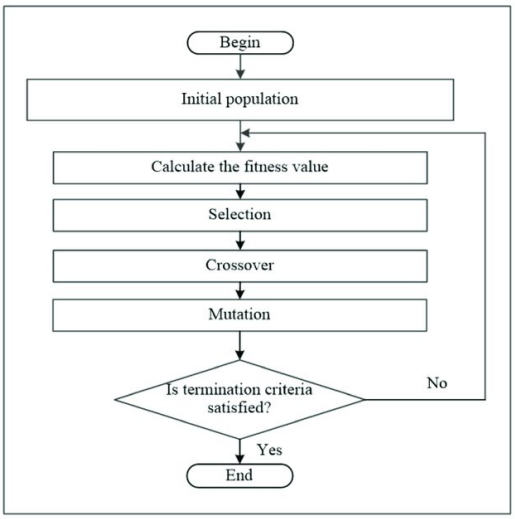

Figure 8.1: Flowchart of a standard genetic algorithm (GA)

is hereby impossible due to the small sample size. However, trends can be identified analyzing the frequencies of techniques used in softwares.

8.3.4 Genetic algorithm

Genetic algorithms (GA) are a class of algorithms based on the biological process of evolution (Whitley, 1994). Yet, the core idea is transferable to many other research areas. Central for this class of analytical procedures is that the evolution of an initial population respectively the development from a given state is an iterative process, improving their fitness.

Figure 8.1 gives the basic structure of the GA according to Höschel et al. (Höschel and Lakshminarayanan, 2019).

8.3.4.1 Initialisation

AFP interprets the observed software as the underlying population where each particular technique is interpreted as a determinant of the fitness.

For the initial population, we use the observed probability of the utilization of a technique. For lowering the computational complexity, the size of the initial population could hereby be lowered. In this sense, the run-time can be decreased without significantly deterioration of the accuracy.

8.3.4.2 Fitness function

The definition of the fitness function is essential for each GA. Within AFP we propose to take three terms into account for describing the fitness function which are the probability term (PT), the correlation term (CT) and the adjustment term (AT).

PT considers the frequency of the utilization of a technique in every observed software as an indicator for their fitness. By doing so, we obtain a probability vector $q \in \mathbb{R}^n$ with the probabilities of all n techniques. The calculation involves multiplying each individual in the respective population $P \in \mathbb{R}^{n \times m}$, where m are the generated softwares with n potential techniques with the vector q.

PT can then be described as follows.

$$\mathrm{PT} = q \cdot P \in \mathbb{R}^m \tag{Exp. 8.2}$$

As some techniques are more likely to appear together or conversely, some techniques are more likely to appear separately, coocurences seem to have an additional information value. This coocurrence of specific techniques can be interpreted as an indicator that they support each other respectively increase the fitness or are even necessary prerequisites. We start with a given generated software vector $p \in \mathbb{R}^n$ with n techniques and the technique correlation matrix $C \in \mathbb{R}^{n \times n}$.

We then substract the unit matrix to remove the self-correlation.

$$\hat{C} = C - E \tag{Exp. 8.3}$$

Furthermore, we multiply p by $\hat{C}$ to identify the correlation value for each technique in the generated software. The resulting vector a with

$$a = \hat{C} \cdot p \in \mathbb{R}^n \tag{Exp. 8.4}$$

is multiplied by the generated software. The sum of correlations k for the used techniques is the result.

$$k = p^T \cdot a \in \mathbb{R} \tag{Exp. 8.5}$$

Lastly, this metric is normalized.

$$\hat{k} = \frac{k}{|p|} \tag{Exp. 8.6}$$

To do this computation for m populations in parallel, we extend the ideas mentioned above as follows.

$$A = \hat{C} \cdot P \in \mathbb{R}^{n \times m} \tag{Exp. 8.7}$$

where $P \in \mathbb{R}^{n \times m}$ is the population matrix of m generated software with n techniques.

Masking the correlation value matrix A again with P results in a $m \times m$ matrix.

$$K = P^T \cdot A \in \mathbb{R}^{m \times m} \tag{Exp. 8.8}$$

CT is then described by the diagonal elements of the matrix $\mathrm{CT} = \mathrm{diag}(K)$.

We furthermore observe that the number of techniques used within a software is relatively stable. Hence, AT is included to account for potential "degenerations" of the predicted software i.e. overfull software vectors. AT thereby takes a key role for determination of the accepted false positive and false negative rates and hence for the accuracy of the algorithm. For this purpose, the mean number of techniques in each software within MITRE ATTA&CK is calculated.

First, we calculate the occurrence $o \in \mathbb{R}^n$ of different techniques for each predicted software from $P \in \mathbb{R}^{n \times m}$ in parallel.

$$o = P \cdot \nVdash \in \mathbb{R}^n \tag{Exp. 8.9}$$

where P is multiplied by $\nVdash$, a vector of ones, so that the occurrence vector o represents the sum of ones within the software vector and thus the number of techniques used per software.

In the next step, the difference between the vector o and the mean value μ is calculated by subtracting μ from all the values in o, which leads to the difference term (DT).

$$\mathrm{DT} = o - \mu \cdot \nVdash \in \mathbb{R}^n \tag{Exp. 8.10}$$

AT penalizes deviations from the mean number of techniques used.

We propose to rely on the following formulation where PST penalizes positive deviations from the mean and BST remunerates negative deviations. PST can thereby be described as follows, where PF denotes the penalty factor.

$$\mathrm{PST} = \frac{d^2}{\mathrm{PF}} \in \mathbb{R}^n \tag{Exp. 8.11}$$

Furthermore, BST can analogously be described as follows where BF is the "bonus factor" for deviations.

$$\mathrm{BST} = \frac{d}{\mathrm{BF}} \in \mathbb{R}^n \tag{Exp. 8.12}$$

Since the differences are negative and the adjustment term AT is subtracted in the final fitness function, the bonus has a positive effect on the fitness. AT thereby works against the tendency of GA to produce overfull software matrices for the predicted softwares. This behaviour happens because it would increase the fitness scores since they are partly based on the occurrence probabilities, and thus more techniques used would lead to higher fitness scores without AT favoring the degeneration of software.

The final adjustment function is now composed of the sum of the two terms *PST* and *BST*.

$$\mathrm{AT} = \mathrm{PST} + \mathrm{BST} \in \mathbb{R}^n \tag{Exp. 8.13}$$

The entire fitness function *FT* consists of the three main terms described above.

$$\mathrm{FT} = (\lambda \cdot \mathrm{PT} + (1 - \lambda) \cdot \mathrm{CT}) - \mathrm{AT}) \tag{Exp. 8.14}$$

Besides the BF and the PF, there is another degree of freedom. The factor λ influences the degree to which the correlation and the relative probabilities of the techniques are included in the fitness function.

8.3.4.3 Selection

The selection function selects from the set of softwares those that should be used within the predictive approach to will most likely be reused and recombinated to generate novel software or in other words that are used by the GA to craft predictions. There are different selection methods. We implemented a roulette wheel selection and a simple tournament selection. The implemented tournament selection is a straightforward method of selection. It involves randomly selecting two individuals from the current population and comparing their fitness scores. The individual with the higher score wins the *tournament* and is included as a so called child in the new generation to be recombinated using the crossover methods described in section 8.3.4.4. This procedure is repeated until an entirely new child generation has been generated.

The roulette wheel selection method considers the relative fitness scores of each individual, which have to be a positive value. Due to the architecture of our fitness function, fitness scores might be negative, owing to which we had to adapt these scores first. We subtract the smallest fitness value in the population from all other fitness values in the respective population. This results in the smallest fitness value becoming zero and all other values correspondingly non-negative. The resulting positive fitness values are then used for roulette wheel selection. This selection method is a fitness proportional selection method, where the individual crossover probability is calculated based on the individual fitness divided by the sum of the fitness of the whole population. The fitness values are normalized so that the sum of the resulting fitness values is one.

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_i} \tag{Exp. 8.15}$$

where p_i is the selection probability of an individual, f_i the respective fitness score and $\sum_{j=1}^{N} f_i$ the sum of fitness in the population.

This selection method can be thought of as a roulette wheel, where each individual takes up an area on the wheel depending on their fitness. The probability of selecting a software depends on the individuals' fitness relative to the rest of the population. Our implementation allows to individually specify the number of softwares that should be selected for crossover.

8.3.4.4 Crossover

The crossover operator is the implementation of recombination in the GA. Pairs of softwares (also called parents) are crossed by exchanging segments of the respective bit strings between the two parent softwares. This creates new softwares (children) based on the genetic makeup of the two parents. The number of crossover points in the software vector is usually one or two, implemented using one-point crossover and two-point crossover (K. A. De Jong and Spears, 1990). However, research has shown that often a more significant number of crossover points can be beneficial (Syswerda, 1989) (K. De Jong and Spears, 1991).

For one-point crossover, two selected softwares are passed and then cut at a random location in the software vector. Two new offsprings are then generated by rejoining the two parent softwares at their intersections. Correspondingly, the two-point crossover method, softwares are cut at two points and then rejoined by combining the respective regions of the parents software vectors to create two new offspring.Uniform crossover, produces on average ($\frac{L}{2}$) crossovers on chromosome strings of length L. We implemented all three crossover operators to test whether the results differ significantly between methods. The uniform crossover function is passed by two parent softwares $A^{(t)}$ and $B^{(t)}$, determined by the selection operator in the previous step, as well as an exchange probability p_s. The function returns two generated children C^{t+1} and D^{t+1} for the next generation of individuals $t + 1$ the exchange of bits is calculated and performed separately for each position of the software vector. First, a random number between 0 and 1 is drawn, resulting in u. If this number is less than or equal to the exchange probability p_s, the respective bit is exchanged (Dominik-Gwiazda, 2006).

$$u \leq p_s \tag{Exp. 8.16}$$

$$c_i^{t+1} = b_i^t \tag{Exp. 8.17}$$

$$d_i^{t+1} = a_i^t \tag{Exp. 8.18}$$

This process is repeated up to the length L of the *parent's* software vector. If $u > p_s$ the bits of the parent softwares are copied to the children without flipping them.

8.3.4.5 Mutation

The mutation function intends to flip bits and thus prevent particular parts of the software vector from being identical in different predicted softwares. This is also to prevent the search for a solution only in a subspace of the original search space. Furthermore, while crossover accounts for the reutilization of software and the recombination of different parts of a software, mutations account for entirely new coding (new developments). Including both, GA seem to come close to real software development processes. Mutations increase the exploratory power of the GA. The probability that a bit mutates is set by a parameter p_m and is usually relatively low.

In our implementation, an array with the shape of the current population is randomly filled with float numbers between 0.0 and 1.0. An element-wise comparison is then performed, and wherever

the values of the randomly generated numbers are below the probability p_m, the corresponding bit is inverted. In the end, a mutated population is returned.

8.3.4.6 Optimizing hyper-parameters

In the previous sections, we defined the parameters BF, PF and p_m. They significantly influence the behaviour of the GA and the trade-off between exploration and exploitation. These hyper-parameters can be optimized regarding better predictive power. A global optimization routine like simulated annealing (Tsallis and Stariolo, 1996) is utilized to produce the results stated in section 8.4.3. While running the optimizer, we fixed the seed for the random number generators to ensure reproducibility. Moreover, we selected the starting values manually and gave them boundaries to restrict search space. The target function of the optimizer is the negative mean of 30 F_1 scores of generated software by the GA.

For comparison, we also tested a local optimization routine. However, the tested limited memory Broyden– Fletcher–Goldfarb–Shanno algorithm (L-BFGS) (Pytlak, 2008) converged to the start values, and stuck in a local minimum.

8.3.5 Generative adversial network

Generative Adversarial Networks (GAN) are a specific type of neural networks (NN) frequently used in deep learning and are trained to produce new instances of objects similar to those they were trained on. This is reached by two neural networks (NN), that compete against each other. The first generates new samples (so called generator), in our case software, out of noise input. The other discovers fake data (so called discriminator) (Goodfellow et al., 2014). After training, the generated software looks like the observed software.

To create a GAN, one need to perform these steps

1. set up the GAN by defining the generator and discriminator,
2. provide samples of real software vectors,
3. create fake software,
4. train the GAN,

which are shown in detail in the following paragraphs.

Due to the small given data set and the exploitative intention, we start with simple models. The generator aims to generate software the discriminator is not able to differentiate from real software. The NN has l input neurons, where l is the dimension of the latent space (we set $l = 50$). There are 150 hidden neurons in the middle layer with relu activation function and 192 neurons in the output layer with linear activation representing all techniques. The discriminator tries to filter out fake software generated by the generator. It is a NN with 192 input neurons, 25 hidden neurons

with relu activation and one output neuron with sigmoid activation. Both networks are using Adam optimizer (Kingma and Ba, 2014).

For training, the GAN is provided by real software samples drawn from the observed data matrix X. Moreover, the generator creates new software out of a multivariate Gaussian noise vector with dimension l. Finally, we train the GAN by training the discriminator on real samples for this batch. Next, the generator creates new software, which the discriminator evaluates. The generator is trained to fool the discriminator. In this setup, we trained batch-wise with the size of 64 and 2000 iterations.

8.4 Evaluation

8.4.1 Database

For evaluation of the proposed methods, we rely on CTI that is presented in MITRE ATT&CK, which is information on 493 malware and 552 unique techniques and sub-techniques these malware rely on.

To be able to train and evaluate AFP, we split all available data on the ratio of 80:20. We therefore sort all data by the date of discovery. 80 % of the malware (the malware discovered first - respectively the oldest malware) is taken for training purposes and 20 % (newest malware) for testing. This results in an evaluation period of approximately one year.

Note the imperfections within the database due to unequal time stamping. This is due to a change in the data structure and represents a burden for evaluation as well as for prediction.

Furthermore, we performed a preprocessing i.e. cleaning of the data. We thereby removed *degenerated* software and techniques, the dimension is reduced by 64.7% from 552 $\times$ 585 to 192 $\times$ 449. Despite that, we kept 95.7% of the information in the data. This shows that the matrix was strongly sparse in the beginning. By reducing the matrix, we kept the information but decreased its complexity significantly.

8.4.2 Experimental setup

8.4.2.1 Evaluation metrics

We measure the performance of the predictions as the harmonic mean of precision p and recall r - commonly known as the F-score. This means, that for evaluation we measure the gap between the actually observed softwares in the one year time period (Y) and those that were predicted with the help of AFP.

$$F = 2 \cdot \frac{p \cdot r}{p + r} \qquad \text{(Exp. 8.19)}$$

Furthermore, we benchmark each module of AFP with a simple simulation of attack evolution (as a basic approach of attack prediction) based on the extrapolation of observed probabilities for the utilization of each technique. For generating the baseline, we generated 100 softwares by random picking of techniques. This basic simulation led to an F-score of 0.42. We take the F-score of this extrapolation as a benchmark.

8.4.2.2 Outline of the experiment

In order to evaluate each implemented module of AFP, we perform a Walk forward evaluation with a time horizon of one year. For every prediction we craft by using AFP relying on the training data set, we test whether there is a real attack within the evaluation data set. We consider the highest parallelism/ correspondence of a predicted software and an observed software as the adequate measure defining the prediction accuracy, where each observed software (software from the evaluation dataset) can only be chosen once. In doing so, we evaluate the extent to which AFP is feasible to predict the development of new attacks within the time horizon of one year which we define as a medium to long term horizon for attack predictions.

We thereby consider the four strategies for crafting attack predictions representing the different modules implemented within AFP.

Algorithm 5 describes the schema of the experiment. Line 1 calls the selected procedure for crafting predictions implemented within AFP. Line 2 to 4 calculates the precision and recall of each prediction for a single prediction and an observed attack. In Line 5, the evaluation metric is calculated for the prediction crafted and line 6 calculates the evaluation metric of the AFP on Y.

Algorithm 5 Evaluation procedure

Input: $DatasetX, DatasetY, AFP\_approach$
Output: $software\_vector, evaluation\_metric$

1: $Predictions \leftarrow GenPredictions(AFP\_approach, DatasetX)$
2: **for** $Pred \in Predictions$ **do**
3: $\quad precision \leftarrow Get\_precision(Pred, DatasetX, DatasetY)$
4: $\quad recall \leftarrow Get\_recall(Pred, DatasetX, DatasetY)$
5: $\quad evaluation\_metric(Pred) \leftarrow Get\_metric(precision, recall)$
6: $evaluation\_metric \leftarrow Mean(evaluation\_metric(Pred))$
7: **return** $evaluation\_metric, evaluation\_metric(Pred)$

Algorithm 6 shows the procedure of AFP. VAR is called and executed with line 1 Line 3 calls and executes the GA, parameters are set, and the respective selection, crossover, fitness and mutation functions are called, compare Figure 8.1. Figure 8.2 shows the best fitness score in each generation and the course of the GA. Last, in line 6 the GAN is called.

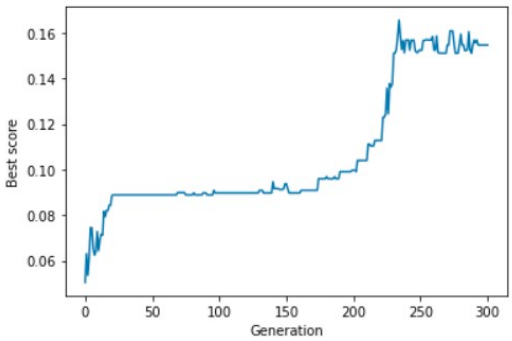

Figure 8.2: Evolution of the fitness scores in different generations of the GA

Algorithm 6 GenPredictions

Input: $DatasetX, num\_predictions, AFP\_approach$
Output: $Predictions$

if AFP_approach = VAR **then**
    $Predicitions \leftarrow GenPredictions\_VAR(num\_predictions, DatasetX)$
else if AFP_approach = GA **then**
    $Predicitions \leftarrow GenPredictions\_GA(num\_predictions, DatasetX)$
else
    $Predicitions \leftarrow GenPredictions\_GAN(num\_predictions, DatasetX)$
return $Predictions$

After initializing the start population , the algorithm executes the GA up to a predefined number of generations. Another essential part of the algorithm is the choice of the replacement strategy. This strategy defines how newly created individuals are selected or what proportion of the *parents* should be replaced by the *children*. The choice is crucial because, in addition to the crossover, mutation and selection functions, it can improve the balance between exploration and exploitation of the algorithm. However, there is no general best replacement strategy, as the choice depends on many problem-specific factors (Wu et al., 2014). Generally, a distinction is made between generational (non-overlapping) GAs and steady-state (overlapping) GAs. In generational GAs, the entire parent generation is replaced by a completely new offspring generation. In contrast, in steady-state GAs the new generation consists of parts of the parent generation and the newly created offspring.

We decided to use a general replacement strategy, where the offspring generation replaces the parent generation. The disadvantage of this method is that possibly good individuals from the previous generation are lost, and thus the average fitness of the total population decreases. However, this also means that the chance of finding less good individuals (local optimum) is lower. In our case, the population's average fitness is secondary, and avoiding being 'stuck' in a local optimum is more critical.

8.4.3 Results

In a first step we applied, hierarchical clustering sorting the software based on the similarity or dissimilarity to one another. The resulting dendrogram is presented in figure 8.3. It becomes evident, that there are clear clusters that can be extracted from the dataset (types of malware). This could enable prediction approaches within clusters.

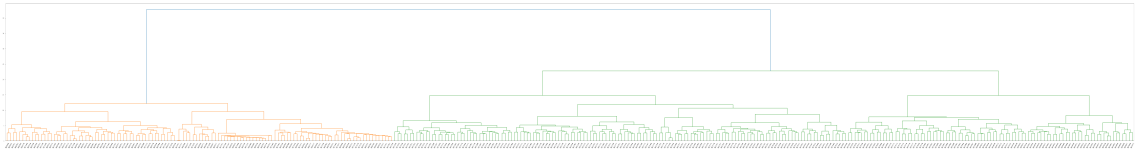

Figure 8.3: Plot of agglomerative hierarchical clustering of software attacks

Thereafter, we used the data of the clusters to apply PCA. In this case, shown in fig. 8.4, it was not possible to recognise a structure of temporal developments, and the data points were distributed randomly.

Table 8.1 presents the results for $n \in \{1, \ldots, 5, 15\}$ lags for the VAR(n) time series model. The mentioned table shows the corresponding average, maximum and minimum F-score. The average ranges from 0.46 to 0.61, peeking at a lag of five. VAR shows with a lag of five a maximal F-score of 0.83 and minimal F-score of 0.4. VAR hence shows in this implementation an improved predictions on future malware for nearly every prediction with only insignificant deterioration even for the worst prediction crafted by the VAR module compared to the baseline.

In addition, for the one-step-ahead forecast, the F-score and the index of the hit entry in the test set are shown. For all lags, the first F-score was higher than average. Therefore, it could be used as a heuristic.

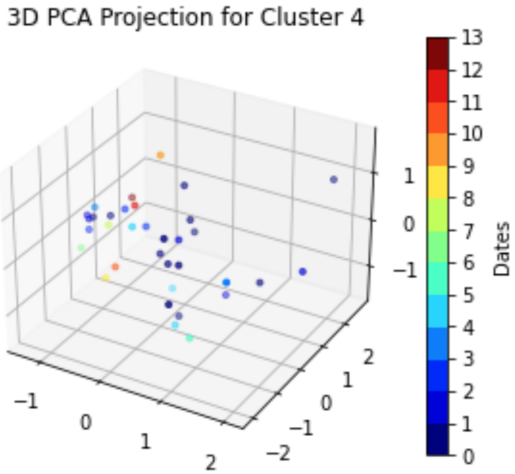

Figure 8.4: Software clusters with 3 dimensions after performing PCA, colored with reference to the respective timestamps

Lag n	**1**	**2**	**3**	**4**	**5**	**15**
F-score avg	0.46	0.52	0.55	0.52	0.61	0.50
F-score min	0.27	0.0	0.29	0.0	0.40	0.0
F-score max	0.73	0.74	0.71	0.74	0.83	0.80
F-score first	0.53	0.58	0.67	0.50	0.62	0.67
# Techniques	20	10	6	4	7	5

Table 8.1: Comparison of results of VAR(n) first prediction for different lag sizes.

Figure 8.5 gives an overview on the performance of the different modules of AFP. It shows a clear superiority of AFP compared with the baseline (simulation based on the extrapolation of observations).

GAN performed with an average F-score of 0.46, slightly higher than the simulation. Since we had 3500 parameters and 350 data points, the risk of over-fitting the model was high.

GA performed best of all methods explored. For evaluation, we executed the GA 50 times, with a maximum generation size of 450 and a population size of 40 individuals each time. Here we used the optimized parameters (see section 8.3.4.6), although these differ depending on the crossover and selection methods used.

In the resulting 50 final generations each with 40 predicted software, we selected the best one, i.e. the one with the highest fitness score. A selection of e.g. the five most successfully predicted software led to equivalent results since the five best software from each run of the GA hardly differed in their binary structure.

We used the reduced matrix with 192 techniques in the GA. With the help of the tournament selection and uniform crossover operator, we achieved a mean F-score of 0.63.

With the roulette selection, against our expectations, a slightly worse F-score of 0.58 was achieved. In addition, the run-time performance of the roulette selection was worse. Fortunately, the variance of the individual F-scores was relatively low, with scores varying only between F-values of 0.52 - 0.8 ensuring better estimations than the extrapolative approach (benchmark) for every prediction. We used uniform crossover and one-point crossover for predictions with the GA, as the results with two-point crossover was — with an average F-score of 0.53 — significantly worse. However, the two-point crossover produced a larger selection of different software, while the predicted software in the other two methods was similar for several runs of the GA.

8.4.4 Discussion

With AFP we present an automated attack predictor able to forecast new malware developments, where each malware is described as a software vector defining the techniques that are used by the novel malware. We showed and validated the predictive power of GA, VAR and GAN based on a Walk forward validation of predictions. AFP can therefore be considered as an efficient means

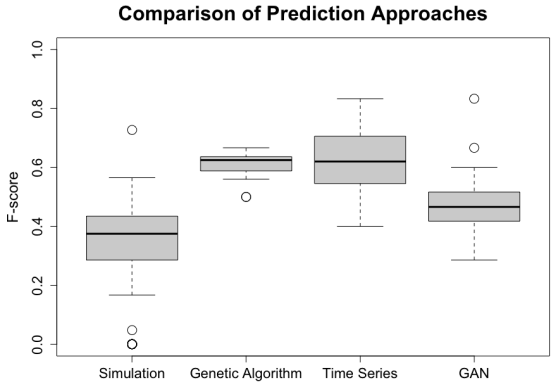

Figure 8.5: Boxplot showing the comparison in terms of the F-score results of different prediction approaches. We used the simulation as a baseline and ran the Time Series with VAR(5).

for predicting software developments within a time horizon of a year. It needs to be noted, that predictions are not exact rather, real software deviates slightly from most predictions. Yet, predictions show considerable descriptive power. Furthermore, AFP helps to increase cyber-security maturity by significantly improving medium to long term predictive capabilities. Furthermore, AFP is completely automated saving (time) resources of security professionals and security operation centers. The results of AFP can be used to prepare proactively for future attacks and improving investment decisions/ research spending of defenders. This is, defenders can better prepare for attacks utilizing specific attack techniques by implementing adequate counter measures. Furthermore, AFP identifies trends in malware development respectively within the utilization of techniques of new malicious software. Security professionals can rely on this information and develop security measures for the most probable predictions even before they where observed *in the wild*. AFP can hence also be considered to be an enabler for engineering secure systems within a dynamically changing threat landscape.

Especially GA seems to be suitable for predicting the development of new malware in the medium to long term (respectively as evaluated in this work on the basis of one year). This is as the GA shows very robust results with little variance in the predictive power as well as the highest mean F-score. Furthermore VAR showed high predictive power. Yet, it suffers from high variance in precision of predictions. Although GAN promises high potential for medium and long term attack prediction, it showed the worst results of the implemented algorithms yet delivering significant improvements compared to the extrapolation of observations suffering most from the limitations of low availability of data on past attacks. This is as AFP implements in its current version a GAN that is trained on a relatively small dataset (X). Yet, the training set of the GAN increases with every newly detected malware. This is, we expect the accuracy of the GAN module to increase significantly with the size of CTI utilized.

Likewise, PCA suffers from the lack of data respectively the structure of the database it was extracted. In the future, when data with different time stamps are available, it will most likely become possible to gain deeper insights with this approach.

8.5 Conclusion, impact & future work

Improving defensive capabilities in cyber-space for improving cyber-security is one of the key challenges that need to be solved to enable resilient societies and modern life, which is increasingly penetrated by information technology.

Understanding the past and predicting the future is an approach being sought in the course of time to develop new security profiles and software to help protect socially sensitive data and critical infrastructure from attackers. Predicting future cyber-attacks can help businesses, individuals and society. Minimising attackers first mover advantage therefore need to be a focal point of research. Yet, it is barely considered or largely based on subjective opinions and biased by individual perspectives of experts.

In this work, we present AFP taking advantage of CTI for automatically crafting predictions on future malware and their attack techniques used. In doing so, AFP leverages on CTI to infer patterns within time series of attacks, making it possible to gain insights to attack evolution and development as well as deriving further relevant information (e.g. the popularity of specific attack techniques) and crafts forecasts based on this information. AFP thereby predicts new attacks as a binary vector of techniques and provides hence actionable insights on the probable development of the cyber-threat landscape. In this way, analysts, researchers and security managers can prepare proactively for threats that are likely to occur in the future. Additionally, cyber-risk managers can perform intelligent and proactive investments relying on AFP, as well as staff training to minimise the attackers' first mover advantage. The ability of AFP to predict the future course and the development of the threat landscape is a critical step towards increasing levels of cyber-defence and security as well as its automation. The development of an automated prediction process is an essential step towards the strategic defence against cyber-attacks and can significantly increase cyber-security maturity for non-specific attacks. In the long run, this anticipation of attacks can be expanded and used for successful attack prevention.

Within our work we identified clusters and patterns that can be used for crafting medium to long term predictions on future attack development empowering security analysts in classifying upcoming software, discovery and reaction to trends. Furthermore, these trends promise useful guidance in strategic actions taken by defenders e.g. security investment decisions. With AFP, we automated the process of crafting attack medium to long term attack predictions and forecasts. The different graph analytical approaches employed by AFP show high potential for attack prediction. Yet, GA and VAR show the most promising results.

Although reaching reasonable good results in medium to long term attack prediction, AFP suffers a lack of data. Hence, it seems to be of utmost importance to gain and share CTI.

References

Abraham, S., Nair, S. (2015). A predictive framework for cyber security analytics using attack graphs. *arXiv preprint arXiv:1502.01240*.

Alhazmi, O. H., Malaiya, Y. K. (2006). Prediction capabilities of vulnerability discovery models. In *Rams'06. annual reliability and maintainability symposium, 2006.* (pp. 86–91).

Al-Shaer, R., Spring, J. (2019). *Automating reasoning of mitre att and ck for predicting cyber attack techniques using statistical machine learning* (Tech. Rep.). Carnegie Mellon University Software Engineering Institute Pittsburgh United . . . .

Boden, P. (2016, Jan). The emerging era of cyber defense and cybercrime. *Microsoft Security*. Retrieved from `https://www.microsoft.com/security/blog/2016/01/27/the-emerging-era-of-cyber-defense-and-cybercrime/`

Chismon, D., Ruks, M. (2015). *Threat intelligence: Collecting, analysing, evaluating* (Tech. Rep.). MWR InfoSecurity.

Contu, R. (2018, Sep). *Forecast analysis: Information security, worldwide, 2q18 update.* Gartner Research. Retrieved from `https://www.gartner.com/en/documents/3889055`

De Jong, K., Spears, W. (1991). On the virtues of parameterized uniform crossover. In *Proceedings of the 4th international conference on genetic algorithms* (pp. 230–236).

De Jong, K. A., Spears, W. M. (1990). An analysis of the interacting roles of population size and crossover in genetic algorithms. In *International conference on parallel problem solving from nature* (pp. 38–47).

Dominik-Gwiazda, T. (2006). *Genetic algorithms reference, volume i, crossover for single-objective numerical optimization problems.* Poland, Tomasz Gwiazda.

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. In *2019 european intelligence and security informatics conference (eisic)* (pp. 40–47).

Endsley, M. R. (1988). Situation awareness global assessment technique (sagat). In *Proceedings of the ieee 1988 national aerospace and electronics conference* (pp. 789–795).

Endsley, M. R. (2017). Toward a theory of situation awareness in dynamic systems. In *Situational awareness* (pp. 9–42). Routledge.

Fava, D. S., Byers, S. R., Yang, S. J. (2008). Projecting cyberattacks through variable-length markov models. *IEEE Transactions on Information Forensics and Security*, *3*(3), 359–369.

Freeze, D. (2021, Apr). *Cybercrime to cost the world $10.5 trillion annually by 2025.* Retrieved from `https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/`

Frost, A., Sullivan, P. (2017). *2017 global information security workforce study.*

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Höschel, K., Lakshminarayanan, V. (2019). Genetic algorithms for lens design: a review. *Journal of Optics*, *48*(1), 134–144.

Husák, M., Komárková, J., Bou-Harb, E., Čeleda, P. (2018). Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Communications Surveys & Tutorials*, *21*(1), 640–660.

Hutchins, E. M., Cloppert, M. J., Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, *1*(1), 80.

Ijaz, S., Hashmi, F. A., Asghar, S., Alam, M. (2018). Vector based genetic algorithm to optimize predictive analysis in network security. *Applied Intelligence*, *48*(5), 1086–1096.

Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.

Mavroeidis, V., Bromander, S. (2017). Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In *2017 european intelligence and security informatics conference (eisic)* (pp. 91–98).

Nguyen, V. H., Tran, L. M. S. (2010). Predicting vulnerable software components with dependency graphs. In *Proceedings of the 6th international workshop on security measurements and metrics* (pp. 1–8).

Oberzaucher, B. (2019, Jan). Spektrum. Retrieved from `https://www.andritz.com/spectrum-en/latest-issues/issue-39/digitalization-as-a-megatrend`

Ozment, J. A. (2007). *Vulnerability discovery & software security* (Unpublished doctoral dissertation). University of Cambridge.

Pytlak, R. (2008). *Conjugate gradient algorithms in nonconvex optimization* (Vol. 89). Springer Science & Business Media.

Qin, X., Lee, W. (2004). Attack plan recognition and prediction using causal networks. In *20th annual computer security applications conference* (pp. 370–379).

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Syswerda, G. (1989). Uniform crossover in genetic algorithms. In *Proceedings of the third international conference on genetic algorithms* (pp. 2–9).

Tsallis, C., Stariolo, D. A. (1996). Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*, *233*(1-2), 395–406.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing*, *4*(2), 65–85.

Wu, Y., Liu, J., Peng, C. (2014). A new replacement strategy for genetic algorithm and computational experiments. In *2014 international symposium on computer, consumer and control* (pp. 733–736).

Yang, S. J., Du, H., Holsopple, J., Sudit, M. (2014). Attack projection. *Cyber Defense and Situational Awareness*, 239–261.

Zhan, Z., Xu, M., Xu, S. (2013). Characterizing honeypot-captured cyber attacks: Statistical framework and case study. *IEEE Transactions on Information Forensics and Security*, *8*(11), 1775–1789.

Zhang, S., Ou, X., Caragea, D. (2015). Predicting cyber risks through national vulnerability database. *Information Security Journal: A Global Perspective*, *24*(4-6), 194–206.

9 Too Stressful to Look Closely? The Information Value of Signal Detection under Cognitive Contraints – A Decision-Theoretic Model for the Case of Phishing Mail Detection[1]

Abstract

In recent years, phishing mail detection gained rising scientific consideration as well as high practical relevance for oranizations and employees. Detecting these "bait-mails" requires not only competence on the part of the user but also a stress-free environment which allows for a permanent level of awareness. In the stressful everyday life characterized by time pressure and information overload, it is a big challenge to keep an eye on these threats without losing the attention and concentration for the own work.

To gain a better understanding of this trade-off between high cognitive effort and the potential risk, we model the phishing detection problem in a simple decision-theoretic framework based on Signal Detection Theory (SDT). The approach enables a distinction between the 'naive' baseline decision, which is made on the basis of the priori-probabilities (subsisiary awareness) and a diagnostic decision based on a so-called Information System (focal awareness). For general occurrence probabilities, damage costs, opportunity costs as well as cognitive costs of the user, the Information Value of a diagnosis is formally derived and analyzed. The focus is on the *depth of diagnosis*, i.e. the question of how many phishing mail indicators ('observables') the user should consult in order to make an efficient decision (in the information economic sense).

We find that consulting multiple observables can lead to an inefficient diagnosis if phishing attacks constitute a low probability - high impact -risk. Moreover, the information value of a phishing detection decision strongly depends on the interaction of the cognitive effort (like stress or

[1] This chapter includes the preprint of the article "Too stressful to look closely? The Information Value of Signal Detection under Cognitive Contraints – A Decision-Theoretic Model for the Case of Phishing Mail Detection" by Marcus Wiens, Frank Schultmann, and myself.

time pressure) and the user's competence in phishing mail detection. Competence can partially compensate the user's cognitive constraints by two different effects – a parsimony-effect and a preservation-effect.

The findings from the model analysis are helpful in understanding the behavior of users who can quickly become overwhelmed with the demands placed on them (high job performance and high risk awareness simultaneously) in their everyday work. Furthermore, they have implications for the question under which conditions organizations should invest in awareness measures (such as anti-phishing training).

9.1 Introduction

Cybersecurity gains in importance for organizations as well as for individuals in today's increasingly connected and digitalized world. This is also acknowledged by the World Economic Forum, which mentioned cyberattacks together with data fraud and theft among the main threats to business performance worldwide (Franco, 2020). Phishing attacks in particular have become a major threat to many companies (Shahbaznezhad et al., 2020). In contrast to technical malware, phishing attacks directly target human characteristics (e.g. pursuit for social affiliation). Phishing therefore can be described as confidence trickery with the aim to get access to personal details or credentials (Tjøstheim and Waterworth, 2020).

Although many phishing mails can often be detected even by less IT-savvy users based on relatively easy-to-identify characteristics, there is often a lack of knowledge and vigilance. Against the backdrop of usually high time pressure and a high number of e-mails in everyday office life, this inattention may well represent rational user behavior.

This study sheds light on the interrelation between the cognitive constraints of the user (e.g. due to time pressure, lack of concentration or inattention due to stress), the user's competence in phishing mail detection and the risk inherent in phishing mails which can imply a significant inconvenience for the user like loss of time, loss of data, reputational loss within the organization or even a threat of losing the job. Our approach is a decision-theoretic model which applies techniques from information economics and is essentially based on Signal Detection Theory (SDT). First, we show under which conditions it is advantageous for the decision maker to make a quick or 'naive' decision based solely on priori probabilities. This decision mode is based on the premise of bounded rationality of the decision maker and corresponds to the condition Polanyi (1965) calls 'subsidiary awareness'. As an alternative to baseline decision making, the decision maker may take the (cognitive) effort to inspect the decision base more closely. This corresponds to signal detection. From an information economics' perspective, the decision maker employs an Information System (IS) that enables her or him - depending on prior knowledge or competence - to evaluate the signals via Bayesian updating and thus arrive at a more precise decision. This decision mode corresponds to the state which Polanyi (1965) calls 'focal awareness'.

Based on this model, we can then derive the information value of the decision, which quantifies the difference between the benefits and costs of information used. We are particularly interested in the question of how many indicators or observables the user will or should include in the decision. Phishing emails can be identified by a number of characteristics, but for which the decision maker needs both some prior knowledge and focused attention.

The model allows conclusions to be drawn as to the parameter constellation for which a high level of competence, i.e. mediated by phishing training (Volkamer et al., 2018, 2016), is particularly effective because it is protective of the decision. Furthermore, the model shows how stress at work, represented in the model by an increase in cognitive costs, can affect the risk of the user. A certain paradox comes into play here: precisely in those constellations where competence could compensate to some extent for the lack of focused attention, the willingness to conduct anti-phishing training will tend to be low - precisely because of a lack of time and stress.

The remainder of the paper is organized as follows. In section 9.2 we briefly outline the state of the art in the overlapping research areas of information economics, signal detection theory and phishing mail detection. In section 9.3 we present the baseline-model and in section 9.4 we derive the information value for the case of one observable and extend it to the general case of multiple observables in section. We summarize and discuss our results in section 9.5.

9.2 State of the art: phishing mail detection as a decision under risk and cognitive constraints

9.2.1 Cognitive models for phishing mail detection

Human cognition modelling incorporates decision theory, the theory of bounded rationality, behavioral economics, and cognitive psychology (Caplin, 2016). Differently to other models where different ways of information processing and information gathering are considered, we rely on rational inattention modelling for describing the bounded rational nature of human information processing (Sims, 2010; Gabaix, 2019; Sims, 2003). Thereby it is assumed, that rational agents just acquire information with a positive information value (Matějka and McKay, 2015).

Taking an information economics perspective, the phishing mail decision of a DM can be modelled as a statistical inference process. The prediction of inductive inference is also central to understanding the effectiveness of phishing attacks and thus quantifying the risks. For this, the clues with which the DM estimates the trustworthiness of a mail (signaling variants). "Sign reading is a fundamental part of deciding whether to trust. Correspondingly, the deliberate use of signs, or signaling, is a fundamental part of making oneself appear trustworthy." (Bacharach and Gambetta, 2001) Whether a DM trusts in a specific mail is not only dependent on individual predispositions and inferred signals but also on incentives (Wiens, 2013). Psychological and sociological research indicates that boundedly rational decision making defined by ineffective processing of information and limited cognitive resources are reasons for individuals to fall victim of phishing attacks

(Vishwanath et al., 2018). Arrow et al. (1949) provided the first scientific study where human decision making was dependent on the sampling of information and hence on the attention users devote to information assessment. Since then many scientific publications elaborated on the role of cognition and rational inattention for decision making (Rogers, 1975; Petty and Cacioppo, 1986; Eagly and Chaiken, 1993).

One notable theory applied to cyber security behavior is the Elaboration Likelihood Model (ELM). Thereby two information processing routes have been identified, the central and peripheral route (Petty and Cacioppo, 1986). With respect to phishing mail detection, authors argue that users who use the peripheral route and focus on simple persuasion cues are particularly vulnerable to phishing mail attacks (Vishwanath et al., 2011). The Heuristic Systematic Model (HSM; (Eagly and Chaiken, 1993)) is another model used to evaluate cognitive processes in the context of phishing mail detection (Vishwanath et al., 2018). Applying this theory, it is possible to account for the "cognitive, preconscious, and automatic processes that potentially leads to phishing-based deception" (Vishwanath et al., 2018).

A model that takes into account the limited interaction between deceiver and recipient in the context of phishing mail detection is the Theory of Deception (Johnson et al., 1992) focusing on cognitive processes of information processing employed when a message is received and processed by the recipient (R. T. Wright et al., 2010). The process of recognizing a deception is divided into four stages: Activation, deception hypothesis generation, hypothesis evaluation and assessment. In the activation phase, individuals pay attention to specific threatening information and try to identify irregularities based on previous experience. In the second step, the target persons try to use their previous knowledge to create interpretation hypotheses to explain the irregularities so that these can be evaluated and compared afterwards. Finally, a subjective evaluation of the established hypotheses is carried out depending on individual prior knowledge (Wang et al., 2012; Johnson et al., 1992; Grazioli, 2004).

In his model of fraud detection, Grazioli (2004) established an extension of the model of deception theory. He suggested that certain clues are crucial for the detection of deception. These include exaggerated claims, implausible scenarios, bad grammar and errors, and other indicators that are not expected by the DM (Grazioli, 2004). R. T. Wright et al. (2010) tested this theory as an explanation for the process of detecting phishing attempts using 446 subjects who were experimentally victims of a phishing attack. In a detailed report, the authors present the results of how the test subjects processed the phishing e-mail and made their decisions, which on overall confirmed the fraud detection model of Grazioli (Grazioli, 2004).

9.2.2 Influences on the precision of phishing mail detection

The efficiency of phishing mail detection is highly dependent on the sophistication of attacks. Highly sophisticated attacks thereby provide significantly higher success rates as they feel more familiar and promise higher incentives to the DM (Alam and El-Khatib, 2016). By this means, attackers target the human nature and exploit human psychology to generate trust (Ferreira and

Lenzini, 2015). Attackers may use personal information that was shared by targeted individuals and are available to them (e.g. social media information that is not protected from public view or is online accessible to everybody).

These sophisticated, targeted attacks are also called spear phishing. Hence, spear phishing can be defined as “a deceptive attack that uses social engineering to obtain confidential information through targeted victimization” (Unchit et al., 2020). The efficiency of a DM in identifying phishing mails is thereby highly dependent on the DMs ability to identify the relevant information, (time) pressure, and the two types of errors in a statistical sense (erroneously trusting a malicious mail as well as rejecting a harmless mail).

According to Halevi et al. (2013) DMs’ information sharing behavior on social media may have an influence on phishing mail efficiency. In this sense, online and publicly available information is used by attackers for specifying phishing attacks (Alam and El-Khatib, 2016). Consistently, posting more information online “may cause them to be more susceptible to privacy attacks” (Halevi et al., 2013). However, the amount of online and oftentimes publicly available information is constantly increasing which may further increase the efficiency and sophistication of phishing attacks.

9.2.3 Effects of pressure on phishing mail detection

When interacting with cyberinfrastructures, users are often asked to act, reminded or informed to behave in a certain way (Chowdhury et al., 2019). An increasing number of contributions address the extent to which certain factors affect people’s cyber security behavior and may lead users to ignore security prompts. Time pressure has been found to be a decisive factor in this area (Chowdhury et al., 2019). Time pressure leads to a higher cognitive load and impairs the user’s ability to concentrate. Wang et al. (2012) find that visual deception indicators such as "now" or "urgently" influence the selective attention of the user and may reduce the ability to process information. In addition to visual cues, trust cues are also indicators that influence the decision-making behavior of users under time pressure (Kirlappos and Sasse, 2011). A study by Kirlappos and Sasse (2011) found that when evaluating the credibility of a phishing e-mail or phishing website, users looked for signs that they believed confirmed the trustworthiness of the site. Rather than focusing on security warnings, participants’ attention was drawn to the trusted designs, marks, seals or advertisements, although none of the participants could explain what protection these indicators actually sought (Kirlappos and Sasse, 2011). Thus, on the one hand, it was found that references or links to other websites evoke positive emotions in online shoppers by inspiring confidence and causing users to ignore security warnings (Kirlappos and Sasse, 2011). On the other hand, especially among employees working under time pressure, researchers also found that employees showed anger or frustration due to interruptions in their primary duties or delays caused by authentication barriers (Beautement et al., 2008). As a result, meeting cybersecurity requirements often holds users back from their primary work task, which can lead to the perception of cybersecurity as a work barrier. (Guo et al., 2011). Accordingly, users tend to perceive the costs of meeting cybersecurity requirements as much higher than the expected benefits, thus inducing them to adopt insecure cybersecurity behavior (Chowdhury et al., 2019).

Time pressure results from the existence of a time limit or time constraint for completing a task and can therefore be defined as "the objective or subjectively perceived limitation of the available time needed to review information or make a decision" (Giger and Pochwatko, 2008). Some researchers have already investigated the effects of time pressure on people's decision-making behavior. Much research has shown that a compromise between speed and accuracy can occur under time pressure and that the speed of information processing is increased under time pressure (which may however also mean to not process information entirely but only superficially) (Edland and Svenson, 1993; Zur and Breznitz, 1981). P. Wright (1974) argues that due to time pressure, decision makers make their decisions on the basis of less information and more emphasis is placed on the negative consequences of the decision. The use of simpler and non-compensatory strategies, including the filtering of information in decision making, is also considered a consistent result of research (Miller, 1960; Payne et al., 1988; Svenson et al., 1990). In the literature, non-compensatory decision strategies refer to decision heuristics that result from a (partial) lack of complete relevant information in the decision and are therefore prima vacie by some researchers considered as less "rational" than compensatory decision strategies that take into account all relevant aspects of options when evaluating decisions (Payne et al., 1988; Zakay, 1985). However, as indicated by rational inattention theory, it may be rational to use decision heuristics or make decisions without reviewing all information.

9.3 Methodological approach

We take an approach from Information Economics (Bikhchandani et al., 2013) and Signal Detection Theory (SDT) (Swets, 2000, 1988; Green and Swets, 1966) and assume that an office employee – the user – daily receives a large number of emails and also has to answer them during the working day. Each of these emails can be a phishing email (email-type B for "Bad") or a normal email (email-type G for "Good"), where the probability for a phishing email is a priori $q_0 \equiv Prob(B)$. For the ease of exposition we suppress the index 0 for the priori probability and just write q. A phishing mail contains a dangerous link or attachment, which – when clicked – causes malware to be installed on the user's computer. This causes damage to the user, which is reflected by the negative utility $-L$ (disutility). Hence, the damage occurs in the case of a classical type-II-error in the statistical sense. However, if the mail is a good one and contains important information or material for the user's job, avoiding to click can lead to delay or otherwise wrong decisions, which corresponds to a statistical type-I-error. In this case, the user suffers a loss in the amount of $-l$. We assume that clicking on a dangerous mail leads to greater losses than not clicking on a harmless mail, $L > l$. Without any further knowledge or detection abilities, the user has to evaluate any mail containing links and attachments based on the priori-probability q alone. Equation 9.1 represents the user's expected loss Eu (or 'disutility' in economic terms) for the decision to click (index K) and not to click (index N) together with the critical threshold for the priori-probability.

$$(i)\ Eu_K = q(-L)\ ; Eu_N = (1-q)(-l)\ \ (ii)\ Eu_K > Eu_N \Leftrightarrow q > \tilde{q}_0 \equiv \frac{l}{l+L} \qquad \text{(Exp. 9.1)}$$

For $q > (\tilde{q}_0)$, the expected loss of clicking is higher than the opportunity cost of not clicking. In this case the user preferred the safe option to ignore the email. However, for $q < \tilde{q}_0$, the user would tolerate the potential threat by a phishing mail because the priori-probability is lower than the critical threshold $\tilde{q}_0$.

A phishing e-mail tends to be recognizable as such on the basis of a number of indicators, whereby the chance of detection depends both on the professionalism of the attacker in terms of designing the e-mail to be as inconspicuous as possible, and on the user's phishing detection skills. From the user's point of view, phishing mail detection is thus a classic *signal detection problem* (SDP).

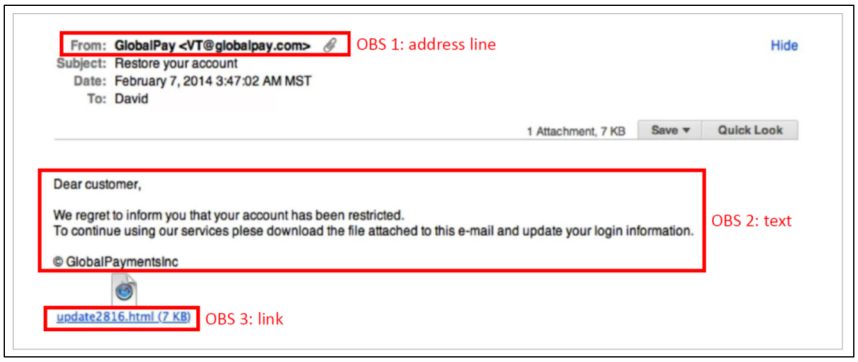

Figure 9.1: Phishing-Mail-Indicators; illustration for three observables (OBS)

Figure 9.1 shows three important indicators of a phishing e-mail: the subject of the e-mail, the wording of the text (especially the salutation), and the link itself. Indicators of this kind are referred to as signals or observables (OBS) in information economics, SDT, and the world of CTI. By looking closely at each of these observables, the user can gain better inference about the actual type of mail and thus make a more informed decision. From a statistical point of view, the evaluation of the observables corresponds to the Bayesian updating process: The decision maker (DM) observes the respective signal and evaluates it based on her prior knowledge about the informative value of the signal using the likelihood function, which finally leads to the posteriori probability, the improved probability estimation about the type of mail (phish versus no phish).

$$(i)\ f_B = f(x_i|B)\ \ (ii)\ f_G = f(x_i|G) \qquad \text{(Exp. 9.2)}$$

Expression 9.2 represents the likelihood density functions (LDF) for observable j and signal expression x for both, a bad and a good email. LDF tells how likely it is to receive a specific signal expression x given that it is either a bad or a good email. The variable $x \in [0, 2]$ determines the expression or "appearance" of the observable and is set by the attacker (designer of the phishing mail). A high x close to 2 represents an extremely suspicious feature, which makes it relatively easy to identify as a phishing mail. A low value of x, e.g. close to 0, corresponds to an almost completely unsuspicious appearance. However, in order for an observable to be evaluable by the user at all, the user needs to be competent in identifying signals. This competence of the user is determined by the parameter $c \in [0, 1]$. A very high c close to 1 represents a user with a great deal of experience, for example because she has already taken part in anti-phishing training courses or works in the IT

security field herself. A very low value of c, on the other hand, represents an extremely low level of competence, in which the user does not know what to do even with very conspicuous observable expressions due to a lack of knowledge. The signal detection capabilities are thus determined by the *competence-difficulty gap*, which in our case can be specified as $(2 - cx)$.

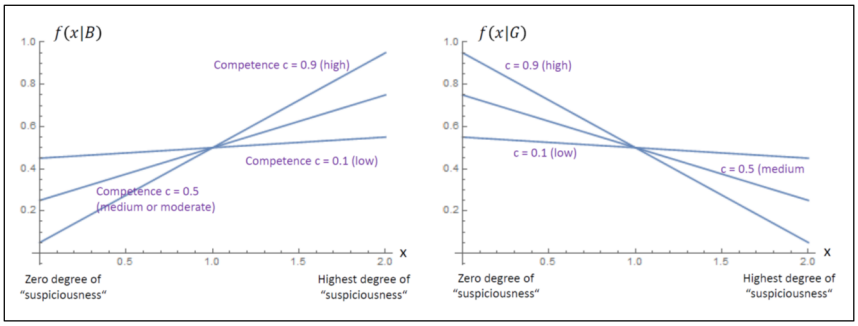

Figure 9.2: LDFs for a bad (B; left) and good (G; right) email and different competence-levels

Figure 9.2 on the left shows three different likelihood density functions for a bad email, each depending on three competence levels of the user. The curve for $c = 0.9$ is steepest and shows that this user is highly capable in identifying suspicious emails. A highly competent user reacts particularly sensitively to conspicuous features along the x-spectrum and concludes with certainty that a feature $x = 2$ indicates a phishing e-mail. With decreasing competence c, this curve flattens out further and further and, in the extreme case of a user who is completely incompetent in terms of phishing mail recognition $(c = 0)$, the curve lies horizontally. This means that a completely incompetent user cannot distinguish any feature characteristics: The probability that any observation x is a phishing mail is always as high in the eyes of this user as if it were not, i.e. the conditional likelihood is 0.5.

The graph on the right side of Figure 9.2 illustrates the analogue LDF for the case of a good email. In this case, there is a negative slope because a harmless mail is only very unlikely to contain suspicious signals (high x -values). By contrast, low x -values (implying no or negligible anomalies) can be expected with a high probability in the case of a good email. Again, the more competent a user is, the more clearly she will be able to establish this connection for herself.

Before we turn to the effects of signal detection, we first need to derive the average expectation of f, $\bar{f}_{B,G}$ for both a good mail and a bad mail respectively. Since our approach is to determine under which conditions signal detection, i.e., more detailed analysis of observables, is worthwhile for the decision maker, we first need the expectation for such a signal. DM cannot know beforehand, which signal she is going to observe later (if she observes a signal, she already decided for signal detection). We get the expectation for the 'average signal' for both types of emails by integration according to expression 9.3 (i). Inserting into f gives the expected or average likelihood for both types of emails 9.3 (ii).

$$(i)\ \bar{x}_j|G = \int_0^2 f_G \cdot x_j\ ;\ \bar{x}_j|B = \int_0^2 f_B \cdot x_j\ \ (ii)\ \bar{f}_G = f(\bar{x}_j|G)\ ;\ \bar{f}_B = f(\bar{x}_j|B) \quad \text{(Exp. 9.3)}$$

9.4 Bayesian updating and information value

The monotonicity condition of the Likelihood-functions from 9.2 ensure that the information system (IS) used here – the system of Likelihood density functions for each type of mails – is efficient in the information economic sense: By using the IS, a decision maker will never get worse, since the IS always contributes to the reduction of the larger statistical error (error of type I or type II). Thus, the informed decision based on the inspection of the observables always generates a positive *gross information value* V. In general, the information value of an activity based on information acquisition results from the comparison of the informed decision with the decision that would be made without the information. Information value is an economic concept because it looks at the extent to which the information changes the decision maker's resulting payoff.

Finally, however, it must also be considered that information acquisition and information processing involve costs. These costs may be material or purely cognitive. In our case, we assume that the inspection of the observables is associated with increased cognitive effort and this is linear in the number of observables considered. The decision maker must use both time and attentional resources to analyze the observables, which she must subtract from other activities. If the total number of potentially analyzable observables is $\theta > 0$, then the DM incurs a cognitive cost $\theta \cdot \kappa$ $(\kappa > 0)$. If the information costs are subtracted from the gross information value, then the *net information value* $v = V - \theta \cdot \kappa$ is obtained. v allows for the judgement whether – after all – the acquisition and processing of information is worthwhile for the decision maker.

In the following, we will derive and graphically illustrate the gross and net information value for the simple case with only one observable.

9.4.1 Information value for one observable

We consider first the constellation $(q > \tilde{q}_0)$.

In this case, due to the high priori-probability for a phishing mail, the decision maker would never click on any attachment or link if the decision is just based on the priori-probabilities. How could an IS improve this decision?

Assume that DM observes a favorable signal x^L (index L stands for 'low'), hinting rather at a good mail than a bad mail. She can now either click (strategy K) or refuse to click (strategy N). The expected payoff resulting from the click-strategy is given by equation 9.4. With probability $(1 - q)$ it is a good mail, in this case clicking was just right and there is no loss (zero payoff). However, with probability q it is a bad mail, which will just be detected with a rather low probability $f_B(x^L)$: The conditional probability to observe a favorable signal x^L in the case of a bad email is rather low, as can be seen on the left side of the left graph in Figure 9.2. Nevertheless, the DM clicked on a link or attachment of a bad email, leading to a loss of $(-L)$. The expected payoff resulting from the no-click-strategy is given by expression 9.4 and can be derived in an analogue way.

$$Eu(x^L, K) = (1-q)f_G(x^L)0 + qf_B(x^L)(-L) \tag{Exp. 9.4}$$

$$Eu(x^L, N) = (1-q)f_G(x^L)(-l) + qf_B(x^L)0 \tag{Exp. 9.5}$$

$$Eu(x^L, K) > Eu(x^L, N) \Rightarrow \tilde{q}_0 < q < \tilde{q}_{1H} \equiv \frac{l}{l + L\frac{f_B(x^L)}{f_G(x^L)}} \tag{Exp. 9.6}$$

To figure out whether we can trust the favorable signal and click on the email, the inequality of expression 9.6 must be fulfilled (clicking should provide a higher payoff, i.e. a lower loss, than refusing to click). It can be seen that this inequality is fulfilled below the critical threshold $\tilde{q}_{1H}$, which constitutes an upper bound on q. It can easily be verified that $\tilde{q}_{1H}$ is indeed an upper bound. The ratio $f_B(x^L)/f_G(x^L)$ is called *positive Likelihood-Ratio* (LR^+) and this ratio is smaller than 1 due to $f_B(x^L) < f_G(x^L)$. Hence, we derived $Eu(x^L)^* = qf_B(x^L)(-L)$ as the optimal, expected outcome for this first constellation.

Next, we assume that DM observes an unfavorable signal x^H (index H stands for 'high'), hinting rather at a bad mail than a good mail. Again, we set up the expected payoffs for the two options K and N and just substitute x^L by x^H (see expressions 9.7 and 9.8).

$$Eu(x^H, K) = (1-q)f_G(x^H)0 + qf_B(x^H)(-L) \tag{Exp. 9.7}$$

$$Eu(x^H, N) = (1-q)f_G(x^H)(-l) + qf_B(x^H)0 \tag{Exp. 9.8}$$

$$Eu(x^H, K) < Eu(x^H, N) \Rightarrow \tilde{q}_0 < q \tag{Exp. 9.9}$$

For our IS to be reliable, the DM should be better off by distrusting an unfavorable signal, which requires $Eu(x^H, K) < Eu(x^H, N)$. It is straighforward to show that this condition holds for $\tilde{q}_0 < q$, which was our initial assumption. For the case of an unfavorable signal we get $Eu(x^H)^* = (1-q)f_G(x^H)(-l)$ as expected payoff, which minimizes the DM's losses.

In a next step, we combine the two constellations. This gives us the loss-minimal, expected payoff for both signal constellations if we use the IS: $Eu(IS)^* = qf_B(x^L)(-L) + (1-q)f_G(x^H)(-l)$.

For the (gross) information value V we need to substract the optimal, expected payoff of the baseline-decision problem without IS according to expression 9.1: For $\tilde{q}_0 < q$, the optimal baseline decision is N (no click) and therefore the expected loss just consists of a potential opportunity cost $(-l)$: $Eu^* = (1-q)(-l)$. The resulting V is shown by expression 9.10.

$$V(\tilde{q}_0 < q) = Eu(IS)^* - Eu^* = [1 - f_G(x^H)](1-q)l - qf_B(X^L)L \text{ for } \tilde{q}_0 < q < \tilde{q}_{1H} \quad \text{(Exp. 9.10)}$$

The (gross) information value V is easy to interpret as it consists of two terms. The first term reflects the upside of the IS, which consists of a reduction of the statistical type-I error: For $\tilde{q}_0 < q$, the user will never click without an IS. This generates a maximum of opportunity cost as many important (and probably also urgent) attachements of harmless mails are ignored by the DM. With an IS, this problem is drastically reduced. However, never clicking on any email-attachments or links reduces the type-II-error to zero. In this regards, an IS inserts some risk into the decision. As can easily be verified, the positive effect of a reduction of the type-I-error more than compensates the additional risk due to the type-II-error.

It remains to look at the opposite constellation $q < \tilde{q}_0$. The expected loss-minimal payoff $Eu(IS)^*$ is not affected by the constellation but just the optimal baseline decision without IS. However, when evaluating the optimal decisions to 'click' or 'no click' for $q < \tilde{q}_0$, we get a new lower bound $\tilde{q}_{1L}$ as a critical threshold according to expression 9.11.

$$Eu(x^H, K) < Eu(x^H, N) \Rightarrow \frac{l}{l + L\frac{f_B(x^H)}{f_G(x^H)}} \equiv \tilde{q}_{1L} < q < \tilde{q}_0 \quad \text{(Exp. 9.11)}$$

Again, it is straightforward to verify that $\tilde{q}_{1L}$ constitutes a lower bound. The ratio $f_B(x^H)/f_G(x^H)$ is called *negative Likelihood-Ratio* (LR^-) and this ratio is larger than 1 due to $f_B(x^H) > f_G(x^H)$.

For $q < \tilde{q}_0$, the optimal baseline decision is K (click) and therefore the expected loss corresponds to the larger loss L due to malware: $Eu^* = qL$. Combining all results, leads us to the (gross) information value according to expression 9.12.

$$V(q < \tilde{q}_0) = Eu(IS)^* - Eu^* = [1 - f_B(x^L)]qL - (1-q)f_G(x^H)l \text{ for } \tilde{q}_{1L} < q < \tilde{q}_0 \quad \text{(Exp. 9.12)}$$

Also this expression is easy to explain. The first term corresponds to the reduction of the type-II-error, which would be maximal without an IS (for $q < \tilde{q}_0$, the DM always clicks). However, the second term represents the drawback of the IS: The fact that the information system enables the decision maker to act more cautiously in this constellation, inevitably leads to a certain degree of the type-I-error. Again, the avoided loss of the first term more than compensates the additional opportunity cost of the type-I-error.

Figure 9.3 shows a plot of V with respect to variations in the priori-probability q. The applied parameter-values for the plot are $l = 1$, $L = 4$, $f_G(x^H) = 0.1$ and $f_B(x^L) = 0.1$. The corresponding critical thresholds are $\tilde{q}_{1L} = 0.03$, $\tilde{q}_0 = 0.2$ and $\tilde{q}_{1H} = 0.69$. It can be seen that V is zero for the leftmost area $q \in [0, \tilde{q}_{1L}]$ and the rightmost area $q \in [\tilde{q}_{1H}, 1]$. In these areas, the priori-probability q is so extremely low (or extremely high) that the naive baseline decision

cannot be improved by the IS. V takes its highest value at the critical threshold $\tilde{q}_0$. Thus, the IS is particularly valuable when DM has the highest uncertainty and is thus indifferent between clicking and not clicking.

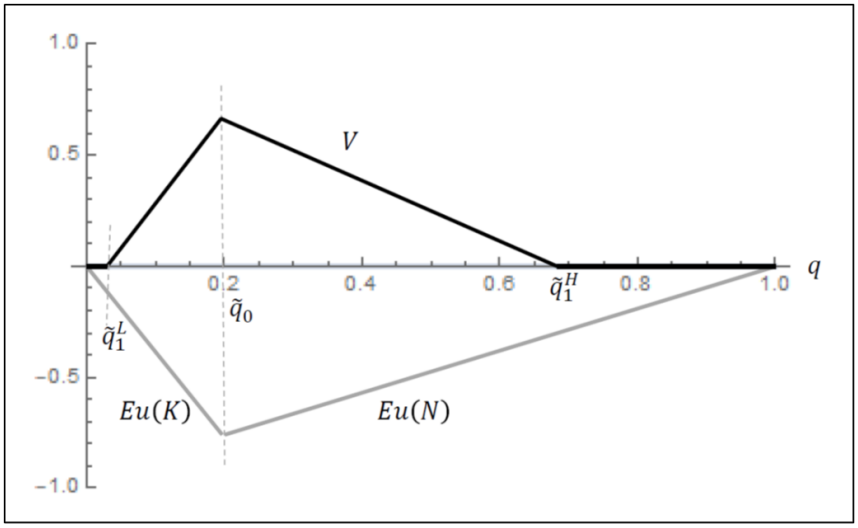

Figure 9.3: Baseline decision (loss area) and Information Value V dependent on q

For this reason, the economic application of SDT is similar in principle to the still frequently applied concept of entropy according to Shannon (1948), but with the important difference that in our model the information value can be derived endogenously from the decision problem itself.

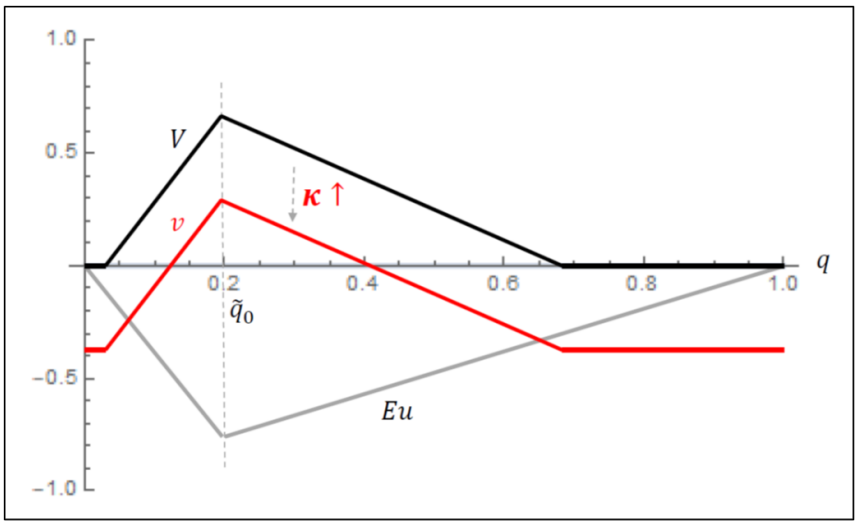

Figure 9.4: Gross Information Value V versus Net Information Value v

As already said, the value of information is determined not only by the potential for improvement with respect to the decision made (this corresponds to V), but also by the costs of information acquisition and processing. If we now assume that cognitive costs of $\kappa > 0$ are incurred by using IS, then the red curve results from a downward parallel shift of the black curve and corresponds to the net information value v (Figure 9.6). Two aspects change, as can be seen in the diagram: First, in the extreme areas (leftmost and rightmost), we now have a negative information value

$(V = 0, v < 0)$. In these areas IS is useless and a costly application just generates costs. From a decision-theoretic perspective, this is the main reason why naive decision routiness and boundedly rational procedures can be highly efficient in a context of lack of time and information overflow. Second, the information value-triangle drops down, which cuts-back the q-spectrum where the IS can improve the baseline decision. Hence, using a IS in the sense of making a decision of high effort and focal awareness (Polanyi, 1965) can be of limited benefit but easily entail substantial costs and thus lead to an inefficient decision on overall.

With this tension in mind, it is no surprise that increasing the user's competence-level c is the most effective lever to reach a positive net information value. If we use the LDFs as introduced in section 9.3, calculate the average expectation $\tilde{f}(c) = \frac{1}{2} - \frac{1}{6}c^2$ and substitute this value in the expression of v, we can express the information value as a function of the user's competence-level c according to expression 9.13. We focus on the area to the left of the critical threshold; however both areas lead to the same result.

$$v(c) = (0.5 + \frac{c^2}{6})qL - (0.5 - \frac{c^2}{6})(1-q)l \qquad \text{(Exp. 9.13)}$$

It can be seen immediately that the competence-level enters the likelihoods quadratically. In addition, it can be easily verified, $\frac{d^2v}{c^2} > 0$, which confirms the overproportional influence of the competence-level (compared to the mere linear influence of the cognitive cost κ, at least according to our assumptions).

9.4.2 Information value for more than one observable

In the previous analysis, the information system consisted of only one observable together with the two LDFs for a good and a bad email. Finally, we want to analyze the case of multiple observables and we assume three observables as an example. As long as the considered observables exert a stochastically independent influence on the information value, their joint effect is multiplicative. Expression 9.14 exemplifies for the case $q < \tilde{q}_0$ the net information value for three observables with stochastically independent signals. The cognitive costs κ are assumed to be identical for all three observables.

$$v(q < \tilde{q}_0) = [1 - \prod_j^3 f_{Bj}(x^L)]qL - (1-q)\prod_j^3 f_{Gj}(x^H)l - 3\kappa \qquad \text{(Exp. 9.14)}$$

Adding one observable is formally equivalent to the multiplication with one further LDF. The multiplicative link increases the overall precision of the IS because it widens the spectrum of the diagnosis. According to expressions 9.6 and 9.11, we can directly derive the new lower and upper bound for the case of multiple observables, which are given by expression 9.15.

$$(i)\ \tilde{q}_{3H} \equiv \frac{l}{l + L\frac{\prod_j^3 f_{Bj}(x^L)}{\prod_j^3 f_{Gj}(x^L)}} \quad (ii)\ \tilde{q}_{3L} \equiv \frac{l}{l + L\frac{\prod_j^3 f_{Bj}(x^H)}{\prod_j^3 f_{Gj}(x^H)}} \tag{Exp. 9.15}$$

The widening of the diagnostic spectrum can also be seen graphically in Figure 9.5. Interestingly, this effect does not occur uniformly on the left and right sides of the critical threshold $\tilde{q}_0$. Since the critical threshold in this example is very low, the left-sided roof of the v-triangle has an increasingly steep slope, while the right-sided roof has an increasingly shallow slope.

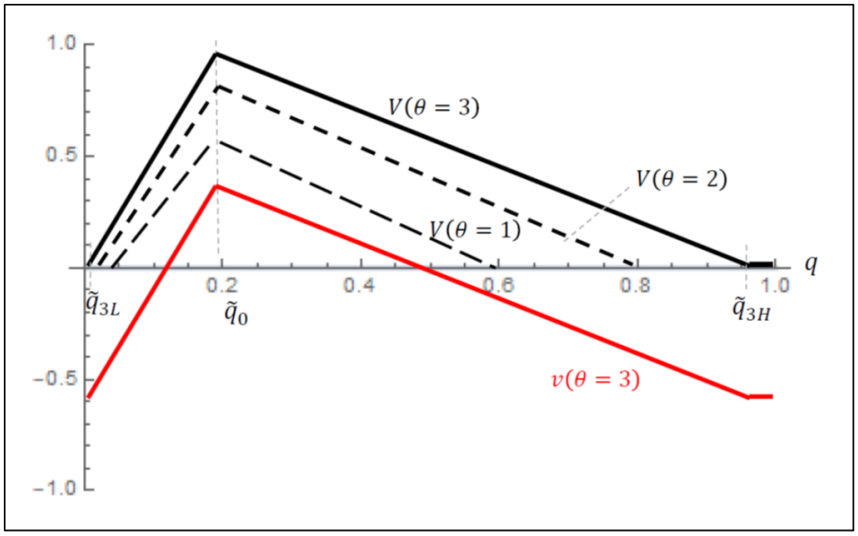

Figure 9.5: Gross Information Value V versus Net Information Value v for multiple observables

This asymmetry also has effects on the shift downside, which takes place to a greater extent in the case of more observables. The diagnosis-spectrum to the left shrinks for two reasons: First, due to the steep slope to the left and second due to the shift downwards. This has implications for the diagnosis-behavior in the area $q < \tilde{q}_0$, which is a risky area because the baseline decision here is to 'click'. Adding more observables shows the effect that the DM no longer differentiates even for larger values of q. For the area to the right oft he critical threshold, the widening-effect and the drop partly compensate each other.

For the presented model it is even possible to derive the optimal number of observables in closed form. To reach this goal, we make use of the fact that the LDFs from Figure 9.2 for a good and a bad email are symmetric. As the DM needs to make an average guess about the expected signal for each type of email, the LDFs for high x-values with a good email and low x-values with a bad email coincide: $f_B(x^L) = f_G(x^H)$. This property can be used for expression 9.16, which represents the net information value for an arbitrary number θ of observables.

$$v(q < \tilde{q}_0, \theta) = [1 - f_B^\theta(x^L)]qL - (1-q)f_G^\theta(x^H)l - \theta\kappa \tag{Exp. 9.16}$$

Setting $f_B(x^L) = f_G(x^H) = f$, taking the first derivate with respect to θ and solving the first-order condition for θ leads to the optimal number of observables θ^* (expression 9.17), which maximizes the net information value.

$$\frac{dv(q < \tilde{q}_0, \theta)}{d\theta} = 0 \Rightarrow \theta^* = -\frac{Log[-\frac{Eu^* \cdot Log[f]}{\kappa}]}{Log[f]} \;\; \theta \in \mathbb{N}_0 \qquad \text{(Exp. 9.17)}$$

It is instructive, to analyse θ^* with a joint look at the parameters c and κ. The effect of increasing cognitive cost κ is intuitive and can be confirmed by expression 9.18 (i).

$$(i)\; \frac{d\theta^*}{d\kappa} = \frac{1}{\kappa \cdot Log[f]} < 0 \;\; (ii)\; \frac{d\theta^*}{dc} \begin{cases} > 0 \vee \kappa > 0.3679 Eu^* Log[f] \\ \leq 0 \vee \kappa \leq 0.3679 Eu^* Log[f] \end{cases} \qquad \text{(Exp. 9.18)}$$

The higher κ, the lower the number of observables the DM takes into consideration for her diagnosis (note that $Log[f] < 0$ due to $0 < f < 1$). However, the effect of the user's competence-level is two-sided according to expression 9.18 (ii). If κ is lower than a critical threshold $\tilde{\kappa} \equiv 0.3679 Eu^* Log[f]$, the optimal number of observables θ^* shrinks in c. The reason for this effect is that the cognitive cost are low enough that even a DM with a rather low competence-level can compensate this deficit by a more differentiated view on the observable phishing mail-indicators. If however, the cognitive cost rise above the threshold $\tilde{\kappa}$, the DM would reduce θ to save cognitive resources. In this case, a higher competence-lecel effectively works against this reduction and allows the DM a more powerful diagnosis despite high cognitive effort. The critical threshold mainly depends on the expected minimal loss of the baseline decision (Eu^*) and the Logarithm of the LDF. Note that $\tilde{\kappa} > 0$ due to $Log[f] < 0$ and Eu^*. For the given calibration of our model $(\bar{f}(c) = \frac{1}{2} - \frac{1}{6}c^2)$, the parameter c has influence on $\tilde{\kappa}$ just within a certain range. If the cognitive cost fall above the upper bound of this range $\kappa > \tilde{\kappa}(c = 1))$, the competence-level has no longer any effect because this parameter is already at its maximum. In this case, the DM will choose the minimal number of observables $\theta^* = 0$ independent from c. In the opposite case, the cognitive cost can be lower than the threshold even for the lowest possible competence-level $\kappa < \tilde{\kappa}(c = 0)$. Also in this constellation, the competence-level has no influence on the optimal decision: The DM picks the maximum number of available observables $\theta^* = 3$, again independent from c.

Figure 9.6 illustrates the joint effect of competence-level and cognitive cost for different parameter constellations. In all four diagrams, the priori-probability for a bad email is lower than the critical threshold $(q < \tilde{q}_0)$, i.e. the user's naive baseline-decision would be to 'click'. The upper-left diagram shows that κ is the more influential parameter on the choice of observables. With increasing κ (starting with three observables at $\kappa = 0$), the DM successively drops the observables and finally falls back to the naive baseline-decision without diagnosis. Nevertheless, a small influence of c can also be identified, which is in line with the findings already discussed above: For lower values of κ, higher user competence can lead to the situation that an observable is no longer needed. This can be seen in the falling contour curves, and we term this effect the *parsimony*-effect of user-competence. However, for higher values of κ, increased user competence rather prevents

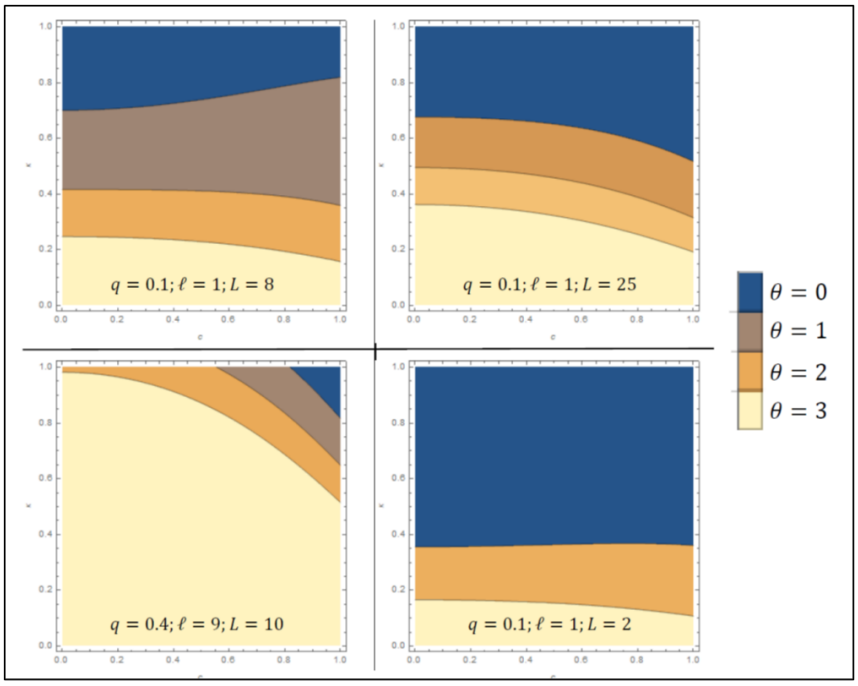

Figure 9.6: Interaction-effects of competence c (horizontal) and cognitive cost κ (vertical)

the loss of single observables and can – albeit to a very limited extent – countervail the effect of high cognitive burden. This can be seen in the rising contour curves, and we term this effect the *preservation*-effect of user-competence.

The diagram at the bottom right has a similar pattern in terms of the nearly horizontal contour lines. However, the loss L due to malware is drastically reduced in this constellation. As in this case, the user faces a significantly lower risk, involving multiple observables in the decision is not worthwile. Hence, the user reacts more sensitive to already small increases of κ and switches back to the naive baseline-decision already at moderate levels of κ. The influence of user competence on this decision is neglegible in this low-risk scenario: There is a very small parsimony-effect and a barely visible preservation-effect.

This clearly changes in the upper right diagram, which is a high-risk scenario. Here, the loss due to malware is set on a very high level $(L = 25)$ compared to the opportunity cost l. In this constellation, the sensitivity of the decision with respect to the competence-level is highest. There is no parsimony-effect but a clear preservation-effect over the entire κ-spectrum. We can conclude that in the context of high cognitive effort and high risk inherent in the phishing mails, user competence can most effectively preserve the use of an information system and thus protect the degree of precision required for an effective screening under risky circumstances.

Finally, the diagram at the bottom left shows medium high values for the two types of losses but a significantly larger priori-probability $(q = 0.4)$. In this constellation, we also have a high risk – this time due to the high probability – but also a high level of uncertainty for the DM because the absolute values of the losses are relatively high and quite close (errors in either direction will result in a relatively high loss). Due to the high level of uncertainty, the IS generates a high information

value which makes the use of a very precise IS worthwhile even for high values of κ. Just in the area of very high κ-values, there is a clear preservation-effect again.

9.5 Results and discussion

Phishing emails are not only annoying, but also dangerous, as they now regularly and with high intensity permeate daily private and professional email communication. Detecting these mails requires not only competence on the part of the user but also the ability to pay attention as well as a stress-free environment to routinely counter this danger. In this paper, a simple decision-theoretic and information-economic approach based on signal-detection theory was used to investigate under which conditions a diagnosis with "focal awareness" is worthwhile for the decision maker and on which factors the information value of the decision depends.

First, model-based critical thresholds were derived by which the diagnostic range is extended. The cognitive costs have a dual effect on the diagnostic decision, which significantly limits its effectiveness: Ranges of negative information value at the extremes and a narrowing of the range of positive information value. The competence level of the decision maker is a very effective lever for increasing the net information value, as it increases disproportionately in c. This has important implications for the question of whether organizations should afford the increased effort to train their employees against phishing attacks. With respect to an individual worker or user, a paradox may arise if they are exposed to increased stress levels: Due to the increased stress level, they overly trust the error-prone baseline decision. The same stress level that prevents informative diagnosis will also prevent the user from taking the time for phishing training, which might just be the most effective antidote to reduce error frequency in the long term. Stress with a very narrow window-width for focal awareness can thus become a double-risk-trap for the respective user.

Multiple observables with uncorrelated signals have a slightly different effect on the decision. First, diagnostic decisions with multiple observables are significantly more time-consuming for the decision maker, which causes the net information value to drop more significantly than in the one-observable case. At the same time, however, there is an asymmetric effect on the left and right sides of the critical threshold: While the right "roof surface" of the information-value-triangle becomes flatter and thus allows for differentiated decisions in the no-click region of the baseline, the left roof surface becomes steeper. This means that the possibilities for more differentiated diagnosis in the click region of the baseline decrease even faster with increasing cognitive costs compared to the case of just one observable. To address this problem, one would need to help the decision maker focus on a few particularly informative observables.

Further simulations for the specific case of maximal three observables (identical in terms of signal effect) showed that under certain conditions a higher competence of the user exerts two different effects for increasing cognitive cost: A *parsimony*-effect, where the higher competence enables the user to make a sufficiently good diagnosis with a fewer number of observables, and a *preseravtion*-effect, where higher user competence perseveres the attained level of diagnostic precision (which would be lost otherwise because the higher cognitive effort would force the user to get along with

fewer observables). The two effects constitute two different ways to compensate a loss in diagnostic accuracy.

In our analysis, we have assumed that the three observables enter the information value stochastically independent of each other. This assumption is quite questionable, since the individual techniques do not differ too much from each other. However, if the observables are sufficiently strongly correlated, then this would have the advantage of significant savings in cognitive costs for the users, since it might then be sufficient to consider only one single observable. In this case, the problem shifts to the question of which observable this is. If there is no one-size-fits-all answer to this question that provides a helpful guide for the baseline decision, the correlation is of little help.

Finally, from an empirical point of view, it would be very important to better understand how the relation of cognitive costs on the one hand and the information value on the other hand looks quantitatively approximate. This is particularly important for risks such as phishing emails, which integrate into our daily routines and can in turn have a strong impact on our attention, stress levels, and thus job satisfaction.

References

Alam, S., El-Khatib, K. (2016). Phishing susceptibility detection through social media analytics. In *Proceedings of the 9th international conference on security of information and networks* (pp. 61–64).

Arrow, K. J., Blackwell, D., Girshick, M. A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica*, *17*(3/4), 213–244.

Bacharach, M., Gambetta, D. (2001). Trust as type detection. In *Trust and deception in virtual societies* (pp. 1–26). Springer.

Beautement, A., Sasse, M. A., Wonham, M. (2008). The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 new security paradigms workshop* (pp. 47–58).

Bikhchandani, S., Hirshleifer, J., Riley, J. G. (2013). The analytics of uncertainty and information. Cambridge University Press.

Caplin, A. (2016). Measuring and modeling attention. *Annual Review of Economics*, *8*, 379–403.

Chowdhury, N. H., Adam, M. T. P., Skinner, G. (2019). The impact of time pressure on cybersecurity behaviour: a systematic literature review. *Behaviour & Information Technology*, *38*, 1290–1308.

Eagly, A. H., Chaiken, S. (1993). The psychology of attitudes. *Harcourt brace Jovanovich college publishers*.

Edland, A., Svenson, O. (1993). Judgment and decision making under time pressure. In *Time pressure and stress in human judgment and decision making* (pp. 27–40). Springer.

Ferreira, A., Lenzini, G. (2015). An analysis of social engineering principles in effective phishing. In *2015 workshop on socio-technical aspects in security and trust* (pp. 9–16).

Franco, E. G. (2020). The global risks report 2020. *World Economic Forum*.

Gabaix, X. (2019). Behavioral inattention. *Handbook of Behavioral Economics: Applications and Foundations 1*, *2*, 261–343.

Giger, J., Pochwatko, G. (2008). Sometimes it is not so bad to decide in a hurry: Influence of different levels of temporal opportunity on the elaboration of purchasing intention. *Polish Psychological Bulletin*, *39*, 209–216.

Grazioli, S. (2004). Where did they go wrong? an analysis of the failure of knowledgeable internet consumers to detect deception over the internet. *Group Decision and Negotiation*, *13*(2).

Green, D. M., Swets, J. A. (1966). Signal detection theory and psychophysics. Wiley, New York.

Guo, K. H., Yuan, Y., Archer, N. P., Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, *28*(2), 203–236.

Halevi, T., Lewis, J., Memon, N. (2013). A pilot study of cyber security and privacy related behavior and personality traits. In *Proceedings of the 22nd international conference on world wide web* (pp. 737–744).

Johnson, P. E., Grazioli, S., Jamal, K., Zualkernan, I. A. (1992). Success and failure in expert reasoning. *Organizational Behavior and Human Decision Processes*, *53*(2), 173–203.

Kirlappos, I., Sasse, M. A. (2011). Security education against phishing: A modest proposal for a major rethink. *IEEE Security & Privacy*, *10*(2).

Matějka, F., McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, *105*(1).

Miller, J. G. (1960). Information input overload and psychopathology. *The American journal of psychiatry*, *116*(8), 695–704.

Payne, J. W., Bettman, J. R., Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 534–552.

Petty, R. E., Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1–24). Springer, New York.

Polanyi, M. (1965). The structure of consciousness. *Brain*, *88*(4), 799–810.

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change1. *The Journal of psychology*, *91*(1), 93–114.

Shahbaznezhad, H., Kolini, F., Rashidirad, M. (2020). Employees' behavior in phishing attacks: What individual, organizational, and technological factors matter? *Journal of Computer Information Systems*, 1–12.

Shannon, C. E. (1948). The mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, *50*(3), 665–690.

Sims, C. A. (2010). Rational inattention and monetary economics. *Handbook of Monetary Economics*, *3*, 155–181.

Svenson, O., Edland, A., Slovic, P. (1990). Choices and judgments of incompletely described decision alternatives under time pressure. *Acta Psychologica*, *75*(2), 153–169.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Swets, J. A. (2000). Enhancing diagnostic decisions. In *Judgement and decision making – an interdisciplinary reader* (pp. 66–81). Cambridge: Cambridge University Press.

Tjøstheim, I., Waterworth, J. A. (2020). Predicting personal susceptibility to phishing. In *International conference on information technology & systems* (pp. 564–575). Springer, Cham.

Unchit, P., Das, S., Kim, A., Camp, L. J. (2020). Quantifying susceptibility to spear phishing in a high school environment using signal detection theory. In *International symposium on human aspects of information security and assurance* (pp. 109–120).

Vishwanath, A., Harrison, B., Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*, *45*(8), 1146–1166.

Vishwanath, A., Herath, T. C., Chen, R., Wang, J., Rao, H. R. (2011). Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, *51*(3), 576–586.

Volkamer, M., Renaud, K. V., Reinheimer, B. M. (2016). Torpedo: Tooltip-powered phishing email detection. In *Ifip international conference on ict systems security and privacy protection* (pp. 161–175). Springer, Cham.

Volkamer, M., Renaud, K. V., Reinheimer, B. M., Rack, P., Ghiglieri, M., Mayer, P., . . . Gerber, N. (2018). Developing and evaluating a five minute phishing awareness video. In *International conference on trust and privacy in digital business* (pp. 119–134).

Wang, J., Herath, T. C., Chen, R., Vishwanath, A., Rao, H. R. (2012). Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Transactions on Professional Communication*, *55*(4), 345–362.

Wiens, M. (2013). Vertrauen in der ökonomischen theorie: Eine mikrofundierte und verhaltensbezogene analyse. *Schriften zur internationalen Wirtschaftspolitik*, *9*.

Wright, P. (1974). The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology*, *59*(5), 555–561.

Wright, R. T., Chakraborty, S., Başoğlu, A. N., Marett, K. (2010). Where did they go right? understanding the deception in phishing communications. *Group Decision and Negotiation*, *19*(4), 391–416.

Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica*, *58*(1), 75–80.

Zur, H. B., Breznitz, S. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89–104.

10 Cyberattacks on Hospitals and their Impact on Medical Service[1]

Abstract

Cyber risks pose a significant threat to modern society and economy as modern production and service provisioning systems (e.g. healthcare providers) are increasingly dependent on information and communication technology. Yet, managing cyber risks proved to be challenging for professionals, analysts, and information security managers as there is no profound and commonly accepted methodology for quantitative cyber risk assessment. Within this contribution, a methodology is presented aiming at forwarding cyber risk assessment. It thereby focuses on quantifying the impact of cyberattacks, which is an important element of cyber risk quantification. The presented methodology represents an alternative and novel perspective on cyber risk quantification by employing a digital twin presenting an actionable approach for quantification taking advantage of cyber threat intelligence. Relying on the digital twin, we simulate the impact of cyberattacks within a case study on the health care industry and use the simulation to evaluate the methodology in quantitative as well as qualitative terms. Our main contributions are (i) the adaption of established techniques for cyber impact assessment, (ii) introducing an alternative, actionable, and novel perspective to cyber impact quantification, and (iii) the exemplary application of the methodology to a specific system as well as its evaluation within this case. The proposed methodology can help analysts, security professionals and managers by forwarding cyber risk quantification, the evaluation of systems' security and proposing new perspectives on the understanding of cyber risk.

Keywords: Cyber risk quantification, cyber threat intelligence, digital twin, naïve Bayesian classification, synthetic health data, medical services

10.1 Introduction

The digitalization of health services promises great opportunities, which are inter alia improved quality of service provisioning and increased accessibility of health care (Kaiser et al., 2021). Yet, increasing dependence on digital systems in healthcare also introduces novel risks to patient safety, security, and privacy. Security of medical services hence needs to be ensured. The importance of

[1] This chapter includes the preprint of the article "Cyberattacks on hospitals and their impact on medical service" by Marcus Wiens, Frank Schultmann, and myself.

cyber security of medical services was long term scarcely considered in organizational decisions over investments in information and communication technology (ICT) and in digitalization respectively. However, recent cyberattacks on medical services highlighted the importance of considering the severe consequences that may emerge (cf. inter alia the WannaCry attack on the NHS, (Ghafur et al., 2019; Mahler et al., 2020). Yet, managing these risks is challenging, as to date, there is no scientifically established and applicable methodology that quantifies cyber risks in a reliable and accurate way (Zeller and Scherer, 2021; Leszczyna, 2021). Especially for the quantification of the consequences of a cyberattack there is a methodological lack (Eckhart et al., 2019).

Within this work, we aim at forwarding cyber risk quantification by presenting a novel method for simulating the effects of cyberattacks on the basis of a digital twin. We thereby rely on the example of medical service provisioning and analyze the impact of cyberattacks based on different dimensions. These different dimensions are inter alia quality of health care services and monetary loss for hospitals. We thereby focus on short, neglecting long term consequences like the effects of delayed medical treatment or reputational damage. We evaluate the presented approach quantitatively for the processes of treatment and diagnosis as well as qualitatively based on the case of the WannaCry attack on the NHS.

10.2 Theoretical foundations

10.2.1 Digital twins

As the complexity of many production systems increases, so does the necessity of detailed models of these systems to enable in-depth analyses. One concept that aims to formalize this detailed mapping of technical and operational systems is that of the digital twin.

Martínez et al. (2018) define a digital twin of a production facility as "a digital replica of the plant's physical assets which contains the structure and the dynamics of how the devices and processes operate" (Martínez et al., 2018, p. 3084). Digital twins are becoming increasingly important as a means of visualizing, and analyzing economic systems and, in particular, production facilities (Ait-Alla et al., 2021). However, digital twins are also increasingly used for the analysis in medical contexts. Simulations represent the standard method for analyzing these systems (de Paula Ferreira et al., 2020). Mourtzis (2020) adds that the digital twinning is still a young but promising approach in manufacturing simulation, which is not yet mature and not yet presented in its full capabilities in the literature. Especially in the field of cyber risk management, where complex interrelations are an important determinant of their effects, digital twins can be considered a helpful means. For this reason, we implement a digital twin of a hospital for analyzing the effects of increasing penetration of the processes with information and communication technology as well as their impact on hospitals susceptibility for cyberattacks. Furthermore, digital twins might be a helpful means for understanding the effect of successful cyberattacks.

10.2.2 Intelligence for specifying digital twins

10.2.2.1 Synthetic health data

Research based on real health data is strictly limited due to privacy concerns, ethical considerations, and their highly confidential nature (Kuo et al., 2022). Yet, analyzing health data is necessary when dealing with health care systems. This is, realistic simulations of medical service provisioning depend on crafting highly reliable and realistic, yet synthetic, data on patients' state of health or efficient anonymization of health records. Synthetic health care generation thereby is superior to anonymization as anonymization falls short in offering rigorous privacy guarantees (Abay et al., 2018). Meanwhile, synthetic data generation and especially the generation of synthetic health data is widely researched. Moniz et al. (2009) review a set of algorithms (inter alia shuffling algorithms) and their ability in generating synthetic health care reports on the basis of reports on real patients. The ability of the different algorithms is evaluated based on similarity metrics (inter alia Jaccard's coefficient and the Euclidean distance) comparing real health records with the synthetic health records. Yale et al. (2019) discuss the use of synthetic health data to overcome limitations in the use of health data and present an approach for generating health records based on a generative adversarial network. Likewise, Kuo et al. (2022) rely on a generative adversarial network for crating synthetic health datasets for patients suffering from acute hypotension, sepsis, and people infected by the human immunodeficiency virus. Abay et al. (2018) propose the use of a generative deep learning method for creating reliable, realistic yet differential private health records. However, scientific findings show that depersonalized electronic health records fall short in providing efficient privacy protection and suffer the potential of re-identification (McLachlan et al., 2016). They present a synthetic health data generator that relies on clinical practice guidelines and health incident statistics. Furthermore, Dube and Gallagher (2013) introduce another approach that neither needs real electronic health records nor authorization or anonymization for generating realistic synthetic health records. Although there is much research on different methods for generating synthetic health care records, most of them fall short in delivering a comprehensive picture of different diseases in the generated synthetic health records as they are often limited to medical issues interesting for specific disciplines diseases. Among different methods for synthetic health care generation, the synthetic patient population simulator represents a method that does not fall short this focusing on a specific domain but offers synthetic information on a broad variety of different issues relevant for medical decision making and medical facility simulation. Synthea is generated for the simulation of disease progression, treatment and health care provision without the risk for real patients and ethical challenges (Walonoski et al., 2018). It relies on the approach presented by Dube and Gallagher (2013) and hence offers the potential of generating realistic health data without posing privacy concerns. Within this work we build upon Synthea and include the approach to the digital twin for producing realistic patient cohorts.

10.2.2.2Cyber threat intelligence

Cyber threat intelligence (CTI) is actionable information on past (observed) cyberattacks. The intelligence can be in different granularity including strategic, tactical, technical and operational CTI (Chismon and Ruks, 2015). CTI has proven to be a powerful means in brought aspects of cyber risk management including threat hunting, intrusion detection and other analytical procedures (Elitzur et al., 2019). Main strengths of using CTI is the ability to gain a holistic picture in the field of cyberattacks by retrieving information on past attacks. This is, CTI can help to gain a comprehensive understanding of the threat landscape by sharing information on past attacks between different victims of attacks (Elitzur et al., 2019).

A major obstacle of CTI is that the information is oftentimes provided in various different formats depending on the preferences of the respective source. To encounter this obstacle, the US Department of Homeland Security's Office of Cybersecurity and Communications and MITRE developed and established a standardized language for sharing CTI (Structured Threat Information eXpression (STIX)). However, other CTI languages exist. These include OpenIOC, Incident Object Description Exchange Format (IODEF) and proprietary languages. Yet, Sauerwein et al. (2017) provided evidence that STIX can be seen as de-facto standard language of CTI sharing. CTI furthermore is shared by many platforms (e.g. Malware Information Sharing Platform (MISP), OpenCTI, Collective Intelligence Framework (CIF), Anomali STAXX, Open Threat Exchange platform (OTX), and many more). These different platforms offer CTI on offensive techniques. However, there are also CTI vendors focusing on the defensive side respectively on vulnerabilities or weaknesses that can enable attacks (e.g. the national vulnerability database -NVD). For an understanding of the specific benefits as well as the weaknesses of these different formats of CTI sharing, respectively the different languages and platforms, the contribution of de Melo e Silva et al. (2020) can be helpful. In their research paper, the authors provide a comprehensive and comparative evaluation of a set of those different CTI languages and platforms.

10.2.3 Knowledge graphs

Relational databases often reach their limits when it comes to representing relationships in data sets. Graph databases are better suited for this purpose. They are optimal for data that is highly interconnected, as they are easily scalable for large and heterogeneous data sets (Elnagar and Weistroffer, 2019). So-called knowledge graphs, which can be generated using graph databases, consist of nodes and edges. Nodes are entities and edges connect two nodes with each other and thus form a relationship between these two entities. The relationships between entities can then be easily read from the triple of the two connected nodes and the edge. This structure provides a simple and intuitive understanding of the information presented and facilitates the inference and interpretation of data, both for humans and for machine systems (Abu-Salih, 2021).

Especially in healthcare, such datasets can be found again, which is why a knowledge graph is very suitable for representing the relationships in these data. Because of this, knowledge graphs find application in the healthcare domain again and again. Probably the best known knowledge

graph in healthcare is the Google Health Knowledge Graph. In the work presented by Rotmensch et al. (2017), a Health Knowledge Graph, which aims to show the relationships between diseases and symptoms, is created from electronic patient data. The information was extracted from 273, 174 anonymized patient data and used to generate a knowledge graph. I. Y. Chen et al. (2019) analyze the knowledge graph presented by Rotmensch et al. (2017) highlighting the potential of knowledge graphs for medical information retrieval. Equivalently, information on cyberattacks can be presented efficiently in knowledge graphs. The works provided by Hemberg et al. (2020) and Elitzur et al. (2019) rely on for example on approaches taking advantage of a knowledge graph for analyzing cyberattacks. Within these works threat knowledge graphs focusing on different aspects are given and the potentials of the approaches highlighted. Elitzur et al. (2019) focus on threat hunting while Hemberg et al. (2020) focus on linking weaknesses and vulnerabilities to attacks in order to extract those attacks that can be executed given the vulnerability.

10.2.4 Cyber risks and their quantification

Setting up models for understanding and quantifying cyber risks is a rather hard problem and in many areas, cyber risks are far from being understood comprehensively (Böhme et al., 2017). To forward the understanding of cyber risk, it is hence necessary to take different perspectives. One perspective that should be taken is the perspective on the key features of technology. This is, information and communication systems do not generate value in isolation. Rather they enable new business models and support processes which vice versa generate value. Hence, information systems provide value indirectly (Wigand et al., 1997). Although this is not a new insight and is moreover known since quite a long time, it is oftentimes not included in cyber risk quantification consistently (e.g. by taking a process oriented approach). This is identified as an essential burden for understanding cyber risks as this fact makes cyber risk quantification more complicated than risk quantification in areas were the value contribution is more direct. Furthermore, the technical understanding of a "cyberattack in progress" need to be considered. Thereby, cyber threat intelligence can take a key role for cyber risk quantification. Last, the strategic nature of cyberattacks should be considered when quantifying cyber risks. Defenders and Attackers are considered to play "a dynamic cat and mouse game" (Elitzur et al., 2019, p. 40). This reflects the adaptive behavior of perpetrators to the (defensive) actions taken by the defenders (e.g. security analysts or business managers) and highlights the importance of risk quantification approaches that allow to consider this processes of strategic adaption (e.g. game theory).

Although there is no systematic formal-mathematical methodology for quantifying risk, there is intensive and interdisciplinary research. However, so far there is no methodology that could solve the problem of cyber-risk quantification in a proper and universal way. Rather, it is common when quantifying cyber-risks, that the impact as well as the probability is estimated qualitatively (oftentimes through the usage of a scale ranging from zero to five). Another example is norm based risk management relying heavily on heuristics and reference values based on assumptions of experts as a valid estimation of cyber-risks (e.g. ISMS Family Standards/ ISO/IEC 27000-series). While experts can play an important role in scenario planning and validation of risk management tools,

their contribution to purely quantitative assessments is rather complementary. Expert decision biases due to cognitive and time constraints, a selective perspective as well as personal interest represent a natural limit for estimation and quantification of risk by experts. Patel and Zaveri (2010) describe a quantitative methodology for assessing cyber-risks. They rely on a statistical approach and divide effects of cyber-incidents into different categories. Thereby they assess the impact of these different categories.

However, they do not introduce neither a business process value analysis nor game theoretical probability quantification. Q. Chen et al. (2015) present a risk assessment methodology for industrial control systems. Thereby they focus on availability. However, the methodology can also be used for other security requirements, like confidentiality, authenticity and integrity. However, their presented methodology can be criticized for its strong dependence on the correct estimation of the importance by stakeholders. Although these stakeholders can be seen as experts in estimating the effects of a cyberattack on the respective requirement, expressing the induced damage in monetary terms may be challenging even for experts. Furthermore, the quantification of probabilities is based on a pure statistical approach and does therefore not consider the strategical decision making of attackers. The work closest to our approach was published by Musman and Turner (2018) who provide a methodology that, "quantitatively identifies cyber security risks and uses this metric to determine the optimal employment of security methods for any given investment level." (Musman and Turner, 2018, p. 127). Within their work, the likelihood of an attack is calculated by using a defender attacker model. Furthermore, the impact is calculated using a business process model. In doing so, they are able to calculate the impact of a cyber-incident on mission fulfillment. However, their approach of process evaluation lacks a market perspective because market outcomes are essential for value determination. Yet, our methodology goes beyond the pure technical modeling of business processes by integration of a market- and customer-perspective on processes. While the approach of Musman and Turner (2018) use expert judgements for process valuation, we derive the process value from customer requirements and their valuations on the market. As it can be concluded from the variety of different and sometimes even contrary approaches for risk quantification, cyber risk management in its current status still seems to be "more art than science" (Woods and Böhme, 2020).

10.3 Method

10.3.1 Schema

Our work comprises the development and implementation of a digital twin of a hospital. The different processes in the hospital are simulated, starting with the arrival of the patients in the hospital followed by a subsequent anamnesis, diagnosis, treatment and release. We thereby rely on a discrete event simulation (DES). For simulation of the diagnostic process we embed a naive Bayesian classifier (NB-C) to the DES accessing data from a medical knowledge graph. The medical knowledge graph consists of synthetic health data and thus represents evidence-based knowledge. After the diagnosis, the treatment of the diagnosed disease follows. Like the diagnosis, the selection

of suitable treatment paths uses evidence-based knowledge extracted from the medical knowledge graph. During the treatment, there is a continuous monitoring of the improvement of the health status. The improvement of disease symptoms is simulated in the digital twin if the treatment is correct. If the diagnosis is correct and the treatment successful, the recovery of the patient occurs after a given time. The recovered patients are released from the hospital if they are recovered.

The schema of the digital twin is shown in detail below (Figure 10.1). In a first step, patients are created. These have a specific medical history as well as acute symptoms. The medical condition as well as their history are simulated by the DES relying on Synthea (Walonoski et al., 2018). In a second step, the patients are admitted to the medical service facility. The patient's history is taken and symptoms are recorded (i.e., anamnesis), as well as an initial diagnosis (based on the symptoms) regarding the patient's condition. This diagnosis is verified, if necessary, by ordering further examinations (e.g., palpation and/or imaging procedures). The resulting observables are recorded as secondary observables in the set of currently recorded observables (i.e., symptoms) on the patient (S). We rely on S and the medical evidence knowledge graph to derive a diagnosis based on a NB-C. Once the diagnosis has been made, the treatment of the patient is initiated. We extract the most feasible treatment (according to evidence) to a disease from the medical evidence knowledge graph. In a last step, recovered patients are discharged and the hospital receives a compensation for the services provided.

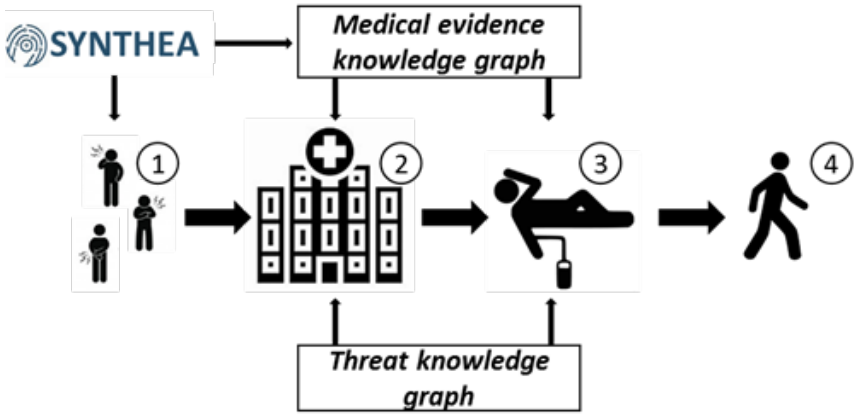

Figure 10.1: Schema of the digital twin of the hospital

Given the technical infrastructure of the hospital, treatments and diagnostic procedures can be hindered by cyberattacks. These can result in certain treatments not being carried out (for example, in the case of a complicated surgery requiring robotic assistance). Diagnostic procedures may likewise be hindered (i.e., unavailability of technology guided imaging) or even lead to misdiagnosis. For example, it is known that diagnostic imaging procedures in medicine can be affected by cyberattacks (Mirsky et al., 2019; Mahler et al., 2022). The effects of various cyberattacks are mapped at the technical level on the basis of cyber threat intelligence and their effects on medical service provision (such as incorrect diagnoses) are simulated in the digital twin. The DES thereby takes advantage of a knowledge graph inspired by the knowledge graph introduced by Hemberg et al. (2020) providing CTI on attacks, their technical impact (extracted from ATT&CK) as well

as vulnerabilities and weaknesses linked to the attacks (extracted from NVD). Furthermore, this threat knowledge graph provides information on linked product configurations. The attacks can have two primary effects. The unavailability, modification or encryption of patient information can make diagnoses more difficult and prevent treatment. In addition, the prevention of certain treatment procedures is also directly influenced by cyberattacks if the necessary procedures (e.g. computers or robotically assisted operations) are not available (e.g., through a denial of service attack). The simulation provides information about the cyber risks of certain medical service facilities and can thus make a significant contribution to their quantification. During the simulation of the digital twin, a network protocol is created that reproduces all processes and can be used for further analysis.

10.3.2 Simulation and operationalization

For simulating medical service provisioning in a hospital, we rely on a DES and combine the approach with a NB-C for simulating the process of diagnosis. The DES takes advantage of two knowledge bases representing medical evidence respectively cyber threat knowledge.

Definition 13 (Cyber threat knowledge graph) *A cyber threat knowledge graph threat* $KG =< A, T, W, V, D, L >$ *is a graph fusing information on attacks where* A *contains attacks that were observed in the past and* T *is a set of nodes representing descriptions on the observed attacks which are in especial offensive (attack) techniques employed by an attack. The set* W *describes weaknesses while* V *denotes vulnerabilities. Furthermore, the set D denotes devices respectively product configurations (that are susceptible to an attack via some vulnerabilities). Last,* L *denotes the set of links between the nodes.*

Definition 14 (Medical evidence knowledge graph) *A medical evidence knowledge graph is a graph medical evidence KG=<S,C,R,*$\dot{D}$*,EHR,L> comprising medical evidence where* S *describes the set of nodes describing observables/ signs of a medical condition* c *and* C *is the set of nodes describing medical conditions (i.e., diseases and injuries). Let furthermore* R *be the set of nodes describing treatment plans observed to contribute to recovery,* EHR *be the electronic health record the evidence is extracted, and* L *be the set of links between the different nodes. Last,* $\dot{D}$ *describes the set of devices that are used in medical service provisioning either for deriving observables (e.g. secondary observables - linked to* s*) or for treatment (linked to* r*).*

The NB-C takes advantage of the *medical evidence KG* for extracting knowledge (graph analysis) and come up with sound estimations on the prior probabilities of diseases as well as on the conditional probabilities for a specific medical condition c_i given S. For medical knowledge extraction, we rely on naïve Bayesian information retrieval based on the count of paths in subgraphs of the *medical evidence KG*. Let X be a node (e.g., condition) and Y be another node in the *medical evidence KG* (node of a different type e.g., treatment). We denote $L\{X - Y\}$ as the set of links between those nodes. The conditional probability of a successful treatment of a patient when relying on a specific treatment $(P_{recovery}(c_i|r_j))$ can consequently be given as follows, where

$|EHR(L\{c_i - r_j\})|$ denotes the number of cases/ studies where c_i was treated relying on the treatment plan r_j and $|EHR(L\{c_{i,recovery} - r_j\})|$ denotes the number of cases/ reports where the treatment was successful (the patient recovered after receiving r_j).

$$P_{recovery}(c_i|r_j) = \frac{|EHR(L\{c_{i,recovery} - r_j\})|}{|EHR(L\{c_i - r_j\})|} \quad \text{(Exp. 10.1)}$$

Equivalently we retrieve information on the prior probabilities of a specific condition $(P(c_i))$ and the conditional probabilities of specific conditions when observing some symptoms $(P(c_i|s_k))$. Where $|EHR(c_i)|$ denotes the number of cases where a patient had a specific medical condition and $|EHR|$ is the number of cases included in the *medical evidence KG*. $|EHR(L\{s_k - c_i\})|$ denotes the number of cases where s_k was observed when the patient had c_i and $|EHR(s_k)|$ denotes the number of cases where s_k was observed.

$$P(c_i) = \frac{|EHR(c_i)|}{|EHR|} \quad \text{(Exp. 10.2)}$$

$$P(c_i|s_k) = \frac{|EHR(L\{s_k - c_i\})|}{|EHR(s_k)|} \quad \text{(Exp. 10.3)}$$

We define the problem of crafting hypotheses on the health condition of a patient (diagnosis) as follows.

Definition 15 (Diagnosis) *Let $P(c_i|S)$ describe the conditional probability of a condition (i.e. disease) i (c_i) given a set of observables (S). The condition with the highest $P(c_i|S)$ serves as the diagnosis for the patient. Let $P(S)$ represent the prior probability of observing S and $P(c_i)$ be the prior probability of a condition i. S thereby consists of j specific observables (s). $P(c_i|S)$ can then be defined as follows.*

$$P(c_i|S) = \frac{P(c_i) * P(S|c_i)}{P(S)} \quad \text{(Exp. 10.4)}$$

with

$$P(S|c_i) = \prod_k P(c_i|s_k) \quad \text{(Exp. 10.5)}$$

For a given diagnosis, we give the treatment with the highest $P_{recovery}(c_i|r_j)$ as the proposed treatment. The recovery of patients is simulated based on this probability. If a treatment plan fails, the next best treatment plan (according to $P_{recovery}(c_i|r_j)$) is selected.

We simulate the effects of a cyberattack based on the technical effects of an attack which can be extracted from the cyber threat knowledge base. If an attack is simulated, we compare the technical equipment of the hospital with the knowledge graph extracting information on which devices are

vulnerable to the attack. If an attack is linked to devices that are not used by the hospital, the attack will not lead to an impact for the hospital. However, if a device/ product configuration used by the hospital is susceptible to an attack, the attack will have an impact on proper medical service provisioning (e.g., denial of service of data encryption). This assumption is based on the work of Allodi and Massacci (2017) giving evidence that if an attack is conducted and a vulnerability exists, which can be exploited by the attack the success is "almost certain". If a device is rendered unavailable through an attack, treatment plans might be impossible (e.g. if a computer guided surgery system is affected) or observables might not be assessed (e.g., if an imaging device is rendered unavailable). These technical effects of an attack are given within the threat knowledge graph (so called impact techniques). The practical effects (e.g., wrong diagnoses or inability to provide feasible treatment to a disease) are extracted from the digital twin relying on the DES. If an attack is successful given a specific product configuration implemented in a hospitals information infrastructure is defined by the existence of links between a specific attack (a_b) and the specific product configuration (d_f) $(L\{a_b - d_f\})$. We give an example of how the effect of an attack is simulated in Section 10.4.2 to guide the readers understanding.

10.4 Digital twin of a fictious hospital

10.4.1 Experimental data

10.4.1.1 AttackKG

Inspired by works on multi-level threat ontologies comprising CTI on attack and defense specific information, we build AttackKG. The structure of AttackKG is inspired by the work of Hemberg et al. (2020) and the knowledge graph presented by them called BRON. AttackKG follows the definition of a cyber threat knowledge graph as presented in Definition 13. We implement AttackKG in neo4j however, any other software for building graph databases could be used.

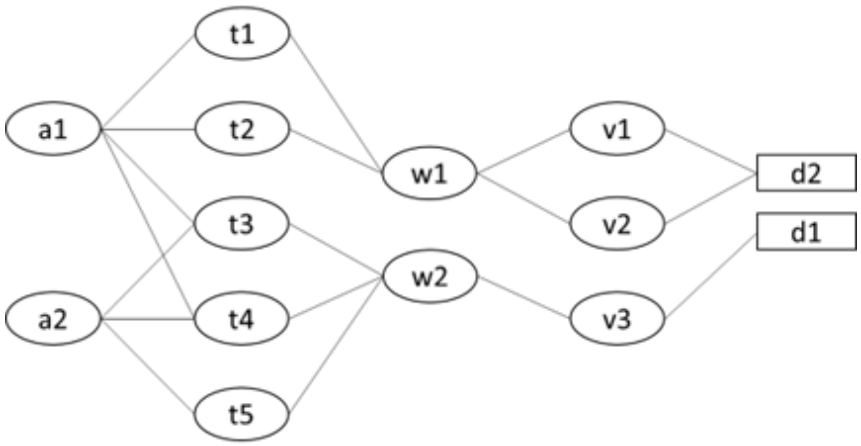

Figure 10.2: Structure of AttackKG

Exemplary, $a2$ would denote to Kwampirs backdoor Trojan, which was observed to be used in attacks against medical imaging devices. $d1$ could here be an X-ray device. Attack $a1$ could represent an attack targeting medical ventilators ($d2$). The linkage between the affected device and the attacks is given by the following logic. An attack takes advantage of techniques that can be used to exploit some weakness that is inherent to a specific device/ product in a specific configuration through a specific vulnerability.

10.4.1.2HealthKG

We present HealthKG as an actionable medical knowledge base fusing medical evidence from patients treated. We therefore generate a set of 600 electronic health records relying on synthea. HealthKG serves as a medical knowledge repository representing a multi-level medical knowledge ontology on these health records. The knowledge base thereby comprises information on (primary and secondary) observables (i.e., symptoms), diseases and treatment plans. It links the observables to the diseases and those wise versa to reasonable treatment plans/ procedures (treatments that have been observed to be able to contribute to the recovery from a disease or other medical condition). HealthKG follows the definition of a medical knowledge base as it is given in Definition 14. The structure of HealthKG is given in Figure 10.3. Like AttackKG, HealthKG is implemented in neo4j.

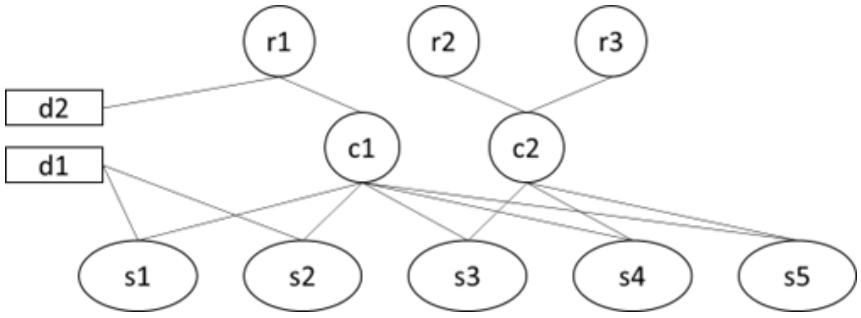

Figure 10.3: Structure of HealthKG

In Figure 10.3, $s1$ and $s2$ represent secondary observables. These secondary observables are linked through a procedure (e.g. medical imaging) to a device (e.g. X-ray - $d1$). $s1$ and $s2$ could therefore be changes in the lung that can be observed trough medical imaging in case of a Covid19 infection ($c1$) while $s3$ to $s5$ could be symptoms like cough, sore throat, or headache which can also be linked to a flu ($c2$). Within this example, $s1$ and $s2$ would be specific and can hence be used to identify $c1$ while $s3$ to $s5$ are nonspecific for the two conditions. Within this example, $r1$ could denote to mechanical ventilation while $r2$ and $r3$ would denote to flu medications which are not useful for treating $c1$.

10.4.1.3Set of patients

In an initial step for evaluation, we construct a scenario where a set of patients arrives at the hospital. We thereby craft a set of 30 patients arriving at the hospital demanding for a treatment/

medical relief. An exemplary electronic health record on a patient is presented in Figure 10.4. We thereby only included the medical findings that are recently acute. We rely on these acute findings to craft a diagnosis for the patient relying on the NB-C that is implemented on HealthKG.

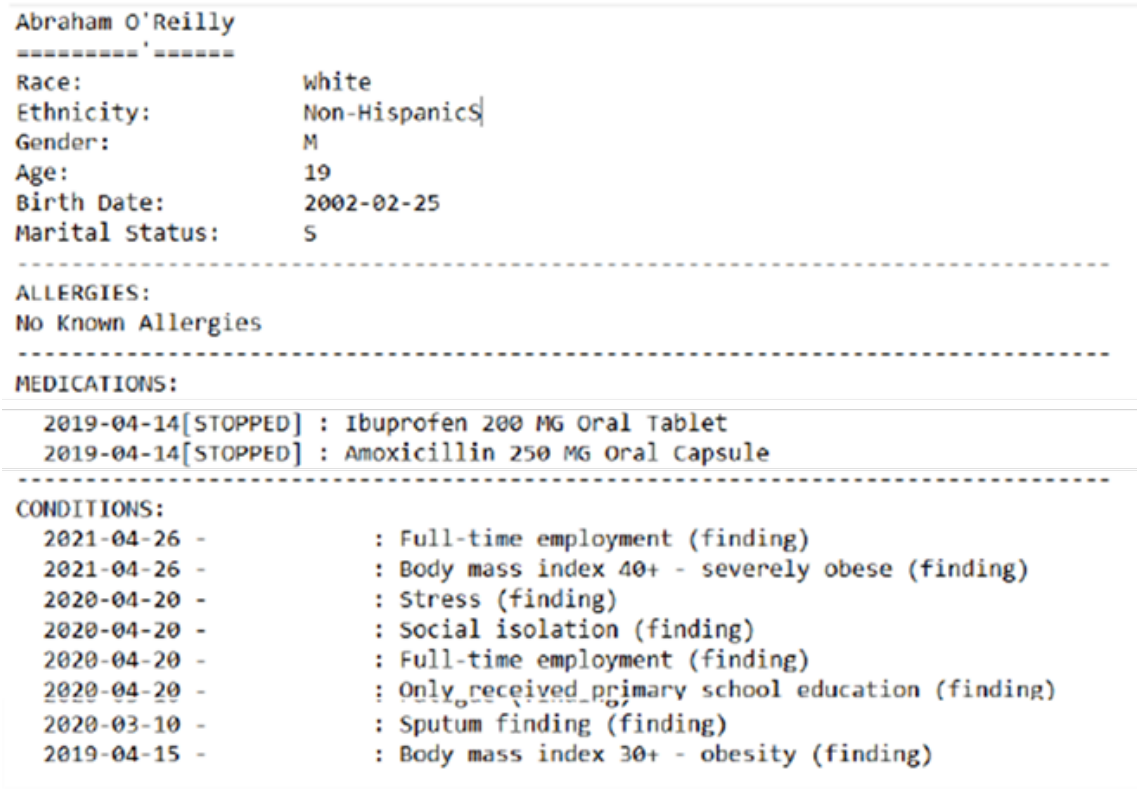

Figure 10.4: Exemplary electronic health record on a patient crafted relying on Synthea

10.4.2 Implementation of the digital twin and evaluation

In a first step, underlying information is made accessible in knowledge graphs. The HealthKG represents the evidence-based medical knowledge of a certain set of past clinical cases. For this purpose, Algorithm 7 uses a set of electronic health records of patients (EHRs) that were created using Synthea. The information about the patient history from the set of patient records is then used to identify clinically relevant symptoms of diseases (co-occurrences) as well as the treatments belonging to the disease (treatments used in the past). In addition, a knowledge graph is created that makes CTI accessible (AttackKG). We extract information from the MITRE ATT&CK database and NVD. Of particular importance is information about the attack, the malware used, and the techniques used as well as affected product configurations, weaknesses, and vulnerabilities linked to attacks. From the techniques, the effects in the system can be derived (e.g. T1489, which means stop service).

Algorithm 7 Craft Knowledge Graphs

Input: EHRs, CTI
Output: HealthKG, AttackKG

procedure $CraftHealthKG$
$Input \leftarrow ExtractInformation(EHRs)$
$CraftHealthKG\ (observations, conditions, disorders, procedures, relationships)$
return $HealthKG$
end procedure
procedure $CraftAttackKG$
$Input \leftarrow ExtractInformation(CTI)$
$CraftAttackKG\ (software, technique)$
return $AttackKG$
end procedure

The digital twin of the medical service facility now maps the following. In a first step, patients who are admitted to the hospital are generated as shown in Algorithm 8. For each patient, a medical record is created using Synthea. By this procedure, the DES simulates a number of patients searching for relief at the hospital for a specific point in time. Synthea thereby generates readily available EHRs for each individual patient. The EHRs are synthetic (Walonoski et al., 2018). The DES thereby profits from Synthea by its ability to craft realistic EHRs for individual patients without leading to data privacy or moral concerns and restrictions in using the data.

Algorithm 8 Generate Patient Sample

Input: Synthea, MeanPatientArrivalRate, PatientArrivalRateVariance
Output: PatientSample

procedure $CraftPatientSample$
$Input \leftarrow CraftPatientEHR(Synthea)$
$CraftPatientSample\ (Synthea, MeanPatientArrivalRate, PatientArrivalRateVariance)$
return $PatientSample$
end procedure

Algorithm 9 represents the decision (probabilistic) on whether or not a cyberattack will occur. The probability of occurrence of an attack can be set arbitrarily depending on the expectations and realistic assumptions of the threat landscape. However, the effect of the attack will depend on the system implemented within the focal hospital (i.e., whether the attack is linked to the product configurations of the devices the focal hospital uses). If no attack takes place, there will be no impact at this point and the hospital can be simulated in normal operation. However, if an attack takes place, the simulation of the impact of the cyberattack will take place and the system will be "affected" in this case. The technical consequences of an attack (e.g. encryption of data) and the consequences for the confidentiality, integrity and availability of data or whether certain services are prevented (availability of services) are extracted from the AttackKG. The effects are dependent on the level of digitalization of the hospital, simulated for the specific scenario (e.g. patient sample) and fed into the digital twin. These effects are inter alia bad data quality e.g. due to non-availability of diagnostic procedures such as medical imaging.

Algorithm 9 Simulate State of Hospital

Input: AttackProb{software}, AttackKG, Impact{technique}, System
Output: HospitalState, AppliedSoftware, Impact{software}

1: **procedure** $SimulateStateofHospital$
2: $SimulateStateofHospital\ (AttackProb\{software\}, System)$
3: **if** Vulnerabilityexists(software, System) **then**
4: $SimulateImpact(AttackKG, Impact\{technique\})$
5: **return** $HospitalState\{"affected"\}, AppliedSoftware, Impact\{software\}$
6: **else**
7: **return** $HospitalState\{"normal"\}$
8: **end procedure**

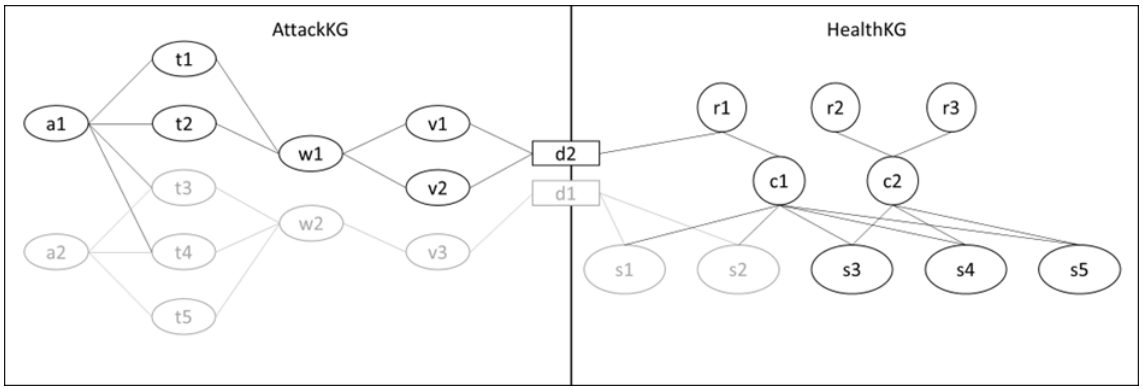

Figure 10.5: Exemplary obfuscation of the knowledge graphs and attack impact simulation

Figure 10.5 shows an exemplary cause-effect chain of an attack with attack $a2$ within the knowledge graphs. We show the effects by obfuscation. Within the example, $d1$ is rendered unavailable causing a loss of $s1$ and $s2$ which causes that $c1$ and $c2$ could not be differentiated anymore relying on the observations ($s3$ to $s5$).

Algorithm 10 describes the process of anamnesis, diagnostics and the resulting treatments. For each patient, current symptoms of the disease are recorded and other relevant information such as risk factors or allergies is requested. Based on the named symptoms of the disease in the further primary observables, a first hypothesis regarding the disease is made. This hypothesis is tested by diagnostic procedures. The output of these diagnostic procedures are secondary observables. The diagnosis is made from the primary and secondary observables. Based on the diagnosis, treatment procedures are derived that offer the patient the best possible recovery perspective. Line 2 to 5 represent possible impediments to the processes due to the unavailability of data and the impediment of medical equipment.

Algorithm 10 Medical service provisioning

Input: HospitalState, Patient Sample, AppliedSoftware, Impact{software}, HealthKG
Output: diagnosis, treatment

1: **procedure** $etStateofSystem(HospitalState)$
2: **return** $diagnosis, treatment$
3: **end procedure**

Algorithm 11 describes the evaluation of the successful course of treatment. It examines the extent to which the selected procedure is the correct procedure. For this purpose, we compare the extent to which the diagnosis and the ground truth (GT) of the disease match and the extent to which the proposed treatment and the standard treatment (GT) match. The time to relief is derived from the order of the proposed treatments (periods until the correct treatment is chosen). This is implicitly based on the assumption that the treatments are always examined for their effectiveness in the same periods and if this is not the case, an adjustment of the treatment strategy is made. Finally, a monetary value is attributed to the simulation, which represents the hospital's revenue. For this purpose, the work relies on the Diagnosis Related Groups (DRG) system.

Algorithm 11 Evaluation Procedure

Input: GT, diagnosis, treatment, DRG
Output: AP_diagnosis, AP_treatment, Earnings, time_to_relief

1: **procedure** $Evaluation$
2: $AP\_diagnosis(GT, diagnosis)$
3: $AP\_treatment(GT, treatment)$
4: $time\_to\_relief \leftarrow time\_to\_relief(GT, treatment)$
5: $Earnings \leftarrow CalculateEarnings(GT, treatment, DRG)$
6: **return** $AP\_diagnosis, AP\_treatment, earnings, time\_to\_relief$
7: **end procedure**

10.4.3 Results and discussion

For the evaluation of the digital twin, a knowledge graph consisting of about $11,000$ different observations for 104 different diseases was mapped (corresponding to a set of 600 patient records). The NB-C achieved a diagnosis accuracy of 80% with this small underlying knowledge graph (HealthKG) in the test run with 30 patients, but 75 observations could not be classified because they were not associated with a disease in the knowledge graph.

Figure 10.6 shows that the diagnosis quality decreases significantly with low data availability. The minimum diagnostic accuracy achieved in the test digital twin for 30 patients is still at an acceptable level of 50% with the presence of only about 5% of the observables. This can be explained by the fact that diseases prevail which have a very high chance to be correctly identified by the symptoms. For example, if a patient reports a cough, the diagnosis COVID-19 (with implementation of high incidences in the digital twin) is often the correct diagnosis because (I) many flu patients do not visit the hospital and (II) other diseases that manifest themselves in this symptomatology (e.g. lung cancer) are very rare. Thus, as long as individual symptoms are still known, the diagnoses are still relatively accurate due to the strong dominance of individual diseases (compared to the low data accuracy). Furthermore, the results presented in Figure 10.5 also reflect the degeneration of average precision of treatments and diagnoses (interpreted as a measure for the quality of care) which emerged due to the simulation of cyberattacks. We thereby simulated different attacks that were observed within the health care system. Our analysis shows that within our model it is of minor difference, whether data availability is impacted through an attack or the availability of

treatment procedures (in particular procedures undertaken for verifying a primary diagnosis such as medical imaging) is impacted through the attack, as both manifest in a decrease of observables or to a disease linked treatment plan. Yet, there are some differences considering the effect on time to relief: If a needed treatment is affected by a cyberattack, the patient would not gain relief/ would not be treated. However, for the undertaken evaluation procedure on a limited sample of patients there were no qualitative differences. As qualitatively the same result is obtained for the effects of cyberattacks on treatments, we restrict the detailed presentation of quantitative results on the diagnostic procedure. Furthermore, we limit the analysis on these quantitative effects without going deeper into the effects on patients' health. These effects on patients' wellbeing are oftentimes hard to assess and need special knowledge of medicines and are thus beyond the scope of this analysis.

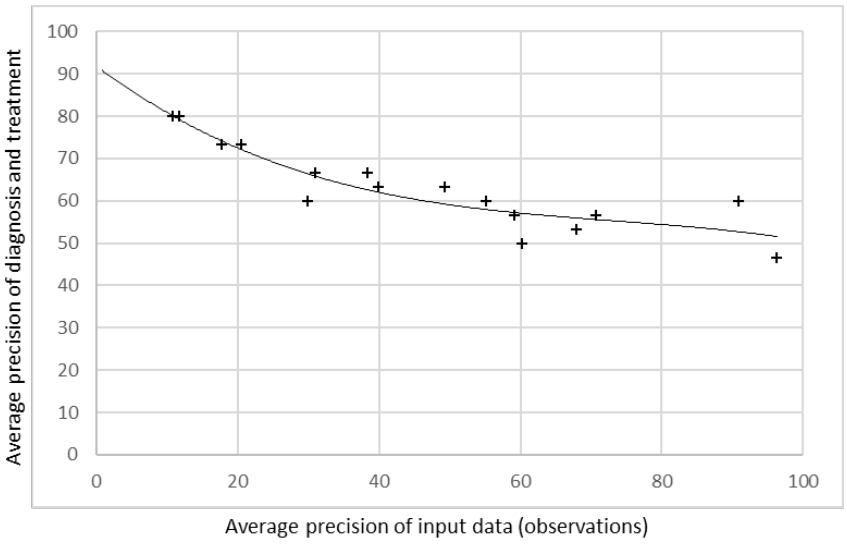

Figure 10.6: Average precision of diagnostic procedure in dependence on the percentage of missing observables

However, the potential of the digital twin is not limited to describing the effects of a cyberattack on diagnostic processes and their quality but can also be used to estimate the effects on unavailability of treatment plans (r_j, e.g., in case of the impairment of a medical device for ventilation or robot-guided surgeries). In the following, the benefits should be demonstrated in a qualitative manner by simulating the attack by WannaCry (S0366) in order to illustrate the effects of prevented services. The choice for WannaCry is due to the fact that the WannaCry attack is one of the best studied cyberattacks on healthcare facilities and the corresponding availability of data is ensured (Ghafur et al., 2019). WannaCry is linked to the impairment of the key systems for service provision which led in the NHS to a revert to manual service, rescheduling of patient appointments and the need for prioritization in service provisioning (Ghafur et al., 2019). All these workarounds observed in reality cannot be understood solely relying on the digital twin which is a major limitation. Yet, prohibited/ unavailable treatment plans are depicted differently in the digital twin. Prohibited treatments (e.g. through the stop/ denial of services) lead to wrong treatments/ suboptimal treatments and hence to an increase in time to relief/ lower probability of recovery as well as to no rewards (foregone compensation in the case of false treatments). Furthermore, the key strength of the presented digital twin lies in its ability to simulate the effects of bad

data quality and unavailability of data. Wannacry was also observed to cause inaccessibility of patient data causing significant obstacles for medical care. Once more, the model thereby does not consider workarounds that might be established to solve the obstacles, yet it predicts high rates of faulty diagnoses, and based on these mistreatments as well as high deprivation through these mistreatments. The qualitative analysis of the model based on the case of Wannacry highlights the strengths and weaknesses of the proposed solution for quantifying the effects of cyber risks based on the digital twin. The methodology helps to quantify the dimension of the problem (instead of problem solving measures), which should be correlated with decision errors, delay and e.g. reputational damage. It should be especially noted, that when analyzing the digital twin, the models boundaries need to be considered (e.g. no consideration of possible workarounds). Yet, the model predicted an economic loss through wrong treatments (which are then not compensated) which was close to the estimated real loss for a hospital effected by Wannacry within the NHS ($4.7m to $9.3m - simulated vs. $4.8m to $11m - observed).

The impact of cyberattacks, in turn, depends on their scale (e.g., there are redundancies or implemented security mechanisms in the system, which means the attack would not impact all areas). The impact of the encryption of data can depend on the severity of the attack as illustrated in figure 10.6. A loss of data implies a lower diagnostic precision and eventually results in a higher proportion of wrong decisions. The risks of some of the attacks recently used in the healthcare sector can be accessed via this effect of data unavailability. For example, the Clop malware (S0611) or WannaCry (S0366) have a negative impact on service provision through encryption (T1486) or the Kazuar malware (S0265) through data destruction (T1485). A special role is played by attacks with the capability of data manipulation (T1565), which may also introduce false evidence. In this case, accidentally correct diagnoses are also prevented, since the implementable indications speak for other diseases than the triggering disease. Besides these techniques, the forced stop of operations/service (T1489) is also considered a key threat for medical services, here the admission of secondary observables as well as the treatment of diseases can be obstructed. Representatives of these attacks include the malware Pysa (S0583) and REvil (S0496), which have also been observed in the healthcare sector. In the digital twin, it can be observed that the encryption of data and the destruction of data have a greater impact on the healthcare sector than the unavailability of treatment paths, as the latter can be used to switch to alternative treatment options in the digital twin. The impact of stopping diagnostic procedures to confirm an initial disease hypothesis can significantly inhibit the availability of diagnostic information, leading to large reductions in service quality. In the simulation of the impact of a cyberattack on the service delivery in a hospital, the malwares acting through encryption and data destruction have a higher impact than attacks directed against individual operations.

The results of this study are intended to quantify the damage to the provision of medical services as a whole. The evaluation of the impact of cyberattacks on surgical processes (e.g., surgical interventions) is not as large as the impact of incorrect diagnoses in this study due to the possibility of using alternative treatment paths. It should be noted here that this corresponds directly to the implementation of the digital twin (here, incorrect treatment of an acute illness is associated with greater negative effects than the use of an alternative treatment method). Long-term effects are not

reflected in the digital twin. An extension of the digital twin to include these effects represents a significant expansion of it and has the potential to more fully understand and quantify the impact of cyberattacks and thus the risk of an attack on medical care delivery. Therefore, this extension can be considered future work. The presented approach is highly flexible in terms of different attacks that can be simulated as it relies on the threat knowledge graph and the technical effects (i.e., impact techniques). The effect-cause chains can hence be tracked for those attacks included in the knowledge graph. Depending on the different impact techniques associated with the attack, different effects can be simulated including unavailability of medical devices causing to obfuscated observations on which a diagnosis need to be crafted but also unavailability of treatment plans and hence the need for medical professionals to postpone medical treatments or resort to alternative treatment plans. In this context, the potential of relying on a digital twin is limited only by the quality of the digital twin (knowledge on the technical infrastructure and devices employed within a hospital) and the quality of the threat knowledge graph (i.e., reliability of information and comprehensiveness). Furthermore, it should be noted that the more complex a system under consideration (e.g., big hospital vs. small medical practice) the more challenging it is to craft a detailed digital twin able to reproduce interrelationships in the proper functioning of devices (e.g. spillover effects of the impairment of one device to another). The simulation approach for assessing the effects of a cyberattack presented within this work is of especial relevance for medical service provisioning for acute cases (e.g. stroke or acute respiratory distress), where diagnoses may be crafted on limited evidence and treatments are hard to postpone (waiting for the medical devices to be restored and normal operations of the hospital is impossible).

10.5 Conclusion

With increasing threats from cyberattacks in the healthcare industry, the analysis and quantification of risks is of increasing importance to ensure high-quality medical care. This is particularly relevant to ensure that the goals of digitization in wide areas of the medical sector, namely improving service quality, are not eroded. The digitization of medical service provision can only be successful in the long term and in terms of improved service quality if the security of the digital devices is ensured. This paper provides decision-makers in the medical sector with a tool for quantifying the impact of cyberattacks in the medical sector. The work is limited to analyzing the impact of a cyberattack over a short time horizon. Cyber risks are considered as a risk vector and include not only monetary values but also non-monetary values such as the quality of medical care. Within the presented qualitative evaluation, the Digital Twin showed reasonable precision of estimations compared to real consequences in the case of a cyberattack (WannaCry). Yet, the evaluation of the model should be substantiated by analyzing more cases. A key obstacle thereby is data availability as well as high uncertainties in reported damage.

Acknowledgements

This work was supported by funding from the topic Engineering Secure Systems of the Helmholtz Association (HGF) and by KASTEL Security Research Labs.

References

Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B. M., Sweeney, L. (2018). Privacy preserving synthetic data release using deep learning. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 510–526). Springer, Cham.

Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, *185*, 103076.

Ait-Alla, A., Kreutz, M., Rippel, D., Lütjen, M., Freitag, M. J. (2021). Simulated-based methodology for the interface configuration of cyber-physical production systems. *International Journal of Production Research*, *59*(17), 5388–5403.

Allodi, L., Massacci, F. (2017). Security events and vulnerability data for cybersecurity risk estimation. *Risk Analysis*, *37*(8), 1606–1627.

Böhme, R., Laube, S., Riek, M. (2017). A fundamental approach to cyber risk analysis. In (Vol. 12, p. 161-185).

Chen, I. Y., Agrawal, M., Horng, S., Sontag, D. A. (2019). Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledgegraph. *Pacific Symposium on Biocomputing*, *25*, 19–30.

Chen, Q., Abercrombie, R. K., Sheldon, F. T. (2015). Risk assessment for industrial control systems quantifying availability using mean failure cost (mfc). *Journal of Artificial Intelligence and Soft Computing Research*, *5*, 205 - 220.

Chismon, D., Ruks, M. (2015). Threat intelligence: Collecting, analysing, evaluating. *MWR InfoSecurity Ltd.*.

de Melo e Silva, A., Gondim, J. J. C., de Oliveira Albuquerque, R., García-Villalba, L. J. (2020). A methodology to evaluate standards and platforms within cyber threat intelligence. *Future Internet*, *12*(6), 108.

de Paula Ferreira, W., Armellini, F., de Santa-Eulalia, L. A. (2020). Simulation in industry 4.0: A state-of-the-art review. *Computers & Industrial Engineering*, *149*, 106868.

Dube, K., Gallagher, T. (2013). Approach and method for generating realistic synthetic electronic healthcare records for secondary use. In *International symposium on foundations of health informatics engineering and systems*.

Eckhart, M., Brenner, B., Ekelhart, A., Weippl, E. R. (2019). Quantitative security risk assessment for industrial control systems: Research opportunities and challenges. *Journal of Internet Services and Information Security*, *9*(3), 52–73.

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. *2019 European Intelligence and Security Informatics Conference (EISIC)*, 40–47.

Elnagar, S., Weistroffer, H. R. (2019). Introducing knowledge graphs to decision support systems design. In *Eurosymposium on systems analysis and design* (pp. 3–11). Springer, Cham.

Ghafur, S., Kristensen, S., Honeyford, K., Martin, G., Darzi, A., Aylin, P. (2019). A retrospective impact analysis of the wannacry cyberattack on the nhs. *NPJ digital medicine*, *2*(1), 1–7.

Hemberg, E., Kelly, J., Shlapentokh-Rothman, M., Reinstadler, B., Xu, K., Rutar, N., O'Reilly, U.-M. (2020). *Linking threat tactics, techniques, and patterns with defensive weaknesses, vulnerabilities and affected platform configurations for cyber hunting.* arXiv.

Kaiser, F. K., Wiens, M., Schultmann, F. (2021). Use of digital healthcare solutions for care delivery during a pandemic-chances and (cyber) risks referring to the example of the covid-19 pandemic. *Health and technology*, *11*(5), 1125–1137.

Kuo, N. I.-H., Polizzotto, M. N., Finfer, S., Garcia, F., Sönnerborg, A., Zazzi, M., ... Barbieri, S. (2022). The health gym: Synthetic health-related datasets for the development of reinforcement learning algorithms. *ArXiv*, *abs/2203.06369*.

Leszczyna, R. (2021). Review of cybersecurity assessment methods: Applicability perspective. *Computers & Security*, *108*, 102376.

Mahler, T., Elovici, Y., Shahar, Y. (2020). A new methodology for information security risk assessment for medical devices and its evaluation. *ArXiv*, *abs/2002.06938*.

Mahler, T., Shalom, E., Makori, A., Elovici, Y., Shahar, Y. (2022). A cyber-security risk assessment methodology for medical imaging devices: the radiologists' perspective. *Journal of Digital Imaging*, 1–12.

Martínez, G. S., Sierla, S., Karhela, T. A., Vyatkin, V. (2018). Automatic generation of a simulation-based digital twin of an industrial process plant. *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 3084–3089.

McLachlan, S., Dube, K., Gallagher, T. (2016). Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 439–448.

Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y. (2019). Ct-gan: Malicious tampering of 3d medical imagery using deep learning. *28th USENIX Security Symposium (USENIX Security 19)*, 461–478.

Moniz, L. J., Buczak, A. L., Hung, L. M., Babin, S., Dorko, M., Lombardo, J. S. (2009). Construction and validation of synthetic electronic medical records. *Online Journal of Public Health Informatics*, *1*(1).

Mourtzis, D. (2020). Simulation in the design and operation of manufacturing systems: state of the art and new trends. *International Journal of Production Research*, *58*(7), 1927–1949.

Musman, S. A., Turner, A. J. (2018). A game theoretic approach to cyber security risk management. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *15*(2), 127–146.

Patel, S. C., Zaveri, J. S. (2010). A risk-assessment model for cyber attacks on information systems. *J. Comput.*, *5*(3), 352–359.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., Sontag, D. A. (2017). Learning a health knowledge graph from electronic medical records. *Scientific Reports*, *7*(1), 1–11.

Sauerwein, C., Sillaber, C., Mussmann, A., Breu, R. (2017). Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives.

Walonoski, J. A., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association : JAMIA*, *25*(3).

Wigand, R., Picot, A., Reichwald, R. (1997). *Information, organization and management: Expanding markets and corporate boundaries*. Wiley.

Woods, D. W., Böhme, R. (2020). Systematization of knowledge: Quantifying cyber risk..

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K. P. (2019). Privacy preserving synthetic health data. *ESANN 2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, *8*.

Zeller, G., Scherer, M. (2021). A comprehensive model for cyber risk based on marked point processes and its application to insurance. *European Actuarial Journal*.

11 Digital Twins and their use for Cyber Risk Quantification - Analyzing the Impact of Cyberattacks on an Automobile Manufacturer[1]

Abstract

The broad penetration of digital technologies in industrial production with the potential of high efficiency gains is a recent challenge keeping managers busy. This is, digitalization does not only bear great potential for companies but high vulnerability to cyber risks. Digital twins can take an important role in modern production systems. They allow analyses and experimentation in a highly dynamic environment, in particular the implementation of tests of new technologies and adaptive production processes, without having to compromise the actual physical production system. Furthermore, they can also be used as a means for quantifying cyber risks through the analysis of different (attack) scenarios. Within this work, we present a digital twin of an automobile manufacturer that is embedded in a virtual market. We introduce the concept of a digital twin, present an implementation based on Python and Blender3D and demonstrate the usability of digital twins for cyber risk quantification.

Keywords: Cyber risk quantification, automotive production, digital twin, 3D-simulation, process analysis, data analysis

11.1 Introduction

The automotive industry is of central economic importance for many countries (Traub-März, 2017). This can be explained by a high preference for individual mobility. However, the automotive industry is facing new challenges and proven forms of value generation are in question. Established companies are in danger of becoming obsolete and need to redefine their role in value generation.

[1] This chapter includes the preprint of the article "Digital twins and their use for cyber risk quantification - Analyzing the impact of cyberattacks on an automobile manufacturer" by Marcus Wiens, Frank Schultmann, and myself.

In order to remain competitive in the future, immense investments will and are already made in technologies. Many of these investments are associated with great risks and uncertainties. Digitalization investments are one of the riskiest investment as they may introduce cyber risks to the manufacturing companies. Quantifying the benefits and risks associated with digitalization investments are thus essential for companies in these times (Urbach and Ahlemann, 2016). However, there is a lack of scientifically grounded methodologies helping professionals within this process. Especially quantifying cyber risks thereby represents a scientific field only small support is given with actionable models (Radanliev et al., 2018).

Simulations have been frequently used in risk management and investment decision making to build a foundation for economic decisions (Schlüter et al., 2019). For digitalization investments, commonly used simulations however oftentimes provide too high levels of abstraction causing a lack of precision of analyses (Schlüter et al., 2019). For analyzing cyberattacks in a quantitative manner as well as for quantifying economic benefits of digitalized production processes, the high level of abstraction frequently represents a burden and causes a lack in precision. Hence, in recent years, conventional, model based simulations have oftentimes been replaced by digital twins based simulations offering higher precision while taking advantage of real world data which leads to high predictive capability (Ermolina et al., 2021; Holmes et al., 2021). This is, digital twin based simulations frequently outperform conventional, model based simulations by including high levels of detail allowing in-depth quantitative analyses on the underlying system. Digital twins have thereby also proven to be able to test investments in digitalization and security technology (Holmes et al., 2021; Zio, 2018). Furthermore, extensions beyond the classical concept of digital twins (e.g. inclusion of market interactions or as proposed in Smogeli (2017) simulated environments) can be considered an interesting option to further increase the predictive capability. In this sense, virtual market modules can be added for enabling predictive pricing and the estimation of the market success in uncertain market areas.

Within this work, we take a perspective on quantifying the impact of successful cyberattacks by means of simulations based on a digital twin. For this purpose, this paper develops a digital twin of a model factory and embeds it into a simplified industrial economic market model (virtual market). The digital twin is thereby crafted as a modular approach. We thereby generate a module focused on the company (manufacturer model) and a module that aims at predicting customers market behavior (customer model). Within the manufacturer model, an automobile manufacturer is described producing vehicles. The company is equipped with modern production technology, where the degree of digitization can be varied. Costs of the production process as well as parameters for describing the production are simulated by the module relying on techno-economic analysis. The customer model includes estimations on potential customers' willingness to pay. For this purpose, the model takes advantage of a quantitative Kano model (Buhl et al., 2007). The model hereby is linked to the real automotive market by open source intelligence in the form of freely available customer reviews. Both modules flow together in the market module (virtual market). The production decision of the company is the result of the customers' willingness to pay for a specific variant of an automobile and the costs of producing the special variant. We use the digital twin to quantify the value of investments in digitalization (technology investments) considering the

effects of cyber risks. The risks are quantified relying on process value analysis. The model can also be used to generate synthetic industry data. This is of particular importance when simulating network data within the model. Here, malfunctions triggered by cyberattacks can be simulated which can be a significant contribution to scientific research in the field of cyber risk management that suffers from a lack of data. The digital twin could at least partially relax the limitations of data availability by generating realistic network protocols that reflect the processes within the digital factory, thereby contributing to the domain of quantitative cyber risk assessment based on digital twins (Eckhart and Ekelhart, 2019), and can be used as a testbed for different methods that are employed in cyber defense (e.g. threat hunting). A schema of the proposed methodology is presented in figure 11.1.

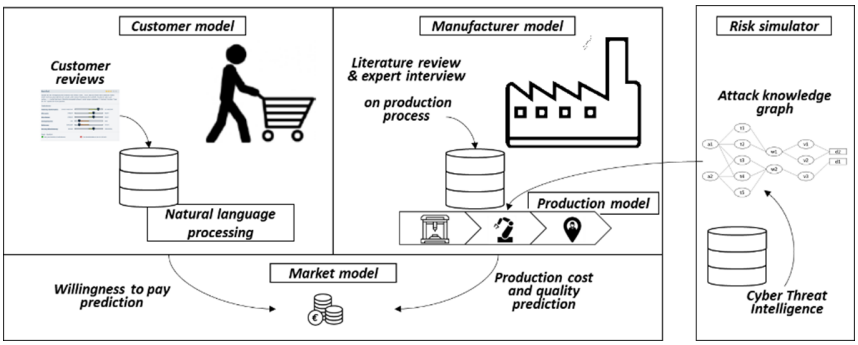

Figure 11.1: Schema of the proposed methodology

11.2 Theoretical background and related work

11.2.1 The automotive value chain

The automotive industry accounts for a large share of value added in many nations around the world and represents a significant economic factor for many nations. In Germany, in particular, 7% of all employees, subject to social insurance contributions, are employed in the automotive industry (Hagedorn et al., 2019). Besides the great importance of automotive production systems nowadays, Staiger and Tantău (2020) forecast a strongly growing international automotive market highlighting the importance of the industrial branch for the future. According to their study, the number of vehicles in operation is expected to nearly double in the coming decades.

According to Kropik (2021), automotive production is a complex, multi-stage production system. Individualization has led to considerable segmentation of the automotive market and to a diversity of vehicles (variants) including different designs, shapes and equipment (Naunheimer et al., 2010). The high number of different variants leads to a correspondingly large number of production processes and testing steps that are required. In addition, Kropik (2021) points out that there are dependencies and feedbacks between the different production processes of an automobile production, which increase the complexity of automobile production. Müller et al. (2021) and Telek

and Bányai (2018) provide a comprehensive overview on the different manufacturing processes included in the production of an automobile.

11.2.2 Digital twins

As production systems get increasingly complex with proceeding digitalization, there is a growing interest in modelling these complex systems with high degrees of detail allowing in-depth analyses of its properties, the benefits that can be reached through digitalization as well as the risks that come with it. A concept that aims to formalize this detailed image of production systems is known as digital twinning. Digital twins were first mentioned by Michael Grieves in 2003. Martinez et al. (2018) define a digital twin of a production plant as "a digital replica of the plant's physical assets which contains the structure and the dynamics of how the devices and processes operate" (Martinez et al., 2018). Digital twins gained in importance as a means to visualize, simulate, and analyze economic systems and especially production plants. However, digital twins are also considered an effective means in risk management (Holmes et al., 2021). In this context, simulation is used as a fundamental method for analyzing complex production systems (de Paula Ferreira et al., 2020). Mourtzis (2020) adds that digital twinning is still in its infancy. However, it represents a promising approach in manufacturing simulation, which is not yet mature and not yet presented in its full capabilities in the literature (Mourtzis, 2020). Here, the real object and its digital twin are connected by shared information, which enables the twin to simulate processes as a spitting image and make predictions for the real world.

Tao, Qi, et al. (2019), Kritzinger et al. (2018), Qi and Tao (2018), and (Negri et al., 2017) conduct literature reviews for digital twins in manufacturing with different focus areas. These are the degree of integration and the application of digital twins (Kritzinger et al., 2018), the definition and role of the digital twin (Negri et al., 2017), its relation to big data (Qi and Tao, 2018), or the similarities and differences between digital twins and cyber-physical systems in manufacturing (Tao, Qi, et al., 2019). Next to these specialized reviews, Rasheed et al. (2020) and Tao, Zhang, et al. (2019) contribute two more general literature reviews. Rasheed et al. (2020) focus on the modelling and construction of a digital twin by clustering into physics-based modelling, data-driven modelling, big data cybernetics and infrastructure and platform and human-machine interface. Tao, Zhang, et al. (2019) reflect the state of the art in terms of the applications of digital twins in industry by clustering the applications into the four phases product design, production, prognostics and health management and other areas. The division of simulation and visualization as well as the realization of the digital twin within the game engine Blender3D followed in this work was motivated by the work of Lind and Skavhaug (2012) describing a similar approach. Thereby an emulation system was implemented for a real material transport and a handling production line using Python. Figure 11.2 gives an overview on a systematic literature review we performed highlighting the increasing research interest in digital twinning and geospatial differences in scientific research efforts related to digital twins.

Within this work, we aim at contributing to the scientific discussion on digital twins and their meaning for managing industrial production systems. We thereby develop a digital twin of an

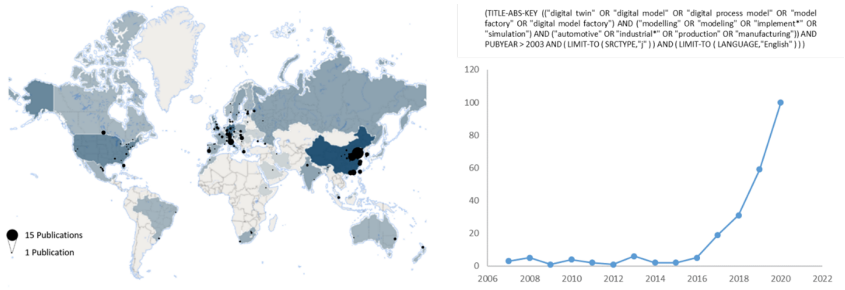

Figure 11.2: Publications by region based on a systematic literature search conducted within scopus ($n = 253$)

automotive manufacturing plant representing a digital replica of a fictious manufacturing plant containing central assets and allows a formalization. The digital twin is able to simulate the dynamics of the operation of the plant.

11.2.2.1 Calibration of digital twins

Calibration of digital twins is essential for enabling analyses with high predictive capabilities for real world systems (Holmes et al., 2021). This is, the process of calibration is essential for the validity of results derived from a digital twin (Brozovsky et al., 2018). Digital twins thereby frequently rely on data intensive/ data driven methods (Semeraro et al., 2021). Big data analytics hence can be considered the technological backbone of digital twins (Tao, Sui, et al., 2019).

Digital twins of production systems can be calibrated relying on data of the underlying production system. Foremost, this information is gathered from the production process (e.g. network artefacts) through a variety of sensors introduced to the system (Brozovsky et al., 2018). Physics simulations can be further helpful elements of improving the precision of analyses based on digital twins (Ritto and Rochinha, 2021; Aivaliotis et al., 2019). The more complex the underlying represented by the digital twin and its operation, the more details need to be considered when calibrating.

11.2.2.2 Methods for generating models of human agents

Relying on customer models, digital twins can reflect customer preferences to assess the customer satisfaction and market success (e.g. predict market shares) by emulating human behavior (Semeraro et al., 2021). Like models simulating industrial production, models aiming at emulating human behavior can be based on databases (Graessler and Pöhler, 2017). Graessler and Pöhler (2017) emulate a human employee interacting with machinery within a production system. They craft a human digital twin representing the employees' preferences and skillsets. Human agents are furthermore the main agents that need to be considered when aiming at simulating the market organizations are operating on. For understanding customers' demands and emulating their behavior, digital twins can be calibrated based on demographic information, artefacts of behavior (e.g. purchase or online search history) or revealed preferences (e.g. product ratings or reviews; Tao, Sui, et al. (2019)).

Customer reviews are a valuable source of information for predicting human behavior and are available on a plethora of different platforms. These platforms allow consumers to express their opinions and criticism on a product, as well as giving ratings to various products and services. The information is widely used by companies as it bears the potential to gain insight into the customer needs and expectation with respect to a product. The information can be acquired at low cost and in an automated manner. As the availability of online documents is increasing day by day, the task of automatic classification of documents and reviews is becoming an essential method for the calibration of customer behavior. The use of customer reviews is thereby motivated by works highlighting the great potential (e.g., Bi et al. (2019)).

Yet, analyzing customer reviews is challenging as only low shares of data are structured while most customer reviews are unstructured (e.g. in a textual form). Gaining insight from this type of data hence needs the employment of advanced analytics (i.e. text mining techniques and natural language processing; (Sun et al., 2017; Hanni et al., 2016)). Natural language processors thereby generally take advantage of various steps which are tokenization, stemming, lemmatization, part-of-speech-tagging, named entity recognition, and chunking (Sun et al., 2017; Hanni et al., 2016). Tokenization is the process of splitting a text into tokens. Here, tokens can be either words, characters, or sub words. In general, it represents the first step in natural language processing. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words. Lemmatization usually refers to the use of vocabulary and morphological analysis of words, aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. Part-of-speech-tagging is the process of marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition. This task is not straightforward, as a particular word may have a different part of speech based on the context in which the word is used. Named entity recognition is the process of detecting the named entities (objects) such as the person name, the company name or the product name. Changing a perception by moving a “chunk”, or a group of bits of information, in the direction of a deductive or inductive conclusion through the use of language is referred to as chunking. Among different methods for text analysis, sentiment analysis is one of the most frequently referred techniques. Sentiment analysis consists in building automatic tools capable of extracting subjective information from natural language texts, in order to create structured and exploitable knowledge that can be used by a digital twin, a decision support system or a decision maker.

11.2.3 Digitalization investment analysis

11.2.3.1 Cyber risk quantification

Cyber risks pose a significant threat to modern society and economy. This is as modern production systems are increasingly dependent on information and communication technology (ICT). However, managing cyber risks proved to be challenging for professionals and information security managers as although there are plenty methods for risk quantification there is no commonly accepted methodology (Leszczyna, 2021; Radanliev et al., 2018). Instead cyber risk management is

mainly based on accepted risk management manuals, norms, and standards which are frequently based on qualitative or semi-quantitative approaches for risk assessment. Furthermore, insuring cyber risks is oftentimes hardly possible, as only limited numbers of polices are available and pricing of cyber risk insurance suffers the same challenges of a methodological lack (Radanliev et al., 2018). Hence, security experts oftentimes rely on their experience, heuristics or qualitative risk assessment (e.g. expert judgements).

Although academic research increasingly acknowledges the impact of these risks on economic performance, there is a lack of a systematic formal methodology for quantifying risk (Radanliev et al., 2018). This is all the more necessary as common risk quantification practices, which are mainly based on statistical approaches, are not suitable or at least insufficient (Elitzur et al., 2019). This is because statistical methods require a large database to derive well-founded statements. However, such a large and comprehensive database (also referred to as Cyber Threat Intelligence; CTI) is often not available in the field of cyber-risks on a strategic level or does not include all relevant data (especially rare is data on the impact of attacks). There are several reasons for this, such as the unwillingness to share data and the fact that there is often low comparability between cases. Furthermore, the field of cyber-security is very dynamic. New attack strategies are constantly being developed so that the accumulation of data to derive statistical forecasts is severely limited by time. Tactical and strategic interactions of mutual adversaries are another reason for the poor suitability of statistical risk assessment. Since the counterparties on both sides (attacker and defender) make an adjustment to the strategy chosen by the other (strategic/ intelligent adversaries; Kaiser et al., n.d.). In risk quantification, it must be taken into account that the opponents are able to adapt and can learn from their experience.

Eckhart et al. (2019) show that the consequences of cyberattacks are also not yet well enough understood. Determining the consequences of an attack is very challenging as ICT does not generate value directly but indirectly by enabling business processes (Krcmar, 2015). These processes vice versa can be matched with attributes that are valued by a customer. Summarized, ICT "does not create value in isolation, but must be a part of a business value creating process" (Kohli and Grover, 2014). Hence, the quantification of impacts caused by manipulated or influenced ICT should draw on process based value creation. Cyber-risk quantification thereby stands in contrast to the quantification of other risks that have a direct impact on values, such as a natural disaster that destroys physical assets or causes environmental damage and therefore needs another approach. Current efforts in research on cyber risk quantifications are given within the systematic literature review provided by (Leszczyna, 2021). For an in-depth review on different methods for risk quantification we refer to their work.

As till now there is no widely accepted methodology for cyber-risk quantification, professionals often need to craft it manually which is prone to personal biases and limited to subjective perspectives on the threat landscape (Riesco and Villagrá, 2019). Consequently, professionals frequently struggle to adequately deal with cyber-risks (Mossburg et al., 2016). This leads to many challenges not only for managers deciding whether to invest in a specific security measure or insurance but also for insurers and providers of security solutions who need to set a price for their products. This is, the methodological lack can be considered a significant burden for efficient and effective cyber

risk management as well as for the establishment of flourishing cyber security related businesses and services. Cyber threat intelligence however is considered to be able to give relief to the problem of cyber risk quantification providing novel opportunities for developing methodological support (Riesco and Villagrá, 2019).

11.2.3.1.1 Simulations in cyber risk quantification Simulations are a means frequently considered when aiming at modeling risk quantitatively (Zio, 2018). Kouril et al. (2014) present a cloud based security testbed that allows to "simulate various security attacks and evaluate their impacts". Kuhl et al. (2007) introduce an approach for simulating cyberattack scenarios. They hereby rely on discrete event simulations. The simulation is based on information known about the attack, e.g., the attack goal, attack steps/ action. The approach presented within their work is thereby restricted on 25 attacks with up to 250 attack steps each (Kuhl et al., 2007). Leszczyna et al. (2008) and Leszczyna et al. (2010) present a malware simulation tool that provides valuable insights "for the qualitative and quantitative evaluation" of cyber risks respectively assessing the security of a system. For simulating the effect of an infection with a specific malware the tool implements attack scenarios and analyzes the impact the attack would have under the specific conditions. Hence the malware simulation tool provides helpful means for security professionals when aiming at quantifying cyber risks, analyzes the robustness of a system, and identifies security measures that can efficiently minimize the risk (Leszczyna, 2019). Xiong et al. (2022) present an approach for simulating cyberattacks based on cyber threat information extracted from the MITRE ATT&CK database. Our work is based on these initial works of simulating the effects of cyberattacks.

11.2.3.1.2 Process value analysis and its use for cyber risk quantification Value creation is a complex construct. A (production) activity always creates added value if the benefit it generates for the customer (or cost savings for the producer) is greater than the expense of the activity. Inversely, value is destroyed if this is not the case (Porter, 1985). On the demand side, the customers' willingness to pay plays a central role. Willingness to pay becomes a function of the fulfillment of the customer's needs and consequently a function of customer satisfaction. The needs of a customer are thereby multilayered and complex. An established procedure for the illustration of these is the quantitative Kano model (Buhl et al., 2007), which forms the basis of a process value analysis (Janatyan and Shahin, 2020). In addition to determining the willingness to pay, the resource input must be quantified. Techno-economic analyses are established for this purpose (Abas et al., 2019; Sakti et al., 2015).

The process value analysis combines these two components and can thus be seen as an instrument of controlling, which assigns to each process within a company the value added as the corresponding share of the product revenues (Orojloo and Azgomi, 1996; Beischel, 1990). The creation of value comprises the producer surplus (entrepreneurial contribution margin) as well as the consumer surplus but of particular importance for the company is the share that remains within the company, i.e. profit. So far, the focused perspective on processes and their contribution to value generation is rarely encountered in practice. Rather, financial analyses in different divisions or with regard to

individual products dominate. In particular, the process value oriented analysis of entrepreneurial activity offers new possibilities for risk management and especially for the management of risks, which have a direct impact on processes and their unhindered flow. Applications of different forms of process value analysis can be found for example in Borgianni et al. (2010) in the field of fashion industry, Borgianni et al. (2011) within biomass production (wood pellets) or in Janatyan and Shahin (2020). We introduce process value analysis in combination with a market model as a means for quantifying risks that are associated with malfunctioning of specific steps of the value creation and hence use measures derived from process value analysis as a means for analyzing the introduced digital twin.

Roumani et al. (2016) propose the use of process value analysis for cyber risk assessment. Henry et al. (2009) use petri nets to analyze cyber risks. Orojloo and Azgomi (2017) model the propagation of attack consequences using process modeling. By using process based cyber-attack impact assessment, it is furthermore possible to account for supply chain attacks. Musman and Turner (2018) provide a methodology that "quantitatively identifies cyber-security risks and uses this metric to determine the optimal employment of security methods for any given investment level" (Musman and Turner, 2018). Within their work, the likelihood of an attack is calculated by using a defender attacker model. Furthermore, the impact is calculated using a business process model. In doing so, they are able to calculate the impact of a cyber-incident on mission fulfilment. However, their approach of process evaluation lacks a market perspective. Market outcomes are essential for value determination. Hence, including market outcomes and a value perspective is assumed to deliver significant benefit. Our methodology unifies the pure technical modelling of business processes with the market- and customer- perspective on processes. Therefore, the business process model is enriched by a process value analysis. In doing so, the value contribution of processes can be assessed. Furthermore, our methodology differs from the approach provided by Musman and Turner (2018) which uses expert judgements for process valuation, as we derive the process value from customer requirements (e.g. assessable through market research or online customer reviews) and their valuations on the market.

11.2.3.1.3 Attack graphs For cyber risk management, the use of attack graphs is considered by many scientists a valuable means in quantifying risks. Salfer and Eckert (2018) present an approach for assessing the exploitability of systems based on attack graphs. As a use case they focus on assessing the exploitability of automotive on- board networks. Gylling et al. (2021) present a method for adapting generic attack graphs to cyberattacks relying on the use of CTI. They thereby take advantage of the ATT&CK database to craft probabilistically weighted attack graphs and make the attack graph based approach for cyber risk assessment actionable. Wu et al. (2017) present an approach for cyber risk assessment based on an approach reasoning on attack graphs. The attack graph they rely on is used for identifying possible attacks (logical attack paths) from the set of attacks included within the knowledgebase (Kaiser et al., n.d.). introduce a game theoretical model based on weighted attack graphs that can be calibrated relying on CTI.

11.2.3.2Quantification of the benefits of digitalization investments

The use of digital technologies in companies is steadily increasing and is becoming a key success factor. In this context, digitization of business processes is associated with increased competitiveness. Digitaization projects are associated with strong uncertainty. This requires a realistic forecast of the revenues associated with the project. However, oftentimes companies lack in a comprehensive understanding of the value that can be created through digitalization (Neumeier et al., 2017). Digitalization investments offer the opportunity to improve operational efficiency by reducing (production) costs, enabling better cooperation with partners (e.g., closer networking along the value chain system), tackle the shortening of product lifecycles (e.g., through computer aided engineering), automated and continuous quality control, and reducing waste (e.g., through additive manufacturing processes; Chan and Ahuja (2015); Barua et al. (2004)). Hence, investments in ICT contributes to the financial performance of firms, business process improvements, customer value creation and satisfaction (Melville et al., 2004; Masli et al., 2011). In addition, ICT can serve as an enabler and open up new business models for companies, thereby creating new sources of competitive advantage (Bleicher and Stanley, 2019).

ICT investments are often associated with a considerable amount of uncertainty, no immediate payouts or political gambling (Weill and Olson, 1989), making them particularly difficult to assess. According to Kim and Sanders (2002), this difficulty is related to the characteristics of the profits pledged by ICT. These characteristics include that ICT does not deliver value directly. Instead, ICT delivers value indirectly e.g. by enabling processes that generate value (Krcmar, 2015). The rapid change and the high dynamics of the development of the underlying technology adds to these challenges. Investments in ICT are usually driven by competitive forces and are therefore closely related to corporate strategy decisions (Weill and Olson, 1989). It has been observed that they bring higher gross margin returns to companies compared to the scenario where firms forego the investment. Thus, choosing the right digitalization strategy is essential, especially when considering the changing (digitalizing) corporate environment. Consistently several studies have revealed that digitization can lead to fundamental changes and can significantly impact a firm's competitiveness and profitability (Pagoropoulos et al., 2017). Thereby Moshiri (2016) reveals within an industry analysis that secondary and tertiary industries have benefited from ICT investments much more than primary sector industries.

According to Porter (1985), the competiveness of a company is shaped by five forces determining the company's position within its sector and competitive environment. These are the rivalry within the sector, the threat through new competitors, the threat through substitute products, the bargaining power of the suppliers, and the bargaining power of the consumers. Hence, ICT investment value is dependent on the strategic decisions of the competitive environment in which they must be evaluated. Hence competition threat has an influence on ICT investment decision making. This is because investments may cause strategic answers under the rivals within a branch (Angelou and a. Economides, 2006).

On the customer side, the increased and heterogeneous demand for individualized products has led to competitive pressure and the need to differentiate between firms and industries. Prior

research has revealed that digitization can especially change the way how firms create and capture value. Reinartz et al. (2019) have identified five new sources of value creation enabled by digital technologies and discussed how these sources can enhance perceived customer benefits and thereby create competitive advantage. Other studies have started documenting several industrial cases of the development towards digital servitization. Firms are shifting their focus from delivering exclusively industrial products to providing combined product and service offerings with advanced functionalities based on technologies (Kowalkowski et al., 2016). Coreynen et al. (2017) proof and investigate how digitalization can enable servitization for manufacturers to offer a higher level of value added services to customers. Further, Cenamor et al. (2017) discuss how a platform approach leverages the value of digital and information technologies for advances service offerings and can therefore help firms by adding services while simultaneously maintaining their cost levels. Subsequently, digital technologies such as embedded devices, analytics or Product Service Systems lead to changes in customer relationships, internal processes and value proposition (Pagoropoulos et al., 2017). In addition, they can as well serve to build resilience for disruptive crises (Rapaccini et al., 2020). Last, consumers invest less in new products why an integrated service strategy can be seen as a necessary solution to make firms more agile and resilient, which is also an advantage in the case of future disruptive crises (Rapaccini et al., 2020). For quantifying the effects (both opportunities and risks) that can be associated with ICT, we rely on process value analysis as it is presented in Section 11.2.3.1.2.

11.2.4 Digital twins in cyber risk management

Digital twins can be used in cyber security for increasing analytical and predictive capabilities and understanding the underlying systems behavior (Holmes et al., 2021). Holmes et al. (2021) introduce the potential of using digital twins for automating the analysis of these systems. The digital twin can thereby be used to understand the impact of different system configurations and continuous validation of security (Holmes et al., 2021). On a more technical level, Lin et al. (2021) provide evidence on the practical use of digital twins for automated software risk analysis. Das and Morris (2018) introduce the use of digital twins for risk evaluation and assessment. The digital twin thereby includes highly detailed descriptions of various elements of an oil terminal including oil pumps of different configurations (Das and Morris, 2018). These different product configurations lead to different exploits linked to the pumps and hence to different vulnerability to cyberattacks. Within their approach, “simulation results involving normal operations and cyberattack scenarios are presented” for an oil terminal (Das and Morris, 2018). The risk is assessed comparing normal operations and operations under a cyberattack. The presented model enables a quantification of the impact of cyberattacks in different dimensions including measures for property (monetary harm) and life (non-monetary harm). The digital twin can therefore be considered a large-scale testbed including multiple interconnected industrial control systems and their interaction demonstrating great potential for cyber security research. Inspired by their work, we apply a similar approach to cyber risk quantification based on a digital twin of an automobile manufacturer.

11.3 Methodology

11.3.1 Implementation of the digital twin

11.3.1.1 Customer model

The customer model describes the willingness to pay of a customer in dependence on the fulfillment of its requirements posed on an automobile. We thereby base the estimation on a quantitative Kano-model (Buhl et al., 2007). We consider the willingness to pay to be a function of customer satisfaction, which is in line with Homburg et al. (2005). Customer satisfaction vice versa is a function of the fulfillment of customer requirements which is consistent with expectancy-disconfirmation paradigm. Yet, disconfirmation with a specific requirement can have different effects. To parametrize the customer model, we rely on natural language processing, especially on a sentiment analysis based approach for analysis of open source customer data (customer reports; Lu et al. (2021)). These feedbacks reflect the degree of the general satisfaction/dissatisfaction experienced by consumers towards a product. From an entrepreneur's point of view, a profound analysis of this data can be an important tool to predict future trends and to understand possibilities that may emerge from introducing novel qualities to the product offered, their perception and the emergence of competitive advantages.

A customer decision starts with the comparison of expectations $e_{l,t,i} \in [0,1]$ for a specific attribute i of a target group l at time point t and performance $k_{t,i,t} \in [0,1]$. There can be confirmation of expectations as well as positive (surprise) and negative (disappointment) disconfirmation (Buhl et al., 2007). Although the preference structures of customers are considered to be relatively constant over time, there may be some variation Furthermore, attributes develop over time. Therefore, attractive attributes will become performance attributes in the long run and may then turn into basic attributes. Especially due to technological progress and increasing innovation cycles, the timeliness of performance evaluation is an important factor, which needs to be considered. Within a quantitative Kano-model the confirmation of customer expectations $x_{l,t,i}$ lies between 1 and -1, where 1 describes extreme positive disconfirmation, 0 describes confirmation, and -1 extreme negative disconfirmation with respect to a specific attribute i (Buhl et al., 2007). According to the expectations confirmation model, expression 11.1 describes the discrepancy between expectations and actual performance of a product with respect to an attribute.

$$x_{l,t,i} = k_{t,i,t} - e_{l,t,i} \tag{Exp. 11.1}$$

This disconfirmation respectively confirmation is the origin of customer satisfaction. The contribution to customer satisfaction of attribute i can be described by $s_{l,t,i} \in [-1,1]$ (Buhl et al., 2007).

$$s_{l,t,i,k} = f(x_{l,t,i,k}) = \begin{cases} f_a(x_{l,t,i}) & \forall \quad k = a \\ f_p(x_{l,t,i}) & \forall \quad k = p \\ f_b(x_{l,t,i}) & \forall \quad k = b \end{cases} \qquad \text{(Exp. 11.2)}$$

The index a indicates attractive attributes, p describes performance attributes, and the index b relates to basic/ threshold attributes. These functions can further be specified with respect to the Kano model (see 11.3, 11.4, 11.5). Of importance is the distinction of different product attributes and their importance for customers (as these attributes are compared with their requirements). Thereby a failure to meet expectations has a stronger effect than meeting expectations, which is analogous to prospect theory (Kahneman and Tversky, 2013). According to Buhl et al. (2007), this aspect can be formally taken into consideration by the following formulation. Where the index x denotes an attribute from class a, y an attribute of class p, and z an attribute from class b.

$$s_{l,t,x} = \begin{cases} 0 & \forall \quad x_{l,t,i} \in [-1,0], \text{and } \underline{u} \in (-1,0) \\ \overline{u}x_{l,t,i} + qx_{l,t,i}^2 & \forall \quad x_{l,t,i} \in (0,1], \overline{u} \in (0,1), \text{ and } q \in (0, 1-\overline{u}] \end{cases} \quad \text{with } |\underline{u}| > |\overline{u}| \qquad \text{(Exp. 11.3)}$$

$$s_{l,t,y} = \begin{cases} \underline{u}x_{l,t,i} & \forall \quad x_{l,t,i} \in [-1,0], \text{ and } \underline{u} \in [-1,0] \\ \overline{u}x_{l,t,i} & \forall \quad x_{l,t,i} \in (0,1], \text{ and } \overline{u} \in [0,1] \end{cases} \quad \text{with} |\underline{u}| > |\overline{u}| \qquad \text{(Exp. 11.4)}$$

$$s_{l,t,z} = \begin{cases} -1 & \forall \quad x_{l,t,i} < c \\ -\underline{u}x_{l,t,i} - \frac{u}{c}x_{l,t,i}^2 & \forall \quad x_{l,t,i} \text{ if } c \leq x_{l,t,i} < 0, \text{ and } \underline{u} \in (-1,0) \\ 0 & \forall \quad x_{l,t,i} \geq 0 \end{cases} \qquad \text{(Exp. 11.5)}$$

The values u, q, and c denote scaling parameters. As it is shown in figure 11.5 the qualities of the functions describing the different attribute class differentiate highly. Attractive or excitement attributes lead to high satisfaction if there is high positive disconfirmation. However, customer dissatisfaction does not decrease in the case of negative disconfirmation (this area is only hypothetical, as attractive attributes are surprising for customers because they do not expect them). Threshold or basic attributes deliver high dissatisfaction when there is negative disconfirmation. However, there is no surplus in customer satisfaction in the case of positive disconfirmation (the customer expects these attributes). Furthermore, there are performance attributes. These are one-dimensional and reflect positive effects on satisfaction if there is positive disconfirmation while there are negative effects on satisfaction (dissatisfaction) if there is negative disconfirmation. However, the negative effect of negative disconfirmation is higher than the positive effect of positive disconfirmation (analogous to Prospect Theory; Kahneman and Tversky (2013)).

We define $y_{l,t,i,k}$ as the importance of attribute i from attribute class k (attractive, performance or basic). $y_{l,t,x}$ would therefore represent the importance of attribute i for customer attractiveness,

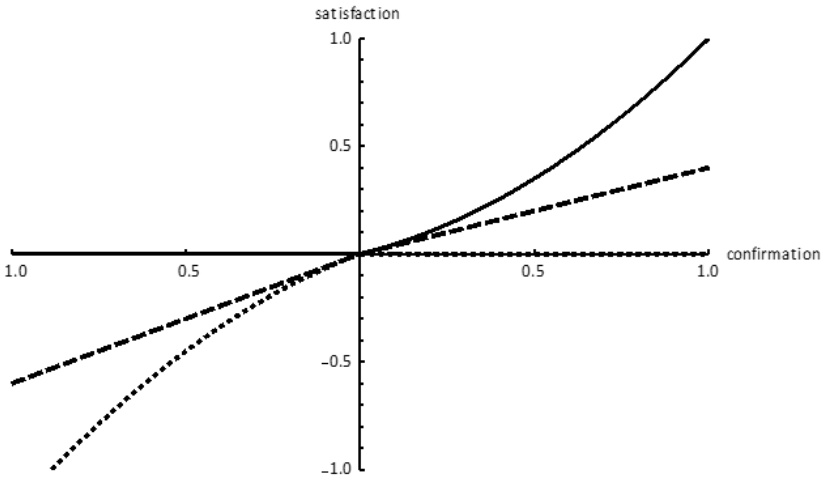

Figure 11.3: Kano Model

$y_{l,t,y}$ coherently represents the importance of attribute i for customer performance evaluation, while $y_{l,t,z}$ represents the importance of attribute i for customer threshold evaluation. Furthermore, A, P, and B are defined as the sum of the weights over all product attributes (Borgianni et al., 2010).

The contribution to customer satisfaction of each attribute i can be aggregated according to the contributions of the individual categories (attractive, performance and basic). The type of aggregation differs according to the specifics of the quality of attributes. For example, performance attributes can be weighted between the attributes, whereas this is not possible for basic requirements (Buhl et al., 2007). Therefore, performance attributes need to be aggregated using a summative formula 11.7 while basic as well as attractive attributes need to be aggregated using a multiplicative formula (see expressions 11.6 and 11.8).

$$\hat{S}_{l,t,a} = 1 - \prod_{x}(1 - \frac{y_{l,t,x} * s_{l,t,x}}{A}) \quad \text{(Exp. 11.6)}$$

$$\hat{S}_{l,t,p} = \frac{1}{n(p)} \sum_{y} \frac{y_{l,t,y} * s_{l,t,y}}{p} \quad \text{(Exp. 11.7)}$$

$$\hat{S}_{l,t,b} = \prod_{z}(1 + \frac{y_{l,t,z} * s_{l,t,z}}{B}) - 1 \quad \text{(Exp. 11.8)}$$

Last, the customer's overall satisfaction $COS_{l,t}$ of a target group l at time t with a product can be calculated according to the following expression by unifying the contributions to customer satisfaction of each attribute class (Buhl et al., 2007).

$$COS_{l,t} = \frac{\overline{s}^{\alpha}_{l,t,a} \overline{s}^{\varphi}_{l,t,p} \overline{s}^{\beta}_{l,t,b}}{0,5^{\alpha}(\frac{\overline{u}+1}{2})^{\varphi}} \quad \text{(Exp. 11.9)}$$

with $\overline{S}_{l,t,a} = \frac{\hat{S}_{l,t,a}+1}{2}$, $\overline{S}_{l,t,a} \in [0.5, 1]$, $\overline{S}_{l,t,p} = \frac{\hat{S}_{l,t,p}+1}{2}$, $\overline{S}_{l,t,p} \in [(\underline{u}+1)/2, (\overline{u}+1)/2)]$; and $\overline{S}_{l,t,b} = \frac{\hat{S}_{l,t,b}+1}{2}$, $\overline{S}_{l,t,b} \in [0, 0.5]$. The variables α, φ and β are weighting parameters of a Cobb-Douglas-Function. The *Willingness to Pay (WTP)* of a specific target group at time t can furthermore be calculated as a function of $COS_{l,t}$ (11.10). For now, this function will not be specified.

$$WTP_{l,t} = WTP_{l,t}(COS_{l,t}) \qquad \text{(Exp. 11.10)}$$

Customer's willingness to pay hence is dependent on the quality as a factor of actual performance of a product within relevant attributes of this product. 'Thus, quality describes an individual matter and is determined by the ability of a product to satisfy a customer's preferences and requirements.' (Schuh et al., 2017) WTP is thereby described by a basic price p_0 and the price premium, which reflects customer satisfaction (Schuh et al., 2017).

$$WTP = p_0 * COS_{l,t} * \pi \qquad \text{(Exp. 11.11)}$$

11.3.1.1.1 Specification and implementation For specifying the customer model, the model described by Buhl et al. (2007) was used. Here, the effect on customer satisfaction is determined for different product attributes (for an automobile, for example, performance or safety) depending on the classification of the attribute in one of the Kano categories (performance, basic or enthusiasm attribute). In order to establish the link between customer satisfaction to willingness to pay, different correlations can be used. For example, willingness to pay can be determined via a stepwise linear regression as suggested by Schuh et al. (2017) or via a polynomial function. Different customers differ in their willingness to pay depending on the importance of product attributes. In this sense, there are customers that value security features of cars most while others value comfort features more. The modeling according to Buhl et al. (2007) represents a modification of the classical formulation according to Kano. Here, the axis definition of the Kano model is adapted from the original representation of a fulfillment degree-satisfaction dependency to an expectation fulfillment-satisfaction contribution dependency. Both axes are normalized to the interval $[-1, 1]$ in order to be able to make quantitative statements. The expectation can be described by the average of the attribute values available on the market. For a detailed account of the model, we refer the reader to Buhl et al. (2007).

Figure 11.4 presents the willingness to pay as a function of expectancy confirmation respectively disconfirmation for an average customer. We vary the different attribute classes (attractive, must-be, and one-linear) assuming confirmation of customer expectation for the other attributes. This enables to see the effects of different attribute classes on the customer's willingness to pay.

11.3.1.1.2 Calibration The model is calibrated as described using online customer reviews. The procedure is as follows. The technical data and price information is extracted from a test set of automobile models. This serves to objectively classify the vehicles according to their

performance parameters. In simplified terms, this can be used to show when a vehicle performs below expectations and where expectations are assumed to be the market average in the selected segment. The information coded in the customer reviews is then extracted (from Kelly Blue Book). This information is used to make the Kano classification. For calibration we rely on an automated analysis of customer reviews by automatically classifying different attributes of a car to the Kano-category. The different product attributes of a car are thereby reflected by technical data. We analyze the data and group the data to the attributes they influence. The analysis consists of two parts:

1. Collecting for each car attribute the relevant technical data and compare it with performance of other cars from the same category (i.e., Sedans). We thereby determine whether the car is performing over or under the average within its category. We assume the average to be a good estimation for the customers' expectations regarding different car attributes.
2. Extracting for each car attribute the expressed opinion in the customer reviews using a text mining tool to estimate whether a customer is satisfied or dissatisfied with the performance of a car with the attribute.

11.3.1.1.3 Preliminary results The quality of the classification depends very strongly on the word similarity considered for the vehicle components. This defines the assignment of the sentiment scores (implemented via Wordnet) to an attribute. Here, synonyms used in natural language usage for the individual attributes are determined (e.g. for the attribute "engine performance" the terms "power", "drive", "dynamics" as well as associated adjectives). Experimentally, the best results were obtained with a similarity score of 2.8.

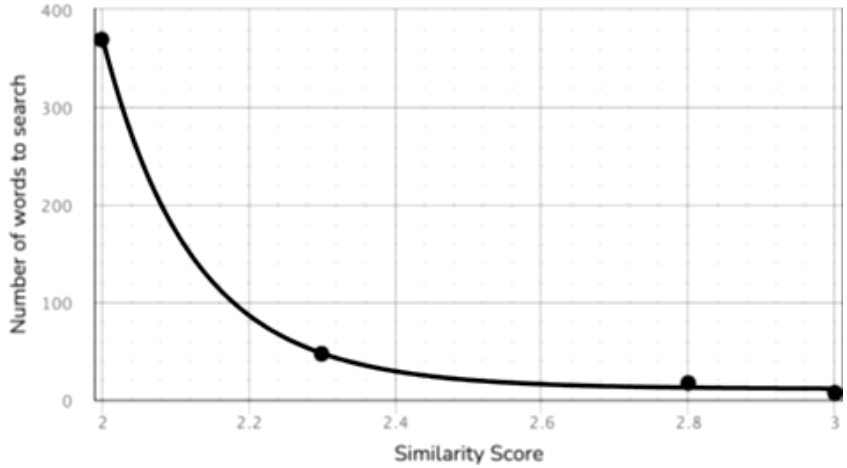

Figure 11.4: Sensitivity of the natural language processor in dependence on the similarity score chosen

For the present analysis, 904 customer reviews of four well-known German car brands were extracted from the Kelly Blue Book, each for one model of the sedan class. According to the analysis, the attributes engine performance, design and consumption are identified as performance attributes. For the named vehicle class, connectivity is a basic attribute, while range, security and

comfort are attractive attributes. Contrary to the current zeitgeist, no significant effect on customer satisfaction can be determined through environmental friendliness (indifference attribute).

For deriving the monetary values of the willingness to pay (figure 11.4), a fitting relying on the price information of the cars was undertaken. Equivalently, to the determination of customers' expectations as the average performance of cars within an attribute, we determined the base price as the mean car price within the set of considered cars of the class sedan ($34,000$€). The price sensitivity of customers is extracted based on a best-fit approach. Note, that the prohibitive price would be determined under the fulfillment of all customer expectations with regard to one-linear, attractive, and base attributes. Hence, this price cannot be extracted from figure 11.5.

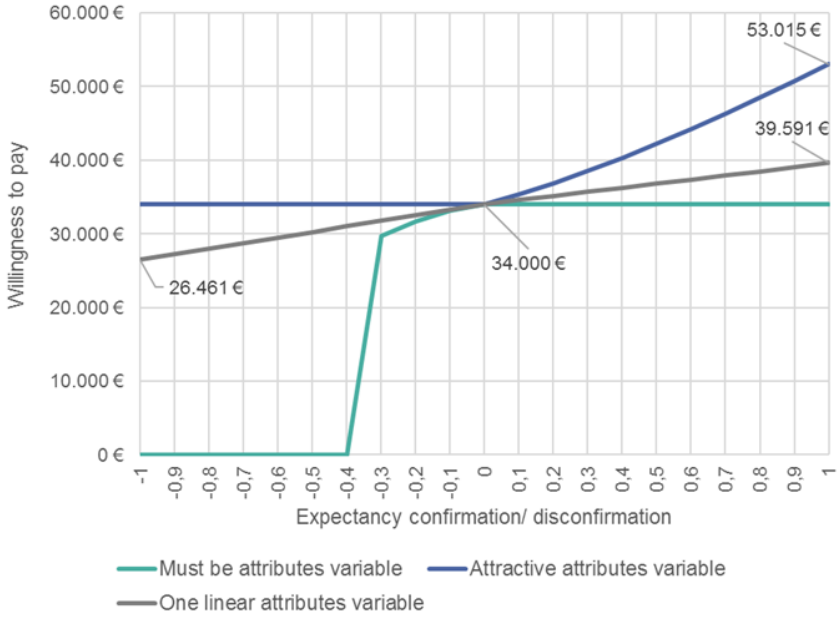

Figure 11.5: Willingness to pay depending on expectancy confirmation/ disconfirmation

11.3.1.2Manufacturer model

The manufacturer model describes the production of the automobile. The different production processes are represented as well as further artifacts that arise during production. These are log files of the machines, product qualities and material usage. The manufacturer model determines the costs at which a product can be brought to market. Furthermore, it is simulated how the production of different product attributes affects the costs of production.

11.3.1.2.1 Specifications The digital twin represents a smart automotive manufacturing plant. It integrates multiple solutions for the industrial internet of things, which are inter alia production robots, driverless transportation systems and surface inspection systems enabling a flexible state of the art production. The digital twin furthermore incorporates the firm's supply infrastructure, which in particular consists of systems for the supply of electricity and compressed

air. In this way, a comprehensive and controllable virtual overall picture is achieved enabling a realistic and rich in details simulation and analysis of the production plant. We develop a process model for an automotive production plant on the basis of a systematic literature analysis. We furthermore validated the process model through expert interviews with production managers from different automobile manufacturers. The product manufacturing process of an automobile involves the assembly of several thousand individual parts (Telek and Bányai, 2018). Central processes of automotive manufacturing take place in the press shop, the body shop, the paint shop and in final vehicle assembly. All these processes are represented in the digital twin with a high level of detail.

11.3.1.2.2 Production model The simulation of the production describes the sequence of different production processes of the manufacturing of the product automobile. The different attributes of an automobile are produced in different processes. The production model begins with the processing of the raw or input materials. The production of the automobile takes place in several substeps (individual production processes) with the participation of different machines (robot-supported production). The production model was generated based on a literature analysis and validated by means of expert interviews. A simplification of the production model is shown schematically in Figure 11.6.

Figure 11.6: Simplified process model

The production process was programmed using Python. Central aspects of the production are simulated. Input of the simulation are in particular throughput times of the individual process steps, transport times, statistical data on quality and production capability. In addition, the input includes the schematic representation of the spatial conditions of the production plant as well as the machines (and their specific configurations) used. The orders of the customers originate endogenously from the model depending on the requirements posed by the simulated customers (where customers differ within the importance of attributes posed on the automobile). The procedure of the simulation is as follows. The input data are read into the production model. This represents the information, which components must be assembled, in order to produce a desired product (order). The necessary components are produced upstream. In this way, the order determines which processes run sequentially.

The simulation of the production quality results from the statistical variables of the quality and follow a normal distribution. This is defined by an expected value as well as its standard deviation. Quality improvement measures, such as rework, are triggered if inferior quality is detected during the quality inspection. Depending on the inferior quality, the different processes of rework or scrapping are triggered. During each production step, log entries are created, representing a realistic synthetic network log of the production plant. The 3D visualization of the production process model was implemented as a 3D animation in Blender3D. Key frames were thus created for the 3D animation at each time point. The key frames are set based on the process protocol

using a Python script. Figure 6 shows an excerpt from the 3D visualization of a fully automated industrial firm.

Figure 11.7: Visual impressions of the implemented digital twin

11.3.1.2.3 Cost model The cost model describes the costs of producing the automobile. In particular, material costs and production costs are considered. Fixed costs are not considered. Thus, the cost model always describes only the operating minimum. For this work, we used the cost estimation as it was given by Burnham et al. (2021) and produced a limited set of customization possibilities for the customer.

11.3.1.2.4 Market model Potential customers and companies meet on the market. It is assumed that the potential customer buys a product (automobile) if his willingness to pay is greater than the manufacturer's offer price. The manufacturer can choose a price, which corresponds at least to the operating minimum. For the analysis at this point it is primarily decisive whether a market can develop i.e. whether the operating minimum lies below the willingness to pay of the potential customers. This is due to the fact that the analysis aims at estimating the willingness to pay of an average customer. Furthermore, the market interaction defines the economic benefit a firm can gain when being active (i.e. offering the product considered). On the other hand, this gives possibilities of estimating the costs of disrupted processes, i.e. in periods a firm cannot be active on the market (e.g. due to impairment of activities in case of a cyberattack).

11.3.2 Simulation of cyber attacks

For simulating cyberattacks, we introduce the cyberattack impact simulator (CAIS), an approach that is inspired by (Xiong et al., 2022). For simulating cyberattacks, the approach chosen within this work relies on known information on the attack gained through forensic analyses of past attacks (CTI). We take advantage of a multi-layered cyber threat intelligence ontology including information from the ATT&CK database. The ontology hereby represents an attack graph. The module draws on known effects of the simulated attacks as well as known artifacts (traces of an attack). For this purpose, a knowledge base is used. CAIS relies on CTI to simulate infections with various malware instances of different families. For this purpose, CAIS simulates the effects of the

tactics and techniques that are linked to the malware. Furthermore, CAIS dynamically interferes with the manufacturer model to produce realistic output (i.e., log files). As for normal operation of the production plant, these files can be considered the ground truth of what happens within the system and can hence be analyzed by different security-modules (e.g., intrusion detection systems, (Dietz et al., 2020)). The log files furthermore are fed into the 3D-simulation providing visual support for understanding the effects of cyberattacks and guide the analyst in deriving insight to cyber risks. We thereby neglect simulations on the probability of attacks and refer for a discussion on the probability i.e. on works on the weighting of attack graphs (e.g., Kaiser et al., n.d.). which can be seamlessly integrated within our approach. Figure 11.8 represents the schematic structure of the knowledge base CAIS takes advantage of for simulating the effects of a cyberattack. Within this schematic overview, a-nodes denote attacks, t-nodes denote attack techniques, w-nodes denote weaknesses (system flaws that may lead to vulnerabilities), v-nodes denote vulnerabilities, and d-nodes denote specific devices respectively machinery that can be affected when an attack is conducted. The knowledge base is thereby given as an attack graph where relationships between the different nodes are considered. The effects that may be caused by a cyberattack are mainly non-availability of machinery and integrity of machinery which may cause quality deviations if e.g., production parameters are altered.

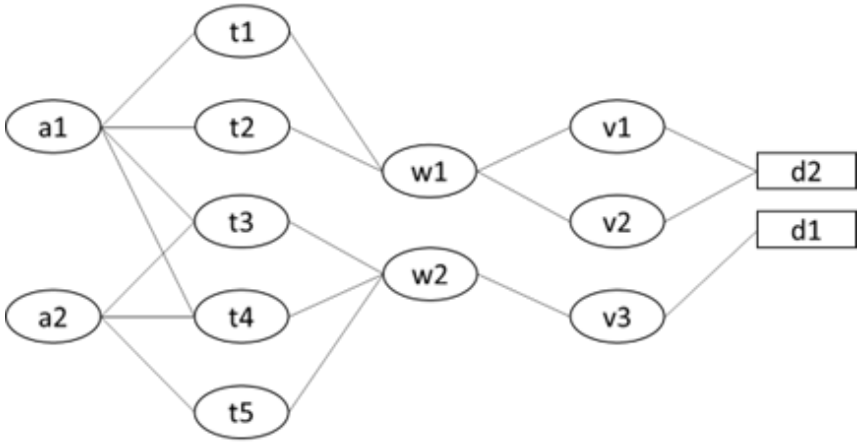

Figure 11.8: Structure of the knowledge base

11.3.3 Value based quantification of the effects of a cyber attack

The quantification of the effects of a cyberattack is based on the Kano-model model as presented within section 11.3.1.1. It is based on the customers' willingness to pay for a produced good and the pricing strategy of the firm. For further determination of the impact, we calculate the net operating profits after taxes.

$$NOPAT_t = p(WTP) * x_s * (1 - T) - C - c_v * x_p \qquad \text{(Exp. 11.12)}$$

To quantify the effects of a cyberattack, we furthermore use the economic value added to include opportunity costs of capital.

$$EVA_t = NOPAT_t - TC_t * WACC \quad \text{(Exp. 11.13)}$$

$$I_t = \sum_t \frac{EVA_{t,0} - EVA_{t,A}}{(1+r)^t} \quad \text{(Exp. 11.14)}$$

EVA_0 is the original economic value added while EVA_A describes the economic value added after being influenced by an attack. Furthermore, r denotes the interest rate.

The contribution of each attribute to customer overall satisfaction as well as dissatisfaction can be calculated, which now serves as a basis for the evaluation of the impact of a cyberattack. The central aspect of this approach lies in a mapping of potentially affected processes to product attributes. Through this effect on attribute generation the functionality of attributes may be influenced. This also includes the case that the attribute cannot be created (e.g. during a production stop). Therefore, the customer's overall satisfaction with attribute i and without/ with an affected attribute $-i$ due to an attack are compared under the ceteris paribus condition (not affected attribute $\bar{i}$ (see 11.12 and 11.13).

$$CS_{l,t,i} = COS_{l,t,a}(s_{l,t,-i,y}) - COS_{l,t,0}(s_{l,t,\bar{i},y}), \forall s_{l,t,i,y} \in [0,1] \quad \text{(Exp. 11.15)}$$

$$CD_{l,t,i} = COS_{l,t,a}(s_{l,t,-i,y} - COS_{l,t,0}(s_{l,t,\bar{i},y}), \forall s_{l,t,i,y} \in (0,-1] \quad \text{(Exp. 11.16)}$$

To link the production processes to value creation it should be stressed that each process partially or completely contributes to product attribute creation (as it was stated above). Therefore $c_{i,j}$ is defined to measure the contribution of process j to product attribute i. For this purpose $c_{i,j}$ can be derived from the statement of work respectively from the process model. Combining $c_{i,j}$, CS_i, and CD_i the process can be evaluated with respect to its contribution to customer satisfaction. Therefore, the process customer satisfaction PCS_j, respectively process customer dissatisfaction PCD_j is determined.

$$PCS_j = \sum_i c_{i,j} * CS_{l,t,i} \quad \text{(Exp. 11.17)}$$

$$PCD_j = \sum_i c_{i,j} * CD_{l,t,i} \quad \text{(Exp. 11.18)}$$

The contribution to customer satisfaction can thereby also be extended to secondary value activities which are not directly oriented to the generation of attributes but which indirectly contribute to attribute and value generation. Consistent with the resource flows within primary value activities,

a process value analysis of secondary value activities can take place where these processes are not product oriented but rather process oriented. This means, that a value allocation should be process requirement based (where the process requirements are derived requirements from the customer requirements). For example, regarding the attributes of a car, acceleration would be a customer requirement which is partially targeted in the primary value process of software integration (right adaption of a 'generic' software to the specifics of the vehicle). However, research and development (secondary value process) contributes as well to the acceleration of a vehicle as it generates technical know-how. Hence, a potential derived requirement of software integration is technical know-how. Therefore, process oriented internal cost allocation could also be included in the process model taking into account intra organizational dependencies.

The overall contribution of process j to customer satisfaction POS_j can be described through a combination of PCS_j and PCD_j. Thereby the combination can be done as a simple addition as the non-univocal relationship between dissatisfaction and satisfaction is taken into consideration when assessing the customer overall satisfaction.

$$POS_{j,t} = PCS_{j,t} + PCD_{j,t} \tag{Exp. 11.19}$$

The market price or economic value respectively can be allocated with respect to the process contribution to customer satisfaction on each process. The process value therefore is calculated according to the following formula where V is the product value. Likewise, the allocation can be carried out for the contribution margin of the process respectively the economic value added.

$$PV_{j,t} = \frac{POS_{j,t}}{COS_{l,t}} * V_t(COS_{l,t}) \tag{Exp. 11.20}$$

Thereby the value of a product is determined by the willingness to pay and the variable costs.

$$V_{max} = p_0 * COS_{l,t} * \pi - c_v \tag{Exp. 11.21}$$

$$c_v = \sum_j c_{m,j} + c_{L,j} + c_{M,j} \tag{Exp. 11.22}$$

The impact of a cyberattack over time (e.g. long-term consequences of lost intellectual property or effects of reputation) can then be described by the change of process value PV_j due to an attack for all periods t where the attack affects the normal functioning of processes.

$$NOPAT_{j,t} = PV_{j,t} - C_{j,j} \tag{Exp. 11.23}$$

Process based quantification of the effects of a cyber incident can thereby be accumulated across all activities j but can also be traced down to a single process j.

$$EVA_{j,t} = NOPAT_{j,t} - TC_{j,t} * WACC_t \quad \text{(Exp. 11.24)}$$

$$I_t = \sum_{t,j} \frac{EVA_{j,t,0} - EVA_{j,t,A}}{(1+r)^t} \quad \text{(Exp. 11.25)}$$

11.4 Evaluation

11.4.1 Simulation of normal production processes

The production process is evaluated on the basis of various criteria. A log file is created during the production process. This can be examined for anomalies. Within a normal production process, no malicious anomalies should occur.

Furthermore, quality (deviations from standards) and efficiency parameters (inter alia lead times) are recorded. The market model can now also be used to estimate the extent to which the product will appeal to the market and thus generate revenue on the market. Figure 11.9 gives the quality of produced cars respectively their components. If the deviation from standard is within the norm, the produced component is accepted. However, if the deviation is high, the component cannot be used within the car. This either causes waste or leads to rework.

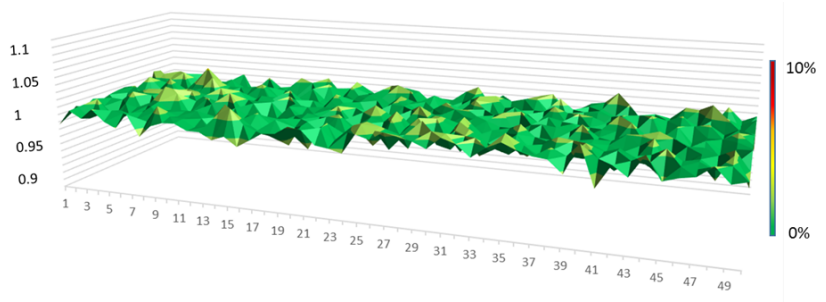

Figure 11.9: Quality of produced products (i.e., size deviations) as a deviation from the standard (normed as 1)

Within the simulated production of 50 cars, there was necessity for rework of 11 components (which accounts for approximately 1.6% of produced components). When the plant was not affected by an attack, the firm hence works with high precision enabling low rework and rejection rates. The cars produced would account for around 1.5 million € in sales and a median price of 30 thousand € per car. The rework costs account for approximately 18 thousand €.

11.4.2 Simulation of cyberattacks within the production system

For the simulation of a cyberattack, we rely on the module implemented to simulate the effects of the attack on the production process and to map the resulting artifacts. Anomalies can be detected in the log file. Figure 11.10 shows an excerpt of an exemplary log file. The line in which an

indication of a cyberattack was discovered is highlighted. A hash is listed that indicates an attack by the 3PARA RAT malware. Figure 11.11 furthermore represents another simulation run showing intentional (by the attacker) quality degradations through a cyberattack (these could be reached by manipulations of control which can e.g., be caused by Industroyer or Stuxnet). In this sense, the figure gives the quality of produced products within the simulated production during the attack. The quality deviations are significantly higher than within normal functioning causing high shares of cars whose components demand for rework. Hence, the efficiency of the manufacturing process was degraded. The degree of degradation can thereby be quantitatively assessed relying on the simulation based on the digital twin.

7.741.722.861.01	[22.10.2022 - 18:54:21]	stf_B1	a72362s2633f213h32n339a12
7.741.722.862.01	[22.10.2022 - 18:54:26]	stf_B2	b13112a1312a0874h1381092
7.741.722.862.02	[22.10.2022 - 18:54:26]	ir_B2	1326afd9843bs8323617h1834
7.741.722.862.03	[22.10.2022 - 18:54:26]	ct_B2	a1874a429845n4671j1728d17
7.741.722.863.01	[22.10.2022 - 18:54:34]	stf_B31	289a27727flbaf8b2e49fd4d98
7.741.722.863.02	[22.10.2022 - 18:54:34]	ct_B31	k11910473n74a4719af492922
7.741.722.863.03	[22.10.2022 - 18:54:34]	stf_B32	9383k16785as17923d1941792
7.741.722.863.04	[22.10.2022 - 18:54:34]	ct_B32	9a1123s418n1781n19ccfaf412

Figure 11.10: Simulated log-file of the automotive production process with a Hash linked with 3PARA RAT

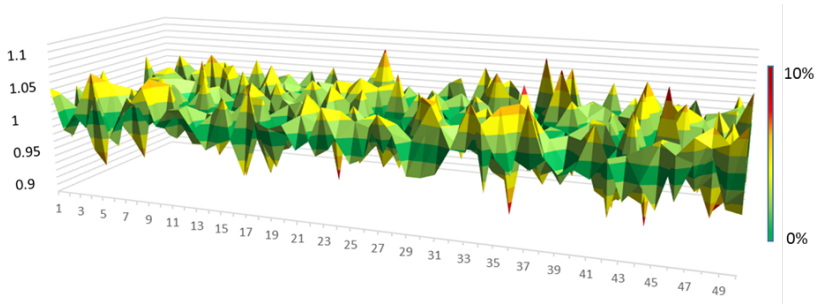

Figure 11.11: Quality of produced products (i.e., size deviations) as a deviation from the standard (normed as 1)

Analyzing the impact of malfunctioning of the automotive production process and hence of the risks that may come with the restructuring of automotive production through the introduction of modern production systems, the digital twin offers multiple perspectives on the risks that may emerge. We consider this to be a special benefit of the method proposed when applied for cyber risk quantification. Hence, the proposed methodology offers to craft a risk vector consisting of different dimensions which are inter alia quality deviations and monetary harm. Comparing Figure 9 with Figure 11, the high variability and therefore lower production ability need to be noted leading to high variability of the quality of produced components. In monetary terms, the impact would account for a monetary burden of 400 thousand € due to the increased costs for rework. Similar to degradations of product quality through malfunctioning of processes, production outages can cause damages within the production system.

11.4.3 Comparison of different digitalization levels

The degree of penetration with technology of the industrial plant is simulated for four conditions which are low level of digitalization, high level of digitalization as well as for attacks on the highly digitalized industrial plant and the sparsely digitalized plant. The results of the simulation are presented within figure 11.12. The results here are quite simple and intuitive. Higher digitalization levels offer promises of higher levels of accuracy with lower rejection rates of produced products during quality assessment cicles. Which leads to low revision rates and can increase the productivity of the company. The variation of the product quality (as deviation from the specification) is lower at high degrees of digitization than at low ones. In the event of a cyberattack, however, the ratio turns and the low-digitized company shows less impairment than the highly digitized company. The core tradeoff in digitization decisions thus results from the comparison of the advantages of digitization (especially increased productivity) with its disadvantages (greater damage in the event of cyberattacks).

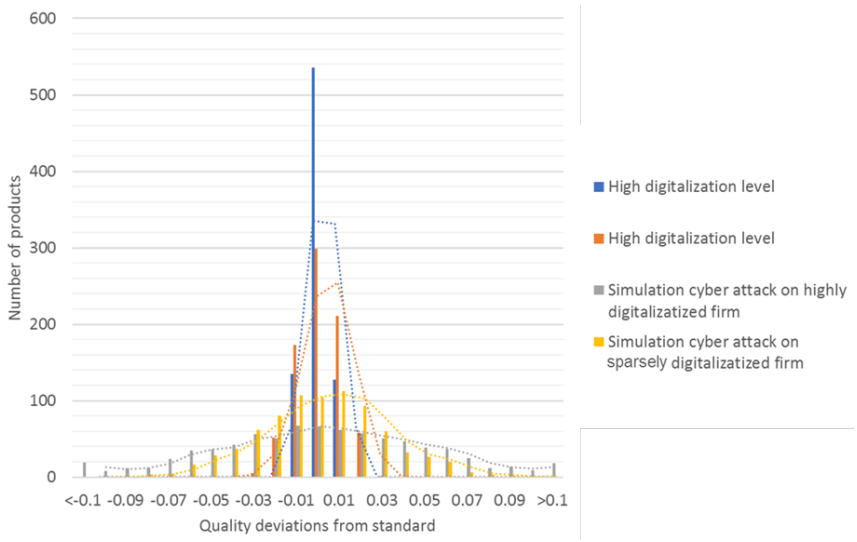

Figure 11.12: Comparison of quality deviations in production (i.e., size deviations) for highly digitalized firms and sparsely digitalized firms under normal production and attacks

11.4.4 Comparison between different processes

A key advantage of the presented approach, as stated, is its ability to drive cyberrisks back to the processes that are impacted. By this, the methods allow to prioritize processes with regard to the prioritization for security investments. At this point, the method relies on the process value contribution of each process. Quality reductions/ poor performance has a different impact in different production areas. It should now be noted that the quality of the products can now increase through digitization if the company is not attacked. According to this line of reasoning, those processes that contribute a great deal of value should be digitized (the exception, of course, is when the product attribute "manual work" influences a large part of the product evaluation). However,

these processes are also those that require special protection. It can be deduced from this that the digitization strategy must always take the risks into account as well.

It should be noted here that the possibilities of digitization can vary greatly. Thus, in addition to the costs of a disruption, the direct costs of digitization (e.g., costs of an industrial robot capable of automating a process) are primarily included in the tradeoff in the digitization decision. However, the results of this study in terms of comparing different impacts of cyberattacks can be used in evaluating the appropriateness of the current protection of systems against cyberattacks, indicating which processes are overprotected in terms of their value contribution/potential to destroy value and which processes have remained underrepresented in previous security investments.

Within figure 11.13 the results of the analysis are presented for an excerpt of processes. A red line is presented representing the average. When relying on conventional cyber risk assessment methods for investment planning that do not take advantage of process value analysis. Wheel production and assembly would thereby be an exemplary process that we consider to be a process that is likely to be overinvested in with regard to cyber security. Conversely, processes on the production and assembly of security features, engines, and automated processes in painting would be subject to underinvestment, if security features would be based on average risk approaches.

Figure 11.13: Impact of cyberattacks in different processes

11.5 Discussion and implications

11.5.1 Implications

Within the work a digital twin is developed, which makes it possible to carry out a simulation of industrial processes. The genericity of the digital twin allows a multifaceted analysis. Thus, the digital twin allows for a holistic analysis of the production process. Among other things, the effects of disruptions in automated production can be analyzed and understood. The possibility of analyzing the effects of cyber risks was demonstrated here as an example.

The applicability of the presented methodology for quantifying cyberattack impacts based on value process analysis is supported by the fact that information technology does not generate value by itself but only in the combination of processes (Krcmar, 2015). Hence, determining the impact of a failure in information technology can also be only understood when investigating the underlying processes to which the information technology contributes. The presented approach thereby focuses on the cash flow. As "the discounted excess profits plus the recorded value will always give the true fair market value" (Buhl et al., 2007), every economic effect of a cyberattack can be quantified relying on the process value analysis methodology. The novel part of the contribution hence, is to give an alternative way of quantifying cyber risks and augment to the rarely researched field of quantification of the costs incurred through a cyberattack (Eckhart et al., 2019). As it was argued, cyber risk quantification thereby should take the same (process based) perspective as value determination. Hence, we propose process value analysis based quantification of cyber threat induced damages. Furthermore, the methodology enables to trace cyberattack impact determination on a technical level by fusing technology-oriented theory of production (Schuh et al., 2017) and process value analysis. Furthermore, this procedure does not only allow ex post quantification of cyber risks but also predictive analyses. Hence, the approach can be used for both, active cyber risk management as well as passive (reactive) cyber risk management on the side of security professionals as well as for cyber risk insurance pricing.

11.5.2 Limitations

The procedure introduced in this work builds upon established methodologies for process evaluation as well as customer requirement determination in market research. However, there are various different methodologies for quantitatively assessing the effects on customer satisfaction. Violante and Vezzetti (2017) show different approaches and their benefits. Thereby they propose different models and analyze their linkages. The authors relied on the approach presented by Buhl et al. (2007) as they believe that it delivers brought benefits while the weakness is that there is no modelling of an indifference area of customer satisfaction. Thereby a key benefit was that the model was applied to the area of information technology (website of a financial service provider) and had therefore proven to be valid for the area of application (Buhl et al., 2007).

Furthermore, the presented methodology poses some challenges for the practical implementation and the embedding into the business processes. Among other things, a high implementation effort would have to be expected in many companies, since the procedure of process value analysis is not a widespread management practice. Moreover, in controlling, a coding and aggregation of cost items to products and divisions or regions is often carried out. This cost aggregation and coding does not allow a breakdown of costs into the individual value-added processes. Switching from one methodology to another can lead to high costs. On the other hand, the implementation costs should be quickly made up for by opposing efficiency gains. The methodology presented requires a strong orientation of management activities towards value creation. In particular, the methodology is thus in the spirit of the increasingly widespread use of value-based management.

Nevertheless, the procedure has some advantages over other methods used. With the presented method, it is possible to determine the risks on a value basis. In addition, a further advantage is the cause-related quantification. This process-based determination of the risks provides a profound understanding of the effects of a cyber-attack. The effects along the entire value chain can be traced. The methodology is thus suitable for tracking cascading effects (also proven by Orojloo and Azgomi (2017)). Thus, far-reaching consequences of a cyber-attack in an economic system can be traced. Hence, our approach makes it possible to understand the consequences of a cyber-attack in a cause oriented and process based way. This cause-related assessment of the damage and the possibility of assessing the damage along the value chain has several advantages over conventional assessment methods.

The attribution of damages to the individual processes allows an assessment and prioritization of IT-security measures. Different value-added processes have different value contributions to the product and hence to the overall value of the company. The company's revenues are therefore dependent on the individual value-added processes in different ways, which is due to the fact that the attributes created by these processes are valued differently by the customer. Furthermore, different companies differ according to the customer's appreciation of the attributes in their assessment of the importance of the processes. This makes it clear that different companies should protect the same processes to different degrees and there is no one-fits-all solution in cyber-security. Rather, investments should be tailored to the individual company and security solutions should be applied according to the company's requirements. Hence, the methodology presented here provides a systematic approach for quantifying the company-specific requirements for implementing optimal security measures based on a quantification.

As a matter of principle, a cause-related damage assessment must be carried out for a large number of different attacks. Value orientation provides new starting points for evaluation. Thus, attacks on values that are difficult to quantify, such as trade secrets, can also be quantified. In this way, the influence of trade secrets on the customer evaluation of the product can be determined. The theft of a trade secret can then be determined according to the effect of the trade secret on the value chain along and for each business process. The value contribution for the individual product can then be multiplied by the expected sales figures over time and discounted accordingly. Thus, in the case of a trade secret, the value would be calculated by perpetual discounting. If, however, only

limited use of the intellectual property can be expected, discounting should be carried out over the period of use.

Furthermore, effects along the value-added system can be traced. Thus the effects of an attack on suppliers and customers can be quantified. Therefore, information flows, monetary flows and flows of goods between business partners are modelled. Accordingly, it can be shown how attacks on a supplier affect the focal company. The attacks can be aimed both at damaging business processes in the sense of disrupting the flow of goods and at a purely economic objective through the diversion of monetary flows. Furthermore, intellectual property of business partners can be stolen by an attack. This can affect the intellectual property of the attacked company as well as intellectual property shared between business partners through the exchange of information. These risks in particular must be taken into account in the cyber-risk management of companies. Nevertheless, this is often neglected in existing approaches. The practical relevance of such assessments becomes apparent in attacks on suppliers of large companies as they are often less secure than their large business partners are.

The presented approach can be criticized for its high demand of various data and high costs of implementation. The approach is hence dependent on the quality of data and the assumptions underlying the specific implementation (Zio, 2018) However, the approach offers the possibility of conducting quantitative risk assessment based on which safe systems can be designed, developed, and operated. Quantitative risk assessment could hereby enable effective decision making (Zio, 2018). A further benefit of relying on the digital twin for assessing the benefits and risks that are associated with digitalization investments is that the method enables risk analysts to craft transparent estimations based on a formalized approach (Zio, 2018).

11.6 Conclusion

In this paper, a methodology for cyber-risk quantification was presented, by which cyber-risks can be allocated to processes. Therefore, process value analysis was used. The process value analysis guarantees a value orientation of risk quantification. Process valuation is based on the attributes generated through these processes. In doing so, the link of attribute performance and customer satisfaction is accessed. Valuation of cyber-risks is thus based on the customer requirements and the customer benefits arising from each business process. Cyber-risk quantification in this sense can be done company and brand specific. Thereby different target groups can be considered. According to their valuation of attributes, different importance can be assigned to the production processes. These differences determine different valuations of cyber-risks between companies and brands. The presented methodology can be used in investment decision making regarding cyber-security. The investment decision making can be based on each production process. Therefore, each component involved in the production process can be secured according to the importance of the production process for customer satisfaction and their valuation of attributes. Furthermore, each process can be evaluated regarding the cyber-risk it brings for the whole company. This information can be

integrated in the valuation of business processes. Moreover, the suggested methodology can be used for business process reengineering and for sourcing decisions.

The article contributes to a more comprehensive understanding and determination of the effects of a cyber-attack, which is still a major problem for cyber risk quantification, insurance pricing and cyber-risk management (Eckhart et al., 2019). In this vein, the work also contributes to the solution of the risk quantification problem with respect to cyber-risks. The main purpose of this contribution is to outline the process value aspects of risk quantification. Furthermore, the work contributes to the establishment of effective cyber-risk management by quantifying the effects of cyberattacks in a comprehensive, formalized and thus, comparable way. Therefore, it can be deployed as the foundation for cyber-risk quantification measures that can be used for engineering secure systems or pricing of security related services or products (e.g. insurances). However, as only the consequences are calculated so far, it is a logical next step to extend our approach by a look into details of the defender-attacker interaction and provide estimations for loss probabilities from a game theoretical model (as it is able to take into consideration the strategic nature of cyber risk decision making on the side of the attacker as well as on the side of the defender).

References

Abas, A. E. P., Yong, J. H., Mahlia, T. M. I., Hannan, M. A. (2019). Techno-economic analysis and environmental impact of electric vehicle. *IEEE Access*, *7*, 98565–98578.

Aivaliotis, P., Georgoulias, K., Arkouli, Z., Makris, S. (2019). Methodology for enabling digital twin using advanced physics-based modelling in predictive maintenance. *Procedia CIRP*, *81*, 417–422.

Angelou, G. N., a. Economides, A. (2006). Ict investments under competition threat. In *Proceedings 8th hellenic conference on operations research.*

Barua, A., Konana, P., Whinston, A., Yin, F. (2004). An empirical investigation of net-enabled business value. *MIS Quarterly*, *28*, 585–620.

Beischel, M. E. (1990). Improving production with process value analysis. *Journal of Accountancy*, *170*, 53.

Bi, J., Liu, Y., Fan, Z.-P., Cambria, E. (2019). Modelling customer satisfaction from online reviews using ensemble neural network and effect-based kano model. *International Journal of Production Research*, *57*(22), 7068–7088.

Bleicher, J., Stanley, H. (2019). Digitization as a catalyst for business model innovation a three-step approach to facilitating economic success. *Journal of Business Management*, *12*.

Borgianni, Y., Cascini, G., Rotini, F. (2010). Process value analysis for business process re-engineering. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *224*(2), 305–327.

Borgianni, Y., Cascini, G., Rotini, F. (2011). Product-driven process value analysis. In *Global product development* (pp. 305–327). Springer, Berlin, Heidelberg.

Abbreviations	Description
$e_{l,t,i}$	customer expectations for a specific attribute i of a target group l at time point t
$k_{l,t,i}$	performance for a specific attribute i for a target group l at time point t
$x_{l,t,i}$	dis-/confirmation with a specific attribute i of a target group l at time point t
$s_{l,t,i}$	customer satisfaction with a specific attribute i of a target group l at time point t
k	attribute classes
a	attractive attributes
p	performance attributes
b	basic attributes
$s_{l,t,x}$	customer satisfaction with a specific attribute x from attribute class "attractive" of a target group l at time point t
$s_{l,t,y}$	customer satisfaction with a specific attribute y from attribute class "performance" of a target group l at time point t
$s_{l,t,z}$	customer satisfaction with a specific attribute z from attribute class "basic" of a target group l at time point t
$\underline{u}$	scaling parameter for the lower bound
$\overline{u}$	scaling parameter for the upper bound
q	scaling parameter
c	scaling parameter
$y_{l,t,x}$	importance of attribute x from attribute class "attractive"
$y_{l,t,y}$	importance of attribute y from attribute class "performance"
$y_{l,t,z}$	importance of attribute z from attribute class "basic"
$COS_{l,t}$	customer overall satisfaction of target group l at time t
α	Cobb-Douglas scaling parameters for attractive attributes
φ	Cobb-Douglas scaling parameters for performance attributes
β	Cobb-Douglas scaling parameters for basic attributes
$WTP_{l,t}$	willingness to pay of target group l at time t
p_0	basic price
π	price premium
T	tax rate
$NOPAT$	net operating profits after taxes
x_s	quantity sold
x_p	quantity produced
c_v	variable costs
C	Fixed costs
EVA	economic value added
A	attack
TC	total capital
$WACC$	weighed average cost of capital
r	Interest rate
$CS_{l,t,i}$	customer dissatisfaction with attribute i at time t
I_t	present value of impact at time t
$CD_{l,t,i}$	customer satisfaction with attribute i at time t
$PCS_{j,t}$	contribution of process j to customer dissatisfaction at time t
$PCD_{j,t}$	contribution of process j to customer satisfaction at time t
$c_{i,j}$	contribution of process j to the generation of attribute i
POS_j	contribution to customer overall satisfaction through process j at time t
$PV_{j,t}$	process value of process j at time t
V_t	product value at time t
$c_{m,j}$	Material costs of process j
$c_{L,j}$	Costs of labor of process j
$c_{M,j}$	Machining costs of process j
I_t	present value of impact at time t

Table 11.1: Short-form expressions of the model's variables and parameter

Brozovsky, J., Haase, M., Lolli, N. (2018). Validation of a digital twin with measurement data. In *Proceedings of the 39th aivc/7th tightvent/5th venticool conference, juan-les-pins, france* (pp. 18–19).

Buhl, H. U., Kundisch, D., Renz, A., Schackmann, N. (2007). Spezifizierung des kano-modells zur messung von kundenzufriedenheit. *8. Internationale Tagung Wirtschaftsinformatik 2007-Band 1*, 897.

Burnham, A., Gohlke, D., Rush, L., Stephens, T. S., Zhou, Y., Delucchi, M. A., ... Boloor, M. (2021). Comprehensive total cost of ownership quantification for vehicles with different size classes and powertrains. Argonne National Lab.(ANL), Argonne, IL (United States).

Cenamor, J., Sjödin, D. R., Parida, V. (2017). Adopting a platform approach in servitization : Leveraging the value of digitalization. *International Journal of Production Economics*, *192*, 54–65.

Chan, Y. E., Ahuja, S. (2015). Digital ecodynamics in small firms: Using information technology to compete. In *Icis*.

Coreynen, W., Matthyssens, P., Bockhaven, W. V. (2017). Boosting servitization through digitization: Pathways and dynamic resource configurations for manufacturers. *Industrial Marketing Management*, *60*, 42–53.

Das, R., Morris, T. (2018). Modeling a midstream oil terminal for cyber security risk evaluation. In *International conference on critical infrastructure protection* (pp. 149–175). Springer, Cham.

de Paula Ferreira, W., Armellini, F., de Santa-Eulalia, L. A. (2020). Simulation in industry 4.0: A state-of-the-art review. *Computers & Industrial Engineering*, *149*, 106868.

Dietz, M., Vielberth, M., Pernul, G. (2020). Integrating digital twin security simulations in the security operations center. In *Proceedings of the 15th international conference on availability, reliability and security* (pp. 1–9).

Eckhart, M., Brenner, B., Ekelhart, A., Weippl, E. R. (2019). Quantitative security risk assessment for industrial control systems: Research opportunities and challenges. *Journal of Internet Services and Information Security*, *9*, 52–73.

Eckhart, M., Ekelhart, A. (2019). Digital twins for cyber-physical systems security: State of the art and outlook. In *Security and quality in cyber-physical systems engineering* (pp. 383–412).

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. In *2019 european intelligence and security informatics conference (eisic)* (pp. 40–47). IEEE.

Ermolina, L. V., Zinovyev, A. M., Melnikova, D. A. (2021). Digital twins as a method of risk management transformation. In *International scientific conference "digital transformation of the economy: Challenges, trends, new opportunities"* (pp. 451–457). Springer, Cham.

Graessler, I., Pöhler, A. (2017). Integration of a digital twin as human representation in a scheduling procedure of a cyber-physical production system. In *2017 ieee international conference on industrial engineering and engineering management (ieem)* (pp. 289–293). IEEE.

Gylling, A., Ekstedt, M., Afzal, Z., Eliasson, P. (2021). Mapping cyber threat intelligence to probabilistic attack graphs. In *2021 ieee international conference on cyber security and resilience (csr)* (pp. 304–311). IEEE.

Hagedorn, M., Baum, M., Eckstein, L., Harter, C., Hartmann, S., Heilert, D., ... Schlick, T. (2019). Automobile wertschöpfung 2030/2050.

Hanni, A. R., Patil, M., Patil, P. M. (2016). Summarization of customer reviews for a product on a website using natural language processing. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2280–2285.

Henry, M. H., Layer, R. M., Snow, K. Z., Zaret, D. R. (2009). Evaluating the risk of cyber attacks on scada systems via petri net analysis with application to hazardous liquid loading operations. *2009 IEEE Conference on Technologies for Homeland Security*, 607–614.

Holmes, D. C. E., Papathanasaki, M., Maglaras, L., Ferrag, M. A., Nepal, S., Janicke, H. (2021). Digital twins and cyber security – solution or challenge? *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 1-8.

Homburg, C., Koschate, N., Hoyer, W. D. (2005). Do satisfied customers really pay more? a study of the relationship between customer satisfaction and willingness to pay. *Journal of Marketing*, *69*(2), 84–96.

Janatyan, N., Shahin, A. (2020). Product value analysis: a developed cost–benefit analysis ratio based on the kano and paf models. *The TQM Journal*, *33*(1), 163–181.

Kahneman, D., Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127).

Kaiser, F. K., Wiens, M., Schultmann, F. (n.d.). Cyber risk quantification - using weighted attack graphs for behavioral cyber game theory.

Kim, Y. J., Sanders, G. L. (2002). Strategic actions in information technology investment based on real option theory. *Decision Support Systems*, *33*(1), 1–11.

Kohli, R., Grover, V. (2014). Business value of it: An essay on expanding research directions to keep up with the times. *Journal of the Association for Information Systems*, *9*, 1.

Kouril, D., Rebok, T., Jirsík, T., Cegan, J., Drasar, M., Vizváry, M., Vykopal, J. (2014). Cloud-based testbed for simulation of cyber attacks. *2014 IEEE Network Operations and Management Symposium (NOMS)*, 1–6.

Kowalkowski, C., Gebauer, H., Oliva, R. (2016). Service growth in product firms: Past, present, and future. *Industrial marketing management*, *60*, 82–88.

Krcmar, H. (2015). *Informationsmanagement* (6th ed.). Berlin.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, *51*(11), 1016–1022.

Kropik, M. (2021). Produktion. In *Produktionsleitsysteme für die automobilindustrie* (pp. 129–235). Springer Verlag, Berlin, Heidelberg.

Kuhl, M. E., Kistner, J., Costantini, K., Sudit, M. (2007). Cyber attack modeling and simulation for network security analysis. *2007 Winter Simulation Conference*, 1180–1188.

Leszczyna, R. (2019). Cybersecurity assessment. *Cybersecurity in the Electricity Sector*, 149–179.

Leszczyna, R., Fovino, I. N., Masera, M. (2008). Maisim: mobile agent malware simulator. In *Simutools* (p. 35).

Leszczyna, R., Fovino, I. N., Masera, M. (2010). Simulating malware with malsim. *Journal in Computer Virology*, *6*(1).

Lin, L., Bao, H., Dinh, N. T. (2021). Uncertainty quantification and software risk analysis for digital twins in the nearly autonomous management and control systems: A review. *Annals of Nuclear Energy*, *160*, 108362.

Lind, M., Skavhaug, A. (2012). Using the blender game engine for real-time emulation of production devices. *International Journal of Production Research*, *50*(22).

Lu, A., Sun, Y. F., Lei, Z., Li, G., Jing, J., Liu, W. P., Hu, C. J. (2021). Kkma - a calculation method for kano classification based on user reviews. *IOP Conference Series: Materials Science and Engineering*, *1043*(2), 22062.

Martinez, G. S., Sierla, S., Karhela, T., Vyatkin, V. (2018). Automatic generation of a simulation-based digital twin of an industrial process plant. In *Iecon 2018-44th annual conference of the ieee industrial electronics society* (pp. 3084–3089).

Masli, A., Richardson, V. J., Sánchez, J. M., Smith, R. E. (2011). The business value of it: A synthesis and framework of archival research. *Journal of Information Systems*, *25*(2).

Melville, N., Kraemer, K., Gurbaxani, V. (2004). Information technology and organizational performance: An integrative model of it business value. *MIS Quarterly*, *25*(2), 283–322.

Moshiri, S. (2016). Ict spillovers and productivity in canada: provincial and industry analysis. *Economics of Innovation and New Technology*, *25*(8), 801–820.

Mossburg, E., Gelinne, J., Calzada, H. (2016). Beneath the surface of a cyberattack: a deeper look at business impacts..

Mourtzis, D. (2020). Simulation in the design and operation of manufacturing systems: state of the art and new trends. *International Journal of Production Research*, *58*(2), 1927–1949.

Müller, M., Lehmann, M., Kuhn, H. (2021). Measuring sequence stability in automotive production lines. *International Journal of Production Research*, *59*(24), 7336–7356.

Musman, S. A., Turner, A. J. (2018). A game theoretic approach to cyber security risk management. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *15*(2).

Naunheimer, H., Bertsche, B., Ryborz, J., Novak, W. (2010). Automotive transmissions: Fundamentals, selection, design and application. Springer Science & Business Media.

Negri, E., Fumagalli, L., Macchi, M. (2017). A review of the roles of digital twin in cps-based production systems. *Procedia Manufacturing*, *11*, 939–948.

Neumeier, A., Wolf, T., Oesterle, S. (2017). The manifold fruits of digitalization - determining the literal value behind. In *Proceedings der 13. internationalentagung wirtschaftsinformatik (wi 2017)* (pp. 484–498). St. Gallen.

Orojloo, H., Azgomi, M. A. (1996). Process value analysis: the missing link in cost management. *Journal of Cost Management*, *6*(3), 4–13.

Orojloo, H., Azgomi, M. A. (2017). A method for evaluating the consequence propagation of security attacks in cyber-physical systems. *Future Generation Computer Systems*, *67*, 57–71.

Pagoropoulos, A., Maier, A. M., McAloone, T. C. (2017). Assessing transformational change from institutionalising digital capabilities on implementation and development of product-service systems: Learnings from the maritime industry. *Journal of Cleaner Production*, *166*, 369–380.

Porter, M. E. (1985). Technology and competitive advantage. *Journal of Business Strategy*, *5*, 60–78.

Qi, Q., Tao, F. (2018). Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access*, *6*, 3585–3593.

Radanliev, P., De Roure, D., Cannady, S., M.Montalvo, R., Nicolescu, R., Huth, M. (2018). Analysing iot cyber risk for estimating iot cyber insurance. in living in the internet of things: Cybersecurity of the iot-2018. In *Iet conference proceedings* (pp. 1–9). London: The Institution of Engineering and Technology.

Rapaccini, M., Saccani, N., Kowalkowski, C., Paiola, M., Adrodegari, F. (2020). Navigating disruptive crises through service-led growth: The impact of covid-19 on italian manufacturing firms. *Industrial Marketing Management*, *88*, 225–237.

Rasheed, A., San, O., Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers from a modeling perspective. *IEEE Access*, *8*, 21980–22012.

Reinartz, W. J., Wiegand, N., Imschloss, M. (2019). The impact of digital transformation on the retailing value chain. *International Journal of Research in Marketing*, *36*(3), 350–366.

Riesco, R., Villagrá, V. A. (2019). Leveraging cyber threat intelligence for a dynamic risk framework. *International Journal of Information Security*, *18*(6), 715–739.

Ritto, T., Rochinha, F. (2021). Digital twin, physics-based model, and machine learning applied to damage detection in structures. *Mechanical Systems and Signal Processing*, *155*, 107614.

Roumani, M. A., Fung, C. C., Rai, S. M., Xie, H. (2016). Value analysis of cyber security based on attack types. In *Itmsoc: Transactions on innovation and business engineering* (Vol. 1, pp. 34–39).

Sakti, A., Michalek, J. J., Fuchs, E. R. H., Whitacre, J. F. (2015). A techno-economic analysis and optimization of li-ion batteries for light-duty passenger vehicle electrification. *Journal of Power Sources*, *273*, 966–980.

Salfer, M., Eckert, C. (2018). Attack graph-based assessment of exploitability risks in automotive on-board networks. *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 1–10.

Schlüter, F., Hetterscheid, E., Henke, M. (2019). A simulation-based evaluation approach for digitalization scenarios in smart supply chain risk management. *Journal of Industrial Engineering and Management Science*, *1*, 179–206.

Schuh, G., Brettel, M., Reuter, C., Bendig, D., Dölle, C., Friederichsen, N., . . . Wolff, B. (2017). Towards a technology-oriented theory of production. In *Integrative production technology* (p. 1047-1079). Springer, Cham.

Semeraro, C., Lezoche, M., Panetto, H., Dassisti, M. (2021). Digital twin paradigm: A systematic literature review. *Computers in Industry*, *130*, 103469.

Smogeli, O. (2017). Digital twins at work in maritime and energy. *DNV-GL Feature*, *1*(7).

Staiger, R., Tantău, A. (2020). *Geschäftsmodellkonzepte mit grünem wasserstoff*. Wiesbaden, Germany: Springer Fachmedien Wiesbaden.

Sun, S., Luo, C., Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information fusion*, *36*, 10–25.

Tao, F., Qi, Q., Wang, L., Nee, A. Y. C. (2019). Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: Correlation and comparison. *Engineering*, *5*(4), 653–661.

Tao, F., Sui, F., Liu, A., Qi, Q., Zhang, M., Song, B., ... Nee, A. Y. C. (2019). Digital twin-driven product design framework. *International Journal of Production Research*, *57*(12), 3935–3953.

Tao, F., Zhang, H., Liu, A., Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics*, *15*(4), 2405–2415.

Telek, P., Bányai, T. (2018). Advanced materials handling processes and devices in the automotive industry. In *Vehicle and automotive engineering* (pp. 315–328). Springer, Cham.

Traub-März, R. (2017). The automotive sector in emerging economies: Industrial policies, market dynamics and trade unions.

Urbach, N., Ahlemann, F. (2016). Digitalization as a risk: Security and business continuity management are central cross-divisional functions of the company. *Computer Science On-Line Conference*, 85–92.

Violante, M. G., Vezzetti, E. (2017). Kano qualitative vs quantitative approaches: An assessment framework for products attributes analysis. *Computers in Industry*, *86*, 15–25.

Weill, P., Olson, M. H. (1989). An assessment of the contingency theory of management information systems. *Journal of Management Information Systems*, *6*(1), 59–85.

Wu, S., Zhang, Y., Cao, W. (2017). Network security assessment using a semantic reasoning and graph based approach. *Computers & Electrical Engineering*, *64*, 96–109.

Xiong, W., Legrand, E., Åberg, O., Lagerström, R. (2022). Cyber security threat modeling based on the mitre enterprise att&ck matrix. *Software and Systems Modeling*, *21*(1), 157–177.

Zio, E. (2018). The future of risk assessment. *Reliability Engineering & System Safety*, *177*, 176–190.

12 Transitions from Threat Hunting and Automated Incident Response[1]

Abstract

Cyber attacks keep states, companies and individuals at bay, draining precious resources including time, money and reputation. Attackers thereby seem to have a first mover advantage leading to dynamic defender attacker game. Automated approaches taking advantage of Cyber Threat Intelligence on past attacks bear the potential to narrow the gap between the capabilities of the actors and hence increase cyber security. Consistently, there has been a lot of research on automated approaches in cyber risk management including works on predictive attack algorithms and threat hunting. Since predictive and descriptive capabilities are becoming increasingly accurate, works on prescriptive capabilities in the field are of high interest for practical use ensuring cyber security. Prescriptive capabilities thereby include the automation of selecting effective countermeasures, which nowadays consumes high shares the scarce time of security professionals. Combining data on countermeasures from “MITRE Detection, Denial, and Disruption Framework Empowering Network Defense” and adversarial data from “MITRE Adversarial Tactics, Techniques and Common Knowledge” this work aims at enabling highly precise and efficient incident responses. We introduce Attack Incident Responder, a toolkit working with simple heuristics to find the most efficient sets of countermeasures for hypothesized attacks. Experimental results are promising high average precisions in the predicting effective defenses when using the toolkit. In addition, we compare the proposed defense measures against a static set of defensive techniques offering robust security against observed attacks. Furthermore, we combine the approach of automated incidence response to an approach for threat hunting enabling a full automation of a security operation center. By this means we define a threshold in the precision of attack hypothesis generation that must be met for predictive defense algorithms to outperform the baseline. The calculated threshold can be used to evaluate attack hypothesis generation algorithms. The presented toolkit for automated incident response may be a valuable support for information security professionals.

Keywords: Cyber security, automated incident response, cyber threat intelligence

[1] This chapter includes the preprint of the article "[Transitions from threat hunting and automated incident response" by Leon Andris, Tim Tennig, Jonas Iser, Marcus Wiens, Frank Schultmann, and myself.

12.1 Introduction

With the increasing penetration of broad aspects of modern life through information and communication technology, ensuring "cyber security has emerged as one of the most pressing issues for society with actors trying to use offensive capabilities and those who try to leverage on defensive capabilities to secure their assets or knowledge" (Kaiser et al., 2021). Yet, resources spent on cyber security for many companies are not high enough and/or measurement selection is ineffective in dealing adequately with the current threats (Frost and Sullivan, 2017). Ensuring a high level of cyber security can push companies to their limits due to budgetary as well as time restrictions of security professionals. This highlights the importance of enabling efficient incident responses and investment support for defensive means providing efficient cyber security as well as the importance of developing means for automation in cyber risk management. Hence, the analysis of data on past attacks (so called Cyber Threat Intelligence - CTI) is of utmost importance within this field.

CTI furthermore reveals the high diversity of methods developed to either secure a system from attacks (defensive cyber capabilities) or to compromise systems (offensive cyber capabilities). This is a result of an arms race between attackers and defenders (Kaiser et al., 2021). Offensive capabilities thereby comprise different techniques and tools. MITRE Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) comprises over 200 different attack techniques which may be employed by different malware. ATT&CK comprises over 650 different malware families. In contrast, MITRE Detection, Denial, and Disruption Framework Empowering Network Defense (D3FEND) comprises over 500 different countermeasures employing around 100 different defensive techniques. To leverage defensive capabilities against the manifold attacks to secure assets and knowledge, decisions need to be taken. Today, low automation levels in Security Operation Centers (SOCs) are prevalent. This leads to current procedures in cyber risk management where human decision makers (security professionals) decide on the incident response. However, security professionals are scarce and available professionals are highly time-constrained. Furthermore, decisions on security spending are prone to human errors due to high computational burdens in selecting the right response (defensive techniques) to an attack, lack of time/ time pressure when making these decisions, and subjective biases due to limited perspectives on the threat landscape (Kaiser et al., 2021), leaving vulnerable systems to hackers. Against this background, automated approaches based on analyses of CTI are frequently considered to be a helpful means for increasing the efficiency of incident response (i.e. identifying the smallest set of defensive means to react on an incident/ defend against all attack techniques employed by an attack) and allocation of scarce resources (i.e., in the sense of time-consumption of handling attacks) (Kaiser et al., 2021). Yet, practical implementations ready to support security professionals are rare. Consequently, the use of such (automated) means is limited. The inefficiencies in incident response can thereby also be observed in long times before detection and responses to data breaches (IBM Security, 2021). Automating incident responses could thereby support the security professionals' work, decreasing the time between breach detection and response and support the process of choosing the most effective solution to defend against attacks.

Yet, till now research on automating incident response and countermeasure selection is scarce. This is as connecting adversarial techniques and defensive techniques responding to them was challenging and cyber threat ontologies combining the attackers view and the defenders view are rare. The recent introduction of the D3FEND matrix addresses this issue and gives a straightforward way of connecting adversarial techniques as well as tools employed with countermeasures. A main benefit of D3FEND is thereby that it is consistent to the Structured Threat Information eXpression (STIX) allowing to link D3FEND to ATT&CK. This offers novel possibilities for research on automating incident responses. We aim at contributing to research on this issue by developing a toolkit to support countermeasure selection based on graph analytics called Automatic Incident Responder (*AIR*). For the evaluation of the approach introduced, we take advantage of D3FEND and ATT&CK and implement automated incident response based on the approach. We test *AIR* and compare the response proposed relying on the toolkit against a static set of smartly picked defense techniques (as a baseline, defending against the most common attacks). Furthermore, we combine the approach of automated incidence response with an approach for threat hunting enabling a full automation of a SOC. We thereby undertake a test on the robustness of the automated process for incident response regarding uncertainty in attack hypothesis generation. By doing so, we also introduce a novel baseline for threat hunting approaches.

The work is structured as follows. Section 12.2 introduces the theoretical foundations and related work. Hereafter, we introduce to the methodology. Section 12.4 describes the evaluative procedure. We thereby describe the process of building the knowledge graph based on which we evaluate the approach, the experimental setup, results derived from the experiment, and give a short discussion of the results.

12.2 Theoretical foundations and related work

12.2.1 Cyber threat intelligence

CTI is structured and actionable information on past attacks gained by intrusion detection systems and other systems for monitoring attacks on information systems (Elitzur et al., 2019). CTI is a data-driven approach that can help improve understanding of the threat landscape, predictions on future attacks or the occurrence of attacks, and adaptation to the dynamical change of attackers behavior. It holds versatile and inferred information about malware and attackers, including common tactics, techniques, tools and procedures (TTP), threat actors and indicators of compromise (IOC) (Elitzur et al., 2019). Since a single security professional will not be able to gain a full understanding of the threat landscape, it is of utmost importance to enable sharing CTI among different security professionals in a machine and human readable manner. This is, sharing CTI could allow security professionals to gain holistic perspectives on the threat landscape a company is confronted with. Sharing thereby necessitates a common language of involved security professionals. With the aim of enabling the process of sharing CTI among different stakeholders, MITRE developed STIX. STIX nowadays represents the de facto standard in sharing CTI (Sauerwein et al., 2017). CTI composes information on adversarial behavior (e.g., ATT&CK) as well as on defensive means

(e.g., D3FEND) and vulnerabilities (e.g., National Vulnerability Database - NVD). Yet, it was for a long time challenging to link the attacker model and the defender model with each other. Pioneering works were thereby undertaken inter alia by Hemberg and O'Reilly (2021). Yet, within their work, CTI describing adversarial behavior was linked to weaknesses and vulnerabilities but not to defensive means that would enable to minimize the cyber risk. Through the publication of D3FEND, linking adversarial behavior to possible countermeasures in response to the techniques employed by the attacker, novel possibilities on the establishment of automated approaches to countermeasure selection are opened.

In our research, we rely on D3FEND and ATT&CK. D3FEND is essentially a collection of publicly known defensive cybersecurity means (safe guards) and defensive techniques employed by these means. Their link to observables (i.e., digital artifacts) is given by adversary techniques provided e.g., in ATT&CK. The measures included in D3FEND are generalized techniques, mainly derived from an analysis of over 500 available patents, as those usually describe the defensive methods they use in detail to obtain sufficient protection and clarify their uniqueness (Kaloroumakis and Smith, 2021). As of now, MITRE declares D3FEND database to be an experimental research project in its infancy with fundamental generalizations and no prioritization or characterization of efficacy. Yet, today to the best of our knowledge, there is no publicly available, high fidelity and structured alternative that matches the informational content of D3FEND. ATT&CK in contrast, links adversarial behavior extracted from real-world observations and threat analysis reports (operational CTI) to techniques and tactics employed (strategic CTI) enabling a mapping of observables to different stages of an attack. These different stages of an attack are frequently described by the model of the Cyber Kill Chain (Lockheed, 2014).

12.2.2 Automation in cyber security

Graph analytics is a frequently used method in automating cyber risk management offering high chances in ensuring efficient cyber security. Kaiser et al. (2021) presents an approach called AFP for predicting the emergence of novel attacks and enables the identification of usage patterns of different attack techniques. Their approach proved high predictive capabilities within a time horizon of one year ahead when employed on the basis data provided by ATT&CK. Also, much research in recent years was undertaken on automating the detection and identification of attacks (Elitzur et al., 2019). Elitzur et al. (2019) propose the attack hypothesis generator (AHG), an approach delivering machine support for human analysts in threat hunting based on multiple link prediction techniques. Kaiser et al. (2022) extend their approach providing solutions for a full automation of the threat hunting process. They thereby add initial hypothesis generators taking over the role of human analysts. Hemberg and O'Reilly (2021) rely on a machine learning approach and links offensive capabilities to weaknesses and vulnerabilities. They thereby introduce a threat ontology called BRON which is also used for various applications in information retrieval, modeling and simulation, and artificial planning (Hemberg and O'Reilly, 2021).

Nespoli et al. (2018) provide an overview on scientific work on optimal countermeasures selection including automated approaches. Countermeasures can thereby either mitigate, or eliminate risks,

prevent a system from an attack or at least harden a system against an attack rendering systems more robust or resilient against cyber attacks (Nespoli et al., 2018). Among the works reviewed by Nespoli et al. (2018), optimality is defined differently. While Dewri et al. (2012), Roy et al. (2012) and Wang et al. (2013) set an economical cost-benefit analysis in the center of their definition of an optimal set defensive means, Viduto et al. (2012) focus on impact minimization by reducing the probability that a specific vulnerability is exploited. Differently to those approaches, Miehling et al. (2015) focus on selecting a set of countermeasures to optimaly counteract to in-progress cyber attacks. Cost-benefit approaches thereby suffer from uncertainty of cost estimations (e.g., outdated estimations extracted from aged works or uncertainties regarding implementation costs) and quantitative estimations of cyber risk minimizations.

12.3 Methodology

In this Section, we describe an analytical procedure for automating the process of selecting optimal cyber security countermeasures given an attack. An attack can thereby be described by the attack techniques it employs. Assume anomalous, suspicious activities within an information system are recorded (observables) leading to alerts of the SOC. The goal of the SOC would hence be to (1) identify the cause of the suspicious activities (threat hunting) by proposing hypotheses on the course of a possible attack and (2) define a set of actions to take in response to the hypothesis generated. An attack hypothesis would thereby consist of a set of attack techniques that are related to the observations. The response can equivalently be described as a set of defensive techniques that are suitable in response to the attack techniques. The problems described above can hence be defined as follows. We adapted the description of the hypothesis generation problem from Kaiser et al. (2022).

Problem 3 (Hypothesis generation problem) *Given a knowledge graph* $KG =< A, D, C, B, R >$*, where A is a set that contains the nodes representing past attacks, D is a set of nodes describing attack descriptions including attack techniques (oT) and observables OBS, and C which is the set of countermeasures that may be employed in response to attacks. Furthermore, B describes the countermeasure (descriptive node) including defensive techniques (dT) employed by C, and R as the set of relationsips (links) between the nodes. Currently observed data generate a hypothesis consisting of a set of oT, denoted as H, related to the currently observed data (COD).*

While there are some works on solving Problem 1 in an automated manner (e.g. Elitzur et al. (2019); (Kaiser et al., 2022)), automating the response against an ongoing attack is rather rarely researched. For enabling a combined hypothesis generation and response on the hypothesized attacks, we take advantage of those approaches presented in the literature and employ *AIR* on hypothesis generation as it is presented in Kaiser et al. (2022).

Problem 4 (Response problem) *Given the knowledge graph* $KG =< A, D, C, B, R >$*, and the hypothesis H, generate a set of defensive techniques dT that are able to successfully defend against H (set of dT related to H).*

The algorithmic approach *AIR* selects a set of cybersecurity countermeasure techniques to defend most effectively against a hypothesized attack by using as few defense techniques as possible combating all attack techniques employed by the hypothesized attack (defense in depth). The definition of an optimal defense technique hereby is similar to the work of Miehling et al. (2015). In this way, an optimal set of defensive measures employed defends against each attack technique employed by an attack in-progress. *AIR* employs a greedy heuristic by selecting one defense node every iteration, including this node to the set of proposed defensive measures, and deleting said defense node and all directly connected attack nodes from the graph. The optimal solution is found when all nodes have been deleted from the subgraph comprising all relevant attack techniques (attack techniques connected to the hypothesized attack). Note that the attack vector thereby might also represent non binary statements on attack techniques (not only binary statements on whether an attack is used or not but statements representing insecurity/ uncertainty on whether an attack technique is utilized in the sense of beliefs on the attack technique utilization/ expectations on the use of an attack technique). The ranking of defense nodes relies on the count of attack technique nodes to which a defense node is directly connected.

We therefore define the number of paths from C to D as follows.

Definition 16 (Count of attack techniques connected) *$|TF(c, oT_a)|$ is the number of paths from a countermeasure c to the set of relevant attack techniques oT_a (offensive techniques employed by a). Let dT denote the defensive techniques. $|TF(c, oT_a)|$ can then be calculated as follows.*

$$|TF(c, oT_a)| = |\{c - dT - oT - a\}|$$

The evaluation of defense nodes described can lead to several nodes with the highest rating, therefore not allowing for a definite node choice. In this case, the algorithm differentiates between three different scenarios:

(i) The defense nodes are connected to the same attack nodes in the current iteration, but the rating in iteration k=0 was not equal.

(ii) The defense nodes are connected to the same attack nodes in the current iteration, and the rating of the nodes was equal in the iteration $k = 0$.

(iii) The defense nodes are not connected to the same attack nodes.

In Scenario i, the algorithm includes the node with the highest initial rating to its optimal solution since it covers more defense nodes according to the definition of the rating. This means that more defense nodes will be covered twice or more by choosing this node over the others. In Scenario ii, the algorithm considers the nodes equivalent to the given attack thus including at least one of the nodes into the optimal solution would be the most efficient way to defend. In this case, the output of the algorithm's optimal solution is no longer a binary vector but becomes a probability vector. This means that value 1, representing the certainty of the node being included in the optimal solution, is distributed equally among the identically valued defense nodes. To simplify the handling within the algorithm and the evaluation, the algorithm deletes all equally valued nodes except one and returns a

list containing all equal nodes at the end. The step of splitting the probability amongst the defenses is thus not directly implemented in the algorithm. In Scenario iii, the algorithm selects one of the defense nodes at random. This means the algorithm is not deterministic. Therefore, according to the algorithm, there would be multiple optimal defense strategies (set of countermeasures for effectively defending against the attack) for a hypothesized attack.

12.4 Evaluation

12.4.1 Multi-level threat knowledge graph

The approaches employed for automating the works undertaken within a SOC rely on graph analytics. To test the novel toolkit *AIR*, we rely on a threat graph consisting of information from ATT&CK and D3FEND whose structure is given in Figure 12.1 (basically the threat graph is a subgraph of the presented graph comprising information on attacks and countermeasures excluding information on observables i.e., IOCs and other technical CTI, e.g., hash values). We therefore web-scraped CTI on every attack from ATT&CK (i.e., descriptions like TTPs associated with an attack) and linked the attack techniques to the according defensive means (countermeasures) and defensive techniques from D3FEND.

For the fully automated approach we furthermore added information on observables linked to the attacks from malware analysis reports. The crafted threat graph hence contains attack descriptions on different levels of the Pyramid of Pain (e.g. observables or IOCs) as well as descriptions on defensive means that may be employed to counter an attack. The structure of the generated threat graph is given in Figure 12.1.

12.4.2 Experimental setup

The experiment on whose basis we test *AIR* comprises two sub-experiments. Within the first setup, we test the ability of *AIR* to come up with reasonable sets of proposed defenses. We thereby only rely on a subgraph of the introduced multi-level-threat knowledge graph ($KG_{sub} =< A, aT, C, dT, R >$) where other attack descriptions besides aT and other defense descriptions besides dT are omitted. For analyzing the proposed set of defensive means, we generate a baseline as a static set of defense techniques. We created this baseline of defense techniques by finding the optimal set of defense technique dT for an average set of attack techniques utilized ($\overline{aT}$) with all observed attacks included in the knowledge graph. ($\overline{aT}$) represents a weighted set of all attack techniques, where the weight represents the probability that a representative attack $\overline{a}$ utilizes aT.

Definition 17 (Weightings of attack techniques in the average attack) *Let a be a specific attack and A the set of attacks in KG. Let furthermore $TF(a, aT)$ be the number of paths between a and aT which is either 0 (no connection) or 1 (a employs aT), and be defined analogous to TF(c,T). Then, the weights w can be calculated as follows.*

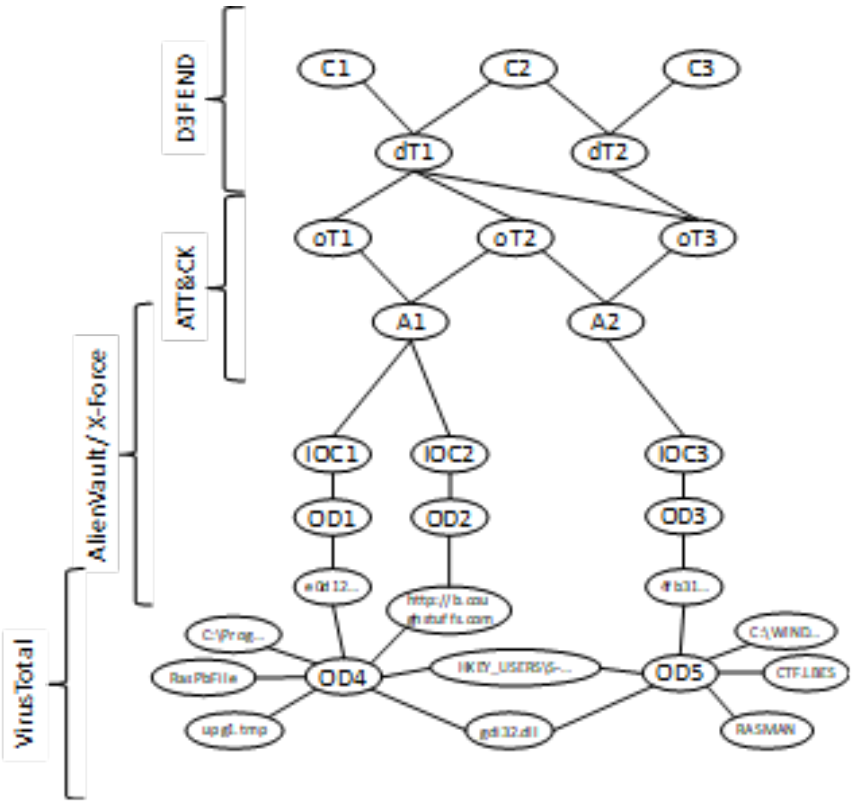

Figure 12.1: Structure of the threat knowledge graph

$$w(aT) = \frac{\sum_{a \in A} |TF(a,aT)|}{|A|}$$

To evaluate the algorithm, we calculated all optimal defense technique vectors (set of means able to defend against each attack technique employed by the specific attack) for every malware documented in the ATT&CK database. We used these vectors as the ground truth for the defense strategies. Then we applied the algorithm to the simulated attack technique vectors. We tested the results against the defense ground truth using the average precision. The second setup takes advantage of the full knowledge graph. Within a first step, we crafted attack hypothesis. The attack hypotheses vary in dependence of the precision of observables detected (false negative rate and false positive rate). Please refer to the works of Elitzur et al. (2019) and Kaiser et al. (2022) for an in depth review on the dependence of the precision of hypotheses and the data input quality. Furthermore, different initial hypothesis generation algorithms and hypothesis generation strategies generate different precisions of hypothesis. We generate hypotheses of different precision and craft propositions on the employment of defensive means. We evaluate the precision of these propositions and the optimality of the set of defenses against each attack. For evaluation, we provide the precision of the proposed set of defensive means as a function of the precision of the attack hypothesis.

12.4.3 Results

The calculated defense baseline shows two defensive techniques rated significantly higher than the others. The degrees of belief that these are helpful against an unknown attack (average attack) are

above 0.9, which can be attributed to the fact that they are connected with most aT in KG. The defensive baseline ($\overline{dT}$) shows high ability to defend against most a in KG. Compared against the ground truths of an optimal defense, against specific attacks, in many cases, a maximal average precision of 1 is achieved. This is due to the nature of some a that use few aT or a set of aT that can all be covered by a small set of dT respectively a single dT. Differently to the automated incident response based in *AIR*, a correlation between the complexity of an attack and a worse performance can be observed for the base protection. We thereby consider the number of techniques used by an attack that cannot be covered by the same defense as a measure of complexity. Hence, the following intuitive circumstance is validated. The higher the complexity/ the higher the maturity of an attack the worse the defensive capability of a static base protection ($\overline{dT}$) and vise versa, the less complex an attack the better the defense through $\overline{dT}$. Yet, for the observed a included in KG, $\overline{dT}$ shows high defensive capability. We calculate an average precision of the baseline of 0.84 where the average precision is calculated on the basis of the optimal set of defensive techniques employed for the given attacks in KG (ground truth).

The results of the combined approach of *AIR* on the top of a hypothesis generation toolkit, where we relied on AHG (second sub-experiment) show that an average precision of circa 0.7 in the attack hypothesis generation process is needed for *AIR* to outperform the static base defense with the average precision of 0.84 (see Figure 12.2). Furthermore, the experiment demonstrates a high robustness of *AIR* on weak attack hypotheses. This is as some defensive techniques are frequently utilized as seen in the first experiment respectively the comparison of the ground truth to the baseline. Looking at the individual malware, it can be observed that for malware with low complexity, the baseline outperforms *AIR* employed on top of AHG, while in the other cases (complex attacks/ attacks with high maturity), the predictive algorithm prevails.

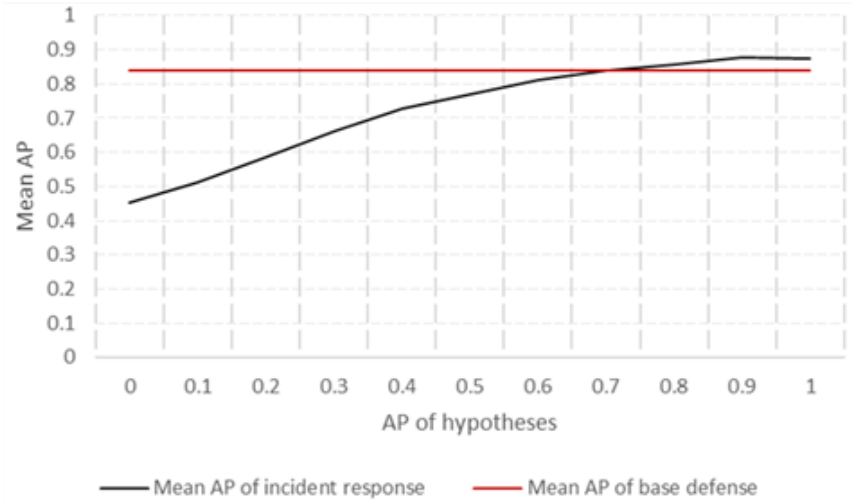

Figure 12.2: Mean AP reached by *AIR* compared to the base defense

12.4.4 Discussion

The results of the first sub-experiment demonstrate the high power of *AIR* to automatically select effective defensive measures against attacks if the underlying attack hypothesis generation is sufficiently precise (where we identified an average precision of circa 0.7 as the threshold where relying on *AIR* in general is beneficial compared to a base defense). Although, relying on *AIR* is only more effective at a hypothesis precision of 0.7 than investing in a base defense, *AIR* may (1) increase the efficiency of security investments, where implementing *AIR* is less costly than investing in the base defense and (2) increase the security of firms frequently attacked with highly sophisticated (complex) attacks (e.g. highly valuable attack targets like military facilities or big companies). Furthermore, robust combinations of base protection and *AIR* might offer highest protective capabilities combining the benefits of both adaptive automated incident response and static base protection in the sense of a bio inspired immune system for information systems. This is, the base protection would offer an effective defense against attacks with low complexity in the sense of a non-specific protection, while *AIR* could offer attack specific protection.

It need to be noted that the knowledge base is limited in content (CTI) and hereby especially in data on defensive measures, as D3FEND is still a relatively novel database and does not claim comprehensiveness, leaving some attack techniques without any possible defense. However, as the graph might become more comprehensive in the future, it is expected that the complexity of optimal defense strategies against attacks will rise. This would mean that the predictive defense algorithm could outperform the baseline even with lower reached average precisions of defensive measures proposed lowering the demand for precisions of attack hypothesis generation. Furthermore, with rising sophistication the base protection will lose in effectiveness while *AIR* significantly outperforms the base defense against those more complex attacks. Additionaly, *AIR* is highly dependent on the data input quality and reliability of the knowledge base it is implemented on. Timeliness of information included and comprehensiveness of information included in the knowledge base are hence essential for efficient automated incident response based on *AIR*. Threat information sharing and automation of attack graph geaneration (Sheyner et al. (2002)) are therefore essential for the approach presented. Within our work, we deliver a novel baseline, which can be used in the evaluation of approaches on attack hypothesis generation, representing a threshold, where automated incident responses in the sense of an (fully) automated SOC deliver more efficient solutions than security investments in a static base protection. However, even below the baseline, automating incident responses may be a valuable option to companies searching for means to increase cyber security. This is as investing in the base protection may lead to high costs - costs that may exceed the companies' monetary resources/ budgets. In this case, the presented approach for an automated response to cyber attacks may increase the companies' defensive capabilities while requiring limited base protection.

12.5 Conclusion

In this work, we present *AIR*, a toolkit for automating incident responses. The toolkit crafts a proposition of defensive means for the optimal reaction on an attack observed within a system. We validate the prescriptive capability of *AIR* when employed based on ATT&CK and D3FEND. We furthermore employed *AIR* on top of an approach for attack hypothesis generation. We thereby set a threshold for the precision that need to be reached for attack hypotheses where automated attack incident response exceeds the defensive capability of a base defense investment. In the future, this means that companies and cybersecurity experts could use automated defense suggestions to protect themselves more efficiently against cyberattacks and enable fully automated SOCs. To prove the ability of *AIR* to increase protective capabilities in practice is considered future work and would provide valuable insights on defensive means/ countermeasure selection and could hence provide valuable insights to further increase defensive capabilities. Furthermore, we combine approaches for automated hypothesis generation and incident response, future research need to focus on integrating other security solutions e.g., anomaly detection or intrusion detection systems to the defense cycle to gain insights on possible interference of algorithms.

References

Dewri, R., Ray, I., Poolsappasit, N., Whitley, L. D. (2012). Optimal security hardening on attack tree models of networks: a cost-benefit analysis. *International Journal of Information Security*, *11*, 167-188.

Elitzur, A., Puzis, R., Zilberman, P. (2019). Attack hypothesis generation. *2019 European Intelligence and Security Informatics Conference (EISIC)*, 40-47.

Frost, A., Sullivan, P. (2017). 2017 global information security workforce study: Benchmarking workforce capacity and response to cyber risk.

Hemberg, E., O'Reilly, U.-M. (2021). Using a collated cybersecurity dataset for machine learning and artificial intelligence. *ACM KDD AI4Cyber: The 1st Workshop on Artificial Intelligence-enabled Cybersecurity Analytics at KDD'21*.

IBM Security. (2021). Cost of a data breach report.

Kaiser, F., Budig, T., Goebel, E., Fischer, T., Muff, J., Wiens, M., Schultmann, F. (2021). Attack forecast and prediction. *Proceedings of the 28th Computer & Electronics Security Application Rendezvous (CE&SAR)*, 77–97.

Kaiser, F., Dardik, E., Elitzur A., Zilberman, P., Wiens, M., Schultmann, F., . . . Puzis, R. (2022). *Attack hypothesis generation based on threat knowledge base.*

Kaloroumakis, P. E., Smith, M. J. (2021). Toward a knowledge graph of cybersecurity countermeasures.

Lockheed, M. (2014). *Cyber kill chain.*

Miehling, E., Rasouli, M., Teneketzis, D. (2015). Optimal defense policies for partially observable spreading processes on bayesian attack graphs. In (pp. 67–76). ACM.

Nespoli, P., Papamartzivanos, D., Gomez Marmol, F., Kambourakis, G. (2018). Optimal countermeasures selection against cyber attacks: A comprehensive survey on reaction frameworks. *IEEE Communications Surveys & Tutorials*, *20*, 1361–1396.

Roy, A., Kim, D. S., Trivedi, K. S. (2012). Scalable optimal countermeasure selection using implicit enumeration on attack countermeasure trees. *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012)*, 1-12.

Sauerwein, C., Sillaber, C., Mussmann, A., Breu, R. (2017). Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017).*, 837–85.

Sheyner, O., Haines, J. W., Jha, S., Lippmann, R., Wing, J. M. (2002). Automated generation and analysis of attack graphs. *Proceedings 2002 IEEE Symposium on Security and Privacy*, 273-284.

Viduto, V., Maple, C., Huang, W., López-Peréz, D. (2012). A novel risk assessment and optimisation model for a multi-objective network security countermeasure selection problem. *Decision Support Systems*, *53*, 599–610.

Wang, S., Zhang, Z., Kadobayashi, Y. (2013). Exploring attack graph for cost-benefit security hardening: A probabilistic approach. *Computers & Security*, *32*, 158–169.

Part III

Discussion, Limitations and Conclusion

13 Implications

> The revolutionary idea that defines the boundary between modern times and the past is the mastery of risk: the notion that the future is more than a whim of the gods and that men and women are not passive before nature. Until human beings discovered a way across that boundary, the future was the mirror of the past or the murky domain of oracles and soothsayers who held a monopoly over knowledge of anticipated events.
>
> Bernstein and Bernstein (1996)

13.1 Methodological support for cyberrisk quantification

Research successes in the quantification of cyberrisks can contribute to provide information for external and internal stakeholders of organizations and can, hence, contribute to making controlling more expressive (Knowles et al., 2015). Quantitative risk assessments can enable the use of meaningful metrics forming great foundations for internal and external reporting. This is communicating cyberrisks to stakeholders who could profit from relying the methods presented. Stakeholders could furthermore be empowered to assess and evaluate whether an organization is compliant with standards, norms, and legal requirements (Knowles et al., 2015).

Cyberrisk quantification can further boost the organizations' ability to engineer secure systems by providing meaningful metrics on the security of a system. Quantitative security analysis can be integrated in the engineering and development process allowing scientific engineering toward more secure systems when quantifications are used for building feedback loops for engineers. In this sense, the set of methods presented can contribute to shaping a more secure future, enabling sustainable digitalization and may, hence, contribute to being able to leverage on the benefits provided by increasing digitalization, automatization, and intelligentization. These metrics can be included within the product development cycle to enable a continuous improvement of security granted (Knowles et al., 2015), to identify weaknesses as well as tracking and monitoring cyberrisks

in the first place (Patel et al., 2008). Furthermore, also for evaluation of the feasibility of different network architectures, the security quantification can be used (Knowles et al., 2015; Abraham and Nair, 2014). The quantitative values can be used for communicating risks to decision makers in an efficient manner and can, hence, be a helpful means for justifying budget demands, insurance premiums, and risk prioritization (Mishra et al., 2020; Eling and Wirfs, 2019; Patel et al., 2008). The set of methods can, be used within budget negotiations. It is essential to be able to base estimations on quantitative metrics as cybersecurity investments are capital intensive and investments are oftentimes subject to great uncertainty (Mishra et al., 2020). The methods for quantifying cyberrisks could, hence, diminish a main obstacle of cybersecurity, which is the bottleneck in investment capital (Mishra et al., 2020).

The set of methods developed for cyberrisk quantification represents a first set of its kind offering methodological support for comprehensive and holistic cyberrisk assessment. Results derived from application of the set of methods can enable the use of established cyberrisk management approaches and, hence, contribute to diminishing misallocations of security resources (Allodi and Massacci, 2017). This is, relying on quantitative metrics and risk management practices taking advantage of quantitative metrics of cyberrisks could increase the effectiveness of security investments. Quantification leverages economic potentials of cyberrisk management (e.g., applicability of insurances and their pricing) and the use of security as a competitive advantage. The lack of quantification approaches was considered a main reason, why the market potentials of cybersecurity (e.g., market for secure products) are under-exploited today (Ruan, 2017). The work provides insights which are helpful for establishing such markets for secure products offering potentials for quantifying cybersecurity and cyberrisks associated with products that can be used by customers to differentiate secure from insecure products. Customers could on such markets for secure products actively decide for products offering the level of security that comes closest to their needs.

One of these market potentials lies in offering cyberrisk insurance. As insurance premium determination is crucially dependent on the highly precise mathematical assessment of cyberrisks, the insights presented within this work can contribute to building a more flourishing market for cyberrisk insurances (Carfora et al., 2019). This is, insurance providers need to set their insurance premiums in a competitive market environment where a too high insurance premium renders the firm not competitive, while a too low insurance premium renders the firm non-lucrative respectively non-profitable. Those companies with high ability in quantifying cyberrisks of their customers would, hence, gain high advantage over their competitors. The present work aims at contributing to a more precise holistic assessment of cyberrisks and, by doing so, can contribute to the development of a market for cyberrisk insurances.

By relying on the digital twin, attack simulation was reached that goes beyond the current state of the art. Hereby, technical data (cyber threat intelligence) was used to simulate the effect of attack within a specific system. The approach gives insights on the effects that a specific attack can have within an underlying system. It increases the capability of quantifying the impact that successful attacks can have.

13.2 Decision support in countermeasure selection and portfolio optimization

Based on the quantification of cyberrisks, a method (chapter 12) was proposed that supports decision makers (e.g., security professionals) in countermeasure selection and the optimization of countermeasure portfolios for *efficient security* (Knowles et al., 2015). The method assesses the effectiveness of safeguards in enhancing security levels and answer the question of "how much security is improved if a specific security enhancement is applied" (Patel et al., 2008). Countermeasure portfolio optimization is focused within this approach on enabling efficient security. Efficient security describes a state of economically optimal security. The method described is thereby unique in the sense, that it rather focused on efficient and not maximal security. Maximal security is defined by what is technically possible and is in the focus of research form information technology research. The decision support system presented within this work is different to these methods for proposing maximal security and is hence, in line with the currently observable trend in security research and practice demanding for a greater introduction of a economic perspective (Bojanc and Jerman-Blažič, 2008).

The work introduces a decision support system that relies on a data analytics based approach for assessing the effectiveness of security controls against attacks. Information on the usefulness of safeguards is extracted from cyber threat intelligence. The method, contributes to the limited understanding of effectiveness evaluations of security means (Knowles et al., 2015) and can be a helpful means for determining which countermeasures are appropriate (Ganin et al., 2020). By doing so, it can be a helpful means for security professionals in the establishment of targeted security levels (Patel et al., 2008). Furthermore, it provides justifications for risk management and safeguard selections based on quantitative estimations (Ganin et al., 2020). The approach for a *bio-inspired artificial immune system* provided can help increasing the security of systems and sophisticate protection against cyberattacks.

Automatic incident response was introduced that can be implemented within a system to enable self protection capabilities. Systems could, hence, automatically estimate cyberrisks in a quantitative manner, predict the course of action of attacks that are currently observed within a system, and react to the ongoing attack in a feasible manner. In this sense, a bio-inspired artificial immune system is built for computer networks. This system protects its users from known and unknown attacks efficiently. It demonstrates its ability to do so within an experimental setting. Through its real time capability, reactions are rapidly deployed. Trusting the system for automated incident responses could hence reduce the need for human intervention and would therefore save scarce time of security professionals. This is, relying on the approach for automated incident responses would give relieve to high workloads of security professionals (Chen et al., 2014).

14 Limitations

The proposed set of methods for cyberrisk quantification is limited by the quality of the knowledge graphs the set of methods takes advantage of and in particular the actuality of cyber threat intelligence. This is, the attack takes into account evidence from past attacks and is, hence, limited to all observed attacks (Patel et al., 2008). Crafting such knowledge graphs is furthermore time consuming. Hence, knowledge graph generation demands for automation (Sheyner et al., 2002). Even more, automation is demanded to keep data bases actual, given the high dynamics in cyberrisk evolution and the complexity of systems (Sheyner et al., 2002).

The approach for decision support is furthermore limited by its generality. This is, efficient security is defined not as an overall optimal state of security, but as an optimal state of security given a defined budget (e.g., budget for a specific number of safeguards that can be employed) that is following the maximization principle (i.e., maximization of the security level under budgetary restrictions). However, the generality of the approach can also be considered as a strength of the approach as it is not valid for a specific branch or industry, but rather presents a generic approach that can be calibrated with industry and organization specific information to fit the special needs within the relevant case.

The introduced approach for cyberrisk management (decision support system) is widely technical in nature and does not draw on the monetary costs of safeguards. This is, as the implementation of safeguards may come at highly varying costs depending on the system it is introduced to. Determinants of these cost are inter alia specifics of the systems (implementation costs) and the caused opportunity costs (usability costs - e.g., if safeguards come at the cost of efficiency of operations; Alsaleh et al., 2017). Nevertheless, for enabling efficient security, an economic perspective on cyberrisk quantification is essential and could provide substantial benefits for cyberrisk management.

15 Conclusion

15.1 Summary

Within this work, a set of methods is proposed that can be used for cyberrisk management. It is based on methods that have been proven successfully to be helpful in cyberrisk quantification and efficient management. Following "the path that builds upon existing methodologies, enforces their strengths and eliminates weaknesses" (Leszczyna, 2021) was thereby followed. By doing so, the work aims at contributing to the closure of research gaps in cyberrisk management research and cyberrisk quantification as a necessary prerequisite for applying established risk management methods.

This work strived for contributing to an effective and highly precise cyberrisk quantification. The work provides a set of methods for decision support that enables cyberrisk quantification for security professionals in an automated manner. The set of methods relies on a multi-level threat ontology representing cyber threat intelligence on past attacks and making this available for graph analytics to tackle the big data analytics problem in cyberrisk quantification. The introduced methods enable a cyberrisk monitoring (chapter 6) as well as a forward looking cyberrisk estimation (chapter 7 and chapter 9 focusing on the human and societal domain). Furthermore, the set of methods comprises a module for predicting the technological development of cyberattacks for augmenting the muli-level threat ontology (chapter 8. Both, the monitoring approach and the forward looking approach, rely on digital twin based cyberattack impact quantification (chapter 10 and chapter 11). Hereby, the degree of digitalization can be varied. Based on this variations, the utility of selecting efficient degrees of digitalization is considered as an efficient means for cyberrisk management expanding the current boundaries of cyberrisk management. Based on the risk estmiations, the work proposes to derive means for managing short term risks and present an approach for a bio inspired artificial immune system for computer networks (chapter 12).

This work contributes to:

(i) Enabling automated cyberrisk quantification and monitoring of cyberrisks
(ii) Alternative means for cyberrisk quantification
    a) building upon methods for automatic information retrieval, data science, and automatic reasoning. Therefore, the work introduces a means for attack hypothesis generation that is based on the analysis of currently observed network artefacts can be used for threat hunting (chapter 7). The approach furthermore enables to quantify the beliefs (probabilities) of attacks given current observations.

b) building upon established approaches by introduction of a game theoretical means for cyberrisk quantification that is actionable and can automatically be calibrated within the highly dynamic field of cyber threats relying on cyber threat intelligence. The work thereby introduces an approach that relies on weighted attack graphs (chapter 7). Furthermore, the work contributes to the scientific knowledge in weighting attack graphs relying on motivations. The game theoretical approach on cyberrisk quantification is holistic focusing on both technical vulnerabilities (chapter 7) as well as human factors in cybersecurity (chapter 9).
c) building upon digital twins for quantifying the effects of successful cyberattacks and simulating their functioning(chapter 10 and chapter 11).

(iii) Tackling currently observed challenges in cyberrisk management by
a) introducing a novel cyber threat ontology fusing data from various cyber threat intelligence vendors (chapter 6).
b) enabling the use of technical cyber threat intelligence to derive more reliable and more meaningful strategic cyber threat intelligence for cyberrisk quantification (chapter 6).
c) introducing means for the analysis of cyber threats in a quantitative manner that take into account the complexities that emerge from interdisciplinary nature of cyber threat analyses (chapter 7 and chapter 9 within probability quantification and chapter 10 and chapter 11 in impact determination).
d) proposing a methodology for capturing the dynamic interactions between defenders and attackers (chapter 7) and the dynamics caused by the human nature of operators (chapter 9).
e) tackling the dynamic evolution of the threat landscape by the development of an approach for predicting and forecasting future attacks (chapter 8).
f) introducing means for simulating cyberattacks and their effects for deriving synthetic yet highly reliable data on the attack consequences (chapter 10 and chapter 11) tackling the challenge of applicability and enabling automated cyber-incident response within a bio-inspired artificial immune system for computerized networks.

15.2 Outlook

In addition to the perception of cybersecurity investments as a cost, cybersecurity can also be seen as an opportunity for companies. Till now, however, the economic evaluation of risks as well as opportunities through cybersecurity is only rarely considered in research. Cybersecurity can also represent a quality feature of a product and can represent a selling point for cutomers. A neccessary prerequisite for this is that customers are empowered in differentiating security of products. This is also as customers awareness regarding the risks that come with cyberattacks are still considered to be low and assessing the quality dimension is oftentimes challenging for customers (differentiating products according to the guaranteed level of security is challenging). Yet, in the same way that negative incidents can affect relationships with suppliers and customers, cybersecurity "as a quality feature" can become a positive competitive factor. Consistently, investments in cybersecurity could

strengthen innovative strength and competitiveness in the market. Cybersecurity, can thus help increase competitiveness within the market, foster customer loyalty as well as benefit in the long term through a positive company reputation (for example, for higher security quality of the products produced). This should be the focus of future risk management methods in order to assess the viability of business models in terms of cybersecurity and to develop resilient business models. Methodically, the developed industrial economic market model based on digital twins can be extended and validated by means of experimental research. In this way, business models could be evaluated according to their long-term prospects of success on the market and their contribution to the welfare of society.

In the area of probability determination, a set of different approaches is tested. Future research should be targeted towards increasing the capabilities. In doing so, further methodological approaches should followed and tested with regard to their ability in quantifying the probability of attacks given observables as well as in a forward looking manner. Also the use of pure decision theoretic approaches should be studied for forward looking cyberrisk quantification based on cyber threat intelligence.

Future research should furthermore include the threats that are posed to the focal company by its suppliers and the threats that may be introduced by them (Poppensieker and Riemenschnitter, 2018). Here, the demand for remote maintenance oftentimes introduces novel attack surfaces that gets scarcely managed by the focal company (Boehm et al., 2017). These unprotected flanks of vulnerability are not characteristic to industrial applications, but also emerge in the private sphere. In this way, refrigerators in smart home environments or pacemakers can be exposed to perpetrators of cyberattacks, leaving their owners unaware (representing blind spots for security professionals) of their existence and the threats they pose (Boehm et al., 2017). Testing the real world application would furthermore close considerable research gaps within the current literature (Leszczyna, 2021).

Bibliography

Abraham, S., Nair, S. (2014). Cyber security analytics: a stochastic model for security quantification using absorbing markov chains. *Journal of Communications*, *9*(12), 899–907.

Allodi, L., Massacci, F. (2017). Security events and vulnerability data for cybersecurity risk estimation. *Risk Analysis*, *37*(8), 1606–1627.

Alsaleh, M. N., Al-Shaer, E., Husari, G. (2017). Roi-driven cyber risk mitigation using host compliance and network configuration. *Journal of Network and Systems Management*, *25*(4), 759–783.

Bernstein, P. L., Bernstein, P. L. (1996). *Against the gods: The remarkable story of risk* (Vol. 383). John Wiley & Sons New York.

Boehm, J., Merrath, P., Poppensieker, T., Riemenschnitter, R., Stähle, T. (2017). Cyber risk measurement and the holistic cybersecurity approach. In *Perspectives on transforming cybersecurity* (pp. 61–74).

Bojanc, R., Jerman-Blažič, B. (2008). An economic modelling approach to information security risk management. *International Journal of Information Management*, *28*(5), 413-422. Retrieved from `https://www.sciencedirect.com/science/article/pii/S026840120800039X` doi: https://doi.org/10.1016/j.ijinfomgt.2008.02.002

Carfora, M., Martinelli, F., Mercaldo, F., Orlando, A. (2019). Cyber risk management: An actuarial point of view. *Journal of Operational Risk*, *14*(4).

Chen, Q., Abdelwahed, S., Erradi, A. (2014). A model-based validated autonomic approach to self-protect computing systems. *IEEE Internet of things Journal*, *1*(5), 446–460.

Eling, M., Wirfs, J. (2019). What are the actual costs of cyber risk events? *European Journal of Operational Research*, *272*(3), 1109–1119.

Ganin, A. A., Quach, P., Panwar, M., Collier, Z. A., Keisler, J. M., Marchese, D., Linkov, I. (2020). Multicriteria decision framework for cybersecurity risk assessment and management. *Risk Analysis*, *40*(1), 183–199.

Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., Jones, K. (2015). A survey of cyber security management in industrial control systems. *International journal of critical infrastructure protection*, *9*, 52–80.

Leszczyna, R. (2021). Review of cybersecurity assessment methods: Applicability perspective. *Computers & Security*, *108*, 102376.

Mishra, S., Anderson, K., Miller, B., Boyer, K., Warren, A. (2020). Microgrid resilience: A holistic approach for assessing threats, identifying vulnerabilities, and designing corresponding mitigation strategies. *Applied Energy*, *264*, 114726.

Patel, S. C., Graham, J. H., Ralston, P. A. (2008). Quantitatively assessing the vulnerability of critical information systems: A new method for evaluating security enhancements. *International Journal of Information Management*, *28*(6), 483-491. Retrieved from https://www.sciencedirect.com/science/article/pii/S0268401208000054 doi: https://doi.org/10.1016/j.ijinfomgt.2008.01.009

Poppensieker, T., Riemenschnitter, R. (2018). A new posture for cybersecurity in a networked world. In *Perspectives on transforming cybersecurity* (pp. 18–26).

Ruan, K. (2017). Introducing cybernomics: A unifying economic framework for measuring cyber risk. *Computers & Security*, *65*, 77–89.

Sheyner, O., Haines, J., Jha, S., Lippmann, R., Wing, J. M. (2002). Automated generation and analysis of attack graphs. In *Proceedings 2002 ieee symposium on security and privacy* (pp. 273–284).