TAL TECH

DOCTORAL THESIS

Advancing Cybersecurity Education through Learning Analytics

Kaie Maennel

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2021

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
29/2021

Advancing Cybersecurity Education through Learning Analytics

KAIE MAENNEL

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science

The dissertation was accepted for the defence of the degree of Doctor of Philosophy on 17 May 2021

Supervisor: Professor Dr Rain Ottis,
Department of Software Science, School of Information Technologies
Tallinn University of Technology
Tallinn, Estonia

Co-supervisor: Professor Dr Stefan Sütterlin,
Faculty of Computer Science
Albstadt-Sigmaringen University
Sigmaringen, Germany

Opponents: Professor Dr Nickolas Falkner,
University of Adelaide
Adelaide, Australia

Professor Dr Cornelius König,
Saarland University
Saarbrücken, Germany

Defence of the thesis: 16 June 2021, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Kaie Maennel _____
signature

European Union
European Regional
Development Fund

Investing
in your future

ISSN 2585-6898 (publication)
ISBN 978-9949-83-701-4 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9949-83-702-1 (PDF)
Printed by Koopia Niini & Rauam

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
29/2021

Küberkaitsealase hariduse parendamine õpianalüütika abil

KAIE MAENNEL

TAL TECH
KIRJASTUS

Contents

List of Publications

The present thesis is based on the following publications that are referred to in the text by Roman numbers.

I K. Maennel, R. Ottis, and O. Maennel. Improving and measuring learning effectiveness at cyber defense exercises. In *Nordic Conference on Secure IT Systems*, pages 123–138. Springer, 2017

II M. Kont, M. Pihelgas, K. Maennel, B. Blumbergs, and T. Lepik. Frankenstack: Toward real-time red team feedback. In *IEEE Military Communications Conference (MILCOM)*, pages 400–405. IEEE, 2017

III T. Lepik, K. Maennel, M. Ernits, and O. Maennel. Art and automation of teaching malware reverse engineering. In *International Conference on Learning and Collaboration Technologies*, pages 461–472. Springer, 2018

IV K. Maennel, S. Mäses, and O. Maennel. Cyber hygiene: The big picture. In *Nordic Conference on Secure IT Systems*, pages 291–305. Springer, 2018

V K. Maennel, S. Mäses, S. Sütterlin, M. Ernits, and O. Maennel. Using technical cybersecurity exercises in university admissions and skill evaluation. *IFAC-PapersOnLine*, 52(19):169–174, 2019

VI M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020: 15th International Conference on Cyber Warfare and Security*. Academic Conferences and Publishing Limited, 2020

VII K. Maennel. Learning analytics perspective: Evidencing learning from digital datasets in cybersecurity exercises. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 27–36. IEEE, 2020

VIII K. Maennel, K. Kivimägi, S. Sütterlin, O. Maennel, and M. Ernits. Remote technical labs: an innovative and scalable component for university cybersecurity program. In *Educating Engineers for Future Industrial Revolutions—Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2021

Author's Contributions to the Publications

I In I, I was the main author, developed the evaluation models, carried out the data collection and the analysis of the results, prepared the figures, and wrote the manuscript.

II In II, I was responsible for the learning evaluation section (paper section: "IV. Assessment"), carried out the data collection and the analysis of the results, prepared the figures, and wrote this section of the manuscript.

III In III, I researched and wrote the literature review, prepared the figures, and wrote the manuscript.

IV In IV, I was the main author and researched and wrote the literature review, conducted the analysis of the results, and wrote the manuscript.

V In V, I carried out the data collection and the analysis of the results, prepared the figures, and wrote the manuscript.

VI In VI, I participated in development of the design model, performed related literature review and contributed to the writing of the manuscript.

VII In VII, I was the main author and researched and wrote the literature review, conducted the analysis of the results, prepared the figures, and wrote the manuscript.

VIII In VIII, I was the supervisor of the second author, supported with the literature review, the methodology and analysis of the results, and wrote the manuscript.

Other Publications

IX K. Maennel, J. Kim, and S. Sütterlin. Team learning in cybersecurity exercises. Proceedings of the 5th Interdisciplinary Cyber Research Conference 2019, [ICR]: 29th of June 2019, Tallinn University of Technology. Osula, A-M.; Maennel, O. (Eds.), pages 17-19. (EXTENDED ABSTRACT)

X K. Maennel, J. Kim, and S. Sütterlin. From text mining to evidence team learning in cybersecurity exercises. Companion Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK20): Cyberspace, March 23-27, 2020. The Society for Learning Analytics Research (SoLAR), pages 107-109. (POSTER)

XI S. Sütterlin, B.J. Knox, K. Maennel, M. Canham, and R. Lugo, R.G. On the relationship between health sectors digitalization and sustainable health goals: A cyber security perspective. In: Flauhalt, A. (Ed.). Transitioning to Good Health and Well-Being. MDPI 2021. (BOOK CHAPTER).

Abbreviations

BT	Blue team
CST	Cyber Security Technologies course in TalTech
CSX	Cybersecurity Exercise
CTF	Capture the Flag
CYBOK	Cyber Security Body of Knowledge [101]
EDM	Educational Data Mining
ENISA	European Union Agency for Network and Information Security
HOTS	Higher Order Thinking Skills
GT	Green team
i-tee	Intelligent Training Exercise Environment [37]
LA	Learning Analytics
LS	Locked Shields, an annual exercise organised by the NATO CCD COE in Estonia
MOOC	Massive Open Online Course
NATO	North Atlantic Treaty Organization
NATO CCD COE	NATO Cooperative Cyber Defence Centre of Excellence
NIST NICE	National Institute of Standards and Technology National Initiative for Cybersecurity Education [87]
RQ	Research question
RT	Red team
SA	Situational Awareness
SITREP	Situation Report
SOLO	Structure of the Observed Learning Outcome
STEM	Science, Technology, Engineering and Mathematics disciplines
VM	Virtual Machine
WT	White team
XS	Crossed Swords, an annual exercise organised by the NATO CCD COE in Estonia
YT	Yellow team

1 Introduction

The importance of cybersecurity has grown and it is now considered an independent discipline [17] or meta-discipline within the computing industry [96]. Cybersecurity is defined as "a computing based discipline involving technology, people, information, and processes to enable assured operations in the context of adversaries" [17]. It involves the creation, operation, analysis, and testing of secure computer systems; and also includes aspects of law, policy, human factors, ethics, and risk management [17]. Cybersecurity is not just about technology and systems but also includes the people who use those systems as a critical component for effective cyber defence. However, the human element (incl. most efficient ways for learning and training) in cybersecurity has traditionally been overlooked and the role of technology overemphasised.

The effective learning, teaching, and skills improvement of cybersecurity students and professionals is a critical research area. This is particularly so, as there is a high demand for skilled professionals and a shortage of suitably-skilled individuals [15]. As the cybersecurity discipline involves multiple aspects, the training programs typically teach a mixture of technical and soft skills.

To teach cybersecurity, a wide range of training programs has been developed, ranging from the traditionally-taught cybersecurity Master programs [17] to other online alternatives. As part of these cybersecurity training programmes, hands-on exercises (both online and classroom) are gaining popularity in both university curricula and professional training paths.

We define a cybersecurity exercise (CSX) as a learning or training event in which individuals or teams implement, manage and defend a network of computers at a tactical, operational or strategic level (Publication VII). CSXs vary significantly in scale and content (from short online or classroom exercises; Capture the flags (CTFs) to large-scale/multi-stakeholder exercises). Following a taxonomy developed by the European Union Agency for Network and Information Security (ENISA) (based on international standard ISO-22398), we categorise an exercise as a CTF, Discussion-based game, Drill, Red team / blue team, Seminar, Simulation, Table-top or Workshop [91].

Such CSXs are generally viewed as an effective and engaging way of teaching a mixture of technical and soft skills in educational and professional settings ([91], Publication VII). In addition to CSXs for learning purposes (e.g., as part of university courses, competitions across universities, etc.), most national and international CSXs also focus on training and providing participants an opportunity to gain knowledge, understanding and skills [91].

Cybersecurity exercises feature a number of common learning design and measurement challenges, despite their different scale or learning content. One such challenge, particularly in exercises, is that when training is offered, there is a lack of evidence that participants have actually learned anything. The literature points out many shortcomings on evidencing that learning objectives have been achieved—e.g., "after-action reports… fantasy documents… few, if any, controls … to verify that … anything has actually been learned" [91], "…evaluation methodologies simply focus on the improvement of one cyber exercise to the next" [5], and "…evidence is often anecdotal and little work to validate learning outcomes has been done" [100].

However, CSXs often leave digital footprints (especially technical training conducted on computers and cyber ranges) of the learning process that allow an application of analytics to advance learning experience and provide evidence that learners have achieved the designed learning objectives. Learning analytics (LA) is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [107].

As a field of research, LA aims to predict and advise on learning by supporting education providers in identification of students' learning needs and improvement of pedagogical strategies, e.g., [106], [114]. It is an emphasis on analytics in order to make sense of the connective structures that underpins this field of knowledge [107]. However, establishing plausible relationships between models derived from quantifiable digital data and the complex socio-cognitive world of "learning" is challenging [63]. Learning analytics is closely intertwined with educational data mining (EDM), which develops, researches, and applies computerised methods to detect patterns in large collections of educational data [93]. EDM is incorporated in learning analytics to provide methods for data analysis processes.

This thesis focuses on the application of LA in cybersecurity training (specifically in CSXs) as a way to provide a more evidence-based and systematic approach for the evaluation of learning impact to enable the design of more effective learning. The research explores how to implement LA to enhance the learning experience or to be used for the skills assessment for cybersecurity students and professionals.

1.1 Problem Statement

Currently, the application of learning analytics approaches and methods using digital datasets in the cybersecurity education is limited. Cybersecurity education as a discipline lacks both consensus and practical implementations of how learning analytics can be applied to enhance the learning experience or be used as skills assessment for cybersecurity students and professionals as part of their learning journey (incl. CSXs).

The implementation of LA is not limited to enhancing the learning experience, but can also assist in the selection/admission process for a job or an academic program (predictive analytics). Verbert et al. [114] identify six objectives in existing LA research: 1) predicting learner performance and modelling learners, 2) suggesting relevant learning resources, 3) increasing reflection and awareness, 4) enhancing social learning environments, 5) detecting undesirable learner behaviours, and 6) detecting affects of learners. All of these objectives are relevant for cybersecurity education (and CSXs), but currently the full potential of LA research is not yet utilised, neither in enhancing the learning experience nor in skills assessment.

Several academic papers (e.g., [97], [77]) and non-academic guides (e.g., [61], [1]) describe CSXs design and evaluation but do not cover the use of learning analytics. Despite digital learning traces being available, many papers on learning in cybersecurity field are based on the experience and interpretation of the authors or based on learner evaluation (e.g., feedback surveys). The evaluation is often anecdotal, e.g., "everyone feels they had learned important lessons [26]" or "exercises are a very effective way of learning the practical aspects of information security" [97]. Typical evaluation methods are score-boards, verbal feedback and after-action reports highlighting conclusions from a manual analysis of data [116], non of which apply an analytical approach using digital datasets. Currently, measuring and improving learning with the help of analytics is an early-stage research area in cybersecurity education, specifically in CSXs.

As many CSXs leave a digital footprint of learning processes, analysis of such evidence-based learning traces can be used to improve the learning experience for cybersecurity students and specialists. However, as the LA research in cybersecurity education is in early stages, there is a lack of both research on the complete cycle of LA processes (such as [107]) and overall guidance how to incorporate LA into CSXs. Also, the analysis of learning data in individual programs including such CSXs can also provide information for an general cybersecurity curriculum or training program design.

This thesis aims to explore the application of LA and support the practical LA implementations of CSXs to assist in achieving planned learning or educational outcomes at the level of individual learners, teams, educators, exercise organisers and organisations.

1.2 Research Questions

This research focuses on how to improve cybersecurity training and potential use of learning analytics from digital traces to enhance the learning experience and skills assessment process. It has a special focus on the CSXs for learning purposes for students or professionals.

This thesis addresses the following research questions (RQs):

1. How to deploy a learning analytics approach into the CSXs in order to improve learning processes and to evidence learning outcomes?

   1.1. How and what data to collect when implementing a learning analytics approach in CSXs (a practical reference model)?

   1.2. What instructional design approaches would enable the connection of raw data from digital learning traces to high-level competencies?

2. How to evaluate effectiveness and efficiency of both individual and team learning with less obtrusive methods in CSXs?

3. Could CSX's LA measures be standardised as performance indicators and be suitable as the performance predictors for cybersecurity technical skills?

The mapping of research questions to the corresponding publication and thesis chapter, where the question is explored and answered, is presented in Table 1.

Table 1 – Mapping of Research Questions and Publications

Research Question	Publication	Chapter Number	Main content
RQ1.1.	Publication VII	3	LA framework
RQ1.2.	Publications VI, III, II	4	CSXs design
RQ2.	Publication I	5	CSXs LA measurements
RQ3.	Publications V, VIII	6	predictive performance indicators

1.3 Research Methodology and Methods

LA research, especially when applied in cybersecurity education and exercises, is a novel research area, and research methodologies and methods are also evolving. The methodological issues in LA need a further focus on addressing inherent trade-offs in learning environments, the clarification of methodological issues, and the scalability of system development [98]. Balancing the diversity and interaction of methodologies is seen as a great challenge in LA [9]. Specifically in cybersecurity educational research, methods vary depending on research focus and there are some studies emerging that start to look into "how" the learner completes tasks (i.e., use of tools, attempts, submission of wrong answers) from the digital datasets, such as [2], [3], [67]. However, validation of applied methods and learning indicators are often limited (e.g., 4 participants [67]), see publication VII.

This work is following the principle that simply measuring what can be measured is not enough, and we should measure what we value [118] and want to know. The metrics used

in the CSXs often focus on easily measurable data (e.g., time spent, number of attacks mitigated, etc.) and individual actions. However, students are “too easily satisfied that a system is secure after identifying only one possible source of security vulnerability for a system rather than seeking to explore the adversarial space more thoroughly” [104]. Thus it is important to understand not only whether the students have found a correct answer but how they found it [117]. Therefore research needs to focus on the development of an approach combining technical and cognitive metrics, while also requiring validation using learning analytics methods.

The novelty of research field, philosophical framework, and fundamental assumptions described above has influenced the choice of methods for this work. As an overall approach, that of "mixed methods" is applied, combining qualitative and quantitative methods [31]. The research follows the rationale [30] that the use of mixed methods is appropriate for most research problems. By giving significant weight to both the quantitative and qualitative evidence, researchers can more easily construct a holistic understanding of the phenomenon by synthesizing the inductive and deductive data [30]. By using this approach, the results of both data sets act as buffer and a check against overstating the conclusions derived from either approach alone [30].

The goal of this research is intended to both explain (quantitative) and explore (qualitative) aspects. This requires hypothesizing and then generalizing or applying an hypothesis to other populations. Simultaneously it aims to gain a more precise understanding of the dynamic interaction and perceptions of the stakeholders (i.e., the learners, the educators, the organisers, etc.) involved. Thus, in the research circle, the quantitative and qualitative methods are combined (see Figure 1). The inquiry can thus move from theory to data and back again, or from data to theory and back again, with inductive and deductive reasoning involving overlapping steps [41].

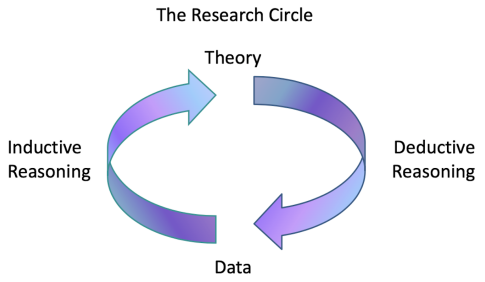

Figure 1 – Research Circle under Mixed Methods adopted from Bachmann & Schutt, 2007 [41].

Building upon interdisciplinary research between different aspects of cybersecurity (incl. technological opportunities and limitations, threat vectors), pedagogy and psychology, the following quantitative and qualitative methods are mixed:

- Conducting exercises on different training platforms;
- Data collection using various social science methods (i.e., surveys, case, interpretive) [10];
- Exploratory data analysis with focus on computationally intensive elements [42];
- Theory construction under a mixed methods research framework [41].

Various cybersecurity exercises and platforms (see Chapter 2) are used as data sources and also as the test-beds for developed theories and models. A qualitative contribution to

theory is particularly important when exploring topics that are difficult to quantify, such as learning, when trying to make sense of complex situations, when answering the broader question (e.g., how are the students learning?) or when attempting to explain how stakeholders (e.g., learners, teams) make sense of their situation [41]. Therefore the traditional social science data collection methods, such as surveys, etc., are used to obtain qualitative insights and also to validate data trends from exploratory data analysis.

This research also focuses on an exploratory element of large digital datasets from the cybersecurity exercises and learning analytics objectives “computationally intensive" elements. LA is data-driven (e.g., data mining, visualization) and grounded on a specific research context [94]. Due to this, a mixture of data mining and visualisation mixed with a range of regression methods [42] is instrumented. For example, this method is applied in Publications V and VIII to search for the best fitting lines and equations to represent relationships in the data and to build models of the phenomena under study.

Theory-building is relevant because it provides a framework for analysis, facilitates the efficient development of the field, and is needed for the applicability to practical real-world problems [41]. Considering that the application of learning analytics in cybersecurity education is currently in its early stages, such research is valuable. In social sciences, the inductive-synthesis model (also referred to as “grounded theory") has developed into a robust and sophisticated system for generating theory across disciplines [40]. However, the mixed methods approach allows switching between inductive and deductive reasoning (see Figure 1) and is thus more relevant in exploratory research of novel aspects. The theory construction under the mixed methods research framework is followed when developing the frameworks or models proposed in publications, I, IV, VI, VII and VIII. The research contributes to the theory bi-dimensionally: (1) originality [incremental or revelatory] and (2) utility [scientific or practical] [29]).

The mapping of research methods to the corresponding research question and publication, where the question is explored and answered, is summarised in Table 2.

Table 2 – Research Methods Mixed in this Research

RQ	Pub	Topic	Methods used / mixed
RQ1.1.	VII	LA framework	literature review, theory construction
RQ1.2.	VI	design	case study (data collection from CSX platform), qualitative survey, data analysis, theory construction
	III	design	case study (data collection from CSX platform), descriptive analysis, quantitative survey
	II	design	case study (experimenting with new tool and data collection from CSX), quantitative survey, data analysis
RQ2.	I	LA measurements	case study (data collection from CSX platform), quantitative and qualitative surveys, data analysis, theory construction
RQ3.	V	predictive analytics	descriptive analysis, quantitative survey
	VIII	predictive analytics	data collection, exploratory data analysis, theory construction

Overall, this research uses methodology and methods that contribute to providing practical knowledge and solutions that can support implementing LA approaches in cybersecurity education.

1.4 Contribution

An opportunity to improve learning would be missed by not considering how users experience learning in different styles and pace, depending on the learning environment. This thesis builds on the view that an application of LA and using digital datasets can allow further analysis and evidence-based improvement. However, with LA and evidence-based measurement, we also need to keep in mind and validate that what we measure (i.e., metrics used) actually helps learners to learn or assesses their knowledge and skills reliably. Note that conducting such measurements does not automatically help learners to learn, but using those measurements is an evidence-based basis for improving learning experience and environment in CSXs.

As learning activities are conducted on the computers and networks, the visual observations of behaviour (e.g., sitting quietly behind computer screen but at the same time mitigating a significant risk or attack) might not provide sufficient information. The observation method should be seen from a different perspective of learning to observe on the network and system-level, i.e., looking into digital footprints. This work promotes and supports the use of several tools and techniques that implement an unobtrusive approach by using the technical dataset. Overall, this thesis is based on eight publications that contribute to the specific research aspects, as described in Chapter 10.2.

The overall contribution of this research is the development of the knowledge and practical approaches of how and what (digital) data should be collected in cybersecurity trainings, and how to connect such learning data to relevant learning theories. The research contributes to this goal by obtaining pedagogical and technological underlying knowledge on how to design and measure the learning process in CSXs. Such understanding can then serve as an input to the exercise design cycle and enable creating interventions during the learning process in CSXs to help learners to learn more effectively.

1.5 Thesis Structure

This thesis is divided into ten chapters. The introductory chapter provides a brief overview of the learning and challenges in cybersecurity trainings and exercises, as well as the research questions and contribution of the thesis. Chapter 2 gives an overview of related work and background in cybersecurity exercises and learning analytics.

Chapter 3 describes the high-level model for implementing a LA approach in cybersecurity exercises, while Chapter 4 focuses on the critical phase of designing the CSXs in a way that data for learning analytics can be collected. In addition, this chapter provides use cases of practical experiences in designing a learning experience and environment for two distinct learning objectives (digital forensics and stealthy red team operations) with learner feedback on the learning experience. Chapter 5 moves on to measuring the learning impact by demonstrating a novel 5-timestamps methodology to capture individual and team learning data for effective feedback in a live Blue/Red exercise.

Chapter 6 moves on to “predictive analytics" and presents a novel idea of using the remote technical cybersecurity exercises as part of the university admission process and using these labs to measure the technical skills of the applicants.

Chapter 7 describes LA and CSXs in the wider context of cybersecurity competency frameworks and education. As the research involves human subjects, Chapter 8 emphasises the privacy and ethics aspects of cybersecurity and LA and also confirms the ethical guidelines followed during this research. Chapter 9 acknowledges the limitations of this research. Chapter 10 summarises the work done, brings out the main findings of this research, and outlines further work.

2 Background and Related Work

This Chapter describes the CSXs used in this research in more detail. We also provide the background, relevant learning theories, current measurement efforts of learning, and discuss the connection to LA and CSXs.

2.1 Cybersecurity Exercises and Learning Aims

Cybersecurity training teaches both technical and soft skills, as the field involves technology, people, information, and processes. A wide range of trainings has been developed by universities [17] and other organisations providing cybersecurity education. As part of such cybersecurity trainings, hands-on exercises (both online and classroom) are gaining popularity in university curriculum and professional training paths.

The author views CSX as "a learning or training event in which individuals or teams implement, manage and defend/attack a network of computers at the tactical or strategic level" (Publication VII). Following the ENISA taxonomy [91], the exercises are categorised as Capture the flag, Discussion based game, Drill, Red team / blue team, Seminar, Simulation, Table-top and Workshop [91]. Exercises with gamified elements are referred to as "serious games" and those with competitive elements as "competition".

2.1.1 Exercise Locked Shields

The Locked Shields (LS) exercise series is the basis for research on the learning measurement framework of 5-timestamps in Publication I.

LS is one of the largest real-time international cyber defence exercises (red team / blue team). It is organised annually by the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE) in Estonia. In 2019, more than 2500 possible attacks were carried out, using over 4000 virtualised systems and involving nearly 1200 participants from 30 nations [82]. The training audience was comprised of national Blue Teams (BT), consisting of computer emergency response specialists. The exercise ran on separate virtualised game-net, which was accessed remotely over VPN [83].

It is a team-based exercise, where individual learning is vital, but an important part is how teams overcome individual shortcomings in skills and knowledge and achieve the best result as a team.

The overall goal of LS is to "train teams of cyber professionals to detect and mitigate large-scale cyber attacks and handle security incidents" [83]. Specific training objectives are defined for IT specialists, including learning the network; system administration and prevention of attacks; monitoring networks, detecting and responding to attacks; handling cyber incidents; teamwork: delegation, dividing and assigning roles, leadership; cooperation and information sharing; reporting/ability to convey the big picture, time management and prioritization [83]. The exercise also includes specialised parts, such as conducting forensic investigation, crisis communication (media play), cyber legal aspects (legal play) [83] and a strategic game.

In the learning design, a game-based approach has been taken, meaning that the participants do not play in their real-life role, and the activities take place in a lab environment. In the exercise, the BTs are playing the role of Rapid Reaction Teams of a fictional country. The primary focus is defence and the BTs are tasked to protect and maintain identical pre-built virtualised networks of fictional, yet realistic organisations against the Red Team's (RT) attacks. The BTs also need to share findings with the White Team (WT) and other BTs; respond to legal, media and scenario injects; and solve forensic challenges designed in accordance to training objectives.

2.1.2 Exercise Crossed Swords

The exercise Crossed Swords (XS) series is used as a case study for designing and implementing visual feedback to assist learning in the exercise and measuring learning impact of such design elements (Publication II).

XS is an intense, hands-on, technical CSX oriented towards penetration testers working as a single united team, accomplishing mission objectives and technical challenges in a virtualised environment. While a technical CSX is commonly aimed at exercising defensive capabilities (i.e., Blue Team), XS addresses unique cyber defence aspects and focuses on training the Red Team members [65].

The main focus is to develop tactical and stealthy execution skills in a responsive cyber defence scenario. Training objectives include practising evidence gathering and information analysis for technical attribution; executing a responsive cyber defence scenario for target information system infiltration; applying stealthy execution and attack approaches; exercising working as a united team in achieving the mission objectives; and developing red teaming skills and effective tool usage, information exchange and situational awareness provision [84].

The exercise is constructed as one mission and each participating team is divided into sub-teams (e.g., network, client-side, web/database, and exploit development) based on the participants specific area of expertise [84].

XS is organised annually by the NATO CCD COE. Note that XS has evolved over the years, and subsequently some design elements have been updated since the study (XS17) in Publication II.

2.1.3 Rangeforce Labs

While the LS and XS focus on team-based activity, the Rangeforce labs used in this research are aimed at individual learning. Rangeforce also facilitates the team-based exercises, however these are out of scope for this thesis. Rangeforce labs are the basis for learning design model in publication VI, for designing learning and assessment in digital forensics (reverse engineering) training in publication III, and for predictive performance indicators for cybersecurity technical skills in publication V and VIII.

The labs have been designed for a wide range of cybersecurity technical topics, including Command Injection, Cookie Security: Secure, Cookie Security: HttpOnly,Cross-Site Request Forgery (CSRF), Defence against CSRF, Insecure Direct Object Reference, Intro Lab, Path Traversal, SQL Injection, Unrestricted File Upload, Reflected Cross-Site Scripting (XSS), Stored XSS and Phishing based on Stored XSS (Publication VI).

A fully automated Cyber Defense Competition platform Intelligent Training Exercise Environment (i-tee) (used also as a basis for Rangeforce[1] platform), is publicly available under MIT license [39]. This architecture allows the creation of new labs and challenges by reusing existing modules such as attacking and assessment scripts and vulnerable targets. The system is designed to start a lab without extra management effort and all game services, routers, networks and scoring bots are allocated on demand (Publication V).

To access the hands-on exercise platform at Rangeforce, a participant needs an HTML5 capable web browser (without any additional plug-ins or VPN). The system uses Virtual Machine (VM) Host platform based on open source tool i-tee [39]. When a lab starts, all VMs, networks, and grading systems are provisioned and personalised for the participant. Also, the automated skill evaluation process is initialised. Each lab may include different VMs (Linux, Windows, BSD, etc.) and different software defined networks. Some VMs are

[1] https://rangeforce.com

accessible for participants (blue systems). Other VMs are dedicated for attack traffic generation (red systems) or for end-user simulation and for network traffic generation. The system architecture ensures network isolation between participants and lab networks. The lab personalisation process creates flags, vulnerabilities, grading for each lab attempt and random IP addresses for attacks and grading. Those IP-s are based on real logs from the servers (fail2ban, sshguard, blacklists). The system provides interactive assistance and guidance for participants using hints, leaked hacker's chat live stream, and media injects using Virtual Teaching Assistant for the learner. Gamification elements such as leaderboards, scoreboards and hackers chat-rooms, are also provided (Publication VI).

By 2019, more than 2 000 learners have used Rangeforce system in more than 15 000 lab sessions and for on-site cybersecurity competitions in 10 countries, in companies and academic training programs (Publication VI).

2.2 Learning and Measuring Learning in CSXs

Learning normally implies a fairly permanent change in a person's behavioural performance [56]. It is a hypothetical construct, i.e., it cannot be directly observed but only inferred from observable behaviour [56]. Learning (behavioural potential) and performance (actual behaviour) are different constructs—thus ultimately the only proof of learning is measuring some kind of performance [56].

Learning can be measured in different ways (e.g., assessment tests/exams, practice and skill checks, self-assessments, etc.), and it takes effort, a structured approach and resources to measure the impact of any training. In the following sections we provide an overview of different schools of thought on learning and learning measurements (both individual and team) and how it links to learning analytics and CSXs.

2.2.1 Individual Learning

While acknowledging different schools of thought on adult learning, it is widely accepted that learning takes place as a result of critical reflection on experiences rather than as a result of formal training in remembering dull theories. In this thesis (and typically also in the design of hands-on exercises), the underlying philosophy relied on is expressed by Kolb, who defines learning as "... the process whereby knowledge is created through the transformation of experience [32]." The experiential learning model presents learning as a two dimensional process: one dimension describes methods of grasping or perceiving information, while the other defines methods of transforming or processing information. The grasping dimension represents two different methods for perceiving (i.e., taking in materials): feeling or thinking, and the processing represents two different methods for transforming materials: doing or watching. While everyone can utilise each learning mode in their learning process, most people favour a particular mode or combination of modes [32]. Hands-on exercises emphasise the "doing" method in the processing dimension.

In many technical fields, hands-on learning ("doing") is very important. This also includes the cybersecurity discipline. Cybersecurity students and personnel are expected to have not only a theoretical understanding of information security concepts but also practical and critical thinking skills to identify security threats, to implement security mechanisms to defend against these and to restore compromised information systems.

The CSXs provide hands-on experimentation that is an effective pedagogy to teach practical skills bas well as higher order thinking skills (HOTS, i.e., skills involving analysis, evaluation and synthesis) as defined within Bloom's Taxonomy [6]. A well-designed, hands-on activity can integrate skills from multiple levels of the taxonomy, thereby en-

hancing both technical and critical thinking skills [102]. Designing learning experiences to teach HOTS is more difficult, but also more valuable and likely to be usable in novel situations (i.e., in situations other than those in which the skill was learned). However, CSXs execution often only focuses on hands-on training (practice) and may omit that combination of theory and practical elements needed to reach higher order thinking. Without clear understanding of a problem and having the essential knowledge of how to address the issue, it is not possible to solve the technical tasks effectively and learning impact is not realised. There are several teaching methods to overcome such shortcoming. For example by providing classroom teaching before the exercise execution (e.g., [53]) or having a virtual teaching assistant incorporated into learning platform (e.g., Publication VI). Therefore, the design of the exercise is crucial to get right, and using digital datasets revealing learning process and experiences can provide evidence-based input to the exercise design process.

In the CSXs context, the participants form their own "rules", that are a result of action, observation and reflection of past learning experiences forming the basis for future learning. Feedback and reflection are thus critical, as from a discovery moment new learning is created that can be applied to different or new situations. As CSXs are often complex and provide large although granular level datasets, it is challenging to provide relevant and sufficient feedback in order to help learners in preparing for unknown future developments. This is where learning analytics has a critical role to play.

2.2.2 Team Learning

Teams are cognitive (dynamical) systems in which cognition (processes or activities such as learning, planning, reasoning, decision making, problem solving, remembering, designing, and assessing situations) that occurs at a team level emerges through interactions [28]. Teams may or may not exhibit intelligent behavior as many instances of team cognition can be readily observed [28].

Definitions of team learning vary considerably across studies. Team learning can be defined as a process in which a team takes action, obtains and reflects upon feedback and makes changes to adapt or improve. According to Senge [105], learning involves collective thinking skills so that groups can reliably develop an ability greater than the sum of individual member talents. Team learning can also be viewed as a dynamic process in which learning steps, environment, individuals in the group, and group behaviors change as the group learns. Some interpretations of group learning, however, confuse levels of analysis by not distinguishing "individual learning in the context of groups" from "group-level learning". If an individual leaves the group and the group cannot access his or her learning, the group has failed to learn—so the other processes like sharing must have happened in a learning context. [119], [88]

In CSXs, teams are usually formed a few weeks before the exercise and dissolve after the exercise execution. For example, in LS the teams can be described as multidisciplinary, as they consist of team and subteam-leaders, IT specialists, legal advisers, forensic analysts, etc. Each team can select their necessary capabilities and members, and wide selection of team members is encouraged (as it has been seen that the teams who were able to assign owners to every system were better at detecting and restoring their systems) [85]. From a cyber operations perspective, the effective collaboration, experience, and functional role-specialization within the teams are important factors that determine a team's success and are important observational predictors of the timely detection and effective mitigation of ongoing cyber attacks [16]. Thus, incident response management requires team-mates to get the right information to the right person at the right time [28].

The objective of team learning is thus “creating a connection between the members", i.e., available knowledge and opinions are shared using clear communications, leading to shared visions and intentions [113]. Van Haar et al. [113] divides team learning needs as both a task-related connection (shared SA, shared mental model of task, shared mental model of team) and a team-related connection (social structure and communication pattern of the team, error management, cooperation (adaptability, flexibility and how to make use of planning), collective orientation. Task mastery and group process are both needed for team learning and “teams learn when they change what they do or how they do it as a group” [113]. Much of this subtle interaction competency is probably not stored in a knowledge repository but is an adaptive response to the interactions of fellow teammates (e.g., who is overloaded and individual work-flow differences) [28]. However, team learning behaviour is expected to be positively related to team performance [48].

To analyse team learning further, it can be broken into sub-processes. Wilson et al. define sharing, storage, and retrieval processes that are intertwined and need to take place for group learning to occur (represented as an equation: GL = Sharing + Storage + Retrieval) [119]. According to Edmonson, team learning behaviours are defined as activities through which team members seek to acquire, share, refine or combine task-relevant knowledge through interaction with one another. Team learning behaviour is viewed as one aspect of a group’s "interaction process" or as an example of a "group action process" [36]. The team learning model by Dechant describes four linked learning processes that explain team learning: framing and re-framing, experimenting, crossing boundaries and integrating perspectives [34].

In the context of CSXs, the incidence response groups can function together for several hours or a few days, but then never meet again. Can we talk about group-level learning for a group that lasts for three hours, disbands, and never meets again? There is not much research on how (often in flux) groups embedded in organisations interact with their external environments, etc. [119]. LS and other cybersecurity exercises’ teams often represent a "flux" and short-term teams that are put together for the purpose of the exercise. However, even if team members are usually in flux, we should not assume that group learning will (practically) never get transferred from the previous team constructs to the current or future teams. Using the memory of a team member who participated in the previous year’s exercise or the internal document (e.g., in-game cyber crisis response/communication manual/plans), the collective knowledge can still be transferred.

2.3 Measuring Learning Outcomes in CSXs

Measuring learning is a complex task, especially without intrusive methods. There are many factors to consider, such as learning impact not being identified; changes being environmental and learning dysfunctional [119]. So far researchers have mostly focusing on a limited set of learning outcomes, mainly learning of simple concrete knowledge. However, cognitive, behavioural, and emotional learning outcomes should also be considered. Developments in education sciences support several alternatives to traditional assessments by leveraging creativity, student involvement, and strategic curriculum development and promoting alternative assessments [54].

An alternative assessment is any assessment practice that focuses on continuous individual student progress and in which the focus is more directly on "performance" [43]. Thus, hands-on CSXs can be thought of as learning and performance tests. In such CSXs the participants are required to perform a complex skill or procedure to demonstrate that they can apply the knowledge and skills they have learned, while there will be evidence left as a digital footprint to evaluate the process and achievement of learning outcomes.

Specifically for CSXs, some general guidance, such as [61, 97], describe how organisers should look at design and performance (training success) measurements in academic literature. Recent research has also attempted to address various evaluation aspects of CSXs. For example, Mäses [80] focuses on evaluating cybersecurity-related competence through simulation exercises proposing a high-level SecTec Window, a conceptual map of cybersecurity-related competencies (categorised as non-technical / technical, not cybersecurity-specific / cybersecurity-specific). In contrast, Ahmad [4] focuses on impacts at an organisational level and investigates how a cyber crisis exercise benefits participants' individual learning and how their experience in the exercises is transferred to their organisation using the four-level Kirkpatrick training post-assessment model.

In recent academic literature, research is emerging on the evaluation of performance in CSXs using the digital data from the learning platforms. For example a novel scoring framework and comparing the participants' scores to a reference (i.e., intended) path [8]. However, the authors acknowledge that the current cyber range infrastructures often lack the ability to monitor the performance of a participant (or team), and only report on exercise completion (or lack thereof) [8]. Therefore, such evaluations for CSXs are in the testing or experimentation phase at the time of writing this thesis. The existing research efforts mainly focus on individual learning aspects, and team learning is often left out of scope for the learning measurements in CSXs. However, some studies also measure team performance and effectiveness, e.g., [47], [53], [68], [58], but mainly use traditional obtrusive methods.

In a wider context, methods to evaluate learning outcomes include meta-analyses, randomised controlled trials, quasi-experimental designs, single case experimental designs (pre- and post test) and non-experimental designs (surveys, correlations and qualitative analysis) [51]. The effect on learning (acquisition of skills or knowledge) can be measured by calculating the difference between pre-test and post-test scores on the questionnaires or cognitive tests and compared to the control group [44].

CSXs are often designed using game-based learning principles. Overall such exercises provide an excellent environment for mixed-method data gathering (i.e., triangulation), including crowd sourcing, panel discussions, surveys and observations, in-game logging and the tracking of hundreds of events and results, including distances, paths, play time and avoidable mistakes, etc. [78]. Questionnaires and complementing these with interviews including probing questions [109] are also used to measure game-based simulations. [27] proposes a model for the evaluation of games for learning that includes motivational variables such as interest and effort, as well as learners' preferences, perceptions and attitudes to games and looking at learner performance. Outcomes not only relate to learning and skill acquisition but also to affective and motivational aspects applicable for measurements in CSXs.

However in learning outcome measurements, not yet explored areas are seamless or "stealth" data-gathering and assessment as well as performance based evaluation [51]. Stealth assessment (i.e,. non-invasive and non-intrusive) could potentially increase the learning efficacy, given that much of the learning remains relatively "implicit" and "subjective" [78]. It should be noted, as in cybersecurity, attack and defence activities are conducted on computers/network—observations of behaviour (sitting quietly behind computer screen while mitigating a significant threat) might not provide sufficient information about learning. Observation should be rather viewed as looking into "digital footprint" by applying non-intrusive measurements (such review of situational reports).

In summary, there are currently no widely accepted and unified methodological evaluation methodologies published and scientifically proven that measure learning impact

or assess cybersecurity skills and/or competencies obtained through CSXs. Such an evaluation model should also incorporate both individual and team learning aspects, as many CSXs are team-based with soft-skills being essential for working in the field of cybersecurity. Considering existing various measurement efforts, using digital datasets can provide the motivation for evidence-based evaluation of learning processes that LA is offering. However, there is an underlying need to ensure that a learning platform design will enable such analytics.

2.4 Learning Analytics

2.4.1 State of Art of Learning Analytics Research

LA has gained an increasing relevance since 2012 [99] and rapidly expanding as a multidisciplinary domain [89]. The research field integrates learning, data sciences and educational technology into a rich socio-technical ecosystem [90]. LA builds upon well established disciplines but contributes further by capturing digital data from students' learning activity and using computational analysis techniques from data science and AI [110].

In past decade, the LA research community has significantly focused in areas like MOOCs and visualisations, performance, assessment [94]. Many new topics (such as natural language processing, multimodal learning analytics, orchestration) are emerging and the field is expanding and following recent technological advancements [94]. However, the field has been criticised for lacking theoretical frameworks which can provide solid common grounds for further development [33]. The LA field has yet to reach its full adoption and utilisation that is related readiness of various stakeholders and institutional readiness [94].

This work adds a dimension of cybersecurity, as LA research needs to be adapted and applied for this domain. The research complements and contributes to LA research maturity and wider adoption by providing theoretical frameworks and practical LA applications in the cybersecurity education.

2.4.2 Learning Analytics Application in CSXs

Note: A comprehensive literature review is presented in publication VII. This section includes a short summary of the key work that frames this thesis.

An overall LA process has been well-established in the existing LA research, such as [107] depicting detailed process steps or as a high-level LA cycle [24], as shown in Figure 2. Regarding CSXs, research is in its early stages and there is a lack of published studies on the complete cycle of LA process and guidance on how to incorporate LA into CSXs.

In cybersecurity education, few examples of LA using an empirical learning data analysis have been completed. Weiss et al. [117] expresses that simply recording the number of correct answers is inferior to in-depth assessment and explore the use of command line history and visualisation. The authors follow the "path" taken by a student in command-line when completing different tasks and levels (for skills level measurement some commands were identified as significant) [117]. Similarly, [67] demonstrates that the assessments of technical skill levels based on indexed similarity (i.e., participants were ranked based on commands usage to achieve objectives) and the classification of actions can be automatically deducted using the clustering of commands. [2], [3] describe techniques for mining the resulting data logs for relevant human performance variables in the exercises. In regards to teamwork and communication, there is some research emerging, such as [7], [68] that have started to explore the use of analytics as evidence for achieving learning in teams. As learning is a complex cognitive process not only technical indicators are relevant but also cognitive indicators, such as [64], [62]. However, the development

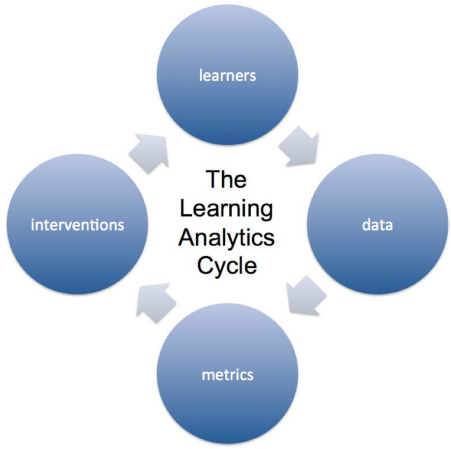

Figure 2 – The Learning Analytics Cycle by [24]

of a combined approach for CSXs with technical and cognitive metrics requires validation and LA methods.

As a distinct research area, LA includes identifying students at risk and predictive analytics, and several other methods for student modelling have been developed (e.g., Knowledge Tracing, Performance Factors Analysis) [45]. Such methods have not been widely researched in cybersecurity (e.g., [18] using log data to predict course grade, [53] predicting team proficiency). Prediction modelling in university admissions is not a new topic, and there are several examples of predictive models [19]). In STEM disciplines, there are several attempts to predict the students' performance, such as [59], [21]. Most of those models assume previous knowledge of past performance or are mainly based on demographic data. However, there is limited work that describes how to assess the applicant's cybersecurity skills using practical tasks in an adequate and scalable manner (i.e., online assessments).

2.5 In Summary: Learning Analytics and Cybersecurity Exercises

This Chapter has provided an overview of relevant learning theories and current measurement efforts for individual and team learning efforts relevant for CSXs and LA. Also an overview of the exercises used as the research subjects in this thesis have been described to provide context and background.

Overall, despite CSXs being seen as a valuable learning and teaching method, there is a lack of research on the complete cycle of LA process and guidance on how to incorporate LA into CSXs using the digital data from the training environments. A few examples of LA usage and empirical learning data analysis are emerging and have been described in this Chapter and in publication VIII. The limited status of implementing LA processes in CSXs has motivated this research, which focuses on overall design and evaluation frameworks and specifically on data collection aspects that are essential for any type of further learning analysis.

3 Implementing LA Approach as an Integral Part of CSX Life-Cycle

Implementation of learning analytics cannot be an after-thought. From the start of an exercise identification and planning stages we need to understand (1) what metrics evidence learning and (2) are they helping the learners to learn or teachers to teach?

This Chapter focuses on RQ1.1., i.e., how to deploy a learning analytics approach into CSXs in order to improve learning processes and to evidence learning outcomes. We propose a high-level Learning Analytics Reference model described in publication VII and discusses how and what data to collect enabling such learning measurements throughout an exercise life-cycle.

3.1 Learning Analytics Reference Model for Cybersecurity Exercises

Applying the learning analytics approach and analysing metrics from digital datasets can provide more detailed and evidence-based input to more comprehensive learning evaluations, such as Kickpatrick or other chosen evaluation model [49]. As those exercises leave an extensive digital footprint of learning processes, it makes them an ideal base to develop the methods within the fields of LA. As a result, using these evidence-based learning traces in learning design can improve the experience for students and specialists. It also helps to investigate the validity of common, yet unsubstantiated claims, such as "everyone feels they had learned important lessons [26]" or "exercises are a very effective way of learning the practical aspects of information security" [97].

Learning analytics should be incorporated into the CSXs' identifying, planning, conducting and evaluating phases (as described by [61]) and seen as an integral part of the exercise design in line with the overall pedagogical approach selected [60]. When starting to implement LA approach into an exercise it is useful to think about the LA process from aspects such as What (Data, Environments, Context), Why (Objectives), Who (Stakeholders) and How (Methods) [20]. The CSX LA reference model shown in Figure 3, builds upon [20] and [61]. The model combines and outlines the key learning analytics considerations to incorporate into the CSXs life-cycle and supporting the model with an extensive overview of existing use cases for a practical implementation. The developers would especially need to consider LA aspects in their design of the cyber ranges, as they lay the technological foundation of instrumenting the exercises enabling LA in the first place.

Asking these learning analytics related questions and finding the answers during an exercise life-cycle will ensure that learning measurements are not simply an after-thought but rather are incorporated from the "identifying phase" (Figure 3). Considering questions, such as "What data can we collect that will help learners to learn?" and collecting only relevant dataset would help with the challenges of storing huge datasets from an exercise and later trying to see what data could be used in providing feedback. When a learning objective of an exercise is improving the incident response process, timestamps that would indicate team communication would be critical data to collect (Publication I). However, when the proficiency of using various forensics tools and command-lines is trained, then capturing bash history or keystrokes is relevant (e.g., [117], [108]). Also, considering how to support learners and instructors by giving feedback is important. Designing automated feedback that takes into account the users' behaviour and predicts their actions and questions now becomes available and can make the learning experience more individualised and effective (e.g., [111], Publication VI).

This reference model for LA in CSXs can be used as a practical guide to the exercise organisers enhancing conceptualisation and the integration of learning analytics into the

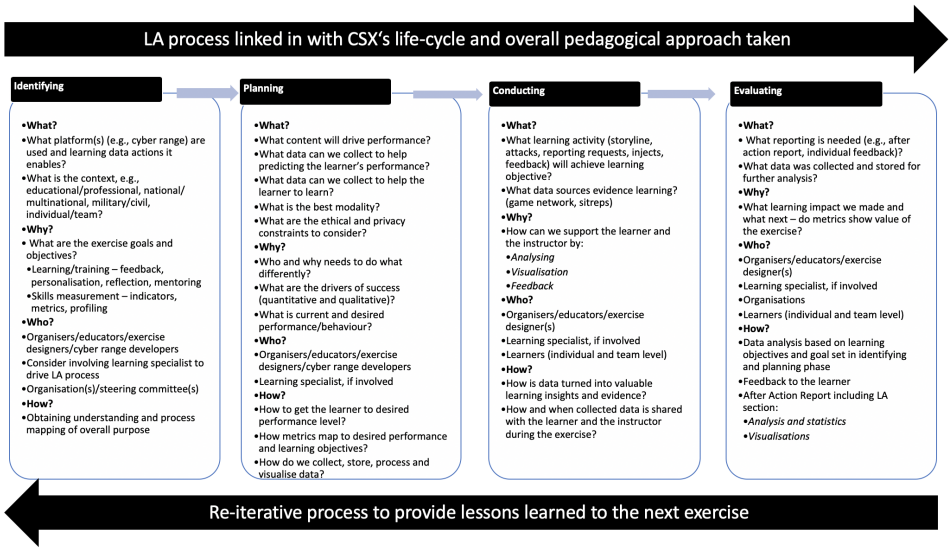

Figure 3 – CSX Learning Analytics Reference Model (Publication VII)

exercise life-cycle as an integral dimension. The considerations presented in the model are supported with an extensive overview of existing uses of learning analytics providing empirical evidence from digital datasets (log files, pcaps) and metrics used in CSXs.

3.2 Combining Metrics from Digital Datasets and Cognitive Indicators

Publication VII summarises learning indicators from digital datasets that have been used in academic research and could be used as a starting point to brainstorm when selecting the metrics to measure how far the training objectives have been achieved. It should be noted that many papers on learning in the CSXs are based on the experience and interpretation of the authors or based on the traditional learner evaluation (e.g., feedback surveys, evaluation forms).

We should focus on measuring what we value or aim to teach as defined in training and learning objectives. The metrics used in CSXs often focus on easily measurable data (e.g., time spent, number of attacks mitigated, etc.) and individual actions. However, the students are "too easily satisfied that a system is secure after identifying only one possible source of security for a system rather than seeking to explore the adversarial space more thoroughly" [104]. Thus, it is important to understand not only whether the students found the correct answer but how they found it [117]. There are some studies that start to look into "how" the learner completes tasks (i.e., use of tools, attempts, submission of wrong answers), such as [2], [3], [67]. However, validation is limited (e.g., 4 participants [67]).

Regarding teamwork and communication, there are some studies emerging, such as [7], [68] that have started to explore the use of analytics as evidence for achieving learning in teams.

As learning is a complex cognitive process, further research should focus on cognitive metrics, such as Knox et al. [64]. From the wider LA field, a similar measure to "cognitive presence" can be applied in the cybersecurity training (e.g., "Active Learning Squared (AL2)" paradigm, which emphasises meta-cognition and uses both active student learning

and machine learning [66], [62]).

Metrics are valuable; however, "being able to report upon a metric does not mean that you should use it, either in the tool, or in reporting its worth [63]". The metrics will depend on the exercise goals, which in turn are guided by different pedagogical principles (e.g., behaviourist, cognitivist or constructivist) [60] and the wider evaluation model chosen [49]. Therefore we need to be mindful of the learner and the learning process. Measurement should move towards the mapping digital traces describing student activity onto interpretable constructs of interest (e.g., Knowledge Components, Q-matrix), which facilitates actionable analytics [70].

3.3 Challenges and Benefits in Implementing LA Approaches in CSXs

The implementation of such approaches is not without challenges. The scientifically valid evidence that learning outcomes have been achieved is difficult to obtain, especially as the exercise design, objectives, technology and learner characteristics vary. These factors make inter-institutional and between exercises comparisons difficult; however sharing the measurement results would enhance comparability and provide further evidence that learning has been achieved and skills obtained across various CSXs. Cybersecurity is a quickly developing discipline and is often chasing moving targets. This also impacts LA, as it needs to keep up with evolving algorithms and learning patterns. In addition, the challenge relates to a large data volume, necessitating efficient and automated data and LA techniques. However, identifying learning patterns is still challenging (i.e., short time of detection, gaps in time-line, etc.) and data is very diverse (e.g., different operating systems, applications, etc.). When combining multiple datasets and different formats, e.g., technical and timing data with self-reported learner data, no detailed descriptions or methods are provided on how data is processed and time-synced. Also, the CSXs are typically events over a short period, but short-term interventions are not particularly effective at affecting behavioural change [52]. Thus, longitudinal studies are needed to evidence learning and behaviour change as result of the exercises and also to separate them from other learning.

In summary, the application of LA and analysis of digital datasets can provide a deeper understanding of learning behaviour and lead to evidence-based improvement. The consideration of LA aspects is also vital for the cyber range developers, as they lay the technological foundation for instrumenting the exercises that enable applying effective LA methods. The LA reference model can assist in implementing LA into the CSXs life-cycle to achieve a more adaptive design and measurement using evidence-based data from the learning environment. In publication VII, we shared an extensive related work overview for the current deployment of LA and analysis of empirical evidence, specifically a summary of metrics (learning indicators) from the digital datasets to assist in implementing the model across all exercise types.

Further work should identify and validate which learning metrics evidence learning process and learning improvement in CSXs. As a first step towards the evidence based measurement, we provided in Chapter 5 a case study using Locked Shields and exploratory research focusing on combining individual and team learning.

4 Identifying and Planning CSXs: Effective Learning Design enabling Data Collection for Learning Analytics

In this Chapter, we discuss the work done and findings related to research question RQ1.2. This investigates which instructional design approaches would enable connecting raw data to high-level competencies and learning experience of the learners in the process. The set-up of an exercise and which data will be collected lays the foundation for the implementation of LA techniques and methods.

The competencies required in the discipline of cybersecurity are varied and extensive as seen in The National Initiative for Cybersecurity Education Cybersecurity Workforce Framework (NICE Framework) [86] or the Cyber Security Body of Knowledge [101]. Thus designing and executing a CSX is a complex task, because it can have many forms, approaches, techniques, objectives, etc. The author has focused on the critical aspects and elements when designing the exercises, in order to be able to later measure their effectiveness. This Chapter explores effective learning design and methods to teach various elements of cybersecurity to a varied and diverse training audience.

Despite the related work addressing many differing issues in CSXs, including skills assessment and evaluation attempts, a practical and scalable model that would provide evidence that high-level competencies can be achieved through analysing the granular data-level of the exercises' raw data is missing. Solving this critical question also helps implement LA approach in CSXs.

Matching such data to competencies models has been proposed in publication VI, while publications III and II provide the different use cases and analysis of effects of such design that allows detailed data collects. Publication III focuses on digital forensics (reverse engineering) and publication II on visualisations and feedback on stealthiness of red team operations. In these use cases we also obtain the qualitative insights from the learners, to evaluate any impact on the learning process.

4.1 Design Approach for Connecting Competencies to Raw Data

In order to automatically select the appropriate learning environment and adapt to the needs of the learner, it is important to have a measurement methodology that is able to accurately capture the capabilities of the user (Publication VI).

However, defining educational outcomes and objectively measuring such outcomes for a CSX is a complex task. The common evaluation approach in the exercises is scoring for completion of various tasks or events. However, tweaking such elaborate scoring systems and story-lines is usually very time consuming and challenging enough to divert attention from the in-depth analysis of the final score.

Our aim is to apply the existing instructional design methods for connecting raw data points to high-level competencies by using an evidence-correlation model. Applying the suggested model for the design and implementation of a cybersecurity exercise would give a structured and automatic feedback of the participants' skills.

In a CSX, different events happen rapidly in a semi-controlled environment. However, the learning experience is neither linear nor predictable, and this makes measuring specific competencies challenging. For example, if the participants have a task to defend a vulnerable web application, they have different correct ways to deal with the attacks: block attacks by implementing intrusion prevention systems, block attacks by using web application firewalls or by fixing the vulnerabilities of the web application (Publication VI). However, there are also different incorrect or insufficient ways to react: removing attacker injected code/content from the system, taking the vulnerable application off-line, break-

ing web application's functionality, etc. (Publication VI).

We describe a model in Figure 4 in which the design of a cybersecurity exercise follows a top-down pattern but with collecting and parsing the relevant raw data laying the foundation.

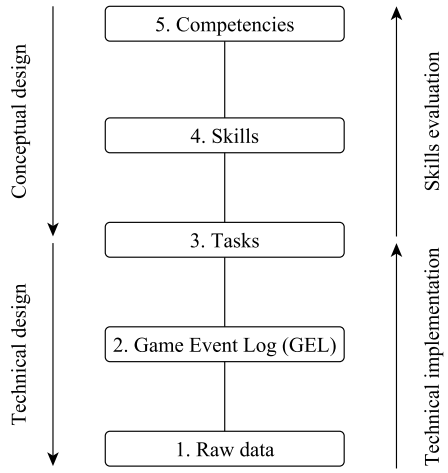

Figure 4 – Suggested Structure for Exercises (Publication VI)

The main idea is to start with the conceptual design. First, targeted competencies that the exercise should teach or assess are defined. Then, those competencies are drilled down to different skills with measurable learning objectives. Based on those learning objectives, specific tasks can be determined that could be measured by evidence—i.e., events happening in the system.

To improve the flexibility and efficiency of the system, Game Event Logs (GEL) are designed to capture all important events. When GEL is designed appropriately, the amount of exercise data needed to analyse decreases significantly. Massive amounts of raw data can be deleted after the exercise while keeping the ability to dynamically change the rules for interpreting the events.

During the CSX, the raw data is used to generate GEL, which is interpreted to evaluate whether a specific task is completed. Commonly, the completion of tasks is used as an input for score calculation, but we suggest to do more than that. The completion of tasks indicates the proficiency level of different skills. Skills, in turn, are gathered into meaningful sets to form competencies.

The model consists of 5 layers. The bottom layer represents the raw data and the highest layer specific competencies that are targeted by the exercise. It is important to note that while the technical data flow happens from the first to fifth layer, the logical design flow should start from the top layer. The layers themselves are connected to each other but at the same time are independent, allowing the use of different formats or processing systems.

The cognitive learning layers based on the revised Bloom's taxonomy [6] are the following: (1) Remembering; (2) Understanding; (3) Applying; (4) Analysing; (5) Evaluating and (6) Creating. The adjustment is made in order to incorporate the attack and defence

aspects of effective cyber defence. In order to defend, learners need to understand and apply the attack technique themselves before being able to creatively avoid such vulnerabilities in the system. The difference between applying and mastering the defence is transferability, i.e., we can assess mastering if the learner is able to defend against this type of attacks in different operating systems, using different tools, etc. Mastering the defence (level 5) in our model usually means that the same skill is measured and mapped over different labs to ensure that the learner is able to transfer and apply the skill using different technologies.

In order to validate the learning impact, the assessment should be performed using various alternative methods and data sources. For Rangeforce labs design evaluation, a thorough hands-on assessment was conducted with 27 participants. This consisted of extensive pre-assessment labs (a set of 13 different labs including XSS, command injection, cookie security, etc.) and post-assessment test (same hands-on test as for pre-assessment). The results show the average completion rate of the pre- and post-assessment labs. We can see that for the "study" group (i.e., those who actively participated in the exercise), the improvement in skills was significant (from 4% to 68%). However, the skills improvement was measured using both pre- and post-exercise assessment by a relatively small group and should be extended as part of future work. After the exercise, additional free-form feedback was collected from the participants and their supervisors, to analyse whether the participants had acquired relevant competencies for their job. Although the feedback gave anecdotal evidence of useful skill improvement, quantitative surveys and structured pre and post assessments could be used for systematic and empirical evaluation.

Further work should continue on defining the rules (i.e., defining GEL, etc.) that can be universally applied in the different exercises and developing the system enabling the exercise' organisers to dive into the learning data more easily. In the future, such data could be used for modelling predictive behaviours, e.g., considering use of learning hints in skills' evaluation [22] and calculating confidence correlation to the skills.

In summary, our structured approach helps to obtain more meaningful learning data from the logged event and to measure relevant competencies. Our approach supports a paradigm shift towards the cybersecurity exercises that by design allows systematic and evidence-based competencies and skills measurement.

4.2 Case study for Teaching and Assessing Malware Reverse Engineering Skills (in Digital Forensics)

Following the overall exercise design concept as described in the previous section and depicted in Figure 3, we discuss a case study of a malware lab design, while putting this approach into practice. Using a fully automated cyber defence competition platform Intelligent Training Exercise Environment (i-tee) [39], our research aim was to design and test a scalable hands-on automated malware reverse engineering lab for university students using open source tools (or free-ware tools) (Publication III).

Mastering reversing engineering could be compared to art. "Reversing equals art" thinking is not new and has stated before in computing education, e.g., Bader has used the same expression with respect to parallel programming [46]. Modern malware can include multiple obstacles [13] for evading detection, which complicates the malware analysis. In the phase when students get graded, they should be familiar with some of them beyond just obfuscation [112].

Studies have found that the human brain constructs models of the world on the basis of past experience, which are subsequently confirmed or denied by experiential input [55].

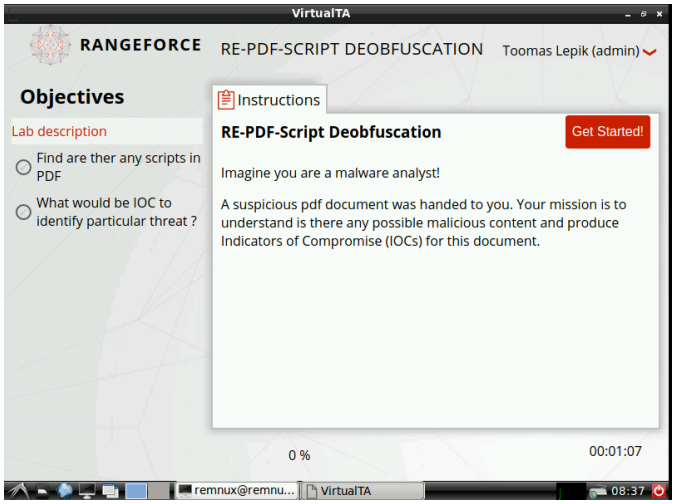

Figure 5 – Malware Lab Instructions. A virtual teaching assistant is guiding the student through the tasks in a step-by-step fashion (Publication III)

When we apply this predictive mechanism of the brain to the cognitive task of problem solving, we see that our brain produces possible solutions for a given problem by predicting solution models based on past experiences [55]. Therefore, it can be argued that a solution to a specific task might be similar, due to students' participation in the same malware course, i.e., similar learning generates inherently predictive solutions to a particular set of problems. However, we can use such predictive patterns, in case we need to understand students ability to use particular tool-set. These types of tool-sets enable the use of different measurement techniques to understand whether the student tackled the multifaceted problem or if the student used an allowed (or not allowed) short cut to a solution.

Our approach was to leave the playing field open and to let the student choose the tools and methods to tackle the problems, using their own tools and programs, which are only limited by the operating system. However, that implies that the problem of assessing skills automatically becomes more complicated, as measuring will depend on reverse engineering the steps that students chose to perform. For example, we do not simply measure the fact of installation of a specific tool by the student, rather whether they understand the reverse engineering process. The concept was tested out on the malicious JavaScript included in the PDFs as part of the course assessment.

The course assessment was designed as a fictional situation where the students acted as malware analysts and their task is to analyse and reverse engineer the malware included in the PDF document, as shown in Figure 5. The malware for each student was different in order to ensure the students were not just copying solutions but actually running the commands and using the tools therefore leaving the evidence on the learning platform.

Specific to the malware lab assessment, some interesting analytics were obtained and evaluated against learning objectives as well. For example, the time taken for resolving first iteration took about 1 hour 54 minutes on average. This was in line with course homework with similar content and complexity, where students reported time spent, vary-

ing from 30 minutes to 8 hours. For assessing the command line usage Snoopy[2] was used. However, due to a deployment error, logs from the student machines were not available and command line usage analysis could only be done manually (using forensic tools). We randomly chose some machines for further analysis and noted that the students used command line tools native to REMUX and discussed in class, for example: pdfid.py6, Peedf.py7. For JavaScript analysis, the students mostly opted for using online de-obfuscators. This was in line with the expectations and tools covered during the course.

To achieve and assess learning outcomes, the educators use frameworks such as Bloom's or SOLO taxonomy. Whatever learning framework is used, it means that higher level of cognitive levels or complexity levels (HOTS) are reached by a student to ensure that they have mastered the art of malware reverse engineering. As teaching malware reverse engineering can be regarded as an art and malware is commonly well-protected to resist analysis, higher cognitive skills are required, which should be evaluated as part of a course assessment. However, with CTFs the common criticism is that they only reach the lower levels of cognition (e.g., [79]). In malware reverse engineering a learner needs to make tool-choices and recreate from different elements of information. This demonstrates comprehension of applying knowledge, analysis and evaluation in the process of dissembling the code, creative thinking and higher-level concepts. Considering Bloom's taxonomy HOTS need to be achieved, and therefore the lab design provides also the reflection questions, to ensure that a student has understood and not only completed learned (memorised) technical tasks without understanding.

Further assessment of the learning impact of using such labs as part of the learning was performed for ten university students and unstructured, qualitative feedback was received. The main takeaway was that using the similar labs in the learning process would increase learning impact when used as part of the course, as when only used in assessment the unfamiliarity with tools causes a negative effect on being able to demonstrate their skills. Overall, the feedback was encouraging, as the students provided positive comments. However, there is some further work required regarding the inclusion of automated labs in course design (i.e., using it already for homework) and technical improvements (e.g., widen the malware types to be analysed—not only malware in PDFs).

Future work should extend the scenario and the lab, to address the elements of download server, compiled executable analysis, etc. The similar lab structure can be easily converted to address a different scenario's learning objective in the malware reverse engineering and thus cover the wide array of malware elements and types. Another future enhancement is incorporating a task of writing up the yara rules[3]. These are descriptions (i.e., rules) of malware families based on textual or binary patterns consisting of a set of strings and a boolean expression which determine its logic. This would help to measure and ensure that Bloom's higher cognitive skills have been achieved—i.e., the student is able to create something new from different elements of information (to create).

This case study illustrates a design for learning and a learning environment based on open source tools to effectively deliver hands-on automated virtual labs to teach malware reverse engineering with partly automated assessment. It shows the importance of connecting raw data to competencies and even further to learning theories in order to ensure the learning objectives are met and the appropriate learning experiences are created. With the right mindset and tool-set the teachers can significantly increase the teaching quality and help students achieve higher levels of cognition and potentially reduce the amount of students who fail to reach the required skill set.

[2]https://github.com/a2o/snoopy

[3]http://virustotal.github.io/yara/

4.3 Case Study for Giving Feedback on Stealthiness in Red Team Operations

Another case study on the concept of connecting low-level raw data to competence while using this data to help participants to learn used a Crossed Swords exercise for practical implementation (Publication II).

XS is an exercise directed at training Red Team members for responsive cyber defence. In the past, feedback was too slow and participants did not understand the visibility of their actions. Therefore, the data analysis needed to be automated as much as possible, in order to improve the feedback (e.g., timely, accurate, detailed) to the learners. To achieve this objective, a Frankenstack tool was developed for XS17 using the open-source tools but added event correlation, a novel query automation tool and a newly developed visualisation solution described in publication II.

Figure 6 depicts the role of Frankenstack in the XS exercise.

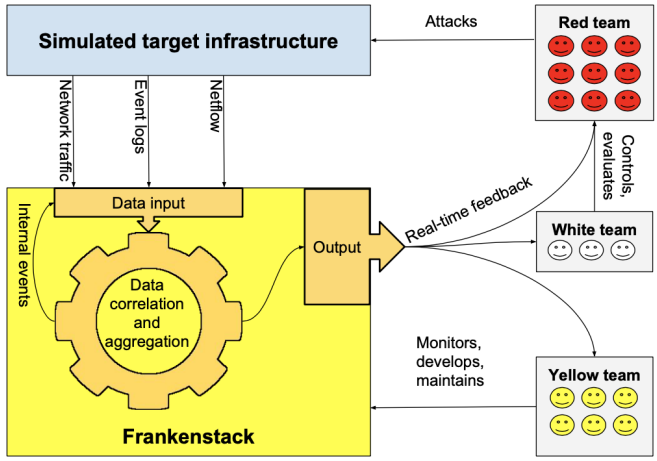

Figure 6 – Frankenstack in XS (Publication II)

An Event Visualization Environment (EVE) was developed for visualizing correlated attacks in relation to the gamenet infrastructure. Probes sent information to EVE listener in JSON format and real-time visualization was achieved using WebSocket technology (Publication II).

There were four large screens in the training room directed towards the RT, displaying Alerta, Grafana, Scirius and Suricata. Alerta served as the primary dashboard to display alerts to the RT and the participants could write their own filtering rules to view information relevant to their current campaign. A fifth screen displaying EVE was only visible to YT and WT members. EVE was shown in replay mode to RT participants after the exercise concluded. This compressed replay was very effective in presenting the most prevalent problems, such as periodic beaconing during otherwise silent night periods and verbosity of particular network sub-team attacks.

From a learning perspective, using the gamenet map makes EVE a very intuitive tool for enabling participants and observers alike to comprehend CSX events on a high level. In order to learn, RT has to receive timely and efficient feedback regarding the detected attacks on the target systems. This feedback is critical to raise the level of stealthiness, identify the gaps of RT coordination and analyse the tools and tactics used for computer

network operations. The effectiveness of our framework was assessed during the main execution of XS17.

Tools and infrastructure are essential for learning; however, human factors, such as how participants perceive and use the tools, have significant impact. One essential part of the assessment was to observe the behaviour of the RT members and their interaction with Frankenstack during the exercise in order to gain further insights into their progress and learning experience.

To estimate their reaction to Frankenstack and their overall learning progress, we carried out qualitative interviews with RT participants. Furthermore, we conducted a quantitative survey consisting of multiple-choice or ranking-style questions with the ability to provide additional comments for each question.

We received 14 survey responses out of 27 participants (52%). As described above there were four large screens in the training room directed towards the RT. The survey revealed that learners did access the monitoring framework on their local computers when attempting new attack vectors. Thus, tools served their intended usage, as 38% participants reported that they checked the screens every 60 minutes or less and another 38% checked the screens every 30 to 50 minutes. Regarding learning impact, 79% agreed that the situational awareness given during exercise was useful for their learning process, and 77% of the respondents considering the speed of feedback to be at the correct level with 57% agreeing that alerts were accurate and sufficient for their learning process. However, several respondents revealed being too focused on achieving their primary objectives, and thus unable to properly switch between their tools and feedback screens.

In relation to visibility, 45% of the participants agreed that they had learned a lot about how their actions can be detected (i.e., it is useful to see simultaneously what attack method could be detected and how), and 30% were more careful with their attacks and thus tried to be stealthier than they normally would have been. However, there were some unintended side-effects: revealing too little or too much to the training audience.

Furthermore, some comments revealed a loss of emphasis on stealth due to exercise time constraints, i.e., RT members knowingly used more verbose techniques closer to the objective deadline.

One of the key training aspects is working as a team to achieve goals. However, feedback concerning the impact of situational awareness tools on team communication and cooperation was mixed: 50% perceived positive impact, whereas 21% were negative and the remainder were neutral. Several respondents acknowledged less need for verbal communication, as they could see relevant information on the screens. Unfortunately, not all RT members were able to interpret and perceive this information correctly. This, combined with the reduced need for communication, meant that not all participants progressed as a team.

Receiving guidance is a critical success factor for learning, especially in a team setting, i.e., when they did not know how to proceed, their team members or sub-team leaders provided guidance. 64% agreeing with the statement on sufficient guidance is a rather disappointing result and could clearly be increased with improved learning design. Some respondents admitted that they did not know how other teams were progressing and wasted time on targets that were not vulnerable. This caused significant frustration and stress, especially when combined with the compressed time-frame of an exercise.

Given the amount of work that goes into preparing such exercises, the level of learning potential needs to be maximised. Our analysis suggests that small learning design changes may have significant impact, e.g., assessment of familiarity of monitoring tools before using those in the exercise, assigning monitoring task to specific team member (for a certain

time period) or other division of tasks between more experienced and novice team members, regular team time-outs for reflection, allowing flexibility in the paths that the RT can take to solve the objectives (i.e., participants should avoid spending too much time on wrong targets), etc.

In summary, as a response to the core challenges such as providing timely and accurate feedback and ensuring participant education without compromising the game scenario, Frankenstack feedback regarding learning impact was mainly positive. However, there are critical questions to answer when designing the RT exercises, such as what is the right balance of information to provide to the RT; does the behaviour change due to monitoring or visual information available (i.e., learners unconsciously limit themselves by not trying out more risky strategies, etc.)? Also, some further learning design changes, and not necessarily only limited to SA, can maximise the return on significant investment into preparing such RT exercises. This emphasises the requirement that methods and metrics for assessing technical RT campaigns have to be incorporated into the game scenario in the initial stages of exercise planning.

5 Evaluating Cybersecurity Exercises for Effectiveness and Efficiency with Less Obtrusive Methods

Learning is such a complex and intractable process that its study and measurement is difficult and contentious. However, methodological measurement is required to determine whether an exercise design was appropriate and effective and whether planned learning outcomes were achieved. This chapter presents how to evaluate effectiveness and efficiency of both individual and team learning aspects with less obtrusive methods in CSXs (RQ2).

CSXs in the current form are often not sufficiently instrumented for learning measurement and existing measurements focusing on scoring do not use learning related metrics. An overall CSX learning measurement approach should bring together pre-exercise, execution and post-exercise phases and individual/team/organisational aspects. The measurement should include a mixture of quantitative and qualitative methods. In this Chapter we discuss a novel method: the 5-timestamp methodology, which focuses on unobtrusive data collection and comparable data analysis linked to learning objectives presented in publication I.

5.1 5-timestamps Model for CSXs Learning Measurement

Learning in CSXs is affected by many variables; however, the basic measurements such as timing and accuracy metrics are still key elements that provide comparable trends in the learning process and benchmarking for the teams. For example, the measurements have shown that when teams took 20 or more minutes to identify an inject's NIST categorisation, they were more accurate [53]. That means an overly time-constrained game-rule may prove to be an unrealistic expectation, which will not contribute to learning. Instead it forces teams to learn, share and store wrong behaviours and later retrieve learned but wrong behavioural models in real-life situations [119]. Such metrics support the development of appropriate exercise learning design.

Furthermore, measuring learning effectiveness and collecting data in order to provide feedback can be combined. The learning potential is not fully realised if the BTs do not know what their weaknesses are and how they progressed in the exercise. Scoring might give some indication of how teams compare; however, without knowing a baseline or standard in more detail, the overall score is worthless from a learning viewpoint. For example, scoring may not take into account how much resistance the BTs put up and how efficient they were in responding.

A novel and simple methodology, called the "5-timestamp methodology", aims to accommodate both effective feedback (including benchmarking) and learning measurement, see Figure 7 for defending (blue) teams. It should be emphasised that this methodology is only a part of the overall learning measurement (including traditional methods, such as surveys, interviews, etc.) and is specifically designed for a team-based cSX, where the main objectives are to train incident management response and team communications. The methodology focuses on the collection of timestamps at specific points during a cyber incident and time interval analysis to assess team performance and argues that changes in performance over time can be used to evidence learning. The timestamps can either be collected non-intrusively from raw network traces (such as pcaps, logs) or using traditional methods, such as interviews, observations and surveys from both the RT and BT. The analysis of time intervals between the proposed five timestamps enables the measurement of technical skills but also soft skills (including leadership, team communications, decision making). The methodology analyses data at a cyber incident/attack vector/target machine

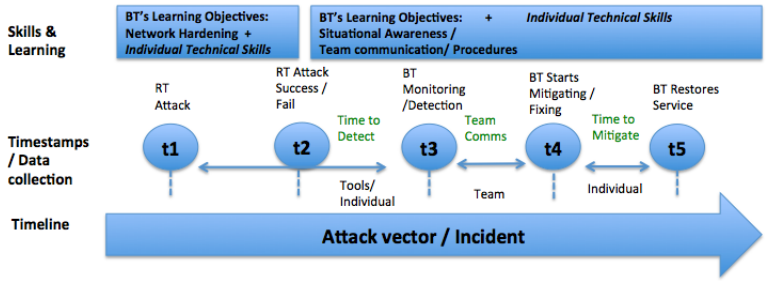

Figure 7 – 5-timestamp Non-intrusive Methodology (Publication I)

level but provides metrics for different learning objectives (Figure 7). For example, the assessment of whether the BTs are effective and achieve incident handling related learning objectives needs basic timing and accuracy metrics. These include how long it takes to respond to an attack, how long teams take to respond to a significant threat vs. minor issues, and the correlation between the teams' detection time and quality of reporting. Further analysis can be carried out to investigate if the most effective strategy from a qualitative aspect was applied by the BTs, but having timing and accuracy metrics will provide input and focus to aid such qualitative analysis and feedback.

The analysis breaks an incident into phases to demonstrate strengths and bottlenecks in individual and team skills in each phase and provides the basis for effective feedback. The model follows the incident time-line and information that can be collected non - intrusively (Table 3) from the game-net or management network[4]. Even when t_1 and t_2 are intrusive for the RTs, data collection is non-intrusive for the BTs. For cross-checking, a sample using intrusive methods should be selected.

Timestamp	Description	Non-intrusive Data	Intrusive (optional)
t_1	RT starts attacking	RT activity reporting	N/A
t_2	RT compromises	RT activity reporting and scoring data	N/A
t_3	BT detects	Possibly by access pattern	BT observation or self-reporting via inject
t_4	BT mitigates	management network (showing traffic activity)	BT observation or self-reporting via inject
t_5	BT restores	scoring, management network (end of session)	N/A

Table 3 – Data Sources for 5-timestamps (Publication I)

In order to fully understand this methodology, it should be noted that there are typically several target machines in a game-network that can be attacked repeatedly using the same attack methods. However, one of the advantages of a live-fire RT/BT exercise is defending against a "thinking" adversary, which implies that the same target can be attacked using different methods.

[4]The exercise runs on separate virtualised machines which are accessed remotely over the VPN [83]. The BTs can reside in their home location and connect via a dedicated management network to the game-environment. The game-net resides typically on a different interface and is where the attacks are happening.

Our experience shows that traditional methods, such as self-reporting, fail in high-speed and complex exercises. The suggested method enhances the feedback loop, allows the identification of learning design flaws and provides evidence of learning value for CSXs and should be incorporated as an integral part of the overall learning measurement framework. Future work should continue with performing the data analysis of an exercise to compile and validate learning metrics and benchmarks for learning patterns. Identification and analysis of the data trends will provide a solid baseline and demonstrate learning improvement achieved in CSXs, both at the individual and team level. This will complement often anecdotal and positive feedback obtained via traditional methods (surveys, interviews) that participants have actually learned. As we demonstrated, incorporating non-intrusive, social and behavioural research methods into the cybersecurity field can give new insights and possibilities in effective training for cyber defence teams in the future.

6 Using Cybersecurity Exercises as Method for Predictive Technical Skills Assessment

This Chapter examines whether CSX's LA measures can be standardised as performance indicators and are suitable as performance predictors for cybersecurity technical skills (RQ3).

The predictive analytics focus on the identification and extraction of specific predictive metrics and correlations that can be included in wider prediction models (together with other validated metrics). Our hypothesis is that completion (and how completed) of hands-on technical exercises can be used as a predictive component in the admission process to identify students who are motivated and likely to succeed in a cybersecurity curriculum. Such exercises leave a digital footprint, which enables the application of LA methods (including predictive modelling) to support learning success in the cybersecurity discipline. Thee dataset used was TalTech admissions using online, hands-on technical exercises as part of the admission process to the university's MSc cybersecurity program over the three years. The full overview of the admission process is provided in Publication V, and evaluation and analysis is presented in Publication VIII.

We specifically addressed the following research questions in the context of our university admission process to the cybersecurity program:

- Can admission procedures that include technical online labs on selected cybersecurity topics predict the students' performance in the technical subjects at the university curriculum? Is it a more accurate predictor than the admission interview component?
- Do more comprehensive and complex cybersecurity technical assessment used at the beginning of the course predict student performance? Is this assessment appropriate as the basis to assign students to courses with different difficulty levels?
- Is more comprehensive assessment necessary at the beginning of the course, or can the results of selected technical labs used during the admission procedure also predict the students' performance?

6.1 Overview of University Admission Process and Exercises Component

In response to the existing and predicted skills gap in the cybersecurity labour market, educational institutions are establishing courses dedicated to producing the cybersecurity specialists of the future. Admissions boards need to select from many applicants with different backgrounds, with the aim of identifying those with the greatest probability of successfully completing the program and identifying those with the appropriate skills to fill the skills gap. Thus, a scalable and predictive admission process that can also be conducted remotely is needed.

In the admission process[5] to the international Cybersecurity Masters program (MSc) at our university, we have used remote technical labs in addition to traditional admission procedures (Publication V). The admission process and how the technical labs fit in are described in Figure 8.

An applicant should exhibit potential for critical thinking and problem solving skills, which we believe can be measured when the candidate is placed into the simulated (gamified) learning environment and actually performing some technical tasks that require putting those skills into practice.

[5] https://www.taltech.ee/en/cyber-msc#p1675

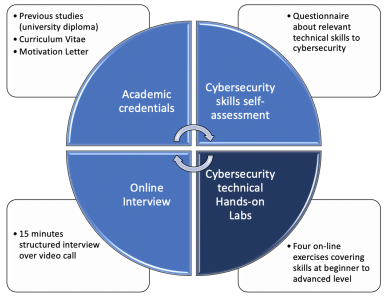

Figure 8 – Elements of Admission Process (Publication V)

The virtual hands-on exercises are based on the i-tee platform [39] that enables on-demand access to a cloud-based virtual environment using a modern (HTML5 capable) web browser. This technical set-up lowers the processing requirements of the applicants' computers. To access the hands-on lab, an applicant needs an HTML5 capable web browser (without any additional plug-ins or VPN). The main requirement is an Internet connection of at least 3Mbps. The system provides interactive assistance and guidance using Virtual Teaching Assistant for the learner.

As measuring all possible skills is not feasible, the exercises represent a mix of selected technical topics. The set of virtual labs is the following:

- Introduction lab—essential command line skills (Git, apt-get, Apache server). Estimated completion time 25 minutes;
- HTTPS Security—basic level skills connected to command line, public key infrastructure, and server administration basics; estimated completion time 45 minutes;
- SQL injection—intermediate level skills connected to attacking SQL databases (SQL, SQL injection); estimated completion time 90 minutes; and
- Botnet—advanced level skills connected to network scanning skills, text parsing (programming skills are beneficial) and SQL injection skills; estimated completion time 45 minutes.

The choice of these different exercises is based on typical attack vectors that the applicants are likely to encounter in their future cybersecurity jobs and require different skill levels (from essential to advanced). This combination of exercises is not only used to determine skill levels but also to cover a variety of different skills. Each lab has a pre-determined skill level, from a basic to advanced.

6.2 Evaluation and Analysis of Using Labs to Assess and Predict Student Performance

The aim of the analysis was to demonstrate that admissions' technical assessment component (hands-on remote labs) (1) is a scalable method, (2) predicts technological skills in cybersecurity as a key component of this international master course and (3) is empirically tested as predictive validity of admission procedure.

We also describe the data and method in detail, to replicate the research and the Cyber Security Technologies (CST) course design that may contribute to the contents of a cybersecurity curriculum and spark interesting in incorporating this type of assessment in other STEM programs.

Data was collected during the admission procedure and the mandatory CST course for 60 first-year MSc students. Specifically, we used "WASE Assessment", a complex virtual lab where an individual has to investigate a website and try to take it back from the hackers. This mandatory assessment is given in the first lecture and used for assigning students to CST1 (introductory) or CST2 (advanced).

We conducted correlation and regression analysis applying the admissions technical labs as a variable to predict the students' performance at the later technical course. The model only uses information from the admission process and assessments at the start of studies, as limited demographic or historic study behavioural data was available about the students. We described the strength and direction of the linear relationship between the variables.

6.2.1 Admission Technical Labs and Later Success in Study Course

We used Pearson's one-sided correlation test method. The variables analysed were course performance, admission interview rank, admission rank and novel admission technical assessment rank. We identified a positive correlation between the input and all variables, while the strongest positive correlation (0.492) is between Course performance and Admission Rank (combining interview and technical lab results). We used admission interviews and technical assessment results as independent variables with course performance as the dependent variable in a regression model (using stepwise regression). In Model 0, the components used are course performance as the dependent variable and admission interview rank as covariate. Model 1 includes admission interview rank and admission technical rank as covariates and course performance as the dependent variable. The interview was a statistically significant predictor ($p<.01$) with an explained variance of $R2=12.6\%$ being a "moderate to large" strength of this effect prediction in social science contexts [25]. Thus, an interview on in its own is already a valid predictor for student performance score, but when adding the virtual lab the R2 almost doubles from 12.6 to 21.5 (considered a "very large" predictor in research standards [25]). Thus, interview and labs are complementary methods that both predict in different aspects the student performance results and are relatively highly correlated (i.e., have common factors). E.g., attitude, eagerness and interest are checked in interview, but are also relevant for technical performance.

6.2.2 Using Complex Technical Assessment (WASE) at Course Start and Later Success

We used Kendall's Tau B testing method and can see an $R=0.502$ positive correlation between WASE assessment and course performance. This indicates that WASE assessment is a valid predictor of the students' performance and the students have been assigned to the correct CST course (introductory vs. advanced).

6.2.3 Using Selected Admission Labs or Complex Comprehensive Lab for Predicting Later Success

We used Pearson's one-sided correlation testing method and a 0.329 positive correlation between the input and admission interview rank was evident. However, the correlation between course performance and WASE assessment rank was stronger with the correlation coefficient 0.548. When using stepwise regression to further explore the relationships, we defined "Admission technical rank" and "WASE assessment rank" as predictors and "Course performance without WASE assessment" as a criterion. In the linear regression model, the admission technical results significantly predict the student performance and explain 8.6% of the student performance variance ($p=.019$). This means that the in-

terview in its own is already a valid predictor for assigning students into the correct CST course. However, when adding the WASE assessment the explained variance R2 is 3 times higher with 27.6%, indicating that the two variables are complementary predictors that both contribute different aspects to the prediction of student performance. Overall, comparing the predictory power of both assessment tools, WASE assessment has a stronger correlation with student performance and thus is a more accurate method to use for skills assessment.

In summary, our main contribution is evaluating the use of remote, technical hands-on labs as a novel part of the university graduate level admissions for cybersecurity programs in predicting students' later success in technical studies. Such an approach would be scalable to evaluate the applicants' technical skills but still incorporates human evaluation to enable a balanced approach for the ethical and evidence-based decision making and assessment. While we acknowledge that this analysis is an initial attempt with a relatively small sample size, it shows some promise, based on regression analysis. As further work, a longitudinal study with a larger sample size should be conducted to evaluate such technical labs' completion data with general intelligence, other cognitive skills, and domain-specific knowledge and the comparison with other prediction modelling methods would be beneficial. While the remote labs used in this research are specific to the technical skills for a cybersecurity program, incorporating this type of assessment may also spark interest in other STEM programs or in the corporate sector.

7 LA and CSXs in Wider Context of Cybersecurity Competency Frameworks

In order to secure cyberspace, we need to educate both users and IT specialists about the dangers, to ensure they have sufficient “cyber hygiene" levels (Publication IV). Cybersecurity training focuses on different levels of skills needed, based on the target audience. While “cyber hygiene” trainings form the basis of understanding, the complexity of the CSXs gradually increases to team-based, highly skilled cyber specialist training events. Exercises vary by scale of complexity, audience types (general users, cybersecurity specialists, individual and team-based exercises) and also complexity and research efforts in LA increase accordingly, see Figure 9.

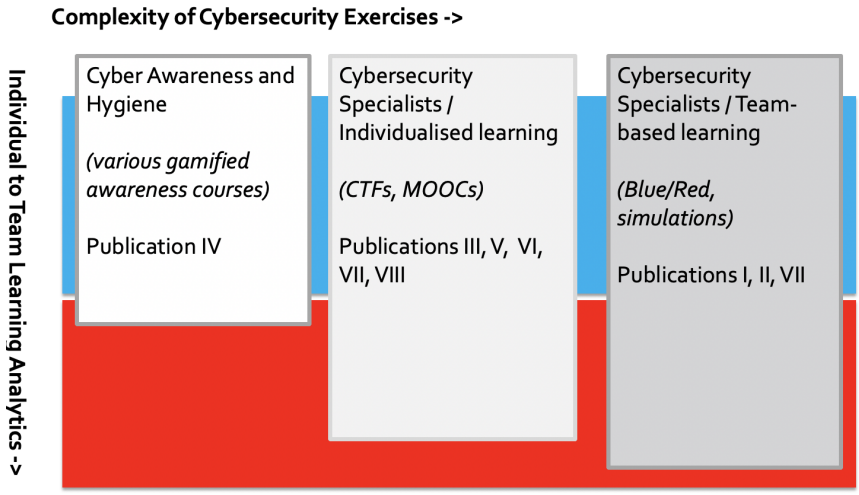

Figure 9 – Research Contribution Mapped to LA and CSX Types

CSXs are a part of wider cybersecurity education and the high-level competency frameworks or curricula act as guidance in developing the exercises and training. One of the well-known frameworks in corporate settings is NIST NICE [87], a workforce-based framework of seven job categories, 33 speciality areas and 52 work roles. But there are many others, including more academic guidelines, such as CSEC2017, developed by a joint task force of ACM and IEEE specialists [12] and CYBOK [101]. A reference with a military focus [81], created by NATO that could be used for defining skill sets of IT specialists in 4 themes—Theme 1: Cyberspace and the Fundamentals of Cybersecurity; Theme 2: Risk Vectors; Theme 3: International Cybersecurity Organizations, Policies and Standards; and Theme 4: Cybersecurity Management in the National Context.

These competency frameworks are inputs to layers 5 (Competencies) and 4 (Skills) in the proposed exercise design model, detailed in Chapter 4. Consequently, the skills and competencies defined in these competency frameworks assume that there is a link to the tasks, game event logs and raw data. Applying learning analytics methods and techniques in the CSXs provides evidence and conclusions to be drawn on skill evaluation in accordance with the competency models used in the industry.

8 Privacy and Ethics

As CSXs are a significant part of curricula and professional training paths, the design and implementation of these exercises and underlying technology (i.e., cyber ranges) will shape the attitudes and ethical limits of those trained. Are the learners allowed to play in the “black box" without any limits for the greater public good? Or is their every move monitored to avoid any wrong-doing? What is the right ethical and privacy balance when designing CSXs and ranges? How does it reflect on implementation of learning analytics in the cyber ranges?

These privacy and ethics-related questions are frequently considered and are essential, as the application of learning analytics aims to improve the learners' behaviours by analysis, developing (automated) algorithms and recommendations. The privacy and ethical issues in the context of LA are tightly intertwined with other aspects, such as trust, accountability and transparency [95]. Incorporating LA, i.e., using the collected information to understand and improve the quality of a learning experience throughout the CSX life-cycle, provides great opportunities to help students learn better. However, this adds another layer of significant considerations of ethical and privacy aspects in relation to the learning processes in the CSXs.

Although understanding ethical dilemmas is imperative, cybersecurity is still an underdeveloped topic in technology ethics [71]. However, all cybersecurity problems have ethical consequences, such as economic damage due to loss of data or physical harm due to critical infrastructure systems breach [23]. Thus many ethical issues are apparent in terms, such as “ethical hacking", the dilemma of holding back “zero day" exploits, weighting data access and data privacy in sensitive health data or value conflicts in law enforcement raised by encryption algorithms [71]. Many of these dilemmas relate to a dual-use issues, when items, knowledge and technology can have both beneficial and harmful applications [103].

Most of the academic work on CSXs covers approaches and recommendations on how to teach ethical behaviour to students, e.g., [11], [50], [14]. When analysing learning analytics in CSXs, the current academic literature often lacks discussion on the privacy and ethical aspects (Publication VII).

The values of exercise designers and range developers are therefore instrumental and will shape both ethical design and learning processes in these exercises. The ethical dilemmas are faced throughout the CSXs life-cycle, in the identifying, planning, conducting and evaluating phases [61]. Thus further research is needed on this topic, specifically on:

- What are the principal ethical and privacy values considered and supported by the cybersecurity exercise organisers and cyber range developers?
- How might these values impact the design and implementation of cybersecurity exercises and cyber ranges?
- How might these values impact the assessment of learning outcomes (incl. learning analytics) of cybersecurity exercises and cyber ranges?

An exercise and cyber range development and execution should follow a value-sensitive design process. The process should allow consideration of ethical and privacy values on the same level as technical and functional requirements [35]. Further studies are needed to obtain deeper insights into the critical aspects of privacy and ethics.

8.1 Ethics, Privacy Data and Data Security as part of this work

Within this work, ethical considerations such as fostering trust, transparency, learner's control over data, right of access and accountability [95] have been followed. The participation was voluntary and consented, as online or digital interactions produce a data trail of a person's activity, and privacy and data security aspects are important. The data is pseudonymised with unique identifiers to ensure the privacy of individuals and teams. The data was stored on the university's server, with access to the research material restricted to only some of the authors and selected university personnel directly associated with this research.

9 Limitations and Future Directions

This thesis addresses learning analytics in the context of the cybersecurity discipline, more specifically the practical application in CSXs. As we focus on the contemporary and evolving research problem, there was little, if any, prior research on the LA topic in CSXs. The lack of research and an empirical or theoretical base to build upon required a rather broad approach, with an initial mapping of the research field. As part of this research, the author conducted an extensive literature review to map the current status, see Publication VII, and discussed both benefits and challenges of implementing LA approaches in CSX, see Chapter 3.3. Findings from the literature review were used as the foundation to achieve the research objectives and propose the LA reference model for CSXs. This, however means that specific questions of potential interest requiring a further deeper investigation can end up beyond the scope of this thesis.

This study informs future research focusing on specific types of CSXs and learning objectives and aims to show how the LA approach can be applied in various learning environments. However, as the research was exploratory in many respects and looked at a variety of exercise types (both individual and team-based learning) and learning objectives, this raises potential methodological challenges.

The empirical studies require replication and sufficient sample sizes to draw generalisations. However, the research time-line and CSXs for data collection was limited. Also the types of exercises (in corporate, military and academic sectors) that this research had access to was a representative mix; however military and corporate could have had more coverage. This limitation mainly relates to Publications I, II, III, VI and VIII. Thus, continuing the research basing the study on larger and more varied sample sizes and performing longitudinal studies could have generated more accurate results. Also, different research methods can be used to investigate certain questions to make the answers more robust. The mixed methods approach, covering both qualitative and quantitative elements, addresses this limitation.

In summary, this research has contributed to the initial mapping of the field and was exploratory in nature covering a variety of exercises. The research can be further improved by replicating and performing longitudinal studies of the published findings.

10 Conclusion

10.1 Summary of Work

The ultimate learning goal is to obtain knowledge and achieve long-term behavioural change. This, however, is a complex task: a wide array of personal or environmental factors such as fear, mood, threat, economic conditions, among others can predispose behaviour in a positive or negative way.

The learners experience learning in different styles and paces depending on the learning environment and their differing needs for learning support. The application of learning analytics and using the un-obtrusively collected digital datasets from the learning platforms can allow further analysis and relevant knowledge to support the learning process and also assess the performance of the learners. However, with LA and evidence-based measurement, we also need keep to in mind and validate that what we measure (i.e., metrics used) actually helps learners to learn.

The following research questions have been addressed:

1. How to deploy a learning analytics approach into the CSXs in order to improve learning processes and to evidence learning outcomes?

   The LA approach should be incorporated in the exercise from the start at the identify and planning stage, and it should not be an after-thought. The LA reference model, see Figure 3, is supported by an extensive literature review, including learning indicators (metrics) collected and analysed in the existing research.

   In order to design a CSX that enables the LA methods implemented, a critical issue is how to connect raw data to competencies. A practical and scalable design model to enable the digital traces to be linked to learning theories (such as Bloom's Taxonomy) has been developed as shown in Figure 4. Such a design approach is supported by the case studies and evaluation in the Rangeforce platform and in the XS exercise (Frankenstack).

2. How to evaluate effectiveness and efficiency of both individual and team learning with less obtrusive methods in CSXs?

   CSXs are believed to be an effective and efficient training method for up-skilling individuals and teams, but often the evaluation is anecdotal and not supported by evidence from the learning environments (i.e., cyber ranges). Traditional evaluation methods (such as observation, surveys) can be supported by collecting and analysing the digital footprint of the learners, as that especially allows less obtrusive and more objective analysis. One of the less researched areas is team learning; however, many exercises are team-based as teaming is integral in cyber operations. A novel and scalable methodology for an incident response learning objective was designed. The method combines elements of individual skills (e.g., network hardening, etc.) but also team skills (e.g., situational awareness sharing and communication, etc.), as shown in Figure 7. This method was tested in the LS exercise.

3. Could CSX's LA measures be standardised as performance indicators and be suitable as the performance predictors for cybersecurity technical skills?

   The predictive analytics is forward-looking by identifying and extracting specific predictive metrics and correlations that can be a scalable and remote method for the evaluation of technical cybersecurity skills. This work specifically looks at the admission process for the cybersecurity program when remote, technical, hands-on

labs were incorporated to the admission procedures and later correlated using regression analysis of the success rate and student performance in the technical cybersecurity course. We see that such remote labs were predictive; however we also identified that both interview and lab elements are complementary methods that both predict different aspects of the student performance results and have common factors, e.g., attitude, interest and eagerness, that are checked in interviews but also relevant for technical performance. While the remote labs are specific for a cybersecurity program, incorporating this type of assessment may also spark interest in other STEM programs or in the corporate sector.

10.2 Contributions

The key contributions from this research and related publications are:

- 5-timestamps model—a relatively simple method using a time-line of the attack and defence to capture performance improvement (learning) in a CSX including individual and team-related learning objectives for a cybersecurity incident handling (Publication I);

- Frankenstack—a tool for sharing instant feedback to RTs in a CSX from the network monitoring tools. While it may appear desirable to share a lot of data from monitoring tools with the learners, in the learning design one needs to consider the cognitive load and how much learners can use that data (Publication II);

- Malware lab—the exercises can be designed and individualised for assessment purposes (Publication III);

- Cyber hygiene—an extensive literature review and analysis was performed to see how the community defines cyber hygiene. This understanding contributes to the analysis of a full spectrum of CSXs (from general end-user audience to more complex exercises). These aim to train a range of cyber hygiene skills to highly specialised technical competencies (Publication IV);

- From scoring to learning—the individual data points do not produce an evaluation of learning impact that can be worked with. There is a need to connect and abstract these events to get meaningful and coherent events for measuring learning objectives (also skills and competencies) (Publication VI);

- LA model—a reference model for how to incorporate LA to the CSXs life-cycle was developed. This model addresses the existing research gap where CSXs researchers have mainly looked at technical set-ups and developed technical and complex exercise and learning platforms (cyber ranges) architectures. However, in terms of learning design and measurement in CSXs, these are not sufficiently evaluated using evidence-based methods and digital datasets and there is lack of guidance on how to incorporate LA in this process (Publication VII);

- Admissions—a novel approach of using the cybersecurity exercises in the university admission process and further evidence of their predictive power were demonstrated. Based on statistical analysis, an applicants' completion of the exercises (remote, technical hands-on labs) used in our university admissions process has a correlation to the student's actual performance in the first semester study course (Publication V and VIII).

The overall contribution is a practical learning analytics approach as well as methods for the cybersecurity exercises and/or awareness/hygiene training that have tested and validated at several training platforms, i.e., LS/XS, Rangeforce, TalTech students. By implementing and incorporating evidence-based learning analytics methods and measurements into the cybersecurity exercises, the cybersecurity community can establish a more evidence-based and systematic approach for the evaluation of learning impact enabling the design of more effective learning experiences.

10.3 Further Work

The author acknowledges that LA in the cybersecurity educational field is evolving and shaping at a fast pace. This thesis only covered some of the aspects, and lays the groundwork for further research and a few potential study topics are described as follows:

- Metrics Mining—research focusing on the metrics mining process (datafying) and machine learning methods of how to extract the metrics from pcap data. The aim is to develop methods for collection and acquisition, storage considerations, cleaning and integration of data, in order to find useful patterns for learning. Although CSXs leave digital footprints, the data extraction faces several challenges. Security challenges, such as intrusion detection, insider threats, malware detection and phishing detection, lack exact algorithmic solutions and the boundary between normal and anomalous behaviour isn't clear-cut, as attackers are continuously improving their techniques and strategies [115]. In addition, as one large exercise can create terabytes of data, the large amount of data generated by automatic logs and sensors necessitates efficient and automated data and LA techniques. There may not be enough traces to identify learning patterns (i.e., short time of detection, gaps in time-line, etc.) and data is very diverse (e.g., different OS, applications, etc.); therefore, identification of the relevant learning traces requires techniques that can deal with such imbalance and diversity.

- Validation of Learning Metrics and Outcomes—research analysing the digital footprint of student activities for selected learning objectives. The aim is to formulate a learning construct as an attribute which often cannot be measured directly (i.e., learning) but can be assessed using a number of indicators (variables), e.g., academic achievement or performance. Currently, there is no large pool of papers on the validation of LA models, methods and metrics. The metrics used in the CSXs literature focus on technical data and simple metrics (such as time). Even for such technical metrics, the articles are limited to the evaluation of a learning tool or method used and rely on feedback surveys. The metrics used are typically tested on small samples without further validation (e.g., experimental set-ups with control groups). However, computer-science education (including cybersecurity) involves "creativity, analysis, and problem solving—not the brute-force regurgitation of examples copied from the Web" [117]. This is challenging to measure with only technical measures, and cognitive metrics to evidence meta-cognition, decision speed and quality, etc. are needed. The existing research focuses on mapping digital traces describing student activity onto interpretable constructs of interest (e.g., Knowledge Components, Q-matrix) which facilitate actionable analytics [70]. The learner-facing approaches aimed at "learning how to learn" require more holistic (including artefacts, surveys, digital traces and physiological factors) validation strategies [63]. Only computational validation methods can lead to valid tools and

approaches being criticised or tools with little educational merit being labelled as well-performing [63].

- Scoring vs. Learning Measurement—research focusing on meaningful learning measurement constructs in CSXs. Currently a point system is often used to evaluate efforts and direct motivation. The scoring provides feedback and options for comparison of participants/teams. When used for learning purposes, the scoring needs to reflect the learning objectives and provide learning insights, not mainly focus on "game" rules aspects. With unclear scoring, a team's effort is unclear and makes identifying their weaknesses challenging. From a learning aspect, scoring and performance results cannot simply be equated. Learning does not necessarily lead to improvement in performance, because the results of learning processes are not the only determinants of behaviour. It is impacted by individual abilities (e.g., skills), personal motivation, tasks, team dynamics, etc. For example, in Cyber Shield exercise the teams that took longer time to identify an inject categorization, were more accurate [53]. Thus, scoring rules need to support appropriate exercise learning design. For example, in a very complex high-risk cyber conflict scenario an expectation (game rule) for teams to respond in 15 minutes to a simulated end-user's complaint about an unavailable website may prove to be unrealistic and will not contribute to learning but instead contradict learning objectives relating to prioritise incidents. It rather "forces" teams to learn, share and store wrong behaviours and later retrieve learned but wrong behavioural models in real life situations.

- Representation and Visualisation—research focusing on what learning metrics (dashboards) are useful for the learners and organisers in CSXs. This exposes the challenging relationship between visualisations and learning. The feedback about low level user actions, such as number of log ins, videos watched, or documents submitted, does not illuminate progress in learning for students or educators [57]. Few papers, such as [92] start analysing the use of visualisations in the CSXs; however visualisations' and dashboards' usefulness and effectiveness is not widely covered in the exercises.

- Team learning—as operational work in cybersecurity often takes place in teams and requires effective knowledge sharing and collaboration between individuals, teams and organisations, a focus on team-based cybersecurity exercises (CSXs) is critical. We have started to analyse and evidence team learning in CSXs, see Publications IX and X. As a novel aspect, we explore which characteristics of situation reports show that a team has shared, stored and can retrieve its collective knowledge. Situation reporting is commonly used in CSXs, and teams report on their collective knowledge of existing situation. Such reports are valuable for post-exercise reconstruction and sense-making of the exercise, as they should capture key incidents and events [47] and thus also show how teams have learned during the exercise. In further work, we plan to apply identified metrics to reflection logs and formalise reporting structures to enhance reflection on perceived individual and team performance. Such metrics form evidence-based foundation for semi-automated SITREPs scoring to ensure unbiased and comparable learning evaluation, and provide feedback to teams.

- Privacy and Ethics—potential for a research idea based on interviews with leading field experts for considerations of privacy concerns in LA application in CSXs, and how this is currently managed in the cyber ranges when collecting the exercise data for monitoring, feedback, etc. See Chapter 8 for details.

Researching how learning experience and performance can be improved among learners participating in CSXs from the digital datasets left in the learning environments highlights the importance of an inter-disciplinary approach. Implementing LA approaches in CSXs in educating professionals in the cybersecurity domain requires multiple perspectives. As future work should continue with varied and longitudinal studies, which will help to obtain an understanding of how to utilise the full potential of the digital footprints left in the learning environments of CSXs.

List of Figures

List of Tables

References

[1] ISO 22398:2013 Societal Security—Guidelines for Exercises.

[2] R. G. Abbott, J. McClain, B. Anderson, K. Nauer, A. Silva, and C. Forsythe. Log analysis of cyber security training exercises. *Procedia Manufacturing*, 3:5088–5094, 2015.

[3] R. G. Abbott, J. T. McClain, B. R. Anderson, K. S. Nauer, A. R. Silva, and J. C. Forsythe. Automated performance assessment in cyber training exercises. Technical report, Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2015.

[4] A. Ahmad. *A cyber exercise post assessment framework: In Malaysia perspectives*. PhD thesis, University of Glasgow, 2016.

[5] A. Ahmad, C. Johnson, and T. Storer. A cyber exercise post assessment: adoption of the kirkpatrick model. *Advances in Information Sciences and Service Sciences*, 7(2):1, 2015.

[6] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives, abridged edition. *White Plains, NY: Longman*, 2001.

[7] D. Andersson, M. Granåsen, T. Sundmark, H. Holm, and J. Hallberg. Exploratory sequential data analysis of a cyber defence exercise. In *Proceedings of the International Defense and Homeland Security Simulation Workshop (DHSS)*, 2011.

[8] M. Andreolini, V. G. Colacino, M. Colajanni, and M. Marchetti. A framework for the evaluation of trainee performance in cyber range exercises. *Mobile Networks and Applications*, 25(1):236–247, 2020.

[9] Y. Bergner, G. Gray, and C. Lang. What does methodology mean for learning analytics? *Journal of Learning Analytics*, 5(2):1–8, 2018.

[10] A. Bhattacherjee. Social science research: Principles, methods, and practices. 2012.

[11] M. Bishop. A constructive build-the-flag contest.

[12] M. Bishop, D. Burley, S. Buck, J. J. Ekstrom, L. Futcher, D. Gibson, E. K. Hawthorne, S. Kaza, Y. Levy, H. Mattord, et al. Cybersecurity curricular guidelines. In *IFIP World Conference on Information Security Education*, pages 3–13. Springer, 2017.

[13] D. Bisson. The four most common evasive techniques used by malware. `http://www.tripwire.com/state-of-security/security-data-protection/the-four-most-common-evasive-techniques-used-by-malware`, April 2015. Accessed: 2021.21.01.

[14] J. Blanken-Webb, I. Palmer, S.-E. Deshaies, N. C. Burbules, R. H. Campbell, and M. Bashir. A case study-based cybersecurity ethics curriculum. In *2018 {USENIX} Workshop on Advances in Security Education ({ASE} 18)*, 2018.

[15] N. G. Brooks, T. H. Greer, and S. A. Morris. Information systems security job advertisement analysis: Skills review and implications for information systems curriculum. *Journal of Education for Business*, 93(5):213–221, 2018.

[16] N. Buchler, C. G. La Fleur, B. Hoffman, P. Rajivan, L. Marusich, and L. Lightner. Cyber teaming and role specialization in a cyber security defense competition. *Frontiers in psychology*, 9, 2018.

[17] K. Cabaj, D. Domingos, Z. Kotulski, and A. Respício. Cybersecurity education: Evolution of the discipline and analysis of master programs. *Computers & Security*, 75:24–35, 2018.

[18] E. Caliskan, U. Tatar, H. Bahsi, R. Ottis, and R. Vaarandi. Capability detection and evaluation metrics for cyber security lab exercises. In *ICMLG2017 5th International Conference on Management Leadership and Governance*, page 407. Academic Conferences and publishing limited, 2017.

[19] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger. Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4):40, 2007.

[20] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs. A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5-6):318–331, 2013.

[21] Y. Chen, A. Johri, and H. Rangwala. Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 270–279. ACM, 2018.

[22] S. Chow, K. Yacef, I. Koprinska, and J. Curran. Automated data-driven hints for computer programming students. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 5–10. ACM, 2017.

[23] M. Christen, B. Gordijn, K. Weber, I. van de Poel, and E. Yaghmaei. A review of value-conflicts in cybersecurity: an assessment based on quantitative and qualitative literature analysis. *Orbit Journal*, 1(1):28, 2017.

[24] D. Clow. The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 134–138, 2012.

[25] J. Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.

[26] A. Conklin. Cyber defense competitions and information security education: An active learning solution for a capstone course. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 9, pages 220b–220b. IEEE, 2006.

[27] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle. A systematic literature review of empirical evidence on computer games and serious games. *Computers & education*, 59(2):661–686, 2012.

[28] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran. Interactive team cognition. *Cognitive science*, 37(2):255–285, 2013.

[29] K. G. Corley and D. A. Gioia. Building theory about theory building: what constitutes a theoretical contribution? *Academy of management review*, 36(1):12–32, 2011.

[30] J. W. Creswell and V. L. P. Clark. *Designing and conducting mixed methods research*. Sage publications, 2017.

[31] J. W. Creswell and J. D. Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[32] E. Crowley. Experiential learning and security lab design. In *Proceedings of the 5th conference on Information technology education*, pages 169–176. ACM, 2004.

[33] S. Dawson, S. Joksimovic, O. Poquet, and G. Siemens. Increasing the impact of learning analytics. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 446–455, 2019.

[34] K. Dechant, V. J. Marsick, and E. Kasl. Towards a model of team learning. *Studies in Continuing Education*, 15(1):1–14, 1993.

[35] H. Drachsler and W. Greller. Privacy and analytics: it's a delicate issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 89–98, 2016.

[36] A. C. Edmondson. The local and variegated nature of learning in organizations: A group-level perspective. *Organization science*, 13(2):128–146, 2002.

[37] M. Ernits and K. Kikkas. A live virtual simulator for teaching cybersecurity to information technology students. In P. Zaphiris and A. Ioannou, editors, *Learning and Collaboration Technologies*, pages 474–486, Cham, 2016. Springer International Publishing.

[38] M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020: 15th International Conference on Cyber Warfare and Security*. Academic Conferences and Publishing Limited, 2020.

[39] M. Ernits, J. Tammekänd, and O. Maennel. i-tee: A fully automated cyber defense competition for students. In *ACM SIGCOMM Computer Communication Review*, volume 45-4, pages 113–114. ACM, 2015.

[40] K. Friedman. Theory construction in design research: criteria: approaches, and methods. *Design studies*, 24(6):507–522, 2003.

[41] B. Gay and S. Weaver. Theory building and paradigms: A primer on the nuances of theory construction. *American International Journal of Contemporary Research*, 1(2):24–32, 2011.

[42] D. Gibson and S. de Freitas. Exploratory analysis in learning analytics. *Technology, Knowledge and Learning*, 21(1):5–19, 2016.

[43] C. Gipps and G. Stobart. Alternative assessment. In *International handbook of educational evaluation*, pages 549–575. Springer, 2003.

[44] C. Girard, J. Ecalle, and A. Magnan. Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, 29(3):207–219, 2013.

[45] Y. Gong, J. E. Beck, and N. T. Heffernan. How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis. *International Journal of Artificial Intelligence in Education*, 21(1-2):27–46, 2011.

[46] P. F. Gorder. Multicore processors for science and engineering. *Computing in science & engineering*, 9(2), 2007.

[47] M. Granåsen and D. Andersson. Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. *Cognition, Technology & Work*, 18(1):121–143, 2016.

[48] P. Guchait, P. Lei, and M. J. Tews. Making teamwork work: Team knowledge for team effectiveness. *The Journal of psychology*, 150(3):300–317, 2016.

[49] H. F. Hansen. Choosing evaluation models: a discussion on evaluation design. *Evaluation*, 11(4):447–462, 2005.

[50] R. Harley, D. Medlin, and Z. Houlik. Ethical hacking: Educating future cybersecurity professionals. In *Proceedings of the EDSIG Conference ISSN*, volume 2473, page 3857, 2017.

[51] J. B. Hauge, E. Boyle, I. Mayer, R. Nadolski, J. C. Riedel, P. Moreno-Ger, F. Bellotti, T. Lim, and J. Ritchie. Study design and data gathering guide for serious games' evaluation. *Psychology, Pedagogy, and Assessment in Serious Games, ed. Thomas M. Connolly, Thomas Hainey, Elizabeth Boyle, Gavin Baxter and Pablo Moreno-Ger*, pages 394–419, 2014.

[52] M. Hendrix, A. Al-Sherbaz, and B. Victoria. Game based cyber security training: are serious games suitable for cyber security training? *International Journal of Serious Games*, 3(1):53–61, 2016.

[53] D. S. Henshel, G. M. Deckard, B. Lufkin, N. Buchler, B. Hoffman, P. Rajivan, and S. Collman. Predicting proficiency in cyber defense team exercises. In *Military Communications Conference, MILCOM 2016-2016 IEEE*, pages 776–781. IEEE, 2016.

[54] J. L. Herman et al. *A practical guide to alternative assessment*. ERIC, 1992.

[55] J. Hohwy. *The predictive mind*. Oxford University Press, 2013.

[56] N. J. Holt, A. Bremner, E. Sutherland, M. Vliek, M. Passer, and R. Smith. *Psychology: The science of mind and behaviour*. McGraw-Hill Education, 2012.

[57] I. Jivet, M. Scheffel, M. Specht, and H. Drachsler. License to evaluate: Preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 31–40. ACM, 2018.

[58] Ø. Jøsok, B. J. Knox, K. Helkala, R. G. Lugo, S. Sütterlin, and P. Ward. Exploring the hybrid space. In *International Conference on Augmented Cognition*, pages 178–188. Springer, 2016.

[59] R. Kabra and R. Bichkar. Performance prediction of engineering students using decision trees. *International Journal of Computer Applications*, 36(11):8–12, 2011.

[60] M. Karjalainen, T. Kokkonen, and S. Puuska. Pedagogical aspects of cyber security exercises. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 103–108. IEEE, 2019.

[61] J. Kick. Cyber exercise playbook, 2014.

[62] K. Kitto, M. Lupton, K. Davis, and Z. Waters. Designing for student-facing learning analytics. *Australasian Journal of Educational Technology*, 33(5):152–168, 2017.

[63] K. Kitto, S. B. Shum, and A. Gibson. Embracing imperfection in learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 451–460. ACM, 2018.

[64] B. Knox, R. Lugo, K. Helkala, S. Sütterlin, and O. Josok. Education for cognitive agility: Improved understanding and governance of cyberpower. In *European Conference on Information Warfare and Security, ECCWS*, volume 2018-June, pages 541–550, 2018.

[65] M. Kont, M. Pihelgas, K. Maennel, B. Blumbergs, and T. Lepik. Frankenstack: Toward real-time red team feedback. In *IEEE Military Communications Conference (MILCOM)*, pages 400–405. IEEE, 2017.

[66] V. Kovanović, S. Joksimović, Z. Waters, D. Gašević, K. Kitto, M. Hatala, and G. Siemens. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on LAK*, pages 15–24. ACM, 2016.

[67] W. A. Labuschagne and M. Grobler. Developing a capability to classify technical skill levels within a cyber range. In *ECCWS 2017 16th European Conference on Cyber Warfare and Security*, page 224. Academic Conferences and publishing limited, 2017.

[68] P. Legato and R. M. Mazza. Modeling and simulation of cooperation and learning in cyber security defense teams. In *Proceedings - 31st European Conference on Modelling and Simulation, ECMS 2017*, pages 502–509, 2017.

[69] T. Lepik, K. Maennel, M. Ernits, and O. Maennel. Art and automation of teaching malware reverse engineering. In *International Conference on Learning and Collaboration Technologies*, pages 461–472. Springer, 2018.

[70] R. Liu and K. R. Koedinger. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*, 9(1), 2017.

[71] M. Loi and M. Christen. Ethical frameworks for cybersecurity. In *The Ethics of Cybersecurity*, pages 73–95. Springer, Cham, 2020.

[72] K. Maennel. Learning analytics perspective: Evidencing learning from digital datasets in cybersecurity exercises. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 27–36. IEEE, 2020.

[73] K. Maennel, K. Kivimägi, S. Sütterlin, O. Maennel, and M. Ernits. Remote technical labs: an innovative and scalable component for university cybersecurity program. In *Educating Engineers for Future Industrial Revolutions—Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2021.

[74] K. Maennel, S. Mäses, and O. Maennel. Cyber hygiene: The big picture. In *Nordic Conference on Secure IT Systems*, pages 291–305. Springer, 2018.

[75] K. Maennel, S. Mäses, S. Sütterlin, M. Ernits, and O. Maennel. Using technical cybersecurity exercises in university admissions and skill evaluation. *IFAC-PapersOnLine*, 52(19):169–174, 2019.

[76] K. Maennel, R. Ottis, and O. Maennel. Improving and measuring learning effectiveness at cyber defense exercises. In *Nordic Conference on Secure IT Systems*, pages 123–138. Springer, 2017.

[77] S. Mäses, B. Hallaq, and O. Maennel. Obtaining better metrics for complex serious games within virtualised simulation environments. In *ECGBL 2017 11th European Conference on Game-Based Learning*, pages 428–434. Academic Conferences and publishing limited, 2017.

[78] I. Mayer, G. Bekebrede, H. Warmelink, and Q. Zhou. A brief methodology for researching and evaluating serious games and game-based learning. In *Psychology, pedagogy, and assessment in serious games*, pages 357–393. IGI Global, 2014.

[79] K. V. Moses and W. M. Petullo. Teaching computer security, 2014.

[80] S. Mäses. *Evaluating Cybersecurity-Related Competences Through Simulation Exercises*. PhD thesis, Tallinn University of Technology, 2020.

[81] NATO. A Generic Reference Curriculum on Cybersecurity, 2016. Accessed: 2021.04.04.

[82] NATO CCD COE. Locked Shields 2019. `https://ccdcoe.org/locked-shields-2019.html`. Accessed: 2020.04.03.

[83] NATO CCD COE. Locked Shields 2016 After Action Report. NATO Cooperative Cyber Defence Centre of Excellence Publication, 2016.

[84] NATO CCD COE. Cross Swords 2017 RT Objectives, 2017. Accessed: 2021.04.03.

[85] NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE). Locked shields 2016 after action report. In *After Action Report*, 2016.

[86] W. Newhouse, S. Keith, B. Scribner, and G. Witte. Nice cybersecurity workforce framework (ncwf): National initiative for cybersecurity education. Technical report, National Institute of Standards and Technology, 2016.

[87] W. Newhouse, S. Keith, B. Scribner, and G. Witte. National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework. *NIST Special Publication 800-181*, 2017.

[88] D. R. Newman, B. Webb, and C. Cochrane. A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology*, 3(2):56–77, 1995.

[89] X. Ochoa and A. Merceron. Quantitative and qualitative analysis of the learning analytics and knowledge conference 2018. *Journal of Learning Analytics*, 5(3):154–166, 2018.

[90] X. Ochoa, D. Suthers, K. Verbert, and E. Duval. Analysis and reflections on the third learning analytics and knowledge conference (lak 2013). *Journal of Learning Analytics*, 1(2):5–22, 2014.

[91] A. Ogee, R. Gavrila, P. Trimintzios, V. Stavropoulos, and A. Zacharis. The 2015 report on national and international cyber security exercises.

[92] R. Ošlejšek, D. Toth, Z. Eichler, and K. Burská. Towards a unified data storage and generic visualizations in cyber ranges. In *ECCWS 2017 16th European Conference on Cyber Warfare and Security*, page 298, 2017.

[93] Z. Papamitsiou and A. A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4), 2014.

[94] Z. Papamitsiou, M. N. Giannakos, and X. Ochoa. From childhood to maturity: Are we there yet? *LAK '20: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020.

[95] A. Pardo and G. Siemens. Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3):438–450, 2014.

[96] A. Parrish, J. Impagliazzo, R. K. Raj, H. Santos, M. R. Asghar, A. Jøsang, T. Pereira, and E. Stavrou. Global perspectives on cybersecurity education for 2030: a case for a meta-discipline. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 36–54. ACM, 2018.

[97] V.-V. Patriciu and A. C. Furtuna. Guide for designing cyber security exercises. In *Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy*, pages 172–177, 2009.

[98] R. Pelánek. Learning analytics challenges: trade-offs, methodology, scalability. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 554–558, 2020.

[99] A. Peña-Ayala. Learning analytics: Fundaments, applications, and trends. *A View of the Current State of the Art to Enhance e-Learning. Edtion ed.: Springer International Publishing*, 2017.

[100] P. Pusey, M. Gondree, and Z. Peterson. The outcomes of cybersecurity competitions and implications for underrepresented populations. *IEEE Security & Privacy*, 14(6):90–95, 2016.

[101] A. Rashid, G. Danezis, H. Chivers, E. Lupu, A. Martin, M. Lewis, and C. Peersman. Scoping the cyber security body of knowledge. *IEEE Security & Privacy*, 16(3):96–102, 2018.

[102] R. Richards, A. Konak, M. R. Bartolacci, and M. Nasereddin. Collaborative learning in virtual computer laboratory exercises. *Network, Security*, 155:9, 2015.

[103] T. Riebe and C. Reuter. Dual-use and dilemmas for cybersecurity, peace and technology assessment. In *Information Technology for Peace and Security*, pages 165–183. Springer, 2019.

[104] T. Scheponik, A. T. Sherman, D. DeLatte, D. Phatak, L. Oliva, J. Thompson, and G. L. Herman. How students reason about cybersecurity concepts. In *Frontiers in Education Conference*, pages 1–5. IEEE, 2016.

[105] P. M. Senge. The fifth discipline, the art and practice of the learning organization. *Performance+ Instruction*, 30(5):37–37, 1991.

[106] G. Siemens. Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 4–8. ACM, 2012.

[107] G. Siemens. Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400, 2013.

[108] A. Silva, J. McClain, T. Reed, B. Anderson, K. Nauer, R. Abbott, and C. Forsythe. Factors impacting performance in competitive cyber exercises. In *Proceedings of the Interservice/Interagency Training, Simulation and Education Conference, Orlando, FL*, 2014.

[109] M. J. Singer and B. W. Knerr. Evaluation of a game-based simulation during distributed exercises, 2010.

[110] Society of Learning Analytics Research. What is learning analytics. https://www.solaresearch.org/about/what-is-learning-analytics. Accessed: 2021.04.03.

[111] V. Švábenskỳ, J. Vykopal, and P. Čeleda. Toward an automated feedback system in educational cybersecurity games. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE'19)*, 2019.

[112] C. Taylor and C. Colberg. A tool for teaching reverse engineering. In *2016 USENIX Workshop on Advances in Security Education (ASE 16)*, Austin, TX, 2016. USENIX Association.

[113] S. Van Der Haar, K. A. Jehn, and M. Segers. Towards a model for team learning in multidisciplinary crisis management teams. *International Journal of Emergency Management*, 5(3-4):195–208, 2008.

[114] K. Verbert, N. Manouselis, H. Drachsler, and E. Duval. Dataset-driven research to support learning and knowledge analytics. *Journal of Educational Technology & Society*, 15(3):133–148, 2012.

[115] R. Verma, M. Kantarcioglu, D. Marchette, E. Leiss, and T. Solorio. Security analytics: essential data analytics knowledge for cybersecurity professionals and students. *IEEE Security & Privacy*, 6:60–65, 2015.

[116] J. Vykopal, R. Ošlejšek, K. Burská, and K. Zákopčanová. Timely feedback in unstructured cybersecurity exercises. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 173–178. ACM, 2018.

[117] R. Weiss, M. E. Locasto, and J. Mache. A reflective approach to assessing student performance in cybersecurity exercises. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 597–602. ACM, 2016.

[118] G. Wells and G. Claxton. *Learning for life in the 21st century: Sociocultural perspectives on the future of education*. John Wiley & Sons, 2008.

[119] J. M. Wilson, P. S. Goodman, and M. A. Cronin. Group learning. *Academy of Management Review*, 32(4):1041–1059, 2007.

Acknowledgements

This research has been financially supported by Smart Specialization project with CybExer Technologies (NSP72; LEP17064), collaborative research project between TalTech and South Korean National Security Research Institute (VA17103), ECHO project (no 8309439), IT Academy and Dora Plus scholarships (supported by the EU Regional Development Fund).

I want to express my gratitude to my supervisors Dr Rain Ottis and Dr Stefan Sütterlin, who helped and guided me through this exciting journey. I also would like to thank Dr Liina Randmann, who started this journey with me also as an supervisor—you helped to lay the foundations of later work.

This work would not have taken place without the open-minded and friendly organisers of LS and XS, Rangeforce, and TalTech MSc admission team, who have allowed me to experiment on their platforms and exercises.

Special thanks to Bernhards Blumbergs, Mauno Pihelgas, Markus Kont, Sten Mäses, Margus Ernits, Toomas Lepik, Kristjan Kivimägi and Joonsoo Kim, who gave me chance and experience in collaborative research and writing together. I have to mention and thank Jussi Jaakonaho who shared his immense experience and insight about Locked Shields and learning in cybersecurity exercises overall.

Thank you, Kimberly Frankenschmidt and Dr Adrian Venables, for being very thorough and taking time to fix my English articles and commas. I learned a lot in the process and this knowledge will help my future scientific publications to comply with US or UK grammar rules.

I thank my husband Olaf Maennel, for endless encouragement and keeping me on track, when I was loosing hope that I will ever finish this thesis. Enormous thanks to my boys, Oliver and Martin, for bringing happiness into every day and making emotional boost to my learning experience. My best friend and role model in the academic studies, my little sister, Tiia Möller—you won the competition who finishes a Ph.D. first in our family but you were source of my inspiration and advice. And last but not least—thanks goes to my mother, who has throughout my life encouraged me to study, study, study…

Thank you all and the academic journey continues…

Abstract
Advancing Cybersecurity Education through Learning Analytics

Effective training and learning for cybersecurity professionals is considered a significant and unresolved issue, especially due to the existing skilled workforce gap. The cybersecurity exercises (CSXs) are believed to be an effective training for all training audiences from top (military) professional teams to individual students. However, evidence of learning outcomes for those exercises are often anecdotal and not supported by evidence from the learning environments (i.e., cyber ranges). Adopting a learning analytics (LA) mindset in cybersecurity trainings can help educators to achieve a more adaptive design and measurement using evidence-based data from the digital learning environment. The study focus on novel aspects of incorporating LA as an evidence-based approach within CSXs and trainings. As CSXs come in a variety of formats, this thesis focuses on technical exercises with both individual and team-based designs. Collecting data from the technical exercises (which forms the basis for LA) is a computer science problem and requires good understanding of the technical aspects of the exercises. However, interdisciplinary approach combining knowledge from cybersecurity, pedagogy and psychology is needed to achieve effective application of LA in cybersecurity education.

This research intends to both explain (quantitative) and explore (qualitative) aspects, and thus mixed methods were applied. This approach requires hypothesising and then generalising or applying an hypothesis to other populations. But it equally aims to gain more precise understanding of the dynamic interaction and perceptions of the stakeholders (i.e., the learners, the educators, the organisers, etc.) involved.

To provide practical methodology to enhance implementing LA approach in CSXs, the LA reference model that should be incorporated to the CSX life-cycle has been developed. The model combines and outlines the key LA considerations, while model itself is supported with an extensive overview of existing use cases for a practical implementation. Especially the developers would need to consider LA aspects in their design of the cyber ranges, as they lay technological foundation of instrumenting the exercises enabling LA in the first place. In order to design the exercises that enable the LA methods, the critical issue is how to connect raw data to competencies. A practical and scalable design model for enabling the digital traces to be linked to learning theories (such as Bloom's Taxonomy) has been described. Such design approach is supported by the case studies and evaluation in Rangeforce platform and also how this raw data can support learning experience. For example, Frankenstack tool in Crossed Swords exercise focused on providing the training audience (Red team) with instant feedback about their actions to ensure effective learning.

When measuring the achievement of learning outcomes and providing effective feedback, traditional evaluation methods (such as observation, surveys), can be supported by collecting and analysing the digital footprint of the learners. Especially as such approach allows less obtrusive and more objective analysis. Many exercises are team-based as this teaming is integral in cyber operations, however this is also less researched area. Top-end and large scale technical exercises are often complex, which makes it hard for organisers and participants to handle. Therefore, both learning design and measurement need careful consideration. The novel and scalable "5-timestamp methodology" for incident response learning objective combines elements of individual skills (e.g., network hardening, etc.) but also team skills (e.g., situational awareness sharing and communication, etc.). This method aims at accommodating for both—effective feedback (including benchmarking opportunity) and learning measurement. The method is capable of assessing team performance, and argues that changes in performance over time equal learning.

The timestamps can either be collected using traditional methods, such as interviews, observations and surveys, but also potentially be obtained non-obtrusively from raw network traces (such as pcap data). The method enhances the feedback loop, allows identifying learning design flaws, and provides solid evidence of learning value for CSXs. This method was tried out at the Locked Shields exercise.

The predictive analytics is forward-looking by identifying and extracting specific predictive metrics and correlations that can be a scalable and remote method for evaluation of technical cybersecurity skills. This work specifically looks at the admission process for the cybersecurity program when remote technical hands-on labs was incorporated to the admission procedures and later correlated using regression analysis to the success rate and student performance at the technical course. We see that such remote labs were predictive, however we identified that both interview and lab elements are complementary methods that both predict in different aspects of the performance and have common factors, e.g., attitude, interest and eagerness that are checked in interview but also relevant for technical performance. While the remote labs analysed are specific for a cybersecurity program, incorporating this type of assessment may also spark interest in other STEM programs or in the corporate sector.

The overall contribution of this research is a practical LA approach and theoretical methods for the CSXs and/or awareness/hygiene training that have been implemented at several training platforms and audiences, i.e., LS/XS, Rangeforce, TalTech students. By implementing and incorporating evidence based learning analytics methods and measurements into the cybersecurity exercises, the cybersecurity community can establish more evidence-based and systematic approach for evaluation of learning impact that will enable designing more effective learning experiences. This work is ongoing, and several research gaps and further work is proposed to advance research in LA and CSXs. Implementing LA approaches in CSXs for improved proficiency for educating professionals in the cyberspace domain requires multiple perspectives, varied and longitudinal studies to obtain the understanding how to utilise the full potential of the digital footprints left in the learning environments of CSXs.

Kokkuvõte
Küberkaitsealase hariduse parendamine õpianalüütika abil

Tulenavalt koolitatud tööjõu puudusest küberkaitses on efektiivne koolitus ja õppimine küberkaitse ametialal oluline kuid seni lahendamata probleem. Küberkaitseharjutusi (ehk küberõppusi) peetakse üheks efektiivseimaks meetodiks erinevate sihtgruppide koolitamisel — see sobib nii professionaalsetele meeskondadele (sh sõjaväelistele) kui individuaalsetele õppijatele. Samas põhinevad teadmised õppustel saavutatud õpitulemustest tihti suulisel infol ja ei toetu digitaalsetest õppeplatvormidelt (nt küber harjutusväljakud) saadud infol. Võttes kasutusele õpianalüütikat rakendava lähenemisviisi küberkaitseharjutuste läbiviimisel, on võimalik luua adaptiivseid koolitusi ja mõõta nende efektiivsust tõenduspõhiste digitaalsetest õpikeskkondadest kogutud andmete baasil.

Käesoleva uurimustöö eesmärgiks on analüüda ja luua lahendusi, kuidas rakendada õpianalüütikat kui tõenduspõhist lähenemisviisi individuaalsetel ja meeskonnapõhistel küberkaitseharjutustel ja -koolitustel. Õpianalüütika aluseks olevate andmete kogumine on arvutiteaduse probleem ning nõuab põhjalikku arusaamist harjutuste tehnilistest aspektidest. Samas selleks, et õpianalüütikat saaks edukalt küberkaitsealases hariduses rakendada, on vajalik interdistsiplinaarne lähenemine, mis ühendab teadmised küberkaitsest, pedagoogikast ja psühholoogiast. Töö eesmärgi saavutamiseks tõstatati järgmised uurimisküsimused:

1. Kuidas rakendada õpianalüütikat küberharjutustel eesmärgiga parandada õppimise protsessi ja õpitulemuste tõendamisel?

    1.1. Kuidas ja milliseid andmeid koguda, kui rakendadakse õpianalüütikat küberharjutustel?

    1.2. Milliseid õppedisaini mudelid võimaldavad ühendada alusandmed (“raw data") digitaalsetest õpikeskkondadest kõrgemal tasemel defineeritud pädevustega?

2. Kuidas hinnata individuaalse ja meeskonnapõhise õppimise tõhusust ja mõjusust mitte-instrusiivsete meetoditega?

3. Kas küberharjutuste õpianalüütika andmeid saab kasutada küberkaitse tehniliste oskuste ennustamisel?

Uurimustöös on rakendatud kombineeritud uuringudisaini (“mixed methods"). Sellise uurimismeetodi kasutamine nõuab hüpoteesi tõstatamist ning seejärel üldistuste tegemist või hüpoteesi rakendamist erinevale populatsioonile. Samuti on eesmärk omandada täpsem arusaamine erinevate huvigruppide (so õppijad, õpetajad, korraldajad, jne.) dünaamilisest koostoimimisest ja arusaamadest.

Praktikas õpianalüütika rakendamisl tuleks lähtuda küberharjutuste õpianalüütika referentsmudelist, mida peaks rakendama küberkaitseharjutuse elutsükli jooksul. Referentsmudelis on esitatud peamised kaalutlused õpianalüütika rakendamisel ning mudel on toetatud ulatusliku ülevaatega näidetest olemasolevas teaduskirjanduses. Eriti küberharjutusväljaku arendajad peaksid arvesse võtma õpianalüütika aspektid, kuna tehniline disain loob põhialused õpianalüütika instrumenteerimisel küberharjutustes.

Selleks, et disainida harjutusi, mis võimaldavad õpianalüütikat koguda ja analüüsida, on oluline ühendada alusandmed (“raw data") kompetentsidega. Uurimistöö tulemusel on kirjeldatud praktiline ja laiendatav mudel, mis võimaldab ühildada digitaalsed andmed õpiteooriatega (nt Bloomi taksonoomia). Seda disainimeetodi kasutati ja hinnati Rangeforce platvormil. Samuti analüüsiti, kuidas kasutada alusandmeid õppimise toetamisel.

Näiteks, Frankenstack lahendus oli kasutusel Cross Swords küberharjutusel, andes punasele meeskonnale (“red team") kohest tagasisidet nende tegevuse kohta toetades efektiivset õppimist.

Õpitulemuste mõõtmisel ja efektiivse tagasiside andmisel on võimalik toetada nö traditsioonilisi hindamismeetodeid (nt vaatlused, küsitlused) digitaalse õpijälje kogumuse ja analüüsimisega. Selline andmete kogumise viis võimaldab mitte-instrusiivset aga samas objektiivsemat analüüsi.

Paljud küberõppused on meeskondlikud, kuna meeskonnatöö on küberoperatsioonides igapäevane. Samas on meeskondlik õppimine ("team learning") vähem uuritud. Suuremahulised tehnilised õppused on tihti keerulised, mistõttu nii korraldajatele kui osalejatele on need harjutused rasked hallata. Seetõttu vajab õpidisain ja tulemuste hindamine hoolikat tähelepanu. Uudne ja skaleeritav õpitulemuste mõõtmise metoodika küberintsidendile reageerimisel on “5-ajatempli metoodika”—võimaldades hinnata nii individuaalsete oskuste (nt võrgukaitse seadistamine) kui meeskondlike oskuste (nt olukorrateadlikkus, infovahetus ja kommunikatsioon) elemente. Metoodika hõlmab nii efektiivset tagasisidet (sh võrdlusvõimalus) kui õpitulemuste mõõtmist. See võimaldab hinnata meeskondade tegevustulemust ja tulemuste muutusena ajas näitab ka õpitulemusi. Ajatempleid saab koguda nii traditsiooniliste meetoditega (nt intervjuud, vaatlused ja küsimustikud), aga ka potentsiaalselt mitte-intrusiivselt võrgulogidest (nt pcap’id). Metoodika aitab parendada tagasisidet, tuvastada õppuse disaininõrkusi ja näidata kübekaitseõppuste õpiväärtust. Seda mudelit katsetati praktikas Locked Shield’ küberõppusel.

Ennustatav analüütika (“predictive analytics") on tulevikku vaatav identifitseerides ja analüüsib ennustavaid mõõdikuid jning seoseid. Sellist analüütikat on võimalik kasutada kui skaleeruvat ja kaugkontakti võimaldavat meetodit tehniliste küberkaitse oskuste hindamisel. Käesoleva töö raames ennustatava analüütika rakendamist uuritud ülikooli vastuvõtuprotseduuridele küberkaitse magistriprogrammi raames. Vastuvõtu protseduurid hõlmavad ka tehnilisi laboreid, mille tulemusi on regressioonimudelit kasutades analüüsitud hilisema edukusega tehnilisel kursusel. Tulemused näitavad, et valitud laborid olid ennustavad, kuid ka seda et nii intervjuu kui labori elemendid on üksteist täiendavad—need ennustavad erinevaid edukuse aspekte ja neil on ühiseid faktoreid (nt suhtumist, huvi ja motivatsiooni hinnatakse intervjuu käigus, samas on need relevantsed faktorid ka tehniliste tulemuste saavutamisel). Uurimistöös kasutatud laborid on spetsiifilised küberkaitse programmile, aga sarnaste laborite kasutamine on rakendatav ka teistes loodus- ja täppisteaduste ainetes (STEM) ja erasektoris.

Käesolev doktoritöö annab üldise panuse praktilise õpianalüütika referentsmudeli ja teoeeritiliste meetodite loomise kujul. Mudelid ja meetodid on spetsiifiliselt rakendatavad küberkaitseõppustel, aga ka küberteadlikkuse ja -hügieeni alastel koolitustel. Uurimistöö raames välja töötud metoodikaid on rakendatud mitmetel õppeplatvormidel ja sihtrühmadel, sh LS/XS, Rangeforce, TalTech üliõpilased. Õpianalüütika meetodite inkorporeerimine ja rakendamine küberharjutustes võimaldab tõenduspõhist ja süstemaatilist õpitulemuste hindamist, mis omakorda võimaldab efektiivsemat õpikogemuse disaini. Uurimistöö on jätkuv ning on rõhutab mitmed edasise uurimistöö suundasid õpianalüütika ja küberharjutuste valdkonnas. Õpianalüütika edukas rakendamine küberkaitseharjutustel eeldab interdistsiplinaarset lähemisviisi, samuti mitmeid ja pikaajalisemaid uuringuid, kuidas õppeplatvormidel olemasolevat õppimise “digijälge"õppimise protsessis paremini kasutada.

Appendix 1

I

K. Maennel, R. Ottis, and O. Maennel. Improving and measuring learning effectiveness at cyber defense exercises. In *Nordic Conference on Secure IT Systems*, pages 123–138. Springer, 2017

Improving and Measuring Learning Effectiveness at Cyber Defense Exercises

Kaie Maennel, Rain Ottis, and Olaf Maennel(✉)

Tallinn University of Technology, Tallinn, Estonia
{kaie.maennel,rain.ottis,olaf.maennel}@ttu.ee

Abstract. Cyber security exercises are believed to be the most effective training for the training audiences from top professional teams to individual students. However, evidence of learning outcomes is often anecdotal and not validated. This paper focuses on measuring learning outcomes of technical cyber defense exercises (CDXs) with Red and Blue teaming elements. We studied learning at Locked Shields, which is the largest unclassified defensive live-fire CDX in the world. This paper proposes a novel and simple methodology, called the "5-timestamp methodology", aiming at accommodating both effective feedback (including benchmarking) and learning measurement. The methodology focuses on collection of timestamps at specific points during a cyber incident and time interval analysis to assess team performance, and argues that changes in performance over time can be used to evidence learning. The timestamps can either be collected non-intrusively from raw network traces (such as pcaps, logs) or using traditional methods, such as interviews, observations and surveys. Our experience showed that traditional methods, such as self-reporting, fail at high-speed and complex exercises. The suggested method enhances feedback loop, allows identifying learning design flaws, and provides evidence of learning value for CDXs.

Keywords: Cyber defence exercise · Training and education · Learning outcomes · Measuring learning

1 Introduction

Cyber security exercises are quickly gaining popularity as a teaching method for cyber-readiness. Globally there are over 200 cyber security exercises and more than 50% have a performance objective focusing on learning [17]. The European Union Agency for Network and Information Security survey describes the state of art: "... after-action reports and 'lessons learned' documents have become increasingly at risk of becoming fantasy documents. There is an increased demand that lessons must have been successfully learned, and that noting such instances of lesson-drawing is all there is to it. Few, if any, controls are actually made to verify that they can even be called lessons by any sensible definition, or that anything has actually been learned" [17]. The evidence of learning outcomes

H. Lipmaa et al. (Eds.): NordSec 2017, LNCS 10674, pp. 123–138, 2017.
https://doi.org/10.1007/978-3-319-70290-2_8

is limited and evaluation methodologies focus on the improvement of one exercise to the next [2]. On one side the literature describes enthusiasm of participants for the knowledge gained and lessons learned [11]. At other end of spectrum, Pusey et al. [19] claim that evidence is often anecdotal and little work has been done to validate learning outcomes.

This paper focuses on cyber defense exercises (CDXs) with the Red (RT) and Blue Team (BT) elements and looks at measuring learning effectiveness from an organizer's perspective. We use the NATO Cooperative Cyber Defense Centre of Excellence's (CCD COE) Locked Shields (LS) as testing platform. LS, that took place 26–28 April 2017 (LS17), is one of the largest and advanced team based live-fire RT/BT technical exercise with nearly 900 participants [14]. The exercise is a hybrid of competition, assessment and complex scenario-based learning event. The training audience comprises of the national BTs that in the exercise context take the role of the computer emergency response teams tasked to defend the pre-built virtualized networks of fictional organizations against the RT attacks. The other teams involved in the exercise are: Green Team (GT) responsible for game network and infrastructure development, White Team (WT) for game scenario development and execution control, and Yellow Team (YT) for monitoring and situational awareness [15]. One additional advantage of such CDXs is that permission to study individual and team performances and learning can be easily obtained by the organizers before the exercise starts.

2 Learning Measurement Dimensions in CDX's

CDXs in the current form are often not sufficiently instrumented for learning measurement and existing measurements focusing on scoring are not using learning related metrics. We recommend an overall CDX's learning measurement approach that brings together pre-exercise, execution and post-exercise phases and individual/team/organizational aspects. The measurement should include mixture of quantitative and qualitative methods. As a novel method we discuss the idea of the 5-timestamp methodology that focuses on unobtrusive data collection and comparable data analysis linked to learning objectives. This methodology is only a part of overall learning measurement (including traditional methods, such as surveys, interviews, etc.).

2.1 5-Timestamp Methodology

Learning in CDXs is affected by many variables, however the basic measurements, such as timing and accuracy metrics are still key elements that provide comparable trends in learning process and benchmarking for the teams. For example, Henshel et al. measurements in Cyber Shield 2015 showed that when teams took 20 or more minutes to identify an inject's NIST categorization, they were more accurate [9]. That means an overly time-constraint game-rule may prove to be an unrealistic expectation, which will not contribute to learning. Instead it forces teams to learn, share and store wrong behaviors and later

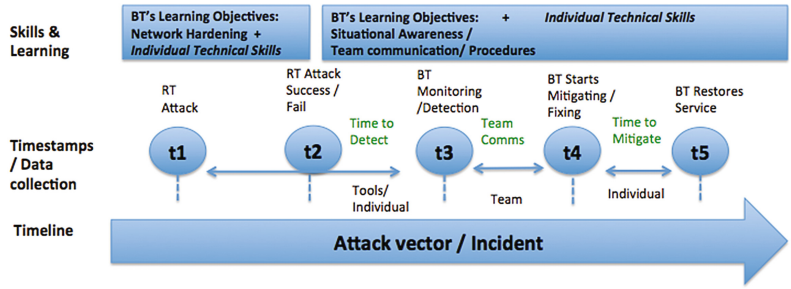

Fig. 1. 5-timestamp non-intrusive methodology

retrieve learned, but wrong behavioral models in real life situations [26]. Such metrics support development of appropriate exercise learning design.

Furthermore, measuring learning effectiveness and collecting data in order to provide feedback can be combined. The learning potential is not fully realized, if the BTs do not know what their weaknesses are, and how they progressed in the exercise. Scoring might give some indication of how teams compare, however, without knowing a baseline or standard in more detail, the overall score is worthless from learning viewpoint. For example, scoring may not take into account how much resistance the BTs put up and how efficient they were in responding.

As a part of solution we propose a non-intrusive methodology to collect and analyze timestamps from both the RT and BT actions from their digital footprint. The analysis of time intervals between the proposed timestamps enables to measure technical skills, but also soft skills (including leadership, team communications, decision making). The methodology is analysing data at a cyber incident/attack vector/target machine level, but provides metrics for different learning objectives (Fig. 1). For example, the assessment whether the BTs are effective and achieve incident handling related learning objectives, needs basic timing and accuracy metrics—how long does it take to respond to an attack, how long did teams take to respond to a significant threat vs. minor issues, what is correlation between the teams' detection time and quality of reporting. Further analysis can be carried out whether the most effective strategy from qualitative aspects was applied by the BTs, but having timing and accuracy metrics will provide input and focus to such qualitative analysis and feedback.

The analysis breaks an incident into phases to demonstrate strengths and bottlenecks in individual and team skills in each phase, and provides the basis for effective feedback. The model follows the incident timeline, and information can be collected non-intrusively (Fig. 2) from game-net/management network[1]. Even when t_1 and t_2 are intrusive for the RTs, data collection is non-intrusive

[1] The exercise runs on separate virtualised machines which are accessed remotely over the VPN [15]. The BTs can reside in their home location and connect via a dedicated management network to the game-environment. The game-net resides typically on a different interface and is where the attacks are happening.

Timestamp	Description	Non-intrusive Data	Intrusive (optional)
t_1	RT starts attacking	RT activity reporting	N/A
t_2	RT compromises	RT activity reporting and scoring data	N/A
t_3	BT detects	Possibly by access pattern	BT observation or self-reporting via inject
t_4	BT mitigates	management network (showing traffic activity)	BT observation or self-reporting via inject
t_5	BT restores	scoring, management network (end of session)	N/A

Fig. 2. Data sources for 5-timestamps, non-intrusive for the BTs (as the training audience). Intrusive methods can be used for cross-checking and validation.

for the BTs. For cross-checking, a sample using intrusive methods should be selected.

In order to fully understand this methodology, it should be noted that there are typically several target machines in a game-network that can be attacked repeatedly using the same attack methods. However, one of the advantages of a live-fire RT/BT exercise is defending against a "thinking" adversary, which implies that the same target can be attacked using different methods.

Collecting Timestamps Non-intrusively From PCAPs. The idea relies on the fact that the organizers are able to collect all raw network traffic (e.g., pcaps) not only from within the game-net, but also from the management network. From those traces it is possible to automatically detect the times of a BT activity for each target machine (e.g., when a BT member is working on a machine or not). This can be done by observing a ssh or remote desktop connection from the BT-network through management network. Even if the traffic is encrypted, and the BT member remains logged-in in the background, simply observing the traffic volume and packet inter-arrival times allows automatically detecting times at which someone is working on a specific game-net target. With traditional methods, this can also be achieved by asking the team member to keep a detailed log about timestamps.

The time intervals between timestamps provide basic learning metrics as shown in Fig. 3. In addition to measuring technical skills, these metrics also give insight to:

Team vs. individual—how long an individual and/or sub-team takes to resolve an issue, e.g., several members connecting to the same machine to work together.

Soft skills (leadership and decision making)—as the teams must make quick decisions (likely to have immediate and significant consequences), teams also learn decision-making. The OODA (Observation, Orientation, Decision, and Action) loop is a decision-making theory where time is the dominant parameter [23], and thus supports this framework using time intervals. The teams need to perform reliably and adapt their responses to mitigate adverse scenarios, and

Timestamps Interval	Description	Learning Objectives	Team vs. Individual
t_5–t_2	incident response time	Overall performance (organizer's objectives=scoring)	team
t_5–t_4	time to mitigate	Responding to attacks (technical skills)	individual, sub-team
t_4–t_3	time between mitigation and detection	Time management and prioritization; Teamwork: delegation, dividing and assigning roles, leadership; Handling cyber incident	team
t_3–t_2	time between compromise and detection	Monitoring networks, detecting of attacks	individual, sub-team
t_2–t_1	time to compromise	Learning the network; System administration and prevention of attacks	individual, sub-team

Fig. 3. Learning metrics from 5-timestamps intervals

that can be measured by t_4–t_3, i.e., time needed for intra-team communication, prioritization, task allocation.

Benefits and Application in Learning Process. The 5-timestamp methodology provides several advantages. Firstly, during a post-exercise debrief, it helps to create a general mental map of the events. For example, letting participants search for events in the BTs of pcaps or logfiles to figure out what happened is not useful for learning. In a similar analogy, where security cameras have become more effective when combined with a motion sensor—logfiles become more easily "searchable" when combined with accurate timestamp annotations. Debriefing an attack from the high-level objectives together with accurate timestamps, facilitates finding the relevant information. As the participants have already been in the situation during exercise, they understand the RT objectives, and are able to "relive" the events. Useful feedback can only be given, if the exercise can be debriefed in a meaningful way, and accurate timestamps are a first critical step towards achieving this.

Secondly, the timestamps can be used in building a baseline for performance or effectiveness. When grouped by the attack methods (not the target systems), those values become comparable. These can be further analyzed in several ways: (1) as an average overall performance against defending against a certain type of attack, (2) viewed over time for the same target machine (e.g., looking at repeated attacks using the same attack method) whether anything has been learned during the exercise—or potentially, even between exercises (if similar team composition returns to an event in which the same attack vector is repeated), and (3) for understanding whether the BTs are able to transfer learned knowledge (e.g., is a BT able to detect and defend the same type of attack against a different target system provided they have learned it earlier).

Thirdly, analyzing the timestamps provides insight into the BTs' strategies. Do the BTs only focus on certain class or difficulty-level of attacks, and maybe miss some more important/unknown challenges? Do they invest time during the exercise to understand the systems? The metrics enable a way of getting some basic baseline and benchmarking for the organizers and participants.

It is important to note that the timestamps themselves only measure effectiveness. However, there is an implicit assumption that measuring changes in effectiveness over time (e.g., repeated comparable events, such as repeated attacks), shows changes in performance. This is an indicator for learning, a dynamic process, together with other qualitative data. The complete exercise data analysis and projections are left for future work, and the scope of this paper discussing the suitability of proposed methodology with the community.

Challenges and Limitations. It is also important to acknowledge the challenges and limitations. The learning measurement process needs to be pre-planned, agreed with the stakeholders, and form an integral part of a CDX organization and evaluation process. Selection of what to measure is a challenging task and depends on training objectives. What learning metrics are "must have", "nice to have" and "wasteful" metrics from learning perspective? Having comparative metrics from several CDXs, would enable developing comparable standardized set of learning metrics.

Data monitoring and collection may fail to capture timing metrics and team actions with perfect reliability. Also a challenge is to develop clearly defined measures that integrate both qualitative and quantitative inputs. Metrics for future evaluation should include appropriateness and quality of responses and actions. Some training goals (such as incident handling procedures) may prove difficult to measure due to teams following different operating procedures, standards, and practices. Separating learning impact from other behavior effects (i.e., learning might not be visible straight away or recognized by participants by themselves, or overestimated and not result in behavior change) will remain a challenging area to assess.

2.2 Data Collection and Sources

The data collected as part of CDXs may vary based on training objectives and software environment, but it should not be an additional burden to the organizers. As shown by the 5-timestamp methodology, often such data is already collected. The learning related data is obtained from several sources:

1. RT reporting—failed attacks, resistance time to the attacks, number of repeated attacks;
2. YT reporting—reporting about situational awareness;
3. Scores—scoring for availability, usability and injects (trends over time);
4. Traffic from game-net and management-net;
5. Surveys—pre-exercise and post-exercise survey with pre- and post knowledge assessment if possible;

6. Injects—can be used to qualitatively verify a data sample from overall dataset (see below) and collecting learning feedback during the exercise;
7. Information from the RT—ratings for resistance level, classification of attack type (this might also be semi-automated by using Cobalt Strike [13] or similar);
8. Observations of the BTs;
9. Communication channels—chat logs, GT management network traffic (volumes and trends);
10. Interviews with participants (and management)—assessing the immediate reaction to exercise and long term impact on the job.

Sample Selection for Qualitative Validation. Due the large volume of virtual machines, attacks, and activities, it is not be possible to confirm all incidents during an exercise qualitatively as it may distract the BTs from learning. However, for a sample of attacks qualitative feedback can obtained from the BTs in order to cross-check the metrics. Such sample should be designed into the exercise as inquiries to the BTs via injects and/or observations.

The sample selection depends on the exercise training objectives, however should cover differing aspects, such as complexity, method of attacking, ease of detecting and mitigating the attack. There is no widely accepted taxonomy that can be applied from learning perspective in CDXs context. In order to measure learning impact, a comparison between easy tasks (potentially nothing learned and knowledge is already existing) and complex tasks (more challenging, more potential to learn) is valuable. As the teams have differing skillset any such criteria classification is somewhat forced and arbitrary, however it provides a comparison and feedback on the appropriate difficulty levels and learning opportunities created by the organizer. We propose the following classification matrix in Fig. 4, when a selection of specific events for learning impact measurement is based on: (1) detection and analysis—some attacks and incidents have visible signs that can be easily detected, whereas others are almost impossible to detect. (2) mitigation and recovery—responding to an incident involves different skillset and actions to be taken containing the damage, eradicating the incident components, and restoring systems to normal operation, and remediating vulnerabilities to prevent similar future attacks.

Easy to Detect and Easy to Mitigate	Easy to Detect and Difficult to Mitigate
Difficult to Detect and Easy to Mitigate	Difficult to Detect and Difficult to Mitigate

Fig. 4. Sample selection matrix

In addition to those two criteria and as an incident is part of whole exercise (scenario, mission), the prioritization of attacks (strategy) needs to be considered.

3 LS17—Learning Measurement

LS provides full experience of managing a major cyber incident to the BTs. The exercise consists of different attacks and tasks based on a scenario over two days and the data set is over 2500 attacks [14]. The measurement plan needs to ensure that the intrusive data collection is not distracting the participants' focus from learning efforts.

LS17 learning measurement included a mixture of quantitative and qualitative methods with focus on gaining some experience using the 5-timestamp methodology, combined analysis of participants' feedback and metrics collected, and identification of plausible learning correlations for further work.

3.1 5-Timestamp Methodology Experience

We illustrate how the 5-timestamp methodology works using the example of LS17. We picked one high-profile RT objective—a Siemens system part of Industrial Control System (ICS) segment for all timestamps to be recorded. The timestamps were obtained from the BTs self-reporting (through Injects), RT attack reports, and scoring data for all teams. Furthermore, those RT members conducting the attack on those systems were asked to keep a detailed log of all events, as accurately as possible. Regarding the pcaps from management interfaces, there was a technical issue and very unfortunately, the GT was unable to record the traffic from management interfaces; leaving analysis of the inter-arrival times for future work in next exercises.

The purpose of this objective was to gain control of the airport fueling station and cause a fuel leak. The BTs had time to mitigate before "all fuel was spilled". Before the exercise the RT had prepared some attack vectors, but which vector would work or not depends on the BT defenses. Starting the fuel spill is a very "noisy" attack, which means even if the initial compromise remained undetected, the BT had some time respond.

Four teams were successfully attacked by the RT (i.e., all fuel was spilled). For two more the RT managed to compromise the systems and start spilling, however, those two BTs managed to mitigate the attack before all fuel was spilled. The remaining 13 teams defended their systems well (e.g. no spilling started).

While all teams were analyzed, for anonymity and clarity reasons only one timeline is presented here. Figure 5 shows detailed timeline of events recorded according to the 5-timestamp methodology for BT Z (Z anonymized).

Before the RT is allowed to attack in the exercise, the respective objective must be opened. For this specific team and objective, this was done at 06:59 UTC, which corresponds roughly to the time first phase of attacks was allowed to start. The objectives are not opened individually in the RT reporting system, but rather for all teams at the same time—and because a RT member might have to "entertain" several teams at a time, the opening of an objective and actual start of an attack might differ. In this example case, BT Z was only attacked at 07:35 UTC (about 1/2 h later), i.e., the timestamp reported from the detailed

Incident Timeline	Time	Description	Data source
t_1 RT starts an attack	06:59	Campaign officially opened	RT reporting system
$t_{1.1}$ RT starts 1st attack	07:35	Attack started	RT members
t_2 RT compromises	07:40	Spilling started	Scoring
t_2 RT compromises	07:43	Spilling started	RT members
$t_{2.1}$ BT mitigates	07:44	Spilling stopped	Scoring
$t_{2.1}$ BT mitigates	07:45	Spilling stopped	RT members
$t_{1.1}$ RT starts 2nd attack	07:58	Attack repeated	RT members
t_3 BT detects	09:00	Suspicious activity noted	BT Inject
$t_{2.1.3}$ RT reporting	09:18	Partial RT objectives scored	RT reporting system
t_4 BT starts mitigating	09:20	Timestamp or interval reported	BT Inject
t_2 RT compromises	09:23	Spilling started	Scoring
$t_{2.1}$ BT mitigates	09:30	Spilling stopped	Scoring
t_5 BT fights back	09:30	Timestamp or interval reported	BT Inject
t_5 BT resolves	09:40	Suspicious user removed	BT Inject

Fig. 5. Example of 5-timestamps reconstructed timeline for an incident

RT member logs. LS has a comprehensive and automated scoring system, which recorded at 07:40 UTC that the attack has been successful and spilling started. However, the RT members reported that spilling started at 07:43 UTC. This small time difference is an artifact of the self-reporting, and understandable, as all teams are very busy during the exercise. It also highlights that self-reporting timestamps should be avoided, if possible. This is not only for accuracy reasons, but also to reduce the work-load for various teams during the exercise. Similarly, the scoring system reported that the BT mitigated the attack at 07:44 UTC, while the RT member recorded a timestamp of 07:45 UTC. Such minor discrepancies were observed throughout. As this attack has only partially been successful, the RT does not give up and manages to gain foothold in the systems again at 07:58 UTC (reported by RT member log), but this time the RT does not manage to cause any fuel spilling. This is not recorded in the scoring (and should not be scored as the BT successfully defended), but it is an important factor that hints at resistance and team performance. Having such timestamps facilitate a reflective team debrief after the event.

However, when analyzing the BT self-reporting then the BT only reports detecting any suspicious activities for the very first time at 09:00 UTC. Clearly, some BT members must have mitigated the attack already before 07:44 UTC, so this points to an intra-team communication/reporting problem. Therefore asking the BTs to self-report accurate timestamps, while defending systems during a "crisis situation", is not going to work (neither observations). The team's internal reporting systems do not capture such information, or at least not accurately enough. It is therefore of vital importance to obtain such timestamps from the management network (e.g., by observing in the pcaps when a BT member logs into the target system, or in case they are already logged in when the activity of system changes by a changed inter-arrival frequency of packets on the management network).

Overall during the exercise, spilling attempts start 7 more times using at least two different attack vectors. The first time spilling was for 3′39″ (3 min and 39 s), the second spilling continued for 6′44″. The next day spilling durations were significantly reduced, in the end only taking 0′07″ (7 s) to mitigate—despite the fact that different attack vectors were used.

The main challenges encountered in the process and assumptions for data quality are:

1. RT scoring timestamps from the system need to be sufficiently accurate—when attacking multiple teams the objectives are started for all teams simultaneously and final scoring is often delayed, so scoring data is not accurate;
2. BTs self-reporting is not reliable and more accurate data collection method is required—this supports the argument that non-intrusive methods for collecting and analysing data from logs (network traffic, log, etc.) is helpful;
3. Traditional observations methods are not possible as in a technical exercise there is nothing to see.

Of course, this is a first attempt to understand the feasibility of proposed methodology. Before drawing any conclusions on learning more data and measurements needs to be obtained in future work, however, such initial tests appear to be promising.

3.2 Discussion and Findings from LS17 Learning Measurement

In addition to the 5-timestamps methodology experience, we also discuss selected findings relevant from other learning measurements, as the 5-timestamps methodology is one integral part of the overall measurement framework. Only aggregated statistics are presented due to the confidential nature and privacy of participants. We used pre-survey, injects, post-survey and interviews to collect the feedback from individuals and teams. Our overall response rate was 21% for individual pre-survey, and team based injects had 89% response rate. Due to the timing constraints, post-survey results have not been included.

Learning in Pre-exercise Phase. We collected information about the participants, their experiences and learning process in the pre-execution phase, team environment, learning expectations about the execution and evidence of long-term learning from previous exercise participation. Our findings confirm that pre-exercise phase is vital part of the overall exercise with 53% of respondents spending 10–50 h preparing. Majority of participants (73%) however report that they prepared individually (over half of the preparation time); whereas sub-teams preparations were taking place either half of the time (35%) or seldom (31%). Despite that the exercise is team-based, whole team preparations were mostly seldom 37% and 22% of participants claim they never attended whole team preparation sessions.

No clear distinction between learning knowledge/skill on technical vs. soft skills learned in pre-exercise phase is visible—on average in each learning

area 40% minor and 13% significant improvement was reported. This links to the fact that training audience consist mainly of the professionals, who assess their knowledge and skills in majority at medium (43%) to high level (37%), related to the similar working experience level (both medium and high 39%). When comparing what the participants have learned in preparation phase and what they expect to learn during the exercise, there is no clear distinction that some training objectives (e.g., teamwork) are more relevant for execution than pre-execution phase or that technical skills are mainly obtained in the pre-exercise phase.

Feedback on the Exercise Design. Feedback was focused on the Industrial Control System (ICS) segment design that has been designed and seen as one of the most complex and technically challenging areas in the exercise. By comparison we can draw conclusions also on other parts of the exercise. The attempt was also made to assess, how teams perceive individual team members/sub teams and whole team learning outcome.

Based on pre-survey 52% of respondents felt they do not have ICS capabilities in their teams, despite of nearly all teams reporting dedicated ICS team member(s). Average self-believed resistance level in the ICS segment was surprisingly low compared to the RT members' assessment—44% believed that their resistance was at medium level, 33% at high level 22% strong. This links positively to an assumption that learning can happen when team acknowledges they lack some knowledge or skills and "sensing more than see" (OODA loop). It is also interesting to see how the teams perceive level of difficulty to defend against those attacks—41% find it easy to detect and easy to mitigate, 39% easy to detect, but difficult to mitigate, 12% difficult to detect, easy to mitigate and 8% find it difficult to detect and difficult to mitigate. In comparison to other attacks in the exercise 44% of teams assessed level of difficulty at same level. Priority for ICS attacks was consistently (78% of teams) at high or critical priority level, as expected in the scenario. 52% of the BT reported that they managed to track the root cause of malicious activity and 42% not (showing missed learning opportunity without proper feedback).

An attempt to evaluate how individual and team learning outcomes are perceived, shows that team learning has quite even distribution (25%) from slight to significant improvement, then individual learning was in majority (59%) assessed as significant. This is somewhat expected, due to the technical specialization but therefore needs further focus of learning transferability within a team.

Furthermore, a question collecting narratives about the teams' learning experience to uncover and understand the big picture was asked. Top 5 expressions that emerged were "successes in learning" ("learning curve"), "challenges in learning", "complexity/variety of system", "preparations" and "team learning". All these themes confirmed the focus of measurement objectives.

Long-Term Learning Impact. We looked at LS16 for long-term impact indicators, based on responses of returning participants, and enquiries with previous LS participants. When asked individually about skills learned and maintained 69%

responded that they recall a skill from participating earlier LS—average for technical training objective is 67% and soft skill related training objective 69%. Sadly survey results were limited in the participants' comments what exactly they learned.

58% agreed that their team has become more coherent, confident and collaborative. Similarly, 64% agreed that their team's knowledge has increased (as a result of individuals sharing). However, as majority (59%) of the teams have changed significantly (less than 50% old team members)—long-term impact need to be interpreted carefully.

Feedback from few participants who participated over five years back tends to indicate long term impact of CDXs on mindset (e.g., "to have an emergency procedure in place, as when you're in the middle of the event there is no time to think, just to act.", "... key is thinking and mindset and learning why something was done, not what.").

Despite of limited evidence, the survey and interview results support the learning value of the LS. Further work needs to be conducted to evaluate long-term impact for specific training objectives.

4 Related Work in Learning Measurement Context

Unfortunately, there are no widely accepted methodological evaluation methods published and scientifically proven measuring learning impact or assessing cyber security skills/competencies obtained through CDXs. Some general guidance, such as [10,18], describe how organizers should look at design and performance (training success) measurements. The related work includes articles published on learning (and other) measurements at cyber exercises, and also interdisciplinary papers, game based learning and team learning, as relevant.

Cyber Exercises. Dr. Ahmad [1] investigated how a cyber crisis exercise benefits participants' individual learning and how their experience in the exercises is transferred to their organization using the four-level Kirkpatrick training post-assessment model (reaction about the exercise, learning skills experienced during the exercise, behavior developed during the exercise, and result, i.e., how the benefits are transferred to their organization). This approach lacks team aspects of learning. U.S. Army Research Institute for the Behavioral and Social Sciences Research [22] measured game-based simulations by different questionnaires and complemented those interviews with probing questions.

Game Based Learning and Serious Games. Connolly et al. [3] proposes a model for the evaluation of games for learning that includes motivational variables such as interest and effort, as well as learners' preferences, perceptions and attitudes to games in addition to looking at learner performance. Outcomes relate to learning and skill acquisition but also affective and motivational aspects. Methods to evaluate learning outcomes include meta-analyses, randomized controlled trials, quasi-experimental designs, single case experimental designs—pre- and post

test, and non experimental designs—surveys, correlation, qualitative [7]. The effect on learning (acquisition of skills or knowledge) was measured by calculating the difference between pre-test and post-test scores on the questionnaires or cognitive tests, and comparison to control group [5]. Game-based learning and serious games provide excellent environments for mixed-method data gathering (i.e., triangulation), including crowd sourcing, panel discussions, surveys and observations, in-game logging and tracking on hundreds of events and results, including distances, paths, play time and avoidable mistakes, etc. [12]. Not yet explored issues are seamless, or "stealth" data-gathering and assessment as well as performance based evaluation [7]. Stealth assessment (i.e., non-invasive, non-intrusive assessment) could potentially increase the learning efficacy given that much of the learning remains relatively "implicit" and "subjective" [12]. These issues are very relevant in the CDXs context, and the cornerstones for the proposed 5-timestamp methodology in this paper.

Team Learning. Measuring team learning is a complex task with many factors, such as learning impact has not been identified (i.e., simply there is no similar event in reality), change can be environmental (i.e., not due to learning) and learning could be dysfunctional (i.e., false connections made between actions and outcome) [26]. Most common methods used are combination of interviews, surveys, questionnaires, observations, and learning maps. Edmondson [4] used observation and interviews (based on "informant sampling approach") to study role of teams in learning and based on her study half of the teams engaged in reflective discussion about process that led to subsequent changes, and would constitute a team learning. Newman et al. [16] measured critical thinking during group learning using a questionnaire and the content analysis method, whereas Hay [8] used concept mapping on the topic before and after. Learning maps or curves at team and organization level were used by Uzumeri et al. and Chiva et al. [24,25]. Two valuable points to note are: (1) the learning is not necessarily consciously accessible, thus asking the team members (survey or interview) what they have learned may not uncover any changes, however there might be learned patterns that members were not consciously aware of [4,26]; (2) measuring long-term learning effect requires detailed and multiple real-time observations of the same group over time [26].

Other Measurements Conducted at CDXs. Some team performance and effectiveness metrics also relate to learning measurement. Study about Baltic Cyber Shields 2010 team effectiveness [6] used different interdisciplinary methods and concluded that a combination of technical performance measurements and behavioral assessment techniques are needed to assess team effectiveness, and cyber situation awareness is required for the defending teams, but equally for the observers and the game control. In Cyber Shield 2015, Henshel et al. [9] attempted predicting proficiency in the teams and to identify the best training and assessment methods by pre- and post-event survey and data collection during the event, and developed proficiency metrics, such as Time-to-Detect, Time-to-apProval, Time-to-End and Category Correct. Reed et al. [20] evaluated

cyber defender situation awareness, and showed that the most pervasive form of competition-based exercise is comprised of jeopardy-style challenges, which compliment a fictional cyber-security event. Silva et al. [21] study considered factors of successful performance in Tracer FIRE exercise with emphasis on the use of software tools and bring out a relevant consideration that speed is often not the main consideration—participants who devoted more time to challenges tended to make more correct submissions (similar finding to [9]).

Findings for CDXs. From CDXs viewpoint, all these methods are applicable. However, similar challenges are faced as by researches so far—i.e., separating learning from other factors and that learning might not be necessarily visible. Also, for the incident response teams activities are conducted on the computers/network—thus observations of behavior (sitting quietly behind computer screen but at the same time mitigating a significant risk or attack) might not provide sufficient information. Observation method should be seen with a different kind of eyes—on the network and system-level and to learn observing at such technical levels.

The key takeaways for CDXs are that learning measurement needs to use mixed-method approach with qualitative and quantitative data, have wide scope, provide comparison ("benchmarking"), consider both individual and team aspects, and ideally be non-intrusive (not distracting participants from main goal of learning).

5 Conclusion

Learning is such a complex and intractable process that its study is difficult and contentious. However, methodological measurement is required to conclude whether an exercise design was appropriate and effective, and planned learning outcomes were achieved.

We presented an idea for non-intrusive data collection and measurement, i.e., the 5-timestamp methodology as an integral part of overall learning measurement framework. Future work should continue with performing the data analysis of an exercise to compile learning metrics and trends benchmark. Identification and analysis of the data trends, will provide solid baseline and demonstrate learning improvement achieved in CDXs. This will complement often anecdotal and positive feedback obtained via traditional methods (surveys, interviews) that participants have actually learned. As we demonstrated, incorporating non-intrusive, social and behavioral research methods into the cyber security field can give new insights and possibilities in effective training for cyber defense teams in the future.

We explored CDXs' learning measurement state of play and presented inter-disciplinary literature review, incorporating relevant findings from team (group) and game based learning studies. The findings support the proposed novel non-intrusive 5-timestamp methodology for mainly timing and accuracy metrics for measuring technical skills improvements, but equally incorporating team aspects

and soft skills. As part of the methodology proposal, we also considered some practicalities of data collection and proposed practical validation approaches with the qualitative measurements.

With work performed in this paper, we have attempted to provide practical steps how the organizers can evidence the learning value and lessons learned at CDXs, and at the same time improve the participants' and teams' learning experience by providing valuable feedback based on such measurement data.

Acknowledgments. This work would not have taken place without the NATO CCD COE open-minded and friendly organizing team of LS17, who allowed the authors to experiment on this large cyber exercise.

References

1. Ahmad, A.: A cyber exercise post assessment framework. Malaysia perspectives. Ph.D. thesis, University of Glasgow (2016)
2. Ahmad, A., Johnson, C., Storer, T.: A cyber exercise post assessment: adoption of the Kirkpatrick model. Adv. Inf. Sci. Serv. Sci. **7**(2), 1 (2015)
3. Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T., Boyle, J.M.: A systematic literature review of empirical evidence on computer games and serious games. Comput. Educ. **59**(2), 661–686 (2012)
4. Edmondson, A.C.: The local and variegated nature of learning in organizations: a group-level perspective. Organ. Sci. **13**(2), 128–146 (2002)
5. Girard, C., Ecalle, J., Magnan, A.: Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. J. Comput. Assist. Learn. **29**(3), 207–219 (2013)
6. Granasen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises-a cross-disciplinary case study. Cogn. Technol. Work **18**(1), 121–143 (2016). Springer-Verlag, London
7. Hauge, J.B., Boyle, E., Mayer, I., Nadolski, R., Riedel, J.C., Moreno-Ger, P., Bellotti, F., Lim, T., Ritchie, J.: Study design and data gathering guide for serious games' evaluation. In: Connolly, T.M., Hainey, T., Boyle, E., Baxter, G., Moreno-Ger, P. (eds.) Psychology, Pedagogy, and Assessment in Serious Games, pp. 394–419 (2014)
8. Hay, D.B.: Using concept maps to measure deep, surface and non-learning outcomes. Stud. High. Educ. **32**(1), 39–57 (2007)
9. Henshel, D.S., Deckard, G.M., Lufkin, B., Buchler, N., Hoffman, B., Rajivan, P., Collman, S.: Predicting proficiency in cyber defense team exercises. In: 2016 IEEE Military Communications Conference, MILCOM 2016, pp. 776–781. IEEE (2016)
10. Kick, J.: Cyber exercise playbook. Technical report, DTIC Document (2014)
11. Mattson, J.A.: Cyber defense exercise: a service provider model. In: Futcher, L., Dodge, R. (eds.) Fifth World Conference on Information Security Education. IFIP – International Federation for Information Processing, vol. 237, pp. 81–86. Springer, Boston (2007)
12. Mayer, I., Bekebrede, G., Warmelink, H., Zhou, Q.: A brief methodology for researching and evaluating serious games and game-based learning. In: Psychology, Pedagogy, and Assessment in Serious Games, pp. 357–393. IGI Global (2014)
13. Mudge, R.: Cobalt Strike. https://www.cobaltstrike.com. Accessed 18 Sept 2017

14. NATO CCD COE: Locked Shields (2017). https://ccdcoe.org/locked-shields-2017.html. Accessed 18 Sept 2017
15. NATO CCD COE: Locked Shields 2016 After Action Report. NATO Cooperative Cyber Defence Centre of Excellence Publication (2016)
16. Newman, D.R., Webb, B., Cochrane, C.: A content analysis method to measure critical thinking in face-to-face and computer supported group learning. Interpersonal Comput. Technol. **3**(2), 56–77 (1995)
17. Ogee, A., Gavrila, R., Trimintzios, P., Stavropoulos, V., Zacharis, A.: The 2015 Report on National and International Cyber Security Exercises. https://www.enisa.europa.eu/publications/latest-report-on-national-and-international-cyber-security-exercises
18. Patriciu, V.-V., Furtuna, A.C.: Guide for designing cyber security exercises. In: Proceedings of the 8th WSEAS International Conference on E-Activities and Information Security and Privacy. World Scientific and Engineering Academy and Society (WSEAS), pp. 172–177 (2009)
19. Pusey, P., Gondree, M., Peterson, Z.: The outcomes of cybersecurity competitions and implications for underrepresented populations. IEEE Secur. Priv. **14**(6), 90–95 (2016)
20. Reed, T., Nauer, K., Silva, A.: Instrumenting competition-based exercises to evaluate cyber defender situation awareness. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2013. LNCS, vol. 8027, pp. 80–89. Springer, Heidelberg (2013). 10.1007/978-3-642-39454-6_9
21. Silva, A., McClain, J., Reed, T., Anderson, B., Nauer, K., Abbott, R., Forsythe, C.: Factors impacting performance in competitive cyber exercises. In: Proceedings of the Interservice/Interagency Training, Simulation and Education Conference, Orlando, FL (2014)
22. Singer, M.J., Knerr, B.W.: Evaluation of a game-based simulation during distributed exercises. Army Research Institute for the Behavioral and Social Sciences, Orlando, FL (2010)
23. Stytz, M.R., Banks, S.B.: Addressing simulation issues posed by cyber warfare technologies. SCS M&S Mag., no. 3 (2010)
24. Svetlik, I., Stavrou-Costea, E., Chiva, R., Alegre, J., Lapiedra, R.: Measuring organisational learning capability among the workforce. Int. J. Manpower **28**(3/4), 224–242 (2007)
25. Uzumeri, M., Nembhard, D.: A population of learners: a new way to measure organizational learning. J. Oper. Manage. **16**(5), 515–528 (1998)
26. Wilson, J.M., Goodman, P.S., Cronin, M.A.: Group learning. Acad. Manag. Rev. **32**(4), 1041–1059 (2007)

Appendix 2

II

M. Kont, M. Pihelgas, K. Maennel, B. Blumbergs, and T. Lepik. Frankenstack: Toward real-time red team feedback. In *IEEE Military Communications Conference (MILCOM)*, pages 400–405. IEEE, 2017

Frankenstack: Toward Real-time Red Team Feedback

Markus Kont, Mauno Pihelgas, Kaie Maennel, Bernhards Blumbergs and Toomas Lepik

(DOI: 10.1109/MILCOM.2017.8170852)

Frankenstack: Toward Real-time Red Team Feedback

Markus Kont
NATO Cooperative Cyber Defence Centre of Excellence
markus.kont[a]ccdcoe.org

Mauno Pihelgas
NATO Cooperative Cyber Defence Centre of Excellence
mauno.pihelgas[a]ccdcoe.org

Kaie Maennel
Tallinn University of Technology
kamaen[a]ttu.ee

Bernhards Blumbergs
NATO Cooperative Cyber Defence Centre of Excellence;
IMCS UL, CERT.LV Laboratory
bernhards.blumbergs[a]cert.lv

Toomas Lepik
Tallinn University of Technology
toomas.lepik[a]ttu.ee

***Abstract*—Cyber Defense Exercises have received much attention in recent years, and are increasingly becoming the cornerstone for ensuring readiness in this new domain. Crossed Swords is an exercise directed at training Red Team members for responsive cyber defense. However, prior iterations have revealed the need for automated and transparent real-time feedback systems to help participants improve their techniques and understand technical challenges. Feedback was too slow and players did not understand the visibility of their actions. We developed a novel and modular open-source framework to address this problem, dubbed *Frankenstack*. We used this framework during Crossed Swords 2017 execution and evaluated its effectiveness by interviewing participants and conducting an online survey. Due to the novelty of Red Team-centric exercises, very little academic research exists on providing real-time feedback during such exercises. Thus, this paper serves as a first foray into a novel research field.**

***Keywords*—automation, cyber defense exercises, education, infrastructure monitoring, real-time feedback, red teaming**

I. Introduction

Cyber defense exercises (CDX) are crucial for training readiness and awareness within the *cyber domain*. This new domain is acknowledged by NATO alongside with land, sea, air, and space [1]. Alliance nations are endorsing the development of both defensive and responsive cyber capabilities. In this context, the paper focuses on further evolving the quality and learning experience of CDX, aimed at developing cyber red teaming [2] and responsive skillset. Crossed Swords (XS) [3], a techical exercise developed by NATO Cooperative Cyber Defense Centre of Excellence (NATO CCD COE) since 2014, is used as a platform to create the proposed framework. The solution is applicable to any other CDX where standard network and system monitoring capability is available.

A. Background

XS is an intense hands-on technical CDX oriented at penetration testers working as a single united team, accomplishing mission objectives and technical challenges in a virtualized environment. While common technical CDX is aimed at exercising defensive capabilities (i.e., Blue Team – BT), XS changes this notion, identifies unique cyber defense aspects and focuses on training the Red Team (RT).

To develop and execute the exercise, multiple teams are involved: rapid response team (i.e., RT); game network and infrastructure development (Green Team – GT); game scenario development and execution control (White Team – WT); defending team user simulation (i.e., BT); and monitoring (Yellow Team – YT).

The RT consists of multiple sub-teams based on the engagement specifics, those being: network attack team, targeting network services, protocols and routing; client side attack team, aiming at exploiting human operator and maintaining access to the hosts; web application attack team, targeting web services, web applications and relational databases; and digital forensics team, performing data extraction and digital artefact collection. These sub-teams must coordinate their actions, share information and cooperate when executing attacks to reach the exercise objectives.

The main goal is to exercise RT in a stealthy fast-paced computer network infiltration operation in a responsive cyber defense scenario [4]. To achieve this, the RT must uncover the unknown game network, complete a set of technical challenges and collect attribution evidence, while staying as stealthy as possible. Note that XS is not a *capture-the-flag* competition, as the RT has to pivot from one sub-objective to another in order to achieve the final mission according to the scenario. Furthermore, Red sub-teams are not competing with each other, and rather serve as specialized branches of a single unit.

B. Problem Statement

Prior XS iterations revealed several problems with RT learning experience. Primarily, the YT feedback regarding detected attacks from the event logs and network traffic was presented at the end of every day, which was not well suited to the short, fast and technical nature of the exercise. The feedback session addressed only some noteworthy observations from the day, but RT participants need direct and immediate feedback about their activity to identify mistakes as they happen. This feedback needs to be adequately detailed, so that the RT can understand why a specific attack was detected and then improve their approach. Finally, to make the feedback faster, the slowest element in the loop—the human operator—needs to be eliminated.

Therefore, manual data analysis by the YT needs to be automated as much as possible. To achieve this, we used the same open-source tools as in the previous XS iterations, but added in event correlation, a novel query automation tool, and a newly developed visualization solution. We decided to call the framework *Frankenstack*. Fig. 1 illustrates the role of Frankenstack in the XS exercise.

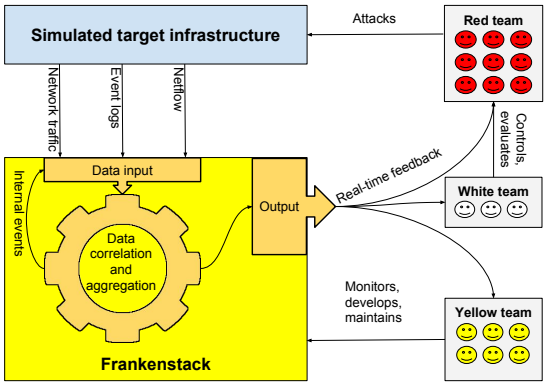

Fig. 1. High-level overview of Frankenstack

The RT has to receive timely and efficient feedback from the YT regarding detected attacks on the target systems. This feedback is critical to raise the level of stealthiness, identify the gaps of RT coordination, and analyze the tools and tactics used for computer network operations. The effectiveness of our framework was assessed during the main execution of XS 2017 (XS17), where the stack provided real-time monitoring feedback to the RT.

The remainder of the paper is organized as follows: section II provides an overview of related work, section III describes our monitoring stack, section IV presents RT feedback results, while section V discusses future work, and section VI concludes the paper.

II. Related work

For teaching purposes, the benefit of exercises and competitions is generally well accepted and documented [5], [6], [7], [8]. Unfortunately, not much research has focused on the perception of feedback which is provided to the training audience, especially in the context of monitoring technical indicators of compromise in realistic environments. Thus, this section presents research related to both measuring and improving the learning experience as well as situation awareness (SA) during cyber exercises.

Dodge et al. discussed CDX network traffic analysis in [9], a practice that is common in modern exercises not only for situational awareness (SA) but also as educational tool, for elaborating attacker campaigns, for training network analysts, etc. However, this early paper focuses on traffic capture and initial profiling, and does not consider distractions such as traffic generation, increasing infrastructure complexity, host instrumentation, data source correlation, or the need for immediate feedback. In [10], Holm et al. correlated network traffic and RT attack logs from Baltic Cyber Shield, a precursor for Locked Shields and Crossed Swords exercises. However, their goal was to improve existing metrics for vulnerability scoring, as opposed to participant education. Likewise, in [11], Brynielsson et al. conducted a similar empirical analysis on CDXs to profile attacks and create attacker personas.

In [12], Arendt et al. presented CyberPetri, a circle-packing visualization component of Ocelot, which was previously presented in [13] as a user-centered decision support visualization. They presented several use cases of the tool, but their main goal was high-level feedback to network analysts based on target system service availability reports. Although the tool was useful for high-level decision making, technical RT members are more interested in immediate effects of their attacks on target systems. Note that any single system is often a supporting pillar for more complex services, and is not noticeable to end-users. Nevertheless, modern security monitoring is built upon instrumentation of these systems, to find RT *footprints* and to trigger notification upon breaching these digital tripwires.

A paper [14] by Henshel et al. describes the assessment model for CDXs based on the Cyber Shield 2015 example, as well as integrated evaluation of metrics for assessing team proficiency. In addition to data collected during the exercise, they also conducted a pre-event expertise survey to determine possible relationships between prior expertise and exercise performance. For future assessments they suggest that near real-time analysis of the collected data is required—they stress that raw data collection is not a problem, but the capability to meaningfully analyze is the limiting factor. Manual methods do not scale with the huge amounts of incoming data. This closely coincides with our observations in section I-B and this is what we aim to improve.

Furthermore, existing academic research commonly relies on monolithic tools, which are often not accessible to the general public, thus, making experiments difficult, if not impossible, to reproduce. We seek to provide an inexpensive open-source alternative to these products. The next section describes our modular monitoring architecture.

III. Frankenstack

Commercial tools are too expensive for smaller cyber exercises, in terms of licensing fees, hardware cost and specialized manpower requirements. Detection logic in commercial tools is also not available to the general public, which hinders YT's ability to provide detailed explanations of detected attacks. Frankenstack is easy to customize as individual elements of the stack are industry standard tools which can be interchanged. Note that we opted to use a commercial tool *SpectX* as an element within Frankenstack for log filtering, due to on-site competency and developer support. However, this function could have been achieved with the open-source Elastic stack [15]. Our stack provides a clear point of reference to other researchers and system defenders who wish to compile the monitoring framework in their particular environments, as the overall architecture is novel.

The data available to us during XS included full ERSPAN (Encapsulated Remote Switched Port ANalyzer) traffic mirror from gamenet switches and NetFlow from gamenet routers. This was provided by the GT. Furthermore, we instrumented

gamenet systems to collect numerical metrics (e.g., CPU and memory usage, and network interface statistics) and logs (e.g., syslog from Linux, Event Logs from Microsoft Windows, Apache web server access logs, and audit logs from Linux command line executions). Such host instrumentations are very difficult to implement in a standard CDX with BT training focus: if the intent is to give BTs full control of a simulated infrastructure, then they also have full volition to disable these tools. However, as the XS training audience is the RT, then we could maintain control of all target systems and ensure a constant stream of monitoring data. Moreover, we complemented the list of BT data sources with various YT honeypots and decoy servers.

Detailed overview of the resulting stack, in relation to data processing pipelines, is presented in Fig. 2. The blue area represents available data sources, the gray area stands for data storage, and the yellow area denotes the YT presentation layer (i.e., visualization tools on five monitors). Blue and green elements represent target systems and all other elements outside colored boundaries are processing tools. Custom tools that we developed are highlighted with a dark yellow circle. Note that some tools, such as Moloch, are designed for both data storage and visualization, but are not presented in these respective areas because only their API components were used for processing automated queries.

We opted against using NetFlow data, as modern packet capture analyzers (e.g., Suricata, Bro, and Moloch) can fill this role, albeit by needing more processing power and memory. Additionally, these tools commonly present their output in textual log format, which we fed back into the central logging and correlation engine. Thus, the problem of identifying and displaying high-priority IDS alerts can be simplified into a log analysis problem.

Frankenstack uses event correlation for integrating various information sources as this field has been well researched in the context of log management [16], [17], [18]. We open-sourced the correlation ruleset in [19]. See Listing 1 for an example raw log entry from Snoopy Logger [20] that was converted into a more universal human-readable security event that could be presented to the general audience on various dashboards while preserving the raw message for anyone wishing to drill down. Note that specific IP addresses have been removed from this example. This generalization is necessary for handling and grouping subsequent log entries that continue describing the same event, e.g., additional commands executed on the same host via SSH.

Listing 1. Event generalization by frankenSEC

```
#INPUT
login:administrator ssh:(SRC_IP 58261 DST_IP 22)
username:administrator uid:1001 group:administrator
gid:1001 sid:6119 tty:(none) cwd:/home/administrator
filename:/usr/bin/passwd: passwd administrator

#OUTPUT
SRC_IP->[DST_IP]: Command execution by administrator
over SSH
```

Post-mortem analysis of available data sources has proven effective during prior CDXs for packet capture (PCAP) analysis, but requires a significant amount of time and manual work. Again, this clashes with the short time-frame of a CDX. Furthermore, search queries are often written ad hoc during investigations and subsequently forgotten, making analysis results difficult to reproduce. Thus, we created *Otta* [21], a novel query documentation and automation tool for periodically executing user-defined queries on large datasets and converting aggregated results into time-series metrics. *Otta* enables trend graphing, alerting, and anomaly detection for stored user-defined queries. This reduces time spent on analysis and ensures reproducibility by documenting the queries that produced the results.

We used various open-source tools for timelining metrics and log data, for displaying alerts, and presenting correlated information. There are slight differences in handling various incoming alerts. While many types of alerts (e.g., CPU and disk usage) trigger and recover automatically based on a set of thresholds, there are some types (e.g., IDS alerts) that lack the concept of a recovery threshold. Thus, the alert will never recover once raised, leading to an overabundance of information on the central dashboard. Furthermore, batch bucketing methods and timelines are lossy, as only the most frequent items are considered. The volatile nature of CDXs and an abundance of generated network traffic can therefore cause these dashboards to be too verbose to follow efficiently.

Attack maps are not usable because they rely on geographical data which is completely fictional in many CDX environments. Therefore, we developed Event Visualization Environment, or EVE, a novel web-based tool for visualizing correlated attacks in relation to gamenet infrastructure. The Alpha version of this tool has been made publicly available in [22]. EVE is a web application that shows attacks carried out by the RT in real time with a custom gamenet network map as background. Probes can send information to EVE listener in JSON format. Real-time visualization is using WebSocket technology—eliminating the need to reload the page for updated results.

EVE supports combining multiple events in a short time window, and that share the same source and destination addresses, into a unified attack. Resulting attacks are subsequently displayed as arrows connecting the circles around source and target hosts on the network map, while detailed attack information is displayed next to the network map. Using the gamenet map makes EVE a very intuitive tool for enabling participants and observers alike to comprehend CDX events on a high-level.

During the exercise EVE was available only to YT and WT members, as it revealed the entire exercise network map that RT had to discover on their own. However, EVE has a dedicated replay mode to display all the attacks condensed into a given time period, allowing participants to obtain an overview of the attacks as well as understand the pace and focus of the previous attack campaign. For instance, attacks from the previous day can be replayed in 15 minutes. EVE

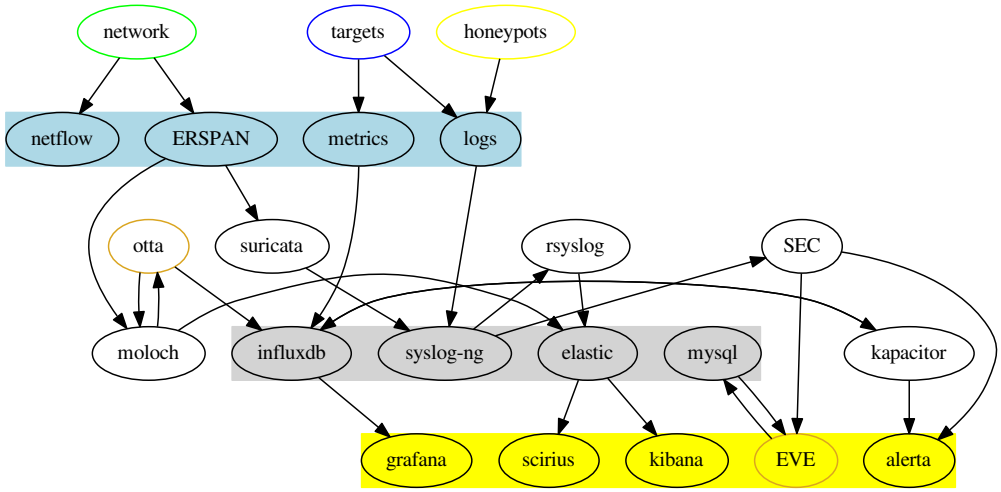

Fig. 2. Data flow between Frankenstack elements during XS17

TABLE I
DEDUPLICATION BY EVENT SOURCE

Event source	Total events	Unique events displayed	Percentage displayed
Apache	1908	35	1.83%
IDS	23790	616	2.59%
Snoopy logger	2962	40	1.35%
Total	*28660*	*691*	*2.41%*

was shown in replay mode to RT participants after the exercise concluded. This compressed replay was very effective in presenting the most prevalent problems, such as periodic beaconing during otherwise silent night periods and verbosity of particular network sub-team attacks.

Alerta [23] served as the primary dashboard to display alerts to the RT. We used the HTTP API for submitting Frankenstack events to Alerta. The RT had direct access to the Alerta web interface and could write their own filtering rules to view information relevant to their current campaign. Finally, we present Tab. I to illustrate how Frankenstack performed in deduplicating the events that were displayed to the RT on the Alerta dashboard. Note that deduplication was primarily based on the generalized event descriptions (see Listing 1).

IV. ASSESSMENT

The tools and infrastructure are essential for learning, but they do not make the exercise successful by default. Often human factors, such as how YT and RT members perceive and use the tools, have significant impact.

One essential part of the assessment was to observe the behavior of the RT members and their interaction with Frankenstack during the exercise in order to gain further insights into their progress and learning experience. We carried out qualitative interviews with RT participants, to estimate their reaction to Frankenstack and their overall learning progress. The interviews took place in casual settings during breaks in execution. Furthermore, we conducted a quantitative survey in the form of an online questionnaire. The survey consists of multiple choice or ranking style questions with the ability to provide additional comments for each question. The survey concluded by asking some general questions about meeting the training objectives and overall satisfaction with the exercise.

A. Feedback combined from interviews, survey and observations

This subsection includes the analysis of participants feedback. Improvement suggestions to learning design are presented in the following subsection IV-B.

We received 14 survey responses out of 27 participants (52%). 46% of participants had attended other exercises, but none of those exercises had attempted to provide SA via a similar toolset. The remaining 54% had not previously attended any exercise.

There were four large screens in the training room directed to the RT, displaying Alerta, Grafana, Scirius, and Suricata. A fifth screen displaying EVE was only visible to YT and WT members. Most RT members preferred to view the main screens displayed in training room, and 38% responded that they checked the screens every 60 minutes or less. Another 38% checked the screens every 30 to 50 minutes. RT members were not restricted from accessing any of the Frankenstack web interfaces. The survey revealed that learners did access the monitoring framework on their local computers when attempting new attack vectors. Thus, tools served their intended usage.

Alerta was considered most useful (46%), followed by Moloch (31%). There was no clear result for the least useful tool. The respondents expressed mixed feelings on the ease of

use of the SA tools: 38% equally agreeing and disagreeing, and the remainder (24%) being neutral.

Regarding learning impact, 79% agreed (of those 57% strongly agreed) that the SA given during exercise is useful for their learning process, while 21% were neutral. In terms of the feedback rate, 77% of the respondents considered the speed of feedback to be at the correct level, 15% considered it too slow and 8% considered it too fast. Furthermore, 57% agreed that alerts were accurate and sufficient for their learning process, while 43% were neutral about this question. However, several respondents revealed being too focused on achieving their primary objectives, and thus unable to properly switch between their tools and feedback screens.

In relation to visibility, 45% of the participants agreed that they had learned a lot about how their actions can be detected (i.e., it is useful to see simultaneously what attack method could be detected, and how), and 30% were more careful with their attacks and thus tried to be stealthier than they normally would have been. However, there were some unintended side-effects. The feedback sometimes provided insight into the network map that the RT was tasked to discover independently. For example, if the RT probes a yet unknown node on a network, the logs generated on the host might reveal the target hostname (e.g., *sharepoint* or *ftp*), which consequently implies the purpose of the system—something that would not be apparent from an IP address. Thus, there is a fine line between revealing too little or too much to the training audience.

Furthermore, some comments revealed a loss of emphasis on stealth due to exercise time constraints, i.e., RT members knowingly used more verbose techniques closer to the objective deadline. To clarify, 64% of respondents confirmed that the SA tools were not distracting them nor had negative impact, while 30% agreed that they were distracted. The remaining 6% were neutral. This confirms the challenges of providing instant feedback, as the learning potential is not fully used. The question is how this learning experience is impacting long-term behavior of the participant.

One of the key training aspects is working as a team in achieving goals. Thus, team communication and cooperation are vital. Overall, 83% of respondents indicated some improvement of the skills for these specific training objectives. However, feedback concerning the impact of SA tools on team communication and cooperation is mixed—50% perceived positive impact, whereas 21% were negative and remainder were neutral. Several respondents acknowledged less need for verbal communication, as they could see relevant information on the screens. Unfortunately, not all RT members were able to interpret and perceive this information correctly. This combined with the reduced need for communication meant that not all participants progressed as a team.

Compared to other CDXs, 50% responded that they needed less information from YT members, as they obtained relevant SA on their own. Guidance, however, is a critical success factor for learning, especially in a team setting. 64% of participants said they had sufficient help for their learning process, i.e., when they did not know how to proceed, their team members or sub-team leaders provided guidance. However, 64% is a rather disappointing result and could clearly be increased with improved learning design. Some respondents admitted that they did not know how other teams were progressing and wasted time on targets that were not vulnerable. This caused significant frustration and stress, especially when combined with the compressed timeframe of a CDX.

B. Learning improvement suggestions

Given the amount of work that goes into preparing such exercises, the level of learning potential needs to be maximized. Our analysis suggests that small learning design changes may have significant impact. This section presents the main recommendations derived from these results.

From the learning perspective, we cannot assume that participants know how to use or interpret the results. Lack of in-depth knowledge of monitoring tools (e.g., where is raw data collected, what is combined and how, what needs to be interpreted in which way, etc.) has a negative impact on learning. A dedicated training session or workshop needs to take place prior to execution. Furthermore, in the light of the survey results, inclusion of various tools into Frankenstack needs to be carefully evaluated to avoid visual distractions for RT participants. There is also a need to reduce prior system and network monitoring knowledge by making the output more self-explanatory.

Given the difficulties in switching between multiple screens whilst also trying to achieve an objective in unfamiliar network, one can easily suggest compressing the amount of presentable information to reduce the number of monitoring screens. However, this cannot be attained without reducing the amount of technical information. The purpose of Frankenstack is not to provide SA to high-level decision makers, but to present feedback to technical specialists. Thus, a better approach would be restructuring each sub-team with a dedicated monitoring station with a person manning it, allowing team members to focus on their objectives and get feedback relevant only for their actions. As such, RT members must be given a *hands-on* opportunity to use monitoring systems.

In RT exercises such as XS, there are several main objectives to be achieved by the whole RT. It is challenging to evaluate reaching objectives, since there are many steps involved in reaching a specific objective. Often the tasks or sub-objectives are divided between sub-teams (network, web and client-side) and between individuals in those sub-teams. The difficulty of a specific exploitation depends on the individual's skillset, which varies widely. Hence, there is a trade-off between assigning a task to an experienced member to increase the chance of success, versus teaching a new member. For example, an experienced network administrator is more effective in exploiting network protocols and is likely less visible while doing so, but may not learn anything new.

Discussions and feedback revealed that several respondents felt they were stuck and working alone. Division of the tasks between sub-teams and individuals also diminishes the learning potential. One training design option to alleviate this issue

would be regular *team timeouts* for reflection. Reflective team sharing is crucial for the learning success of each individual, and would overcome the project management approach where each team member focuses only on personal objectives. Higher emphasis should be on offering tips and helping those stuck on an objective to move forward whilst also keeping track of the feedback provided by Frankenstack. The coaching could also be handled in the form of a *buddy system* where RT members are not assigned a sub-task individually, but in groups of two or three. They would then have to share their knowledge and can benefit from different individual backgrounds.

Finally, it is important to have better time-planning during the execution. While it is certainly appropriate to allow for flexibility in the paths that the RT can take to solve the objectives, participants should avoid spending too much time on wrong targets. Nevertheless, the learning impact of the exercise in this format (i.e., with real-time feedback) is very positive. Only 13% of all participants' responses reported no significant change in their skills, while an overwhelming 87% perceived an improvement in their skill level, and 93% agreed that they were satisfied with exercise.

V. Future work

We encountered several unforeseen problems, as methods for assessing technical RT campaigns have to be incorporated into the game scenario itself. However, most XS17 targets had already been developed before the initial stages of this research. We plan to increase information sharing between Red and Yellow teams to improve RT progress measurement. Thus, we can develop better assessment methodologies for RT skill levels and YT feedback framework.

Development of a new dynamic version of EVE is already underway for the next XS iteration. In addition to the network map view, it can draw the network map dynamically as RT compromises new targets. Currently, EVE can only be used after the end of the exercise. However, in addition to providing more actionable alerting, the new version can also reduce RT work for mapping new systems and allow them to focus on the technical exercise.

VI. Conclusion

In this paper, we have presented the core challenges in organizing a CDX with Red Team emphasis, such as timeliness and accuracy of feedback, and ensuring participant education without compromising the game scenario. We compiled a novel stack of open-source tools to provide real-time feedback and situational awareness, and conducted surveys among the RT members to assess the effectiveness of this method.

Frankenstack feedback regarding learning impact was mainly positive. However, there are critical questions to answer when designing the RT exercises, such as what is the right balance of information to provide to the RT, does the behavior change due to monitoring or information visible (i.e., learners unconsciously limit themselves by not trying out more risky strategies, etc.). Also, some further learning design changes, and not necessarily only limited to SA, can maximize the return on the significant investment into preparing such RT exercises. We hope to spark a discussion on improving these problems.

VII. Acknowledgments

The authors would like to thank Mr. Risto Vaarandi, Mr. Hillar Aarelaid and Prof. Olaf M. Maennel for their valuable contributions. This work has been supported by the Estonian IT Academy (StudyITin.ee).

References

[1] T. Minárik, "NATO Recognises Cyberspace as a Domain of Operations at Warsaw Summit," Available: https://ccdcoe.org/nato-recognises-cyberspace-domain-operations-warsaw-summit.html.

[2] P. Brangetto *et al.*, "Cyber Red Teaming - Organisational, technical and legal implications in a military context," NATO CCD CoE, Tech. Rep., 2015.

[3] "Crossed swords exercise," Available: https://ccdcoe.org/crossed-swords-exercise.html.

[4] P. Brangetto *et al.*, "From Active Cyber Defence to Responsive Cyber Defence: A Way for States to Defend Themselves Legal Implications," Available: https://ccdcoe.org/multimedia/active-cyber-defence-responsive-cyber-defence-way-states-defend-themselves-legal.html.

[5] B. E. Mullins *et al.*, "The impact of the nsa cyber defense exercise on the curriculum at the air force institute of technology," in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, Jan 2007, pp. 271b–271b.

[6] A. T. Sherman *et al.*, "Developing and delivering hands-on information assurance exercises: experiences with the cyber defense lab at umbc," in *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.*, June 2004, pp. 242–249.

[7] R. C. Dodge *et al.*, "Organization and training of a cyber security team," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 5, Oct 2003, pp. 4311–4316.

[8] G. H. Gunsch *et al.*, "Integrating cdx into the graduate program," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 5, Oct 2003, pp. 4306–4310.

[9] R. C. Dodge and T. Wilson, "Network traffic analysis from the cyber defense exercise," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 5, Oct 2003, pp. 4317–4321.

[10] H. Holm *et al.*, "Empirical analysis of system-level vulnerability metrics through actual attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 825–837, Nov 2012.

[11] J. Brynielsson *et al.*, "Using cyber defense exercises to obtain additional data for attacker profiling," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, Sept 2016, pp. 37–42.

[12] D. Arendt *et al.*, "Cyberpetri at cdx 2016: Real-time network situation awareness," in *2016 IEEE Symposium on Visualization for Cyber Security (VizSec)*, Oct 2016, pp. 1–4.

[13] D. L. Arendt *et al.*, "Ocelot: user-centered design of a decision support visualization for network quarantine," in *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, Oct 2015, pp. 1–8.

[14] D. S. Henshel *et al.*, "Predicting proficiency in cyber defense team exercises," in *MILCOM 2016 - 2016 IEEE Military Communications Conference*, Nov 2016, pp. 776–781.

[15] "Elastic stack," Available: https://www.elastic.co/.

[16] R. Vaarandi *et al.*, "Simple event correlator - best practices for creating scalable configurations," in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2015 IEEE International Inter-Disciplinary Conference on*, March 2015, pp. 96–100.

[17] R. Vaarandi, "Platform independent event correlation tool for network management," in *Network Operations and Management Symposium, 2002. NOMS 2002. 2002 IEEE/IFIP*, 2002, pp. 907–909.

[18] ——, "Sec - a lightweight event correlation tool," in *IP Operations and Management, 2002 IEEE Workshop on*, 2002, pp. 111–115.

[19] "Frankensec," Available: https://github.com/ccdcoe/frankenSEC.

[20] "Snoopy Logger," Available: https://github.com/a2o/snoopy.

[21] "Otta," Available: https://github.com/ccdcoe/otta.

[22] "Eve - event visualization environment," Available: https://github.com/ccdcoe/EVE.

[23] N. Satterly, "alerta," Available: http://alerta.io/.

Appendix 3

III

T. Lepik, K. Maennel, M. Ernits, and O. Maennel. Art and automation of teaching malware reverse engineering. In *International Conference on Learning and Collaboration Technologies*, pages 461–472. Springer, 2018

Art and Automation of Teaching Malware Reverse Engineering

Toomas Lepik( ), Kaie Maennel, Margus Ernits, and Olaf Maennel

Tallinn University of Technology, Tallinn, Estonia
{toomas.lepik,kaie.maennel,margus.ernits,olaf.maennel}@ttu.ee
http://cybercentre.cs.ttu.ee/en/home/

Abstract. The threat environment is rapidly changing and the cyber security skill shortage is a widely acknowledged problem. However, teaching such skills and keeping professionals up-to-date is not trivial. New malware types appear daily, and it requires significant time and effort by a teacher to prepare a unique, current and challenging courses in the malware reverse engineering. Novel teaching methods and tools are required. This paper describes an experience with an automated hands-on learning environment in a malware reverse engineering class taught at Tallinn University of Technology in Estonia. Our hands-on practical lab is using a fully automated Cyber Defense Competition platform Intelligent Training Exercise Environment (i-tee) [1] combined with typical Capture-The-Flag competition structure and open-source tools where possible. We describe the process of generating a unique and comparable reverse-engineering challenge and measuring the students' progress through the process of analysis, reporting flags and debugging data, recording and taking into account their unique approach to the task. We aim to measure the students' using the Bloom's taxonomy, i.e., mastering the art of malware reverse engineering at the higher cognitive levels. The presented teaching and assessment method builds foundation for enhancing the future malware reverse engineering training quality and impact.

Keywords: Higher education teaching · Cyber defence exercises
Malware reverse engineering

1 Introduction

The subject of reverse engineering is multi-faceted and difficult to teach. Modern malware has become powerful and complex—it has evolved from simple replicating viruses to highly evasive and adaptable applications. To understand how the malware is working, a cyber security specialist needs essential knowledge how to dissemble and reconstruct the code. Reverse engineering is "a cyber defense task used to investigate malware, construct functionality of compiled software, and identify vulnerabilities from closed-source software code already being used in operational contexts" [2].

P. Zaphiris and A. Ioannou (Eds.): LCT 2018, LNCS 10925, pp. 1–12, 2018.
https://doi.org/10.1007/978-3-319-91152-6_35

With the predicted cyber security skill shortage, we are facing a problem that there are not enough qualified teachers to deliver high-quality courses (e.g., as qualified experts earn more in industry than in academia). Malware reverse engineering is even more challenging due to complex and ever changing natures of threat and changes in malware landscape. Developing a malware reverse engineering course requires a significant amount of effort, as coming up with the unique code that would mimic malware for disassembly task is time consuming [3]. As the industry requires more cyber security specialists with up to date skillset (and consequently universities have a growing number of students), the teaching must respond with relevant and engaging learning methods. This is where automated learning environment becomes a necessity and we are facing a problem that we need to design novel teaching techniques where motivated students can self-learn. Our aim is to automate the process in order to allow a larger number of students to benefit from the course, but ensure in the automation process a unique learning experience is created, which enables the students to master the art of malware reverse engineering.

A practical problem-based, hands-on learning is considered an efficient way to study Information and Communications Technology (ICT) subjects, as it provides an exciting learning experience [4]. Mahoney *et al.* express that reverse engineering is much more of an art and much less of a skill—the students either get it or they do not [3]. On the other side, the human brain constructs the models based on past experiences and based on those experiences produces possible solution for a given problem by predicting solution models based on past experience [5]. Teaching methods in the reverse engineering should provide such "foundational" experiences to solve such complicated problems.

2 Problem Statement and Research Design

In response to the high-demand of the qualified cyber security experts with malware reverse engineering skills, we propose a solution as combination of the automated learning environment and the novel teaching techniques. As the traditional learning methods (e.g., lectures, videos, quizzes) do not allow the scalable hands-on experience that is needed for mastering the art of reverse engineering. Our objective is to create a teaching and assessment platform that allows the teachers to help the students learn the art of malware reverse engineering. The motivation for automation is very desired by the teachers of such courses and would increase the quality of the learning experience.

Our research design was to build and test a scalable hands-on automated malware reverse engineering lab for the university students using the open source tools (or free-ware tools). This paper reports our experience with initial steps towards such a platform, in order to receive feedback from the community on our efforts. We describe the auto generation of malware-like samples with different layers of difficulty and corresponding flags, usage and scaling of Intelligent Training Exercise Environment (i-tee) system [1], and measuring the student behavior during the multi hour final challenge. In this initial stage, we have

tested the system using the malware embedded into PDFs. However, the architectural concepts are not limited to PDFs. and we will add more to our github[1]. In the github, the configuration for the lab described in this paper is available. Note that the i-tee platform is publicly available under MIT license [1].

3 Related Work

Our proposed methodology and system focuses on the teaching and assessing advanced reverse engineering skills. As an educational tool, this work builds upon the other developments in teaching, testing the latest tools and competition-based systems in the cyber defence education.

The Intelligent Training Exercise Environment i-tee [1] is an open source virtual cyber simulator that enables hands-on, practical learning. It allows to simulate realistic cyberattack situation in virtual and sandboxed environment and can be integrated into existing curricula or used to create a new subject or a competition event. A student needs only a web browser and a remote desktop protocol client to start exploring the system [6].

There are magnitude of different Capture-The-Flag (CTF) style competitions that include reverse engineering, the most up to date CTF listing at CTF-Time.org web site [7].

Burns *et al.* [8] analyzed the solutions of about 3,600 Capture The Flag (CTF) challenges from 160 security competitions and describes the security issues that are most concerning to industry and academia. The paper furthermore enumerates the security tools and techniques that are used by the players. In "reverse" engineering challenges, flags are usually obfuscated and embedded in executable programs. Static analysis and dynamic analysis are mostly used to solve the "reverse" challenges.

The paper describes the use of Virtual Box (VBox) image, the students can download and run it in their own computers. However, some challenges are met with such design. For example, the tablet computers do not support VBox and as many reverse engineering tools and libraries are Linux oriented, or the students who use Windows and Mac OS X do not have enough skills to install and setup the needed tools and libraries in their own computers. As result about 13% of students gave up on the exercises due to such issues [8]. In the proposed solution described in this paper, many such challenges can be avoided.

One of such the CTF exercises that includes reverse engineering challenge (understanding the behavior of compiled, obfuscated, or cryptic program code) is picoCTF [9], however targeted at high school students, not university students. The picoCTF challenges are hosted on a Linux server, which students can access either with SSH or through a web client. The system is supported with an IRC-based chat room for students to discuss challenges with the organizers. No detailed overview is provided about the system architecture.

Taylor *et al.* [10] created automated obfuscated challenge for students with randomly generated program in C language used similar obfuscation and auto

[1] https://github.com/toomasl/RE-PDF-DO.

generated virtual machines and made students after completion to upload the virtual machines with answers. This tool be also used in teaching code reverse engineering techniques as described in the paper, and is relevant as it describes the possibilities how to understand that the students have completed de-obfuscation task. Our proposed system address several issues raised in this paper, such as automatic assessment of whether a student used known tools that they were used in the class, randomised malware sample allocation to a student, etc.

Mahoney *et al.* [3] describes a course design however the course design did not describe automated tools. Our paper focuses on advancing such courses with automated teaching assistant methods and tools.

There are limited publications on teaching the reverse engineering using the automated open-source system architectures via hands-on CTF style learning methods. This is the gap we are aiming to fill with the work described in this paper.

4 Teaching the Art of Malware Reverse Engineering

4.1 Mastering the Reverse Engineering

Mahoney *et al.* writes that the "hand holding" did not help, some students never picked up the skill set. [3]. "Reversing equals art" thinking is not new and stated before in computing education, e.g., Bader has used same expression with respect to parallel programming [11]. Modern malware can include multiple obstacles [12] for evading detection, which complicates the malware analysis. In phase when students get graded they should be familiar with some of them beside obfuscation [10].

The studies have found that the human brain constructs the models of the world on the basis of past experience, which are subsequently confirmed or denied by experiential input [5]. When we apply this predictive mechanism of the brain to the cognitive task of problem solving we see that our brain produces possible solution for a given problem by predicting solution models based on past experience [5]. Therefore, it can be argued that a solution to a specific task might be similar due students' participation at the same malware course, i.e., similar learning provided generates inherently predictive solutions to particular set of problems. However, we can use such predictive patterns in case we need to understand students ability to use particular tool-set. This type of tool-set enables the use of different measurement techniques to understand whether the student tackled the multifaceted problem or if the student used an allowed (or not allowed) shortcut to a solution.

Our approach is to leave the playing field open and let the student choose the tools and methods to tackle the problems—they can use they own tools and programs that are only limited by an operating system. However, that implies that the problem of assessing skills automatically becomes more complicated, as the measuring will depend on reverse engineering steps that the students should perform. For example, the we do not simply measure the fact of installation

of a specific tool by the student, rather whether they understand the reverse engineering process.

4.2 Assessing the Learning Outcomes

To achieve and assess learning outcomes the educators use frameworks such as Bloom's or SOLO (Structure of the Observed Learning Outcome) taxonomy. Whatever learning framework is used, it means that higher level of cognitive levels or complexity levels are reached by a student to ensure that he/she masters the art of malware reverse engineering. For example, in SOLO the learning outcomes are classified in terms of their complexity, enabling us to assess students' work in terms of its quality not of how many bits of this and of that they have got right [13]. Under Bloom's taxonomy, the learning objectives describe six progressive levels of learning: knowledge, comprehension, application, analysis, synthesis, and evaluation [14]. In this paper we refer to the updated version of Blooms's taxonomy by Anderson *et al.* [15], who explain the the levels as follows: 1. Remembering: Learner's ability to recall information 2. Understanding: Learner's ability to understand information 3. Applying: Learner's ability to use information in a new way 4. Analysing: Learner's ability to break down information into its essential parts 5. Evaluating: Learner's ability to judge or criticize information 6. Creating: Learner's ability to create something new from different elements of information.

As teaching malware reverse engineering is rather an art and malware is commonly well protected to resist analysis—thus higher, i.e., elevated cognitive skills are required and should be evaluated as part of a course assessment. However with many Capture the Flags (CTFs) the common criticism is that they only reach the lower levels of cognition. For example, Moses *et al.* analyzed CTF competition of Cyber Security Awareness Week (CSAW) Conference of the New York University Polytechnic School of Engineering and concluded that the vast majority of challenges met objectives corresponding to levels 1–3, the challenges with the lowest completion rates typically involved multiple learning objectives at levels 3–4, and there is a complete absence of challenges mapping to level 5–6 [15]. The learning design for the malware reverse engineering course, needs to overcome such challenges and therefore our system allows teachers to consider what cognitive levels they are aiming to teach and assist in achieving learning outcomes for higher levels.

In malware reverse engineering a learner needs to make tool-choices and recreate from different elements of information. Thus, demonstrate comprehension of applying the knowledge, analyze and evaluate in the process of dissembling the code, creative thinking and higher-level concepts. Considering Bloom's taxonomy higher cognitive levels need to be achieved, the lab design provides also the reflection questions to ensure that a student understood and not only completed learned (memorized) technical tasks without understanding.

5 System Architecture

In this section, we describe the architecture of the automated system built on i-tee platform going over how it is integrated, i.e., how malware is integrated to the lab environment, how malware analysis is enabled/supported by the system (i.e., what tools the student has available) and how the automated scoring/assessment is built in.

5.1 Integration to i-tee System

Figure 1 shows when a student connects over the internet to the i-tee platform and is prompted with the choices to connect the lab. The lab connection is supported over the multiple environments, i.e., it does not matter what equipment (e.g., Windows, Linux, Mac OS tablet) the students are using. Note that i-tee is a completely isolated environment, where only the remote desktop viewing part is connected to the Internet. None of the malware run within this framework can "escape", and thus making the system suitable experimentation platform even for the inexperienced students.

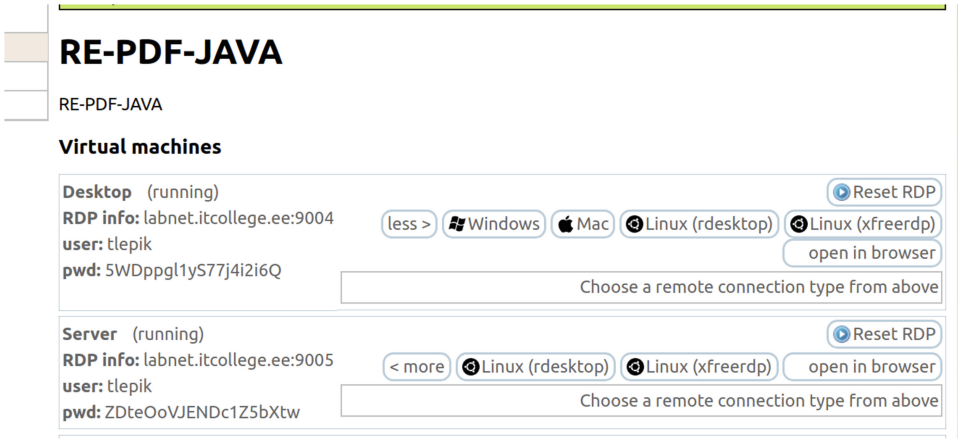

Fig. 1. Automated setup of an completely isolated reverse engineering lab using i-tee [1]. This is the view the student is first confronted with, and allows them to setup the labs and start the exercise. Systems are provisioned on demand and thus save resources.

Once connected to the lab, the student will receive instructions on the challenge (shown in Fig. 2).

The students are prompted with a situation when they act as malware analysts and a user in their organization has received an email with a suspicious attachment. This attachment has been now forwarded for their analysis.

The student's task is to reverse engineer the malware included in the PDF document. The malware for each student is different in order to ensure students

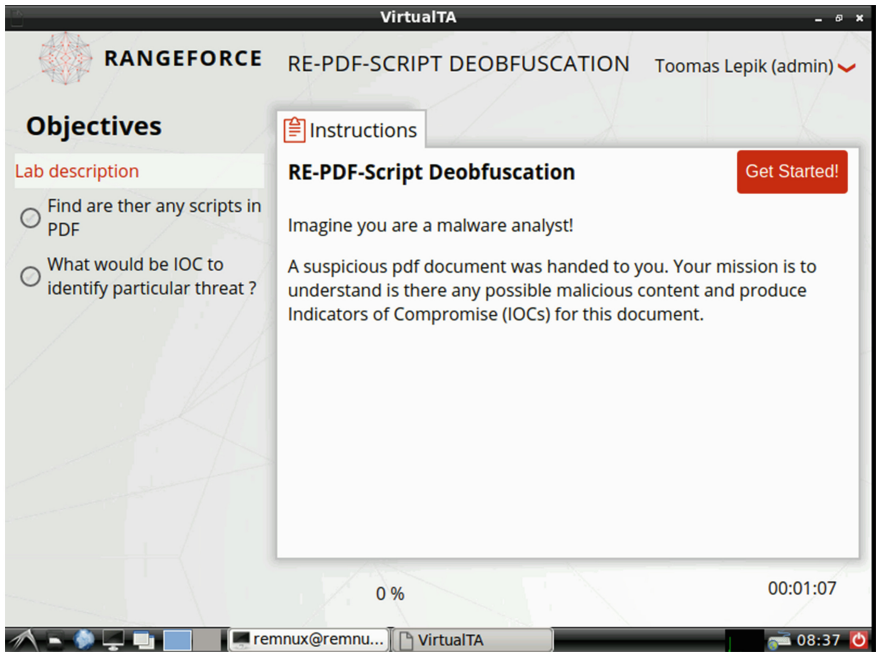

Fig. 2. Lab Instructions. A virtual teaching assistant (VTA) is guiding the student through the tasks in a step-by-step fashion. The VTA component is not an open source tool, however it can be replaced by other learning management systems (e.g., Moodle, etc.).

are not just copying solutions, but actually running the commands and using the tools—even if they get some help from their friends.

The ultimate objective is a discovery of an adversary infrastructure, including download and command and control servers (C&C servers) and channel, i.e., computers/servers that are used to remotely send the malicious commands to a botnet, or a compromised network of computers.

5.2 Malware Analysis Steps

Once the student has been provided a PDF document, the analysis will need to take place. The basic steps that the student needs to perform are:

1. collect the evidence—calculate file hash,
2. identify that PDF contains possible malicious elements including JavaScript extract JavaScript/scripts,
3. de-obfuscate JavaScript layers.
4. understand PDF content and propose hypothesis who might be the target

This is achieved by finding the following planted evidence:

1. connection test—usually the malware is trying to connect to well-known site, such as youtube.com, google.com, etc. (that is a first layer—simple obfuscation base64)
2. Downloads at different levels of obfuscation, such as:
   - website (i.e., second layer)
   - uri (i.e., third layer)
   - browser header (i.e., forth layer)
3. Other more complex obfuscations.

The de-obfuscation works as peeling the onion, breaking apart the from outside to the inside. Firstly after an extraction in one of the JavaScripts extracted from the PDF, a student might find something like this: YXBwLmxhdW5jaFVSTCgiaHR0cDovL3d3dy5nb29nbGUuY29tLyIsIHRydWUpOwo=.

This string needs to be evaluated, de-obfuscated or decrypted by a student. The learning curve can be monitored and identified in what stage (i.e., layer) the student gets stuck.

5.3 Malware Generation

We use the open source tools in order to auto-generate the PDF with malicious script(s). For the first iteration we used real malware samples from the github users, such as from Jodesva[2]. Scripts are in different maliciousness categories and in the real world, when the victims would open PDF documents with such scripts, their machines might be get abused. In the isolated i-tee environment, compromised machines can simply restored from the latest clean snapshot. For achieving the learning impact, using latest and real malware gives additional feeling of realism and allows to keep the learning and assessment up-to-date and unique.

As the specific objective for this assessment was malware incorporated into PDF documents, we used make-pdf-javascript.py[3] that allows to create a simple PDF document with embedded JavaScript that will execute upon opening of the PDF document. It is essentially a glue-code for the mPDF.py module, which contains a class with methods to create headers, indirect objects, stream objects, trailers and XREFs. The created PDFs were further modified and combined with existing random pdf with PDFftk[4].

When a student starts a lab, a set of machines will be provisioned: a lab router (a Linux system that also forwards packets, where students do not have access to), Server and REMNUX desktop. A unique PDF document is generated for each student in time when they first start a lab environment. The PDF creation takes place on the lab router that student does not have assess to and from there are distributed to the student's desktop by a simple script.

[2] https://github.com/jodevsa/malicious-pdf-javascript.
[3] https://blog.didierstevens.com/programs/pdf-tools/.
[4] https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/.

Before PDF creation, the learning environment is synced with a git pull command from the teacher's closed git environment. Through that the teacher can push additional changes to the lab as needed.

5.4 Assessment Scenarios

The scenario presented to the student is more realistic compared to simple obfuscated code analysis, because it involves plausible storyline typical for a good CTF-s. During the test for the first iteration of the lab, the students had 3 h time to engage with the tool. The actual amount of time taken for resolving first iteration was varying between 42 min (fastest) and 2 h 48 min (slowest) student. Average time is about 1 h 54 min. During the reverse engineering course, the students are given homework with similar content and complexity. We also asked from the students how much time they spent for their homework, and the answers were varying from 30 min to 8 h.

For assessing the command line usage the Snoopy[5] was used. However, due to a deployment error, logs from the student machines were not available and command line usage analysis was only be done manually (using forensic tools). We randomly chose some machines for further analysis. We noted the students used command line tools native to REMUX and discussed in class, for example: pdfid.py[6], Peedf.py[7]. For JavaScript analysis, the students mostly opted for using online de-obfuscators. This was in line with the expectations and tools covered during the course.

5.5 Student Assessment

The labs are not only used as supporting hands-on teaching material accompanying the lectures, but we also used them in the final exam to assess the learning objectives of the course. The students were expected to demonstrate the use of PDF analysis and JavaScript de-obfuscation skills using tools that are native or easy install on REMNUX[8] Linux distribution.

For successful completion of the task, the students need to find specific Indicators of Compromise (IoC), that would describe particular PDF and embedded code such as network evidence domain names to be resolved, exploits that could be used, and so on. The students also had to answer a questionnaire regarding the discovered IoCs, the exam questions were also auto-generated and aligned with the hands-on malware reverse engineering tasks to be completed.

For an automated student's progress checking, the system uses strings generated at the PDF creation time and the checks are done regularly from the lab router that is not accessible for the student. The lab router remotely checks the directory that the students are instructed to use as working directory and where

[5] https://github.com/a2o/snoopy.
[6] https://blog.didierstevens.com/programs/pdf-tools/.
[7] http://eternal-todo.com/tools/peepdf-pdf-analysis-tool#releases.
[8] https://remnux.org/.

the PDF is located, and provides automated assessment of the completion stage of the task.

6 Learning Impact and Evaluation

The assessment was performed for 10 university students and unstructured qualitative feedback was asked.

Some examples of the student experiences: One student said that first (s)he did not know how to begin, then started analyzing the basics from the file, the hashes, learning that it had JS in it, then the next step was to extract the code, once extracted the next step was to de-obfuscate it and try to learn what it does. The assessment allowed to apply several tools and methods learned along the course within the tools provided in the lab environment necessary to do the tasks. On other hand, some see that virtual environment is not necessary and rather be given an injected PDF by link and analyze it locally. This however does not scale, as more complex topology may be required.

In regards of learning impact the students pointed out that "is interesting as it points out more similar to a real world scenario", "we can apply every tool and method we have learned", "however the instructions could include some tips". During the assessment the student noted that there was no time to "experience all the REMNUX utilities hidden inside but during exam however I have analyzed everything before time was not enough to analyze PDF and encrypted and obfuscated Java script and shell script code." So it should be considered that using the similar labs in the learning process would increase learning impact when used as part of the course, as when only used in assessment the unfamiliarity with the tools is having negative effect on demonstrating their skills.

Overall, the feedback was encouraging as the students provided positive feedback, however there are some further work required in regards inclusion of the automated labs in course design (i.e., use it already for homeworks) and technical improvements (e.g., widen the malware types to be analyzed—not only PDFs).

6.1 Limitation and Future Work

The concept described in the paper was only tested out on the malicious JavaScript included in the PDFs, therefore future work includes extending the scenario and the lab to address the elements of download server, compiled executable analysis, command and control communication and command control structure discovery. The similar lab structure can be easily converted to address different scenarios learning objective in the malware reverse engineering, and thus cover the wide array of malware elements and types.

Another future enhancement we are planning to incorporate is a task of writing up the yara rules[9], the descriptions (i.e., rules) of malware families based on textual or binary patterns consisting of a set of strings and a boolean expression which determine its logic. This would help to measure and ensure that the

[9] http://virustotal.github.io/yara/.

Bloom's higher cognitive skills have been achieved—i.e., the student is able to create something new from different elements of information (to create).

7 Conclusion

This paper presented the initial work and case study for design and architecture of the learning environment based on open source tools to effectively deliver hands-on automated virtual labs to teach malware reverse engineering.

The contribution of this paper is an improved hands-on lab architecture with partly automated assessment. The proof of concept was presented for using open source tools in teaching malware reverse engineering, and combining that with learning theories to ensure the learning objectives are met and the appropriate learning experiences are created. Teaching the malware reverse engineering is an art, but with the right mindset and tool-set the teachers can significantly increase the teaching quality and potentially reduce the amount of students who never pick up the required skill set.

Acknowledgements. This work is partially supported by the European Regional Development Fund (Tallinn University of Technology project VOSA - 2014–2020.4.01.16-0038)

References

1. Ernits, M., Tammekänd, J., Maennel, O.: i-tee: a fully automated cyber defense competition for students. ACM SIGCOMM Comput. Commun. Rev. **45**, 113–114 (2015)
2. Sisco, Z.D., Dudenhofer, P.P., Bryant, A.R.: Modeling information flow for an autonomous agent to support reverse engineering work. J. Def. Model. Simul. **14**(3), 245–256 (2017)
3. Mahoney, W., Gandhi, R.A.: Reverse engineering: is it art? ACM Inroads **3**(1), 56–61 (2012)
4. Eagle, C.: Computer security competitions: expanding educational outcomes. IEEE Secur. Priv. **11**(4), 69–71 (2013)
5. Hohwy, J.: The Predictive Mind. Oxford University Press, UK (2013)
6. Ernits, M., Kikkas, K.: A live virtual simulator for teaching cybersecurity to information technology students. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2016. LNCS, vol. 9753, pp. 474–486. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39483-1_43
7. CTF Time. Accessed 02 Jun 2018. https://ctftime.org/ctfs
8. Burns, T.J., Rios, S.C., Jordan, T.K., Gu, Q., Underwood, T.: Analysis and exercises for engaging beginners in online CTF competitions for security education. In: USENIX Workshop on Advances in Security Education (ASE 17), USENIX Association (2017)
9. Chapman, P., Burket, J., Brumley, D.: PicoCTF: A game-based computer security competition for high school students. In: 3GSE (2014)
10. Taylor, C., Colberg, C.: A tool for teaching reverse engineering. In: USENIX Workshop on Advances in Security Education (ASE 16), Austin, TX, USENIX Association (2016)

11. Gorder, P.F.: Multicore processors for science and engineering. Comput. Sci. Eng., 9(2) (2007)
12. Bisson, D.: The four most common evasive techniques used by malware, April 2015. Accessed 01 Jan 2018. http://www.tripwire.com/state-of-security/security-data-protection/the-four-most-common-evasive-techniques-used-by-malware
13. Biggs, J.B., Collis, K.F.: Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome). Academic Press, New York (2014)
14. Buchanan, L., Wolanczyk, F., Zinghini, F.: Blending Bloom's taxonomy and serious game design. In: Proceedings of the 2011 International Conference on Security and Management, vol. 11, SAM, July 2011
15. Moses, K.V., Petullo, W.M.: Teaching computer security (2014)

Appendix 4

IV
K. Maennel, S. Mäses, and O. Maennel. Cyber hygiene: The big picture. In *Nordic Conference on Secure IT Systems*, pages 291–305. Springer, 2018

Cyber Hygiene: The Big Picture

Kaie Maennel( ), Sten Mäses, and Olaf Maennel

TalTech University, Tallinn, Estonia
{kaie.maennel,sten.mases,olaf.maennel}@taltech.ee

Abstract. Cybercrime is on the rise and it's widely believed that an appropriate cyber hygiene is essential to secure our digital lives. The expression "cyber hygiene" appears in conversations, conferences, scientific articles, legal texts, governmental publications and commercial websites. However, what cyber hygiene is, what is appropriate or optimal cyber hygiene, or what is really meant by this expression and related practices—that is often varying and even somewhat contradicting. We review and analyze selected academic papers, government and corporate publications with the focus on implicit and explicit definitions of what cyber hygiene means to the authors. We also draw parallels and contrast the expression in cyber security context and terminology (cyber awareness, behavior and culture). We present a conceptual analysis and propose a definition to assist in achieving a universal understanding and approach to cyber hygiene. This work is intended to stimulate a clarifying discussion of what appropriate "cyber hygiene" is, how it should be defined and positioned in the wider cyber security context in order to help changing the human behavior for achieving a more secure connected world.

1 Introduction

Human factor is increasingly targeted by cyber criminals. A lot of work is being done to improve "cyber hygiene"—a term that can be broadly perceived as creating and maintaining online safety. Unfortunately, the definition of "cyber hygiene" and its related practices are often varying, and sometimes even somewhat contradicting, therefore hindering the efforts to protect the information assets. The lax use of the term can lead to situations where efforts to improve cyber hygiene are not considering the context and have either too mild or too strong effects. For example, some phishing awareness trainings can create so much fear in employees that they do not open any e-mail attachments anymore, including legit ones from paying customers, which has a negative impact on a company's productivity [1].

The expression "cyber hygiene" appears in the academic publications, advertisements of commercial cyber security products, and everyday news. However, it is not used consistently. For example, Wikipedia [5] indicates that cyber hygiene relates to an individual, whereas the European Union Agency for Network and Information Security (ENISA) refers to the organizational health (i.e., their

N. Gruschka (Ed.): NordSec 2018, LNCS 11252, pp. 291–305, 2018.
https://doi.org/10.1007/978-3-030-03638-6_18

study [15] focuses on cyber hygiene programs targeted at businesses). In popular media, the importance of cyber hygiene is often stressed, e.g., "For citizens the most important thing they have to understand is cyber hygiene." [9], or used ambiguously, e.g., "[the organization] could have protected itself with proper patching and better cyber hygiene" [20].

In this paper, we aim to provide a definition for "cyber hygiene" based on literature review. We analyze selected academic papers, government and corporate publications with the focus on implicit and explicit definitions of cyber hygiene. We aim to gather existing knowledge on cyber hygiene and learn its current use and positioning in information security. The objective is to stimulate a discussion within the community. The paper intends to be an initial step towards a commonly accepted understanding of cyber hygiene. To the best of our knowledge, this is the first work on a deeper dive on cyber hygiene meaning.

2 Results of Literature Review

The underlying research design consists of two phases. Firstly, conducting literature search. We put our focus on research papers in 2001–2018 in major scientific databases. In addition, we also consider papers and brochures published by governmental and corporate organizations. Secondly, the identified literature is manually reviewed and analyzed for the purpose of definition cleaning and applying this knowledge in cyber security context.

2.1 Cyber Hygiene in Academic Literature

In our research design we focus on the term "hygiene", to see how this word has embedded itself into academic literature in the cyber security context. The search is limited to academic journals and book chapters, as peer reviewed and credible academic content. The list of pre-defined search terms and databases is shown in Table 1. The table presents total search results per database as of February 2018. The manual review is limited to first 200 results in each database, as relevance of the papers diminishes and likelihood of finding another topical article is found to be low. The numbers in brackets indicate those papers, where the term is used in cyber security context.

There are several attempts to define "cyber hygiene", but in many instances the term is used in different contexts without clearly defining it. We firstly look at the full definitions provided, followed by implications from the context analysis. As often no clear definition is provided, it has resulted in the various forms of interpretations and uses of the expression.

Kickpatrick [44] quotes an industry expert who defines cyber hygiene as "implementing and enforcing data security and privacy policies, procedures, and controls to help minimize potential damages and reduce the chances of a data security breach." The definition is broad, in essence aiming to incorporate procedures and controls of cyber defense in the organizational setting. The main focus of the article is to give an overview of the market for cyber insurance. Practicing

Table 1. Scope of literature review search on cyber hygiene and similar terms

Database hygiene	Cyber hygiene	Cyber(-) hygiene	Cyber(-) security hygiene	Digital hygiene	IS hygiene	Internet hygiene	Online hygiene
GoogleScholar	493 (9)	24 (3)	41 (0)	104 (0)	11 (0)	41 (0)	26 (0)
Scopus	16 (6)	7 (6)	2 (2)	431 (0)	102 (0)	0 (0)	539 (0)
ACM Digital	7 (0)	0 (0)	0 (0)	14 (0)	49 (1)	11 (0)	1 (0)
EBSCOHost	4 (4)	4 (4)	4 (4)	6 (1)	1 (0)	15 (0)	24 (0)
IEEEXplore	69 (2)	2 (2)	90 (4)	610 (0)	208 (1)	267 (0)	436 (0)
ScienceDirect	195 (13)	1 (1)	171 (0)	41 (0)	8 (0)	39 (0)	76 (0)
SpringerLink	25 (1)	25 (1)	3 (0)	1 (0)	0 (0)	1 (0)	2 (0)
Taylor & Francis	16 (3)	1 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)

cyber hygiene is brought out also in other cyber insurance related articles, e.g., that "cyber-hygiene is important, but this needs to be proven" [25].

Pfleeger *et al.* [57] define security hygiene as "ways to encourage users of computer technology to use safe and secure behavior online" and discuss how to persuade individuals to follow simple, fundamental processes protecting themselves and others. The term is used more widely, not only in the cyber context. The main focus of the article is user awareness and training. Similarly with training focus, Kiely *et al.* [13] say that in information security management people "must not only practice fundamental security "hygiene"—that is, implement security processes and procedures such as strong and frequently changed passwords, separation of duties, and so on—but also receive added training for securing enterprise data, communications, and so on (especially in more complex enterprise systems)." Also, others use the term in training context, e.g., "cyber hygiene that trains an educated workforce to guard against errors or transgressions that can lead to cyber intrusion" [35].

O'Connell [52] describes that a good cyber hygiene is "an essential step in maintaining a good cyber defence is applying best practices and educating everyone legitimately using the Internet on good network hygiene." The author says that due to increased cyber risk, the "standards for cyber hygiene have elevated, especially for those who have access to vital information" [52]. This paper does not define the cyber hygiene, but attributes it to the individuals by indicating that a good hygiene can be taught and through this cyber hygiene base line increased. The main focus is "identification and application of rules with a far better chance of keeping the Internet open and safer for all" [52]. Almeida *et al.* [23] say that "cyber hygiene initiatives aim at using cybersecurity best practices to appropriately protect and maintain systems and devices connected to the Internet".

Dodge *et al.* [31] describe "cyber hygiene" as a cybersecurity role of each employee with computer, equal with employee responsibility to safeguard his or her door keys or access codes (comparison to physical world). Singer [63] uses

the expression "to observe basic cyber hygiene" and brings an example of an organization getting compromised via a memory stick left in the parking lot (only defined by example). In other cases, the authors only provide an analogy, e.g., "best practices starting at an early age, potentially equating good cybersecurity citizenship with good hygiene such as the importance of washing hands" [61].

Sheppard *et al.* [62] see it more as a perception, i.e., employee's "cyber-hygiene mentality" to prevent the spread of a cyber-attack caused by people opening infected email links or organizations having lax password security processes. They say that cyber hygiene extends to an organization's supply chain and that the lack of cyber hygiene hampers the organization's ability to respond. Thus, cyber hygiene and adequate protective measures are seen as an approach to mitigate the consequences of cyber-attacks. The authors bring out inter-company scope that is not usually mentioned in other articles.

Maybury [49] classifies fostering cyber hygiene (e.g., encrypting data at rest/in motion, effective identity management, passwords) as part of asymmetry principle under operations and maintenance. The author also points out that much of today's cyber hygiene efforts are toward human element and predicts that soon they need to focus more on design and architecture [49].

Kerfoort [43] says "companies fail to practice basic cyber hygiene" and cyber hygiene is mentioned in the context of adopting best practices and standards. Mouradian says security awareness and training "should also have the goal of cleaning up cyber hygiene across the board" [51]. Sanders discusses the creation of cyber security practices in the organization's culture, including the impact of a good cyber hygiene to an organization, the role of senior executives (C-suite) in responding to cyber attacks, and the employees understanding of cyber security standards [58]. The organizational view is also taken by Beris *et al.* [24], who say that when the organization has ensured security hygiene, this can contribute to the behavior towards compliance. The security hygiene is defined "as process of identifying and re-designing high-friction security" [24]. The hygiene in these examples rather implies organizational policies and culture.

Dobbins [30] claims that attackers mostly exploit poor "online hygiene". The good online hygiene practices include, among others, avoiding malicious email attachments, compromised websites, or infected media; employing antivirus and antispyware scanners; updating applications, software, and operating systems within 48 h of patches becoming available, etc. [30] This use of expression combines behavioral and technical measures.

However, many authors simply focus on technical measures. [48] refers to an industry expert: "...how you've configured your firewall or do you have a firewall and how is it configured? Do you have AV? Do you have a patching regime in place? It's all good stuff: it's all good cyber-hygiene!". Some other uses in technical context include [59], who says that"...security controls describe basic cyber hygiene, such as maintaining accurate asset inventories and limiting network ports and protocols, and will have limited effect against advanced cyber tactics or even insider threat where there are many more unknowns", and [32], who writes that the best way to mitigate the threat is "just ordinary hygiene:

downloading the patch to keep your software up to date, and making sure your firewalls are operating".

Furthermore, it is commonly claimed that cyber hygiene is a protective measure, e.g., "proper cyber hygiene would prevent most hacking attempts; however, cyber hygiene is not properly implemented in most organizations" [66], "such attacks are made possible because organizations are not doing things like basic cyber-hygiene around patching and understanding where their weaknesses lie" [47], "poor hygiene is a risk factor" [28], and "adapt to better cyber hygiene that will make phishing harder to achieve" [27]. The failures are blamed on the bad hygiene—"the WannaCry attack were criticized for failing to observe basic cyber hygiene" [19],"users avoid patching regularly or practice weak operational security (i.e., cyber hygiene)" [39].

Several authors aim to classify user behaviors and incorporate "hygiene" into their models. Kelley *et al.* [42] classify user security behaviors in two categories—cyber hygiene and threat response behavior. Stanton *et al.* [64] developed a six-element taxonomy of security behavior that varies along two dimensions: intentionality and technical expertise. The lowest level of their categorization is "basic hygiene (novice and benevolent user)"—whose "behavior requires no technical expertise but includes clear intention to preserve and protect the organization's IT and resources." Another example is by Wang *et al.*, who propose e-hygiene model in which human factor is the major vulnerability of the information security; and "Awareness, Capitals and Abilities form the three dimensions that information users must act to minimize the risks of information malice" [65].

Some authors use the term in combination with activity, e.g., "cyber hygiene scans of Internet-facing systems" [16]. This indicates that hygiene can be separated from the person and considered as service, i.e., "the underlying infrastructure is maintained for you, including all patches and required cyber-hygiene" [50].

In Internet of Things (IoT) context, Oravec *et al.* [53] suggest that "Cyber hygiene" strategies may soon expand from current computing technologies and there is need for designing instructional materials in establishing cyber hygiene routines. In [54], Oravec describes that "individuals engage in some minimal cyberhygiene routines". Fabiano [33,34] similarly refers to the need of establishing expert consensus concerning "key risky user behaviors that may undermine cyberhygiene in IoT environments".

Overall, we note that the expression is finding its way into academic literature in the cyber context. However, the "cyber hygiene" has various meanings and used in many differing contexts in the academic literature. There is no common approach whether hygiene has behavioral or technical implications, or whether it is seen at individual or organizational level.

2.2 Cyber Hygiene in Non-Academic Use

For the non-academic publications, we use Google search engine and apply the same keywords as for academic literature. However, as the Internet content is extremely varied and rapidly changing, our research design focuses on finding the main use cases in the United States of America (USA) and European Union

(EU), by international organizations and in the industry guidelines. We use judgment to assess the reliability and relevance of the source and content for our research purpose. We present our findings, starting from the governmental and legal publications as they are in the capacity to set the standards followed by corporate publications. In the cyber security standards and legislation the term "cyber hygiene" is rather implicit by establishing set of baseline practices of safeguarding (controls) to protect against cyber intrusions.

Examples of the USA and EU: In the USA, the "cyber-hygiene" term was brought into public attention in the five-step National Hygiene Campaign in April 2014, that was organized by the Center for Internet Security (CIS) and the Council on CyberSecurity to help preventing hack attacks on computer systems [45] and promote cyber security as a public "health" issue [12]. The five steps [12] were simply expressed as: (1) Count, (2) Configure, (3) Control, (4) Patch, (5) Repeat [45]. An explicit use of the terminology can be found in The Good Cyber Hygiene Bill [18] that was introduced in June 2017—it is still to become a law but the draft suggests the National Institute of Standards and Technology (NIST) to establish a set of baseline voluntary best practices for safeguarding against cyber intrusions that would be updated annually. NIST Special Publication [11], provides a catalog of security and privacy controls to protect organizational operations, organizational assets, individuals, other organizations, and the state from a diverse set of threats including hostile cyber-attacks, natural disasters, structural failures, and human errors. Awareness and training is one of the security controls. There is also a small companies special publication [10] that provides basic recommendations without forcing the business to implement a specific technology. NIST itself offers no definition of cyber hygiene in the glossary [7].

For the EU, ENISA has issued an overview document about cyber hygiene [15]. The Interactive Terminology for Europe promotes the definition of CIS [12]: protecting and maintaining computer systems and devices appropriately and using cyber security best practices [9]. ENISA uses analogy that cyber hygiene should be viewed similarly to personal hygiene and, once properly integrated, it would consist of simple daily routines, good behaviors and occasional checks to make sure the organizations' online health is in optimum condition [15].

Despite all the Member States having developed their national cyber security strategies, such strategies have rarely (only in the United Kingdom (UK), France and Belgium) translated into direct cyber hygiene programs that would provide guidance around what constitutes good practice, according to [15]:

- The UK has Cyber Essentials guidance to identify the basic technical controls required to defeat the vast majority of cyber attacks. There are only 5 control areas and the emphasis is very much on physical infrastructure controls [4];
- France has set 40 Essential Measures for a Healthy Network, produced by ANSSI10. The foundation guide covers 13 control areas and suggests in-depth approach. Those controls are focused around standard office systems (separate

guidance is available for SCADA/ICS systems) [2]. Because of the size and perceived complexity of the 40 rules, there is a cut down version of 12 rules to assist small to medium size enterprises [8];
- Belgium has a high level Cyber Security Guide that is split into two parts: (1) 10 Key Security Principles which should be adopted by every business, and (2) 10 "must do" security actions which look to turn the principles into more accessible guidance. It also includes a self-assessment questionnaire [3].

All these initiatives focus largely on the organizational cyber hygiene from a perspective of technical controls of the organization's IT system. The human aspects are considered in various degrees (mainly with focus on awareness) and various levels of emphasis, e.g., Belgium's guidance first principle is "implement user education and awareness" compared to UK 5 cyber essentials that include none. France recommendation list includes "RULE 39 - Make users aware of the basic IT rules." ENISA emphasizes need for a standard approach to cyber hygiene across all the EU [15]. The new voluntary certification process suggested in September 2017 by the European Commission will shape the standardization of cyber hygiene in the EU over coming years.

International Organizations: CIS [12] defines cyber hygiene as a means to appropriately protect and maintain IT systems and devices and implement cyber security best practices. Developed by leading experts in the field of security, the CIS Critical Security Controls (CSCs) are a prioritized, consensus based set of twenty security controls designed to reduce the risk of cyber attack [12]. Controls CSC 1 through CSC 5 are considered essential to success. These are referred to as "Foundational Cyber Hygiene"—the basic things that one must do to create a strong foundation for your defense: inventory authorized and unauthorized devices; inventory authorized and unauthorized software; develop and manage secure configurations for all devices; conduct continuous (automated) vulnerability assessment and remediation; and actively manage and control the use of administrative privileges. In addition, the CSC control 17 "Security Skills Assessment and Appropriate Training to Fill the Gaps" [12] addresses the awareness training by analyzing employees' skills and behaviors. Periodic testing can be used to monitor the awareness level among employees as well to measure the training impact in time. Tripwire report [21] examines implementing security controls that CIS refers to as "cyber hygiene" and reports that many issues stem from a lack of basic cyber hygiene and the organizations need to improve their fundamentals such as addressing known vulnerabilities, ensuring secure configuration, and monitoring systems for changes. The CIS Controls align with top compliance frameworks such as NIST, PCI, ISO, HIPAA, COBIT and others [12].

Industry Initiatives: Payment Card Industry guidelines involve different levels of content for different types of organization roles, e.g., IT administrators, developers, and management. The approach is mainly technical and focuses on

educating the users about security standards and best practices [14]. In cloud and mobile environment, VMware [17] uses cyber hygiene definition when referring to the basic things that an organization should have in place for cyber defense. They propose five core principles of cyber hygiene (1. Least Privilege; 2. Micro-segmentation; 3. Encryption; 4. Multi-factor Authentication; 5. Patching) as a universal baseline. They also note that mandatory education process should be in place for everyone.

3 Analysis and Discussion of Findings

Our literature review demonstrates that there is no commonly accepted cyber hygiene framework and definition. Two themes emerged from the literature: cyber hygiene as standard (set of practices), and cyber hygiene as behavior. Both themes were represented both in individual and organizational context. The literature brought out the interdisciplinary side of the cyber hygiene—it is about both human behavior and technology. Based on the standards, the cyber hygiene aspects are often seen as technological, and human side focuses more on cyber security awareness. What makes finding a common approach more challenging, is that the concept of cyber hygiene is highly subjective. It is a human and business problem, not only an IT problem, and no two individuals or organizations will implement it the same way—that makes it very challenging to implement or measure it consistently. Nevertheless, having a solid foundation and at least somewhat similar understanding will help to create a common baseline.

3.1 Origins, Existing Definitions and Use in Other Disciplines

To start with, it is interesting to define the components of the term "cyber hygiene": (1) Cyber—relating to or characteristic of the culture of computers, information technology, and virtual reality, (2) Hygiene—conditions or practices conducive to maintaining health and preventing disease, especially through cleanliness [6]. As combined and adapted, a simple definition could be as "conditions or practices to stay secure and prevent attacks related to the information technology". When comparing this to the definitions in the dictionaries, then Wikipedia [5] offers the following definition: "Cyber hygiene is the establishment and maintenance of an individual's online safety. It is online analogue of personal hygiene, and encapsulates the daily routines, occasional checks and general behaviors required to maintain a user's online "health" (security)." The further explanation emphasizes that cyber hygiene relates to individual, rather than a group or an organization. Collins Online Dictionary [6] proposes (approval is pending as of August 2018): "Cyber hygiene refers to steps that computer users can take to improve their cybersecurity and better protect themselves online. Cyber hygiene habits need to be inculcated by users while using computing tools."

In order to find a definition for cyber hygiene that aligns with common understanding, it is helpful to understand origins of the word "hygiene". It originates

from New Latin *hygina*, from Greek *hugieina*, from *hugis* healthy [6]. Curtis [29] defines hygiene as "the set of behaviors that animals, including humans, use to avoid infection." The humans appear to have hygiene instincts (reactions that people find hard to explain). Curtis hypothesizes that the disgust is the urge to avoid disease (stimuli) and "the perception of a disgusting cue should almost automatically produce a hygienic reaction" independently from conscious decision making [29]. How can we use this knowledge in cyber hygiene context? The problem is that most people do not see Internet as harmful, so hygiene reaction simply does not kick in. In relation to metaphors used in mental models for security, Camp [26] describes health and hygiene as one of the metaphors in security context and specifies that "different examples and metaphors currently used as inchoate mental models all indicate different responses by the user".

Looking at the ways how the word "hygiene" has been adopted in other disciplines, we use "occupational hygiene" as a comparison. The occupational hygiene definitions include the anticipation, recognition, evaluation and control elements, and as a discipline it aims separating people from unpleasant, hazardous situations or exposures [38].

3.2 A Definition for Cyber Hygiene

We propose the following definition: "Cyber hygiene is a set of practices aiming to protect from negative impact to the assets and human life from cyber security related risks." Therefore, secure behavior (in cyber security context) means implementing cyber hygiene. It should be noted that commonly it is implicitly indicated that the set of practices named "cyber hygiene" are relatively easy to perform. Following basic cyber hygiene should be considered as normal as washing hands before eating (example of traditional hygiene). Nevertheless, similarly to different general hygiene standards in different contexts (e.g., hospital, restaurant, coal mine) cyber hygiene is highly context dependent. The basic level of cyber hygiene depends on security requirements.

In wider context, the cyber hygiene is an outcome of creating and maintaining online safety of individual and organization based on their risk assessments and taking different forms considering the technology they are using. The activities are same, but performed in the different context (or level). Imagine a university and a bank—the organizational type and culture provide different hygiene context. For example, an organization can perform or take responsibility for some of the individual tasks (e.g., patching and software updates are automated and pushed down to employees by an IT department).

We think of cyber hygiene as set of practices performed to protect from cyber harm and usually, it is also implied that such practices are relatively simple to perform. The cyber threats connected to cyber hygiene are mainly focusing on human factor—whether directly (e.g., phishing email inviting to insert sensitive data) or indirectly (e.g., people being not motivated to use long and complex passwords). The "practices" part in the definition indicates behavior that has technological and psychological aspects. There are many models used to explain behavior, see Fig. 1.

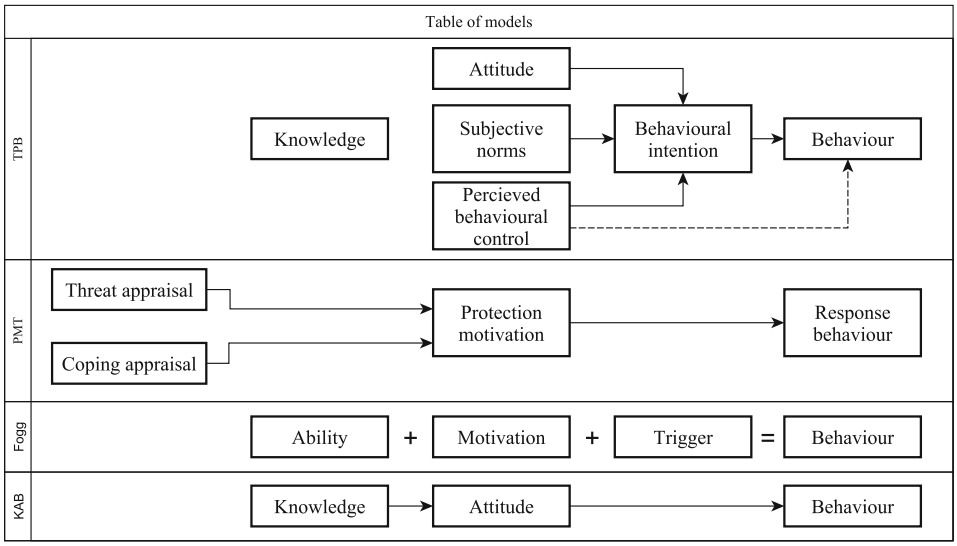

Fig. 1. Overview of selected behavior models to position cyber hygiene

The summary on Fig. 1 presents in comparative way the key elements of Theory of Planned Behavior (TPB) [22], Protection Motivation Theory (PMT) [36], Knowledge Attitude Behavior (KAB) [60] and Fogg's Behavioral Model [37]. The cyber awareness campaigns are aiming to improve the attitude and motivation for a more secure behavior. The security trainings take a step forward and are aiming to increase the knowledge and skills related to the secure behavior.

The cyber hygiene is not the process itself, but the set of practices. Therefore, a cyber hygiene measurement would map out the current practices of the individuals at a timing of the hygiene level evaluation attempt. It is important to note that the behavior depends on context and therefore the set of practices (i.e., cyber hygiene) can be very different in personal and in organizational settings. Different context can limit the set of possible behaviors—e.g., an organization can enforce its security policy by deleting all suspicious emails that are caught by their firewall.

3.3 Related Terminology and Context

The cyber hygiene should be seen in wider context of cyber security and it is helpful to compare and contrast it to some other close terms. We consider in relation to the cyber awareness, behavior and culture to encompass cyber security framework from individual to organization. Figure 2 illustrates the connections between related terms. It uses the KAB (knowledge, attitude, behavior) model described by [56] as the basis to illustrate how cyber hygiene and related terms are connected to each other.

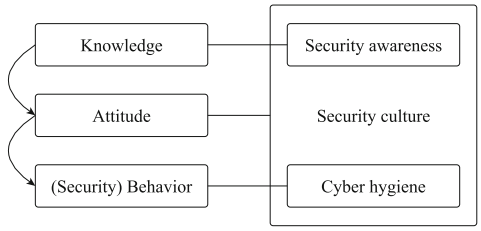

Fig. 2. Illustration of cyber hygiene and related terms

Cyber Security Awareness. Hänsch *et al.* [41] aimed to clarify the term "security awareness" as it also lacked concise definition. They claim that since there is no agreement on the term, different (and sometimes not compatible) ways of raising and measuring security awareness exist—that is a very relevant observation also for cyber hygiene. They analyze the existing literature and conclude that "there is no 'right' or 'wrong' security awareness" and when talking about it, researchers need to express what they mean by it. They conclude that there are at least three ways of interpreting the term—perception, behavior and protection [41]. The awareness brings focus attention on security, and allows individuals to recognize IT security concerns and respond accordingly [7].

Often cyber hygiene and awareness are used interchangeably. Based on our suggested definition, the cyber hygiene is a set of practices while security awareness is commonly used connected to security knowledge. Having good cyber hygiene can be an outcome of awareness, training efforts, individual's attitudes, peer pressure, motives, opportunities, etc. However, the awareness does not necessarily translate into behavior or "good" cyber hygiene practices. The focus of cyber campaigns (e.g., Cyber Security Month, Cyberstreetwise, Stay Safe Online, etc.) is on awareness raising that is a cornerstone for achieving cyber hygiene.

Security Behavior. Security behavior is closely related to cyber hygiene. When cyber hygiene is the set of protective practices, then security behavior shows whether those practices are followed. According to Fogg, the behavior is a product of motivation, ability, and triggers and to perform a target behavior, the person must be sufficiently motivated, have the ability to perform the behavior, and be triggered to perform the behavior at the same moment [37]. From information security viewpoint, Guo [40] proposes a framework for conceptualizing security-related behavior, as there are the divergent conceptualizations and classifies security-related behavior into four categories: security assurance behavior, security compliant behavior, security risk-taking behavior, and security damaging behavior [40]. The taxonomies such as [55] help to determine "good" and "bad" behaviors related to cyber hygiene, i.e., represent desirable and undesirable behavior and are helpful in determining also cyber hygiene levels.

Cyber Security Culture. Security culture based on Mahfuth [46] is "integration process of beliefs, perceptions, attitudes, values, assumptions and knowledge

that guide, direct and manage employees' perceptions and attitudes to influence employees' security behavior or to find an acceptable behavior for employees when they are interacting with the information assets in their organizations." Cyber security culture is a wide term encompassing cyber security awareness, secure behavior and cyber hygiene. Cyber security culture is also often mentioned (e.g., [46]) to affect individual attitude regarding security measures.

4 Conclusion

In order to secure cyberspace, we need to educate every user about the dangers. For an average internet user, "cyber hygiene" trainings will form the basis of understanding. However, in order to make this first line of defense most effective it is important to have a common and solid definition to start from. In this paper, we provided a definition for the term "cyber hygiene" based on extensive academic literature review and selection of corporate and governmental publications in 2001–2018. We analyzed the current usage of expression "cyber hygiene" in different dimensions to provide the comprehensive understanding of how this term is used and positioned in the wider information security context. The results show that cyber hygiene has made its way into the academic and non-academic use, but the meaning and context varies significantly. Our proposed definition is aligned with the common and historical use of the word hygiene and aims to unify the understanding and approaches to support minimizing cybersecurity-related risks. We hope that this paper can spark some discussions within the community to build a solid foundation for a proper and secure cyber hygiene culture in the future.

Acknowledgment. The authors would like to thank Archimedes SA and CybExer Technologies for their support.

References

1. NIST (2018). https://www.nist.gov/video/youve-been-phished
2. Essential Measures for a Healthy Network, ANSSI. https://www.ssi.gouv.fr/en/actualite/40-essential-measures-for-a-healthy-network/
3. Belgian Cyber Security Guide, ICC Belgium, FEB, EY, Microsoft, L-SEC, B-CCENTRE and ISACA Belgium. https://www.b-ccentre.be/wp-content/uploads/2014/04/B-CCENTRE-BCSG-EN.pdf
4. Cyber Essentials-Keeping UK Businesses Safe, CREST. http://www.cyberessentials.org/index.html
5. Cyber hygiene. https://en.wikipedia.org/wiki/Cyber_hygiene
6. Cyber hygiene. https://www.collinsdictionary.com/submission/1930/Cyber+hygiene
7. Glossary of Key Information Security Terms, NISTIR 7298, Revision 2, nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7298r2.pdf
8. Guide Des Bonnes Pratiques De L'informatique, CGPME / ANSSI. https://www.ssi.gouv.fr/uploads/2015/03/guide_cgpme_bonnes_pratiques.pdf

9. IATE: Term of the Week-Cyber Hygiene. http://termcoord.eu/2017/10/iate-term-of-the-week-cyber-hygiene
10. Small Business Information Security: the fundamentals, NIST. http://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.7621r1.pdf
11. Special Publication 800–53 - NIST Computer Security Resource Center. Version 5, August 2017. https://csrc.nist.gov/publications/drafts/800-53/sp800-53r5-draft.pdf
12. The CIS Critical Security Controls for Effective Cyber Defense. Version 6.1. http://www.cisecurity.org
13. Systemic security management. IEEE Secur. Privacy **4**(6), 74–77 (2006). https://doi.org/FEC0FD8D-A181-4AFD-BEA7-AEADF75DEE82
14. Information Supplement: Best Practices for Implementing a Security Awareness Program, Security Awareness Program Special Interest Group PCI Security Standards Council (2014). https://www.pcisecuritystandards.org/documents/PCIDSSV1.0BestPracticesforImplementingSecurityAwarenessProgram.pdf
15. Review of cyber hygiene practices. ENISA, Heraklion (2016). http://publications.europa.eu/publication/manifestation_identifier/PUB_TP0217008ENN
16. US officially accuses Russia of DNC hack while election systems come under attack. Netw. Secur. **2016**(10), 1–2 (2016). https://doi.org/10.1016/S1353-4858(16)30092-7
17. Core Principles of Cyber Hygiene in a World of Cloud and Mobility, VMware, August 2017. https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/vmware-core-principles-cyber-hygiene-whitepaper.pdf
18. The good cyber hygiene bill (2017). https://www.congress.gov/bill/115th-congress/house-bill/3010/text
19. The WannaCry ransomware attack. Strateg. Comments **23**(4), vii–ix (2017). https://doi.org/10.1080/13567888.2017.1335101
20. The week that was, 29 October 2017). https://www.thecyberwire.com/issues/issues2017/October/WTW_2017_10_29.html
21. Tripwire state of cyber hygiene report, August 2018. https://www.tripwire.com/misc/state-of-cyber-hygiene-report-register/
22. Ajzen, I.: The theory of planned behaviour: reactions and reflections (2011)
23. Almeida, V.A.F., Doneda, D., de Souza Abreu, J.: Cyberwarfare and digital governance. IEEE Internet Comput. **21**(2), 68–71 (2017). https://doi.org/10.1109/MIC.2017.23
24. Beris, O., Beautement, A., Sasse, M.A.: Employee rule breakers, excuse makers and security champions: mapping the risk perceptions and emotions that drive security behaviors. In: Proceedings of the 2015 New Security Paradigms Workshop NSPW 2015, pp. 73–84. ACM, New York (2015). https://doi.org/10.1145/2841113.2841119
25. Bradbury, D.: Insuring against data breaches. Comput. Fraud Secur. **2013**(2), 11–15 (2013). https://doi.org/10.1016/S1361-3723(13)70020-4
26. Camp, L.J.: Mental models of privacy and security. IEEE Technol. Soc. Magaz. **28**(3), 37–46 (2009). https://doi.org/10.1109/MTS.2009.934142
27. Chaudhry, J.A., Rittenhouse, R.G.: Phishing: classification and countermeasures. In: 2015 7th International Conference on Multimedia, Computer Graphics and Broadcasting (MulGraB), pp. 28–31. IEEE (2015)
28. Craig, J.: Cybersecurity research-essential to a successful digital future. Engineering **4**(1), 9–10 (2018). https://doi.org/10.1016/j.eng.2018.02.006
29. Curtis, V.A.: Dirt, disgust and disease: a natural history of hygiene. J. Epidemiol. Commun. Health **61**(8), 660–664 (2007). https://doi.org/10.1136/jech.2007.062380

30. Dobbins, J., et al.: Choices for America in a Turbulent World: Strategic Rethink. Rand Corporation (2015)
31. Dodge, R., Toregas, C., Hoffman, L.J.: Cybersecurity workforce development directions. In: HAISA, pp. 1–12 (2012)
32. Emerson, R.G.: Limits to a cyber-threat. Contemp. Politics **22**(2), 178–196 (2016). https://doi.org/10.1080/13569775.2016.1153284
33. Fabiano, N.: Internet of things and blockchain: legal issues and privacy. the challenge for a privacy standard. In: 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 727–734, June 2017. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.112
34. Fabiano, N.: The internet of things ecosystem: the blockchain and privacy issues. the challenge for a global privacy standard. In: 2017 International Conference on Internet of Things for the Global Community (IoTGC), pp. 1–7, July 2017. https://doi.org/10.1109/IoTGC.2017.8008970
35. Farwell, J.P., Rohozinski, R.: The new reality of cyber war. Survival **54**(4), 107–120 (2012)
36. Floyd, D.L., Prentice-Dunn, S., Rogers, R.W.: A meta-analysis of research on protection motivation theory. J. Appl. Soc. Psychol. **30**(2), 407–429 (2000)
37. Fogg, B.J.: A behavior model for persuasive design. In: Proceedings of the 4th International Conference on Persuasive Technology, p. 40. ACM (2009)
38. Gardiner, K., Harrington, J.M.: Occupational Hygiene. Wiley, Hoboken (2008)
39. Gartzke, E., Lindsay, J.R.: Weaving tangled webs: offense, defense, and deception in cyberspace. Secur. Stud. **24**(2), 316–348 (2015). https://doi.org/10.1080/09636412.2015.1038188
40. Guo, K.H.: Security-related behavior in using information systems in the workplace: a review and synthesis. Comput. Secur. **32**, 242–251 (2013)
41. Hänsch, N., Benenson, Z.: Specifying it security awareness. In: 2014 25th International Workshop on Database and Expert Systems Applications (DEXA), pp. 326–330. IEEE (2014)
42. Kelley, D.: Investigation of attitudes towards security behaviors. McNair Res. J. SJSU **14**(1), 10 (2018)
43. Kerfoot, T.: Cybersecurity: towards a strategy for securing critical infrastructure from cyberattacks (2012)
44. Kirkpatrick, K.: Cyber policies on the rise. Commun. ACM **58**(10), 21–23 (2015)
45. Magnuson, S.: New cyber hygiene campaign seeks to curtail attacks. Nat. Defense **98**(726) (2014)
46. Mahfuth, A., Yussof, S., Baker, A.A., Ali, N.: A systematic literature review: information security culture. In: 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), pp. 1–6, July 2017. https://doi.org/10.1109/ICRIIS.2017.8002442
47. Mansfield-Devine, S.: The death of defence in depth. Comput. Fraud Secur. **2016**(6), 16–20 (2016). https://doi.org/10.1016/S1361-3723(15)30048-8
48. Mansfield-Devine, S.: Meeting the needs of GDPR with encryption. Comput. Fraud Secur. **2017**(9), 16–20 (2017). https://doi.org/10.1016/S1361-3723(17)30100-8
49. Maybury, M.T.: Toward principles of cyberspace security. In: Cybersecurity Policies and Strategies for Cyberwarfare Prevention, pp. 1–12 (2015)
50. Mears, J.: The rise and rise of id as a service. Biometric Technol. Today **2018**(2), 5–8 (2018). https://doi.org/10.1016/S0969-4765(18)30023-7

51. Mouradian, A.: Employees are lax on cyber fundamentals. Comput. Fraud Secur. **2017**(8), 17–18 (2017)
52. O'Connell, M.E.: Cyber security without cyber war. J. Conflict Secur. Law **17**(2), 187–209 (2012). https://doi.org/10.1093/jcsl/krs017
53. Oravec, J.A.: Emerging "cyber hygiene" practices for the internet of things (iot): professional issues in consulting clients and educating users on IOT privacy and security. In: 2017 IEEE International Professional Communication Conference (ProComm), pp. 1–5. IEEE (2017)
54. Oravec, J.A.: Kill switches, remote deletion, and intelligent agents: framing everyday household cybersecurity in the internet of things. Technol. Soc. **51**, 189–198 (2017). https://doi.org/10.1016/j.techsoc.2017.09.004
55. Padayachee, K.: Taxonomy of compliant information security behavior. Comput. Secur. **31**(5), 673–680 (2012)
56. Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., Jerram, C.: Determining employee awareness using the human aspects of information security questionnaire (HAIS-Q). Comput. Secur. **42**, 165–176 (2014)
57. Pfleeger, S.L., Sasse, M.A., Furnham, A.: From weakest link to security hero: transforming staff security behavior. J. Homeland Secur. Emerg. Manage. **11**(4), 489–510 (2014)
58. Sanders, R.: Embedding cyber-security into your company's DNA. People Strategy **39**(1), 8–9 (2016)
59. Savold, R., Dagher, N., Frazier, P., McCallam, D.: Architecting cyber defense: a survey of the leading cyber reference architectures and frameworks. In: 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 127–138. IEEE (2017)
60. Schrader, P.G., Lawless, K.A.: The knowledge, attitudes, & behaviors approach how to evaluate performance and learning in complex environments. Perform. Improv. **43**(9), 8–15 (2004). https://doi.org/10.1002/pfi.4140430905
61. Shackelford, S.J.: Business and cyber peace: we need you! Bus. Horiz. **59**(5), 539–548 (2016). https://doi.org/10.1016/j.bushor.2016.03.015. THE BUSINESS OF PEACE
62. Sheppard, B., Crannell, M., Moulton, J.: Cyber first aid: proactive risk management and decision-making. Environ. Syst. Decis. **33**(4), 530–535 (2013). https://doi.org/10.1007/s10669-013-9474-1
63. Singer, P.W.: The 'Ocean's 11' of cyber strikes. Armed Forces J. (2012)
64. Stanton, J.M., Stam, K.R., Mastrangelo, P., Jolton, J.: Analysis of end user security behaviors. Comput. Secur. **24**(2), 124–133 (2005)
65. Wang, C.P., Snyder, D., Monds, K.: A conceptual framework for curbing the epidemic of information malice: e-hygiene model with a human-factor approach. Int. J. Inf. Comput. Secur. **1**(4), 455–465 (2007)
66. Winkler, I., Gomes, A.T.: Chapter 5 - how to hack computers. In: Winkler, I., Gomes, A.T. (eds.) Advanced Persistent Security, pp. 41–46. Syngress (2017). https://doi.org/10.1016/B978-0-12-809316-0.00005-1

Appendix 5

V

K. Maennel, S. Mäses, S. Sütterlin, M. Ernits, and O. Maennel. Using technical cybersecurity exercises in university admissions and skill evaluation. *IFAC-PapersOnLine*, 52(19):169–174, 2019

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 52-19 (2019) 169–174

IFAC Papers Online CONFERENCE PAPER ARCHIVE

Using Technical Cybersecurity Exercises in University Admissions and Skill Evaluation

Kaie Maennel * **Sten Mäses** * **Stefan Sütterlin** **
Margus Ernits *** **Olaf Maennel** *

* *School of Information Technologies, Tallinn University of Technology, Estonia, (e-mail: firstname.lastname@taltech.ee)*
** *Faculty of Health and Welfare Sciences, Ostfold University College, Norway (e-mail: stefan.sutterlin@hiof.no)*
*** *RangeForce.com, (e-mail: margus.ernits@gmail.com)*

Abstract: Cybersecurity is a fast growing domain. The supply of workforce entering the labour market can not match the current demands. Due to this currently existing and predicted future skills gap in the labour market, educational institutions attempt to minimize dropouts and study times. As a direct consequence, the relevance of valid admission and selection procedures has grown in recent years. However, there is a mismatch between the increased demand for high-quality admission procedures and the still existing lack of established methods and routines to conduct these. In this paper we discuss our experience from running admissions in one of the oldest European master level cybersecurity curricula in Europe. We argue that cybersecurity skills assessment cannot simply be traditional knowledge-based assessments as this may exclude suitable candidates, who have not had the opportunity to learn the subject matter or are joining from different fields. Also selection decision cannot be done purely based on previous grades, because decomposing school subjects into cybersecurity skills is challenging due to the domain's interdisciplinary nature. We present a technical skills assessment method using cloud-based virtual labs that can be done by the candidates remotely. Those labs focus on assessing the technical competencies of a candidate and leave the assessment of non-technical skills (which are at least equally important) to a human interviewer. Also identifying cheaters, who do not prepare their labs themselves, will be left for the human interviewer. Such on-line exercises show potential as scalable option to evaluate the cybersecurity technical skills, motivational levels and cognitive strategies applied for problem-solving in a complex, novel task when being under performance pressure. The lessons learned are shared; feedback obtained from the applicants and possible technical metrics for predicting their success in a cybersecurity program are explored. As further work, we plan to conduct full data analysis and time-delayed interviews to generate hypothesis that can be further empirically tested with appropriate designs to detect causal relationships.

Keywords: Cybersecurity, university selection process, skills assessment, virtual labs

1. INTRODUCTION

The field of cybersecurity encompasses a large variety of job profiles demanding various degrees of technical or non-technical skills. Cybersecurity is now considered an independent discipline in accordance to Cabaj et al. (2018) or meta-discipline by Parrish et al. (2018). Cybersecurity is "a computing based discipline involving technology, people, information, and processes to enable assured operations in the context of adversaries"—it involves creation, operation, analysis, and testing of secure computer systems; and also includes aspects of law, policy, human factors, ethics, and risk management (Cabaj et al. (2018)).

The current best practices vary, however the best selection systems look for some combination of credentials (incl. commercial certifications and academic credentials), knowledge, and skills (Campbell et al. (2015)). Traditional, knowledge-based assessments may exclude suitable applicants both in university admissions or job market, who have not had the opportunity to learn the subject matter but provide large potentials. Applicants with non-technological skills may acquire cybersecurity knowledge later, but would not be distinguishable from the pool of applicants.

In this paper we focus on assessing the technical competencies of an applicant through hands-on technical exercises and leave assessment of non-technical skills (which are at least equally important) to human interviewers and different methods (e.g., aptitude tests). To determine technical skills and knowledge, hands-on technical exercises (both virtual and live events) can provide insight into the currently existing skill-set, but also provide insights into the problem-solving strategies, motivation, perseveration, and thus the learning potential of a candidate. The appliance of hands-on tasks allows not only to assess existing technological knowledge and performance under pressure, but also provides with a possibility to obtain performance data

2405-8963
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.12.169

in less stressful and judgmental environment (in contrast to a personal interview). For example, it allows to control confounding factors, such as social or performance anxiety. The applicant's skills will still be detected, when using online labs. Also, as the existing skills shortage in the job market is widely known, many candidates apply to universities without knowing what to expect from the program. The labs are expected to give some exposure to digital skills to an applicant who may not have a computer science background, but wants to study cybersecurity. Thus completion of the labs shows motivation and attitude—e.g., is a student "afraid" of trying or is (s)he attempting to complete the labs. Note that the labs have a virtual teaching assistant, that is describing the objectives for the virtual labs, gives helping hints in easier labs, and gives near real-time feedback about the successful task completion.

The hands-on technical exercises are gaining popularity in universities cybersecurity curricula for assessing skills and in teaching process. However, there is limited research that describe how to assess the applicant's cybersecurity skills using practical tasks in adequate and scalable manner (i.e., online assessments) and potential use of online hands-on technical labs as a validated method for skills measurement and predicting future learning success of a student.

This paper focuses on following aspects in the students' admissions to cybersecurity program while using component of hands-on practical exercises:

- Selection of hands-on technical skills to be included in the admission process to cover a wide area of cybersecurity as part of overall admission process;
- Technical, cultural and emotional factors that may impact the inclusion of (remote) hands-on technical skills assessment as part of university a admissions;
- Future research outline to validate what exercise information/metrics can be useful when giving the applicants hands-on technical tasks, and whether such analytical data can be used to estimate future learning success.

Our contribution is sharing the lessons learned from implementing combined approach of skills self-assessment, hands-on technical exercises and on-line interviews; and attempting to translate these findings into practical implications and further research questions. The results are based on ongoing work in the admission process to the cybersecurity curriculum at MSc (graduate) level and feedback collected over three years. We use a conceptual framework of combining credentials, self-assessment, interviews and hands-on labs. This is helpful in scaling up the assessment practice, and also analyzing whether measuring technical skills would indicate future learning potential or "practical" intelligence needed to succeed in cybersecurity career. The further work will analyze digital data collected from the exercises and a questionnaire with quantitative and qualitative questions to obtain further insight to the applicants' experience.

2. ADMISSION PROCESS

Many factors (such as an applicant's maturity, motivation, employment history, writing skills, work experience and other accomplishments) influence the success of a potential graduate student. Existing skills level and practical hands-on experience, will also contribute to success or failure of applicants' future learning. Namely, in addition to understanding what knowledge is needed for executing tasks and responsibilities, how to execute and implement in practice is more challenging. Thus it is not easy to evaluate whether an applicant to a cybersecurity curriculum is suitable based solely on their motivation letter, transcripts, grade point averages (GPAs), CV, etc. Even having short video interviews is not sufficient to have a deeper understanding of the applicant's skills. Therefore, we decided to add some practical hands-on technical labs to the application process where the applicants can demonstrate their technical skills. This approach is scalable compared to additional interview by the admission staff. Labs are optional and a human interviewer assesses skills comprehensively (both soft and technical) and has final decision.

The current admission system in our university, depicted on Figure 1, uses combination of traditional procedures (such as academic credentials and online interview with admissions team), but in addition also novel components of cybersecurity skills self-assessment and completion of hands-on online technical exercises [1].

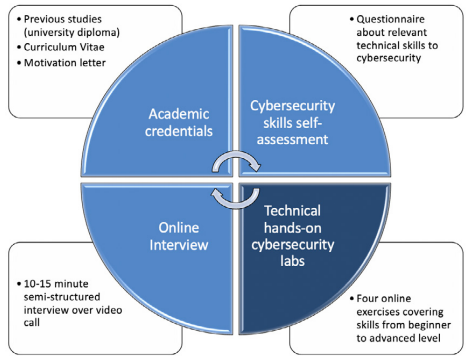

Fig. 1. Elements of Admission Process

So far, the applicants have been scored for motivation letter, online interview and supporting information, however completing questionnaire and technical exercises have not been scored as part of application process. As cybersecurity is an interdisciplinary field and the applicants come from many different study fields (e.g., IT, law, social sciences, etc.) these two components rather provide the opportunities for applicants to demonstrate their motivation and skills or give insights of their planned study field. This information also provides input for curriculum design based on the applicants' skills profiles.

We describe cybersecurity skills self-assessment questionnaire, hands-on technical exercises and online interview, as these components are specific to the cybersecurity program and relevant to scalable selection challenges addressed in this paper.

[1] https://www.ttu.ee/?id=175198

2.1 Self-Assessment Questionnaire

Questionnaire (mostly multiple choice answers) is an online form. The applicants rate their knowledge and skills in Likert scale for computer languages, public key infrastructure (PKI), cryptographic hashes, reverse engineering, network monitoring, risk assessments, firewalls, pentesting, security policy, system administration, finance, psychology, laws and regulations, data mining, cryptography, digital forensics, security and management.

The questionnaire has been included to encourage applicants to assess their existing knowledge and skills in the context of preferred specialization (cybersecurity, digital forensics or cryptography). The self-assessment was selected over knowledge test, as knowledge test would not be feasible or practical for scalable and flexible format of the application process to cover all knowledge domains. In addition, the definition of relevant knowledge in a rapidly developing field of cybersecurity may also be controversial and change over time. Furthermore, knowledge questions can always be "googled" and thus only invites cheating during the process. Also, use of previous grades is not necessarily relevant, as despite few subjects have more direct relations to cybersecurity (e.g., cryptography relates to mathematics)—de-composing school subjects into cybersecurity skills is challenging as cover diverse spectrum (organisational, psychological, mathematical-technical, social/societal, and other aspects). Also in cybersecurity the applicants come from many different study fields (e.g., IT, law, social sciences, etc.).

2.2 Technical Hands-On Cybersecurity Labs

An applicant should exhibit potential for critical thinking and problem solving skills, which we believe can be measured when the candidate is placed into the simulated (gamified) learning environment actually performing some technical tasks that require putting those skills into practice.

The virtual hands-on exercises are based on i-tee platform (Ernits et al. (2015)) that enables on-demand access to cloud-based virtual environment using a modern (HTML5 capable) web browser. Such technical setup enables lowering requirements to the applicants' computers. For accessing the hands-on lab, an applicant needs HTML5 capable web browser (without any additional plug-ins or VPN). The main requirement is a decent (at least 3Mbps) Internet connection. The system provides interactive assistance and guidance using Virtual Teaching Assistant for the learner.

As measuring all possible skills is not feasible, the exercises represent a mix of selected technical topics. The set of virtual labs is the following:

- Introduction lab—essential command line skills (Git, apt-get, Apache server). Estimated completion time 25 minutes;
- HTTPS Security—basic level skills connected to command line, public key infrastructure, and server administration basics; estimated completion time 45 minutes;
- SQL injection—intermediate level skills connected to attacking SQL databases (SQL, SQL injection); estimated completion time 90 minutes; and
- Botnet—advanced level skills connected to network scanning skills, text parsing (programming skills are beneficial) and SQL injection skills; estimated completion time 45 minutes.

The choice of these different exercises is based on typical attack vectors that the applicants are likely to encounter in their future cybersecurity jobs and require different skill levels (from essential to advanced). The combination of exercises is used to determine skill levels, but also to cover variety of different skills. Each lab has pre-determined skill level from basic to advanced level.

2.3 Online Interview

The interview is conducted after the academic credentials are checked, and the applicant has also completed the skills self-assessment and technical labs. We use a point system for each candidate covering areas published on the admissions website. This allows the admissions team to build upon the information about the applicant from these components, and obtain more detailed picture of applicants' knowledge, skills and motivation.

Due to the fact that our applicants apply from all over the world, 10-15 minute online video interviews are conducted with following structure:

- Welcome and explanation of the interview scope;
- Opening question: For statistical purposes the admission office collects information about reasons for selecting the university and program, and information channels the applicants found this program.
- Transition to technical questions with encouragement that if applicant does not know the answer—they should not be concerned and interview can proceed. Time allows for 5-6 questions in total.

  Example question 1: "If you had to both encrypt and compress data during transmission, which would you do first, and why?" This is a very popular interview question [2], because it sounds almost like a security question, however is actually an algorithmic computer science question. The aim of the question is to see if an applicant has knowledge about how compression algorithms work. Depending on answer, the follow-up question explicitly asking fundamental principles of compression algorithms.

  Example question 2: "In public-key cryptography you have a public and a private key, and you often perform both encryption and signing functions. Which key is used for which function?" The aim to see logical thinking—high temptation is to encrypt with a private key, which is incorrect.
- Thesis topic discussion: As part of the motivation letter the applicants are asked to identify two topics they consider for their thesis. The interview allows the student to elaborate further on their study and research interest.
- Closing: Explanation of further steps in the admission process.

[2] https://danielmiessler.com/study/infosec_interview_questions/

3. LESSONS LEARNED

3.1 Evaluation of Process and Data Sources

Firstly, we have obtained analytical data from the hands-on exercises from the online platform. Data includes metrics such as labs completed, completion time, percentage of lab tasks completed, etc. Secondly, we have conducted a quantitative survey to obtain input and the applicants' experience and motivation levels. The surveys were sent to all cybersecurity program's applicants and completion of survey was on voluntary basis. The survey informed respondents that "the individual survey responses are treated with strictest confidentiality, and only aggregated and anonymized results are considered as part of research publications". The survey also included explicit consent for future structured interviews.

The selected analytical data from the exercise platform and survey responses are presented in this section, however planned full-scope analysis is described in Section 5. The selected data from hands-on online system were correlated to the survey responses in order to identify possible correlations and relevant metrics to predict future success in the studies. As this is work in progress, we will continue with interviews to obtain further insight on the identified correlations. We intend to collect data after their studies are complete, to identify any other learning correlations.

We assume that our selection procedure as explained in Section 2 is reflective of the skills and levels needed for succeeding in the university studies of cybersecurity field. We also assume the applicants will be responding truthfully (despite of being rejected).

This study has been conducted on sample of 177 international students, with 38% completing the voluntary questionnaire. The applicants who did not respond, may have responded differently, thus imposing possible limitation. Also, the study covers short term period, i.e., up to two years from admission test, another longitudinal study should be conducted to obtain information and correlation over longer time period.

3.2 Self-Assessment Questionnaire

The meta-cognitive skills self-assessment can be useful, as the better one knows himself, the more likely the person will apply the right strategies as a student, as more likely she/he will mitigate her/his weaknesses (e.g., one who knows that he/she postpones difficult tasks will try to find a group to work with or ask for deadlines and seek early supervision).

However, the skills self-assessment is not representative, as such self-evaluation has several limitations. For example, expert level bias or cultural differences.

The plan is to use time-economic questionnaires to detect social desirability tendencies to remove the statistic influence on self-assessments. By checking for correlations with other self-assessment scores, we can further flag those for further investigation.

3.3 Online Interview

The questions were selected to mainly see applicants' reaction and logical thinking process. As in the cybersecurity, the person will face many unknown factor and it is important how they approach it. Does this have impact on their future study success?

We have many applicants who believe "WinRAR" is a compression algorithm, and cannot even admit they do not know how the computer algorithm works that makes the file-size smaller; some even do not know what "algorithm" means. Also, many applicants would share their private key in public key cryptography question with the interviewer by sending it by e-mail.

Furthermore, questions need to be adjusted to take into account the quality of the network connection. It is important that the candidate can understand the question despite a potential poor network connection. Technical problems happen and a candidate must not be punished for an issue that is caused by the video conferencing software. However, we are also aware of attempts to misuse this for cheating. We even have cases where the interviews were conducted by someone else. For this reason a "not working" webcam is problematic and raises suspicion.

3.4 Technical Hands-on Cybersecurity Labs

The following general patterns are emerging:

- The successful applicants complete all labs 100%. Any student who does not complete all labs, will get questions in the respective subject areas in the interview. The objective is to figure out why the student did not attempt all labs, because a lack of motivation in the labs could indicate a systematic lack of motivation in the study program later. There are some exceptions where the current workload of a prospective student does not allow the candidate to find the time to complete all the labs. Yet again, this maybe indicative of how they deal with future studies (it should be noted that most of our students are working at least part-time while studying).
- The applicants who need several tries and take many hours for the labs demonstrate high motivation. As they might not have been exposed to the technical subjects (e.g., not much prior Linux or programming/scripting experience, etc.), taking long time is not considered a problem.
- However, those applicants who quit labs after 5 minutes are likely going to be rejected. The rationale here is that, if they give up that easily in the admissions test, their motivation to study can be questioned and a similar behaviour might be expected during their studies.
- As a result, applicants who achieve 100% do not necessarily get accepted (e.g., labs are done at home and cheating might be suspected by the interviewer). However, also the applicants with legal, economic background have a good chance of getting in (assessed as good cybersecurity students), but this does not reflect in the technical assessment.

This section presents an overview of the initial analysis, in order to derive potential predictive metrics from the col-

lected data from in the admission process to be subjected for further research.

Time and Other Completion Activity Patterns Time is one of the common metrics that covers the same grounds in learning, gaming and other analytical evaluation. We hypothesize that applicants who spend longer time than estimated for admission lab completion may not have necessarily pre-required technical skills level, however demonstrate higher potential for learning success as they exhibit motivation. On other hand, the skilled applicants who spend nominal time have pre-required technical skills and therefore demonstrate learning success potential. Our dataset includes information about time spent on individual four labs and total time, also timestamps (start and end times) and number of attempts. In addition, we have asked for self-assessment for motivation level and learning success information through the surveys.

Other metrics (such as number of tutorials hints given, system access patterns, timing, length of sessions, etc.) are collected as part of the online platform, and could also be relevant, however they are not yet analysed in detail, and not used for selection decision so far.

Lab Completion Percentage The lab completion percentage (i.e., the level of tasks completed) is another readily available metric. We expect this metric to correlate with skill level of the applicants, and also potentially correlate to success in their later studies. Our exercise dataset includes information about lab completion percentage spent on individual four labs and number of attempts. In addition, we have asked for self-assessment of skills level and length of experience for programming languages through the surveys. The applicants with longer programming experience or higher skills assessments, are expected to perform better in the labs (especially in Botnet lab).

Emotional, Cultural and Social Factors Are there any common concerns raised by the applicants? For example, one of the top reasons why new students enter and stay is the quality of human interaction between faculty and existing students (Biggers et al. (2008)). However, when using remote labs reduces such interactions and rather replaces them with human-computer interaction, is this approach potentially negatively impacting the acceptance and entry decisions by the applicants?

One underlying question that has not been answered is whether use of such hand-on labs can create unfair exclusion and bias in the admission process. The collected data includes the qualitative feedback from the candidates to obtain insight whether use of such hands-on technical exercises is appropriate and does not create bias, exclusion or emotional stress. For example, creating an impression to the applicants that technical skills are more valued compared other relevant knowledge areas. Also, as cybersecurity is interdisciplinary domain and the applicants come from different study fields, can inclusion of such labs potentially scare off potential successful students.

When analyzing the qualitative feedback obtained as part of the surveys, the overall positive impressions weight out negative impressions among the applicants, see details in Table 1. The words used indicated mainly positive reaction (words link interesting, excited, fun, attractive, etc.), it also indicated motivation (words like challenging, difficult/easy, etc.). The word "scared" was mentioned few times but mainly in context of being scared in the beginning, and few mentioning they were "scared off" by labs.

Table 1. Feedback summary on students' experience on use of hands-on labs in admission process (i.e., scare off, time constraints, etc).

Feedback comments	Admitted	Not started	Rejected
Positive	78%	29%	57%
Neutral	17%	71%	38%
Negative	6%	0%	5%

Some examples of feedback include comments such as: *"I wasn't sure that I will be able to pass any of them, at start. But then I saw that they contain documentation, and it was quite easy to read documentation and complete the lab."*, *"Really exciting to stimulate"*, *"I strongly support the university having this as part of the admission process and using it to admit those from non-traditional educational backgrounds."*, *"At first, I was scared and unsure of what to expect but on getting to see the introductory texts, I noticed it wasn't just a blank screen with getting to see whats not on the screen...In all, it was a good experience and I loved it."*, etc.

From negative aspects, cheating is major concern, i.e., the labs are completed by someone else than the applicant. We currently do not assess cheating. We are aware of a few cases where someone else did the labs, and/or took the interview, only in order that a friend could obtain a Shengen-visa. Luckily such cases are very rare and the state authorities are making a careful check on incoming students to reject those who likely only apply in order to obtain a visa.

Overall use of the labs across the applicants supports that the applicants rather appreciate it as a positive experience than view it has negative experience.

Technical Limitations While implementing the virtual labs we learned that the labs are blocked in certain countries (e.g., Iran). The applicants typically know their way around by using VPNs, but they might face issues with latency during their labs. Therefore ensuring fair chance to complete needs to be considered.

4. RELATED WORK

Prediction modeling in the university admissions (including in Science, Technology, Engineering and Mathematics (STEM) disciplines) is not a new topic, however not been done in cybersecurity specifically. Most of the models assume previous knowledge of past performance or are mainly based on demographic data (e.g., Campbell et al. (2007), Kabra and Bichkar (2011), Chen et al. (2018)). The analytical metrics include the learners' individual characteristics, such as socio-demographic information, personal preferences and interests, responses to standardized inventories, skills and competencies, prior knowledge and academic performance, as well as institutional transcript data (Loh et al. (2015)). However, many of such metrics may not necessarily be available and admission decisions

need to be made with limited data. Also, the existing work does not address the question whether gamified hands-on exercises can be used as relevant predictor for future learning success.

In cybersecurity field, Augustine et al. (2010) describes a Computer Science freshman recruiting tool that provides an eight hour cyber training and competition framework designed to be attended by Computer Science candidates. However, this approach is not scalable in case of international admissions as assumes attendance and significant time commitment from existing faculty and students. Campbell et al. (2015) propose a model for predicting cybersecurity aptitude beyond a general-intelligence approach. They suggest that tasks, work roles, and people can be represented along the same set of axes to match job requirements to person attributes. These constructs can then be used to create assessments of potential for cybersecurity applicants, including the Cyber Aptitude and Talent Assessment as proposed by Campbell et al. (2015). However, the challenge is that the applicants are still exploring different career paths and the admissions process should allow such flexibility, and also accept different student profiles. Also such aptitude tests do not reveal technical base-level skills.

5. FUTURE WORK

This paper focused on sharing our experience and lessons learned from implementing skills self-assessment questionnaire and technical hands-on labs, as novel part of the university graduate level admissions for cybersecurity program. Based on the full data analysis and further evidence to be collected from the interviews with the students, we plan to evaluate the admission process by:

- Assessing academic performance throughout the studies and correlate it with the lab performance, the interview and self-assessment questionnaire scores to identify which part of the current admission process—or which interaction of parts—has most predictive power for academic performance.
- Validation of the hands-on technical exercise tasks by correlating it with general intelligence, other cognitive skills, and domain-specific knowledge to improve our understanding for what is tested in this task (construct validation).

6. CONCLUSION

In this paper we described our university cybersecurity MSc (graduate level) program admissions experience, combining cybersecurity skills self-assessment, hands-on technical online labs and online interviews. Online technical labs are scalable and easy-to-implement, however selection of technical skills is critical and conclusions from analytical data need further validation. The feedback from the applicants has shown that such labs are seen very positive, as they enable the applicants to demonstrate their skills or gain insights what to expect during their studies. As part of lessons learned, we discussed selected analytical data from the technical hands-on exercises in the context of scalable technical skills assessment. We hypothesize that the analytical data can be used as a predictive component in the admission process to identify the applicants who are motivated and likely to succeed in the cybersecurity curriculum. However, the further analysis via interviews and other methods will be conducted to confirm these correlations as causative, as part future work.

The further plan is to combine cognitive aptitude and team skills measurements with technical element, to improve the selection process. As part of selection process we want to predict their ability to learn new knowledge and acquire new skills, and meta-cognitive abilities are known as such predictor.

ACKNOWLEDGEMENTS

The authors would like to thank Rangeforce [3] for allowing us to use their system.

REFERENCES

Augustine, T.A., DeLooze, L.L., Monroe, J.C., and Wheeler, C.G. (2010). Cyber competitions as a computer science recruiting tool. *Journal of Computing Sciences in Colleges*, 26(2), 14–21.

Biggers, M., Brauer, A., and Yilmaz, T. (2008). Student perceptions of computer science: a retention study comparing graduating seniors with cs leavers. In *ACM SIGCSE Bulletin*, volume 40, 402–406. ACM.

Cabaj, K., Domingos, D., Kotulski, Z., and Respício, A. (2018). Cybersecurity education: Evolution of the discipline and analysis of master programs. *Computers & Security*, 75, 24–35.

Campbell, J.P., DeBlois, P.B., and Oblinger, D.G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4), 40.

Campbell, S.G., ORourke, P., and Bunting, M.F. (2015). Identifying dimensions of cyber aptitude: the design of the cyber aptitude and talent assessment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, 721–725. SAGE Publications Sage CA: Los Angeles, CA.

Chen, Y., Johri, A., and Rangwala, H. (2018). Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 270–279. ACM.

Ernits, M., Tammekänd, J., and Maennel, O. (2015). i-tee: A fully automated cyber defense competition for students. In *ACM SIGCOMM Computer Communication Review*, volume 45-4, 113–114. ACM.

Kabra, R. and Bichkar, R. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications*, 36(11), 8–12.

Loh, C.S., Sheng, Y., and Ifenthaler, D. (2015). Serious games analytics: Theoretical framework. In *Serious games analytics*, 3–29. Springer.

Parrish, A., Impagliazzo, J., Raj, R.K., Santos, H., Asghar, M.R., Jøsang, A., Pereira, T., and Stavrou, E. (2018). Global perspectives on cybersecurity education for 2030: a case for a meta-discipline. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 36–54. ACM.

[3] https://rangeforce.com

Appendix 6

VI

M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020: 15th International Conference on Cyber Warfare and Security*. Academic Conferences and Publishing Limited, 2020

From Simple Scoring Towards a Meaningful Interpretation of Learning in Cybersecurity Exercises

Margus Ernits [1], Kaie Maennel [2], Sten Mäses [2], Toomas Lepik [2] and Olaf Maennel[2]
[1]RangeForce.com, Tallinn, Estonia
[2]School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia

margus.ernits@rangeforce.com
kaie.maennel@taltech.ee
sten.mases@taltech.ee
toomas.lepik@taltech.ee
olaf.maennel@taltech.ee

Abstract: To overcome the current skills shortfall in cybersecurity, a broad range of IT professionals and users should be educated in the fundamentals of protecting computer systems and the data they contain. This requires novel and scalable teaching methods. The main contribution of this paper is to introduce an approach of how to create cybersecurity exercises that can measure relevant competencies. We demonstrate how technical event logging can be linked to learning outcomes and skills measurement by defining intermediate abstraction layers. These take raw forensic data from the game-system and network, and gradually group them into events and abstract measurements until they can be mapped to learning theories. The suggested approach enables deeper insights into learning. This approach has been applied for developing labs using our cloud-based open-source tool. The labs have been used by more than 2 000 learners in over 15 000 sessions. A thorough hands-on skills assessment was conducted before and after a set of exercises for 27 participants. Results show that the suggested method can be used for creating and improving cybersecurity exercises.

Keywords: cybersecurity education, exercises, skill assessment, learning

1. Introduction

Cybersecurity exercises are becoming increasingly popular for educating and evaluating security specialists (Ogee et al, 2015). This comes as no surprise when our society is fundamentally dependent on IT systems and vulnerabilities are subject to exploitation by threat actors. To overcome today's cybersecurity problems, a very broad range of IT professionals should be educated and everyone should understand the fundamentals of cybersecurity.

Therefore, novel and scalable teaching methods need to be developed. Realistic attacks and complex simulated systems in virtualized environments can provide engaging and practical hands-on learning experiences using fully automated training that utilizes adaptive learning methods. The goal is to have a training environment that would detect the skill-level of the learner and automatically select the most appropriate learning tasks for the user. Thus, it is important to have a measurement methodology that is able to accurately capture the capabilities of the user. However, current research suggests (e.g., Fulton et al, 2012) a lack of defined educational outcomes. This might be due to the overall difficulty of designing and implementing a complex defence oriented gamified cybersecurity exercise. Specifically, constant adjustments to the scoring system and storyline are usually very time consuming and may divert the attention from in-depth analysis of the final score.

Our aim is to apply the existing instructional design methods for connecting raw data points to high level competencies using an evidence-correlation model. Specifically, we suggest a method for the design and implementation of an exercise that would give a structured and automatic feedback of the participants' skills and competencies. This is implemented on i-tee, an open-source software platform (Ernits and Kikas, 2016) developed from experience gained in several large-scale exercises including Locked Shields and Cyber Security Challenge UK. Nevertheless, the suggested model is itself platform independent and can be implemented using other environments such as OpenStack.

Over 2 000 learners have used the system in more than 15 000 various lab sessions on a wide range of topics. Of those, the skills of a pilot group of 27 IT developers were measured before and after a set of training exercises.

The set consisted of 13 different labs: Command Injection, Cookie Security: Secure, Cookie Security: HttpOnly, Cross-Site Request Forgery (CSRF), Defence against CSRF, Insecure Direct Object Reference, Intro Lab, Path Traversal, SQL Injection, Unrestricted File Upload, Reflected Cross-Site Scripting (XSS), Stored XSS and Phishing based on Stored XSS.

2. Related work

Performance measurement is crucial for mastering a skill as argued in *Understanding by Design* by Wiggins (2005). There is also a growing body of work on this topic in the field of cybersecurity exercises. This section focuses on existing conceptual approaches and evaluation models in cybersecurity training, with a focus on training through exercises.

2.1 Design approaches to gamified cybersecurity training

Katsatonis et al (2017) provide a concept map of cybersecurity game-based approaches' key elements that include also learning objectives and assessment. Learning objectives should be based on performance, proficiency and be connected to game-play. Vykopal et al (2016) suggest decomposing the training activity into individual levels that learners have to accomplish for satisfying specific learning objectives. The data collected include metrics such as the start and end of each game and level within it. More detailed data include submission of incorrect flags and their content, hints used, skipping a level, displaying a level's solution and game ID (Vykopal et al, 2016). Vykopal et al (2017) state that setting learning objectives based on learners' skills before the actual exercise is a challenging undertaking.

Clark (2015) proposes a 4-level model, where in order to create a broader understanding of the security elements the impact on the target-host, server, firewall and intrusion detection systems is highlighted at each level. Nicholson et al (2016) suggest that learning should be individualised to fit the higher skill or competency level the subject is aiming for. These should consider the learners' profiles, content brokering, experience tracking and competency network. With these high-level conceptual approaches our proposed model—experience and competency tracking allows the learner profiles to be updated based on the progress in competencies and also need to be easily adjusted.

2.2 Evaluation models used in cybersecurity exercises

There are multiple frameworks looking at classification of cybersecurity-related competencies. The National Initiative for Cybersecurity Education Cybersecurity Workforce Framework (NICE Framework), published by the National Institute of Standards and Technology (Newhouse et al, 2017) lists knowledge, skills and abilities required to perform tasks in specific work roles. Rashid et al (2018) introduce the Cyber Security Body of Knowledge project aiming to codify the foundational and generally recognized knowledge on cybersecurity. Nevertheless, those high level frameworks provide only generic ideas instead of specific tasks for measuring skills. For example, task T0349 in NICE framework is described as "collect metrics and trending data" (Newhouse et al, 2017) leaving space for various interpretations about the task specifics.

Abbott et al (2015a) provide a quantitative evaluation of techniques for student performance assessment. This uses an automated mechanism for parsing log entries into blocks of time during which participants are focused on specific high-level objectives. Abbott et al (2015b) also describe an exercise instrumentation that enables automated performance assessment by capturing students' computer-based transactions in a log. This is time-synced with the game-server to deliver challenges and registers student responses. Labushange and Grobler (2017) describe assessing technical skill level based on indexed similarity of the commands used to achieve the specified objectives from which the level of participant's practical knowledge could be inferred. The paper classifies learner's actions that can automatically deducted using the clustering of commands but does not go into details. The similar underlying idea is implemented in our model at raw data to Game Event Logs (GEL) transition level.

Many articles address different issues in cybersecurity exercises including skills assessment and evaluation attempts. Scoring is often seen as a tool to provide evaluation and feedback to exercise participants. However,

scoring systems are not necessarily connected to learning objectives. What is missing is a practical and scalable model that would provide evidence that high level competencies can be achieved through analysing the granular data level of the exercises' raw data.

3. Connecting competences to raw data

In cybersecurity exercises, different events happen rapidly in a semi-controlled environment. However, learning experiences are not linear or predictable. For example, in case the participants have a task to defend a vulnerable web application, they may have different correct ways to deal with the attacks. These include block attacks by implementing intrusion prevention systems by using web application firewalls or by fixing the vulnerabilities of web application. There are also different incorrect or insufficient ways to react such as removing the attacker's injected code/content from the system, taking vulnerable applications offline or by breaking web application's functionality. These make measuring specific competencies challenging.

The primary focus when designing an exercise is to start with the conceptual design. First, targeted competencies that the exercise should teach or assess are defined. These competencies are decompiled to different skills with measurable learning objectives. Based on those learning objectives, specific tasks can be determined that could be measured by evidence—i.e., events happening in the system.

To improve the flexibility and efficiency of the system, Game Event Logs (GEL) are designed to capture all the important events. When GEL are designed appropriately, the amount of the exercise data needed to analyse decreases significantly. Massive amounts of raw data can be deleted after the exercise while maintaining the ability to dynamically change the rules for interpreting the events. The raw data is used to generate GEL that are interpreted to evaluate whether a specific task is completed. Commonly, the completion of tasks is used as an input for score calculation, but we suggest doing more than that. The completion of tasks indicates the proficiency level of different skills. Skills in turn are gathered into meaningful sets to form competencies.

Fig. 1 illustrates the suggested design model for exercises enabling the evaluation of participant competencies. The model consists of 5 layers with the bottom layer representing the raw data and the highest layer being specific competencies that are targeted by the exercise. It is important to note that while the technical data flows from the 1st to the 5th layer, the logical design flow should start from the top layer. The layers themselves are connected to each other but at the same time independent, allowing the use of different formats or processing systems. In the following sections, we explain each layer in detail and use Cross-Site Scripting (XSS) related competency for illustration.

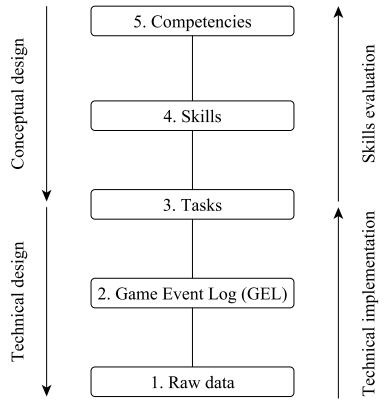

Figure 1: Suggested structure for exercises

3.1 Layer 5 - Competencies

The creation of every exercise should start with question "Why?". In a cybersecurity context, it is usually the need to protect the systems or build more secure systems and to achieve this goal, organisations need people with specific competencies. The term competency is somewhat ambiguous as shown by Le Deist and Winterton (2005). In the context of this paper, we define competency as a set of knowledge, skills and abilities. These focus mainly on practical skills that are relevant in the field of cybersecurity and that can be measured by performance in virtual hands-on exercises.

The competency can be taken as a general theme for a set of skills that it encompasses. For example, defending against Cross-Site Scripting (XSS) was one of the competencies used in our labs.

3.2 Layer 4 - Skills

Using competency as the general theme, we can look at skills as sub-topics of an area. As different jobs require groups of skills, it is possible to define the set of skills for a particular position and then evaluate an individual accordingly. However, it should be noted that dividing job qualifications (professionalism) into different skills (and sub-skills) is a subject of debate (Rigby and Sanchis 2006). In cybersecurity, the relevant skills and sub-skills can be further analysed based on types of attack vectors. In the XSS lab for example, the types of XSS attacks can be defined as reflected XSS, stored XSS, and DOM based XSS (Gupta and Gupta (2017). In our example, the exercise focuses on developing skills to defend against the reflected XSS.

The layer of skills defines both high- and low-level skills. High level skills are more meaningful and are used in daily conversations. Installing WordPress to a web server from scratch is an example of a higher-level skill. Lower level skills are usually not meaningful or useful by themselves. Being able to remotely log in to a server using SSH is an example of a sub-skill. It is not that useful by itself, but it is a needed step in a larger process. Note that skills and sub-skills can be used to specify different proficiency levels. Some sub-skills can be considered necessary for some higher-level skill and different learning objectives can indicate different skill proficiency levels.

Specifying a skill in a measurable way is not an easy task. For example, a learning objective can be 'Learner can fix a web application with a XSS type vulnerability'. An undesired fix, which would fulfil the task could be to take the site offline. This means that appropriate learning objective should be: 'Learner can fix a web application with a XSS type vulnerability without disturbing service or breaking functionality'.

Skills are closely related to learning outcomes. Learning objectives define the expected goal of exercise in terms of demonstrable skills or knowledge acquired by a participant as a result of exercise (Malan, 2000). The skills and knowledge can be analysed using different models such as Bloom's, SOLO, etc. Figure 2 illustrates how in our cybersecurity exercise training model we follow and map the task with a range of cognitive learning and skills layers. These are Not graded (0), Remembers (1), Understands risks/attacks (2), Applies attacks (3), Applies defences (4) and Masters defences (5).

Name	Web - Client Side Attacks						Web - Server Side Attacks							
	Reflected XSS	Stored XSS	DOM based XSS	Session hijacking	CSRF	CSRF + XSS	Command Injection	Privilege escalation	SQLi authentication bypass	Path Traversal	Insecure Direct Object Reference	Union based SQLi	File Upload and Inclusion	Blind SQL injection
User 1	4	4	2	4	3	3	2	3	3	2	2	4	4	3
User 2	2	3	3	2	2	2	3	0	0	0	0	0	0	2
User 3	0	1	0	0	1	0	0	0	0	0	0	0	2	1

Figure 2: Skills report

The cognitive learning layers based on the revised Bloom's taxonomy (Krathwohl and Andersen 2001) are remembering, understanding, applying, analysing, evaluating, and creating. The adjustment is made in order to incorporate the attack and defence aspects. In order to defend, the learner needs to understand and apply the attack technique themselves before being able to creatively avoid such vulnerabilities in the system. Difference in applying and mastering the defence is transferability such as assessing if the learner is able to defend against this type of attacks in different operating systems or tools. Mastering the defence (level 5) in our model usually means that the same skill is measured and mapped over different labs. This ensures that the learner is able to transfer and apply the skill using different technologies.

In our lab example, the learning objectives are as follows:

- Learner understands impact of the reflected XSS.
- Learner mitigates the attack (e.g., applies HttpOnly flag).
- Learner fixes XSS vulnerabilities (using PHP in this lab).
- Learner performs reflected XSS.
- Learner recognises reflected XSS.

3.3 Layer 3 - Tasks

Tasks represent assignments that can be clearly defined and measured. Different tasks should represent different proficiency levels for the connected skills. For example, performing a blind SQL injection can be considered more advanced than a simple SQL injection.

In the reflected XSS lab, the tasks mapped to the cognitive learning layers are the following:

- Learner finds reflected XSS from unfamiliar target (Linux, php, web application).
- Learner uses reflected XSS found to retrieve session cookie of emulated computer user.
- Learner uses session cookie to login.

Tasks can be arranged in groups or in hierarchies. Fig. 4 shows an example from our XSS lab displaying a task of finding a vulnerable field in a web form. This task has two sub-tasks that are evaluated based on user input.

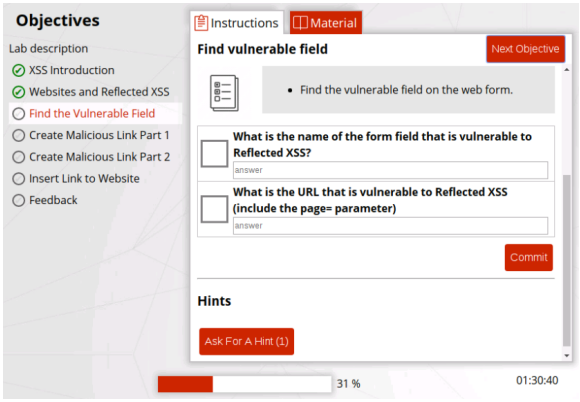

Figure 4: Example of tasks in the XSS lab

Tasks can be measured differently—by using direct input from the learner or with the help of some automation (Mäses et al, 2017). In case of automated task evaluation, it is useful to have a dedicated abstraction level for GEL.

3.4 Layer 2 - Game Event Log (GEL)

As seen in Fig. 1, the conceptual design of an exercise finishes with defining tasks. The exercise organisers do not need to go deeper into technical design. Therefore, starting from Layer 2, the focus moves from design to measurement. The question is how to quantify a particular task. This could be done by observing system logs or using specific scoring scripts such as a script that pings a web server and registers responses.

GEL enables flexible and meaningful rules for task evaluation. GEL include may references to the system log. For the XSS lab, the following events have been defined:

- The lab is personalised correctly (learner has successfully initialised the lab).
- Target website in the lab works correctly (based on user emulation checks).
- Learner has found the vulnerable form field (submitted correct field name without submitting all/lots of them).

In the following example there are two events described in GEL. The first one is stating that the simulated user (searching for the link with XSS injection) was active. The second example comes from a lab dealing with drones, where drone number 10 is functional, has operational backdoor (backdoor vulnerability not fixed), has XSS vulnerability (status false, because vulnerability is not fixed).

```
Apr 30 09:24:58 GEL index.js[926]: debug: Stored XSS: Simulating manager user

Apr 30 09:55:59 GEL index.js[1763]: info: [Drone 10] Status: functional=true,
backdoor=false, xss=false
```

GEL allows new rules to be added to higher levels of the model. These define new competencies, skills and tasks based on already existing game events, enabling rules to be redefined and fine-tuned more dynamically.

We use event correlation to evaluate objectives based on event logs. The attack evidence, such as successful or unsuccessful attacks, is collected and correlated with evidence that emulated computer users are able to use a full functionality of the simulated system.

3.5 Layer 1 - System state, raw data

The system state can be found from system logs or by a specific script checking such as whether the ping packet to the server gets a reply. Based on the system states, the event logs are created.

4. System architecture

The system architecture gives a more holistic view for understanding the suggested approach. However, the proposed method that links log events and raw data to learning outcomes does not depend on the underlying software stack. **The main goal is to measure the cyber skills in a scalable way.** Such measurement is a prerequisite for individually adaptive learning and enables to measure training effectiveness.

In our architecture, to access the hands-on exercise platform, a participant needs HTML5 capable web browser without any additional plug-ins or Virtual Private network(VPN). The system uses a Virtual Machine (VM) Host platform based on the open source tool i-tee (Ernits and Kikas, 2016). When a lab starts, all VMs, networks, and grading systems are provisioned and personalised for the participant. Also, the automated skill evaluation process is initialised. Each lab may include different VMs (Linux, Windows, BSD, etc.) and different software defined networks. Some VMs are accessible for participants (blue systems), whereas other VMs are dedicated for attack traffic generation (red systems) or for end-user simulation and for network traffic generation.

Fig. 4 illustrates the general architecture. The system provides network isolation between participants and lab networks. Lab personalisation process creates flags, vulnerabilities, grading for each lab attempt and random IP addresses for attacks and grading. Those IPs are based on real logs from our servers (fail2ban, sshguard,

blacklists). The system provides interactive assistance and guidance for participants using hints, leaked hacker's chat live stream, and media injects using Virtual Teaching Assistant for the learner. Gamification elements such as leader-boards, scoreboards and hackers chatrooms are provided.

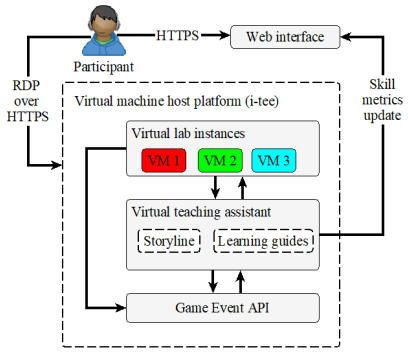

Figure 4: System Architecture

This architecture allows the creation of new labs and challenges by reusing existing modules such as attacking and assessment scripts and vulnerable targets. The system is designed to enable a lab to start without extra management effort. All game services, routers, networks, scoring bots are allocated on demand.

Our system scales as follows:

- First, we can easily teach small groups of students (1-30 lab sessions per server). When using cloud resources, we can run defence-oriented exercises with 300 participants without building a new expensive cyber range.
- Second, automatic assessment system demands no human red team (attacking team) or assessors.
- Third, the system is remotely accessible using web browser and can be used from home or in a classroom at any time.

5. Initial results

More than 2 000 learners have used the system in more than 15 000 lab sessions. We have used our platform for on-site cybersecurity competitions in 10 countries, in companies and academic training programs. A more thorough hands-on assessment was conducted with 27 participants as follows:

- Participants completed hands-on pre-assessment labs of approximately a 4 hours long assessment covering wide range of cybersecurity skills.
- Participants were assigned 13 different labs including XSS, command injection, cookie security.
- Participants completed hands-on post-assessment labs.

The results in Fig. 5 show average completion rate of pre- and post-assessment labs. The results are presented in three sub-groups. "Skilled" (high pre-assessment score, successful lab completions and post-assessment), "study" (low pre-assessment score, successful lab completions and post-assessment) and "passive" (low pre-assessment score and unsuccessful in lab completion or post-assessment). These are subjective sub-categories to analyse participants based on pre-assessment results and completion of the labs. It can be seen, for example, that for the "study" group, the improvement in skills was significant (from 4% to 68%).

Group description	Group composition	Pre-assessment	Post-assessment
Skilled	15%	82%	100%
Study	67%	4%	68%
Passive	19%	0%	0%
Total	100%	22%	76%

Figure 5: Skills improvement

After the exercise, additional free-form feedback was asked from the participants and from their supervisors to analyse whether participants had acquired relevant competencies. As anecdotal evidence, the organisation identified a participant who lacked cybersecurity-related experience and skills before training program but was able to identify 5 security vulnerabilities from their internal system after training. Such unstructured feedback has given subjective evidence of skill improvement, quantitative surveys could be used to gather feedback systematically.

6. Discussion

There are multiple initiatives towards more competency-driven computer science education that aim to focus more on competencies instead of primarily covering topics (Sabin et al, 2018). There are also discussions on topics such as integrating hands-on cybersecurity exercises into the curriculum (Weiss et al, 2019), and what core cybersecurity skills students should learn (Jones et al, 2018). The big question is: how to measure those skills?

Presenting learning outcomes without a clear mapping to relevant measurement points makes it difficult to evaluate individual skills. For example, a student who completes a course with a passing rate of 82% might be equally above average in all the relevant skills or have mastery of most skills and complete lack of others. Having a clear line of thought between different abstraction levels helps to provide evidence for the achieved competency.

From the learner's perspective, the main added value of our competency-driven approach is automated (therefore timely) individual feedback regarding the learner's skills. The individual skills report (such as illustrated by Fig. 2) enables the learners to find their current strengths and weaknesses and modify their future learning activities accordingly. For the evaluator, the skills report enables to track learners' progress focusing on their actual skills. At the same time, automated and scalable approach helps to reduce the evaluator's workload.

Following the competency-driven design can seem like a challenging task at first. However, the effort pays off because it helps to keep a clear overview when modifying or scaling up the system. Also, such competency-driven structure helps to remove the tendency to design too simple and easy-to-create tasks and potentially not covering intended core skills. Exercise design is an iterative process (Mäses et al, 2018) and technical design can influence task selection (e.g., technical implementation might be too complex or just unfeasible). Additionally, there are ethical considerations, which might force to abandon measurement of some initially planned skills. Exercises are often aimed to be as realistic as possible (Fox et al, 2018), but in the design phase a balance should be struck between what is realistic and what is measurable. In measuring a skill, the realism of a task is not the ultimate goal and it can be tailored to make the task quantifiable.

A more universal competency evaluation forces the educators to focus on how to measure higher level skills in a scalable way.

7. Future Work

Our current work focuses on defining the rules that can be universally applied in different exercises irrespective of the platform. This enables the organisers to dive into the learning data more easily. This data could be used for modelling predictive behaviours, such as considering the use of learning hints in skills' evaluation (Chow et al, 2017) and calculating confidence correlation to the skills. Our method can be used for connecting raw data

to widely used competency frameworks such as NICE Framework. Potentially, it could be also used for comparing the learners' profiles to current job market requirements.

6. Conclusion

The opportunity to benchmark the competencies in a comparable way provides more insights that simply offering scores from cybersecurity exercises. Our structured approach helps to obtain more meaningful learning data from the logged events instead of counting points. Our main contribution is demonstrating how to create cybersecurity exercises that measure relevant competencies. We have applied this method for developing labs and assessing the skills of 2000 learners (including 27 having a more thorough assessment). Initial validation results show that connecting cybersecurity exercise raw data to the skills and competencies is a promising way forward.

Our approach supports a paradigm shift towards the cybersecurity exercises that by design allow the systematic and evidence-based competencies and skills measurement. The open-source platform, described design and evaluation process with ultimate aim to measure competencies, form together a step to achieve this change.

References

Abbott, R.G., McClain, J., Anderson, B., Nauer, K., Silva, A. and Forsythe, C., 2015. Log analysis of cyber security training exercises. Procedia Manufacturing, 3, pp.5088-5094.

Abbott, R.G., McClain, J.T., Anderson, B.R., Nauer, K.S., Silva, A.R. and Forsythe, J.C., 2015. Automated Performance *Assessment in Cyber Training Exercises*. Sandia National Lab, Albuquerque, NM.

Chow, S., Yacef, K., Koprinska, I. and Curran, J., 2017, July. Automated data-driven hints for computer programming students. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 5-10). ACM.

Clark, D.J., 2015, November. An onion approach to cyber warfare training. In *2015 Military Communications and Information Systems Conference* (pp. 1-4). IEEE.

Ernits, M. and Kikkas, K., 2016, July. A live virtual simulator for teaching cybersecurity to information technology students. In *International Conference on Learning and Collaboration Technologies* (pp. 474-486). Springer, Cham.

Fox, D.B., McCollum, C.D., Arnoth, E.I. and Mak, D.J., 2018. Cyber Wargaming: Framework for Enhancing Cyber Wargaming with Realistic Business Context.

Fulton, S., Schweitzer, D. and Dressler, J., 2012, October. What are we teaching in cyber competitions?. In *2012 Frontiers in Education Conference Proceedings* (pp. 1-5). IEEE.

Gupta, S. and Gupta, B.B., 2017. Cross-Site Scripting (XSS) attacks and defense mechanisms: classification and state-of-the-art. *International Journal of System Assurance Engineering and Management*, *8*(1), pp.512-530.

Jones, K.S., Namin, A.S. and Armstrong, M.E., 2018. The core cyber-defense knowledge, skills, and abilities that cybersecurity students should learn in school: Results from interviews with cybersecurity professionals. *ACM Transactions on Computing Education*, *18*(3), p.11.

Katsantonis, M.N., Fouliras, P. and Mavridis, I., 2017, September. Conceptualization of Game Based Approaches for Learning and Training on Cyber Security. In *Proceedings of the 21st Pan-Hellenic Conference on Informatics* (p. 36). ACM.

Krathwohl, D.R. and Anderson, L.W., 2009. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Labuschagne, W.A. and Grobler, M., 2017, June. Developing a capability to classify technical skill levels within a Cyber Range. In *ECCWS 2017 16th European Conference on Cyber Warfare and Security* (p. 224). Academic Conferences International Limited.

Le Deist, F.D. and Winterton, J., 2005. What is competence?. *Human resource development international*, *8*(1), pp.27-46.

Malan, S.P.T., 2000. The 'new paradigm' of outcomes-based education in perspective. *Journal of Consumer Sciences*, *28*(1).

Mäses, S., Hallaq, B. and Maennel, O., 2017, October. Obtaining better metrics for complex serious games within virtualised simulation environments. In *European Conference on Games Based Learning* (pp. 428-434). Academic Conferences International Limited.

Mäses, S., Randmann, L., Maennel, O. and Lorenz, B., 2018, July. Stenmap: framework for evaluating cybersecurity-related skills based on computer simulations. In *International Conference on Learning and Collaboration Technologies* (pp. 492-504). Springer, Cham.

Newhouse, W.D., Keith, S., Scribner, B., Witte, G., 2017. *Nice cybersecurity workforce framework: National initiative for cybersecurity education* (No. Special Publication (NIST SP)-800-181).

Nicholson, D., Massey, L., O'Grady, R. and Ortiz, E., 2016. Tailored Cybersecurity training in LVC environments. In *MODSIM World Conference, Virginia Beach, VA*.

Ogee, A., Gavrila. R., Trimintzios, P., Stavropoulos, V. and Zacharis, A. [n. d.]. The 2015 Report on National and International Cyber Security Exercises.

Rashid, A., Danezis, G., Chivers, H., Lupu, E., Martin, A., Lewis, M. and Peersman, C., 2018. Scoping the cyber security body of knowledge. *IEEE Security & Privacy*, *16*(3), pp.96-102.

Rigby, M. and Sanchis, E., 2006. The concept of skill and its social construction. *European journal of vocational training*, *37*, p.22.

Sabin, M., Alrumaih, H. and Impagliazzo, J., 2018, April. A competency-based approach toward curricular guidelines for information technology education. In *2018 IEEE Global Engineering Education Conference* (pp. 1214-1221). IEEE.

Vykopal, J. and Barták, M., 2016. On the design of security games: From frustrating to engaging learning. In *2016 {USENIX} Workshop on Advances in Security Education*.

Vykopal, J., Vizváry, M., Oslejsek, R., Celeda, P. and Tovarnak, D., 2017, October. Lessons learned from complex hands-on defence exercises in a cyber range. In *2017 IEEE Frontiers in Education Conference* (pp. 1-8). IEEE.

Weiss, R., Mache, J., Taylor, B., Kaza, S. and Chattopadhyay, A., 2019, February. Discussion of Integrating Hands-on Cybersecurity Exercises into the Curriculum in 2019. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 1245-1245). ACM.

Wiggins, G. and McTighe, J., 2005. Understanding by design. Alexandria VA: Association for Supervision and Curriculum Development.

Appendix 7

VII
K. Maennel. Learning analytics perspective: Evidencing learning from digital datasets in cybersecurity exercises. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 27–36. IEEE, 2020

2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)

Learning Analytics Perspective: Evidencing Learning from Digital Datasets in Cybersecurity Exercises

Kaie Maennel
School of Information Technologies: Department of Software Sciences
Tallinn University of Technology
Tallinn, Estonia
kaie.maennel@taltech.ee

***Abstract*—Cybersecurity exercises are gaining in popularity in university curricula and professional training paths and are seen as an effective teaching method. Such exercises provide digital datasets to facilitate a learning analytics approach such as by using the traces that learners leave behind to improve the learning process and environment. While there are various learning measurement efforts from digital datasets in the existing literature, a holistic learning analytics approach incorporated into cybersecurity exercises is still lacking. We propose a practical reference model for incorporating a learning analytics approach into the cybersecurity exercise life-cycle. To facilitate this application, we have performed an extensive review of existing academic research on applying learning analytics in the context of cybersecurity exercises. We specifically focus on the learning indicators used to measure empirical impact and training effectiveness that could indicate achievement of defined learning outcomes. This reference model and overview of existing learning analytics use cases and learning metrics in various types of exercises can help educators, organisers and cyber range developers. This results in more adaptive exercise design and measurement using evidence-based data and connects digital learning traces to skills and competencies.**

***Index Terms*—cybersecurity, learning analytics, learning metrics, training, exercises**

1. Introduction

Effective learning, teaching, and skills improvement of cybersecurity students and professionals is a critical research area. As there is a high demand for skilled professionals and a shortage of such individuals [1], the development of scalable and effective teaching methods is critical. We focus on the application of learning analytics in the cybersecurity training, specifically in cybersecurity exercises, as a way to provide a more evidence-based and systematic approach for the evaluation of learning impact and to enable the design of more effective learning. This is a critical aspect to consider for educators, organisers and cyber range developers.

Learning analytics (LA) is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [2]. As a field of research, LA aims to predict and advise on identifying students' learning needs and improve pedagogical strategies based on analytical approaches [2]. However, establishing plausible relationships between models derived from quantifiable digital data, and the complex socio-cognitive world of "learning" is challenging [3]. LA is closely intertwined with educational data mining (EDM) that develops, researches, and applies computerized methods to detect patterns in large educational data sets [4].

Cybersecurity training teaches both technical and soft skills, as the field involves technology, people, information, and processes. A wide range of training methods have been developed by universities [5] and organisations to provide cybersecurity education. As part of such cybersecurity trainings, hands-on exercises (both online and classroom) are gaining in popularity in university curricula and professional training paths. Cybersecurity exercises (CSXs) are viewed as an effective and engaging way of teaching both technical and soft skills in addition to CSXs for learning purposes (e.g., as part of university courses, competitions across universities, etc.). Most national and international CSXs (47%) also focus on training and providing participants an opportunity to gain knowledge, understanding and skills [6]. The CSXs can vary significantly in scale and content, ranging from short online or classroom exercises, Capture the Flags (CTFs) to large-scale/multi-stakeholder exercises, etc. However, most share common aims and challenges with respect to learning. We view CSX as a learning or training event in which individuals or teams implement, manage and defend/attack a network of computers at a tactical or strategic level.

As those exercises always leave an extensive digital footprint of learning processes, it makes them an ideal base to develop the methods within the learning analytics field itself. As a result, using these evidence-based learning traces in learning design can improve the experience for both students and specialists. It also helps to investigate the validity of common, yet unsubstantiated claims, such as "everyone feels they had learned important lessons [7]" or "exercises are a very effective way of learning the practical aspects of information security" [8].

We propose a reference model for LA in CSXs in Section 2 offering a practical guide to the exercise organisers to enhance the conceptualisation and integration of learning analytics into the exercise life-cycle. We support the proposals presented in the model with an extensive overview of existing uses of learning analytics by pro-

DOI 10.1109/EuroS&PW51379.2020.00012

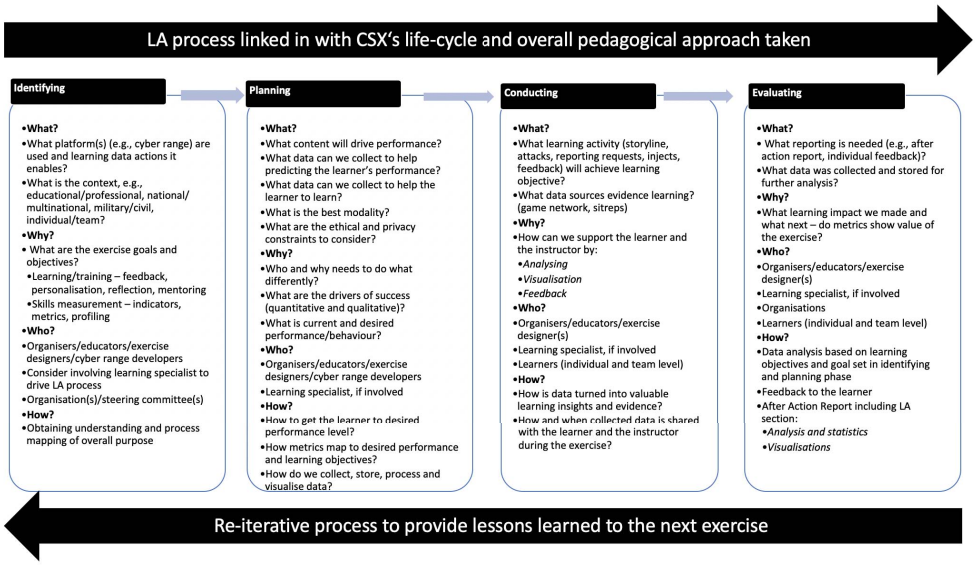

Figure 1. CSX Learning Analytics Reference Model

viding empirical evidence from digital datasets (log files, pcaps) and metrics used in CSXs.

When implementing LA measurements into exercises we need to understand (1) what metrics evidence learning and (2) are they helping the learners to learn or teachers to teach? The metrics (i.e., indicators of learning success) collected and analysed provide technical data (e.g., time, command-line, tools used, etc.), and there appears to be an overemphasis upon what we can measure, instead of measuring what we value—a longstanding concern in educational assessment [9]. However, applying a learning analytics approach and analysing metrics from digital datasets, can provide a more detailed and evidence-based input to more comprehensive learning evaluations, such Kickpatrick or other chosen evaluation models [10].

2. Reference Model for LA in CSXs

Learning analytics should be incorporated to the CSXs' identifying, planning, conducting and evaluating phases (as described by [11]) and be seen as an integral part of the exercise design in line with the overall pedagogical approach selected [12]. When starting to implement a LA approach into an exercise it is useful to think about LA process from aspects of What (Data, Environments, Context), Why (Objectives), Who (Stakeholders) and How (Methods) [13].

We propose a practical tool: the CSX LA reference model, Fig. 1, that builds upon [13] and [11]. Our contribution is to combine and outline the key learning analytics considerations to incorporate into the CSXs life-cycle and support the model with an extensive overview of existing use cases for a practical implementation. Developers would need to consider LA aspects in their initial design of the cyber ranges when they incorporate the technological foundation of instrumenting the exercises.

Asking these learning analytics related questions and finding the answers during an exercise life-cycle, will ensure that learning measurements are not simply an afterthough but rather are incorporated in the "identifying phase" (Fig. 1). Considering questions, such as "What data can we collect that will help learners to learn?" and collecting only that relevant dataset, would help with the challenges of storing huge datasets from an exercise and later trying to determine what data could be used to provide feedback. For example, if the learning objective of an exercise is to improve the incident response process, then timestamps that would indicate team communication would be critical data to collect [14]. However, when the proficiency of using various forensics tools and commandlines is exercised, then capturing bash history or keystrokes is relevant (e.g., [15], [16]). Also consideration should be given how to support instructors in giving feedback. Designing automated feedback that takes into account the users behaviour and predicts their actions and questions becomes available and can make the learning experience more individualised and effective (e.g., [17], [18]). Depending on the purpose, scale and type of a CSX, it may be recommended to include a learning specialist in the organising team to coordinate the LA implementation throughout the CSX life-cycle.

2.1. Comparison to Other Frameworks in CSXs

There are various frameworks that have been developed for CSXs and cybersecurity education. Several academic papers (e.g., [8], [56]) and non-academic guides (e.g., [11], [57], [58]) describe the overall CSXs de-

TABLE 1. MAPPING LA PROCESSES BY TYPES OF CSXS. NOTE: ONE PAPER MAY COVER VARIOUS TOPICS, WE MAP HERE MAIN LA THEME.

Exercise Type/LA Process	Collection	Storage	Cleaning	Integration	Analysis	Visualisation	Action
Capture the flag	[19], [20], [21], [22], [23], [24], [25]	[23]		[22]	[19], [20], [21], [22], [26]	[27], [28]	
Discussion based game							
Drill							
Red team / blue team	[29], [14], [30], [31]	[32]			[14]	[33], [33], [34]	
Seminar							
Simulation	[35], [36], [37], [38], [39], [40], [41], [42]			[35]	[35], [36], [43], [44]	[38], [39]	
Table-top	[45], [46]					[45]	
Workshop	[47], [48]						
Exercise/lab	[15], [49] [50], [16], [51], [52], [53], [8], [54]		[50]	[52], [53]	[50], [55]	[15], [49], [32]	[17]

sign and evaluation process. Also, more general approaches to exercises are proposed: [59] that describe an extended competence development and assessment framework and [56] suggests specific metrics in complex simulated CSXs. For designing competition based exercises [60] describes a mindmap, while for CTFs [61], [62] describe 5 steps when designing an evaluation (purpose, frame, questions, information needed and systematic collection method). At a higher level conceptual level, models for a multidisciplinary cybersecurity training methodology [63], a pedagogical framework [12], a framework incorporating cognitive aspects [64], and a holistic model of professional competence in the cyber domain [65] have also been developed.

Despite some of these models providing elements of analytical approaches, no framework has been developed that would explicitly include the use of LA methods in the CSXs when looking at the LA in each phase of exercise life-cycle or what would be the appropriate metrics or learning indicators to measure when considering the pedagogical approach taken.

3. Supporting Reference Model with Existing Research Results and Practical Considerations in Implementing LA Approach in CSXs

To gain an insight about what empirical evidence and metrics have been collected and analysed in the scientific literature, we conducted extensive related work review. We searched Google Scholar, a widely used and available scientific database and limited the search to empirical studies published in a peer-reviewed journals and conferences in English. We used the keyword "learning" in combination with "cyber(-)security" for different exercise types. For the exercise type classification we followed the European Union Agency for Network and Information Security (ENISA) taxonomy: Capture the flag, Discussion based game, Drill, Red team / Blue team, Seminar, Simulation, Table-top and Workshop [6]. In some cases, exercises with gamified elements are referred as "serious games", or with competitive elements as "competition", and thus included these in our search strings for completeness.

We reviewed the abstracts of 200 articles for each exercise type, as this was considered sufficient to encompass all relevant material. The academic papers identified as covering CSXs for learning purposes were manually reviewed to identify LA topics and all empirical/analytical learning data collected or analysed. It should be noted that even though there is a large number of articles describing the CSXs, these do not necessarily include empirical data from the digital datasets to evidence learning. We used feature mapping [66] in which the content is analysed and recorded in a standardized format documenting the key features of a predetermined aspects (i.e., LA model [2]) to produce a summary of the topic. Related work was mapped to an overview matrix in Table 1 by the ENISA exercise types [6] and LA model [2] consisting of collection, storage, data cleaning, integration, analysis, representation and visualization and action.

3.1. Uses of Digital Datasets to Evidence of Learning Effectiveness in CSXs

An overview matrix that shows the application of learning analytics by exercise types and LA process steps is presented in Table 1.

3.1.1. Capture the Flag. A CTF is typically a challenge designed to help sharpen cybersecurity skills and provide hands-on learning taking various styles, such as jeopardy, attack-defence and a mix of the two. However, participating in CTFs does not necessarily ensure future success, and participants rarely receive a detailed critique of their performance, which is essential in learning [67].

There are few studies that provide empirical evidence from digital game-play traces for learning and skill acquisition in CTFs. Clothia and Novakovic [19] show that Jeopardy-style challenges with automatic marking of flag submissions complemented by manual marking of detailed written answers provided students with instant feedback during an exercise with an improved student satisfaction with the academic course. The ability of students to acquire the flag is highly correlated with their overall marks (written assignment), and flag-based marking effectively assesses a student's basic skills and understanding of cybersecurity topics [19]. However, acquiring flags and a student's deeper understanding of the underlying issues is much less-correlated [19]. Cheung et al. [20] describe

combining logging results from log servers along with a key-logger to track participant sessions. The authors gathered statistics about login times (e.g., the environment was most often used on Saturdays between 3:30pm and 8pm) and command usage to see what commands students were still having trouble with after the lectures, which helped determining what to spend more time on [20]. Chapman et al. [21] evaluate PicoCTF based on survey responses and user interaction logs to explore the effectiveness of design choices (e.g., younger students prefer a game interface compared to older students, and a general dislike for challenges requiring learning new tools). The system kept track of answers submitted by every team, both correct and incorrect—recording the time, content and relevant problem identifier, as well as the IP address of the submission [21]. In order to evaluate student engagement, the authors determined periods of time during which teams were most active (i.e., time interval at submissions) [21]. Tseng et al. [22] also focused on data collection from heterogeneous environments and proposed an ontology to represent concepts (consisting of teams profile, skills (tools), test items and environment) within exercises and their relationships (including linking heterogeneous logs to participants' intentions). The authors focused on problem-solving behaviours and applied a modified a priori algorithm, analysed frequent item-sets and identified learning behaviours such as that novices keep guessing keys and better performing students focused on particular items [22]. The above papers describe metrics applied and the authors research selection of patterns and relations in the digital learning dataset. [23] describe possible metrics, such as teamwork, challenge difficulty, challenge strategies, tools, and general problem solving techniques. Whereas [27] and [26] propose, apply, and experimentally evaluate data analysis and machine learning techniques to obtain interactions from the in-game data and provide learners who progress differently with individualized feedback. Vykopal et al. [25] suggest decomposing the exercise training activity into individual levels to achieve specific learning objectives, and collecting timestamps (or events) such as start and end of the game, start and end of each level, submission of incorrect flags and their content, hints used, skipping a level, displaying a level's solution, game ID [24].

Dark and Mirkovic [61] bring out an aspect that to measure learning we may need to rely on proxy indicators (e.g., identifying reasonable and observable indicators of adversarial thinking). Overall, we can see that analytical approaches are emerging to evaluate learning from digital dataset(s) that are good starting points when incorporating the learning analytics process into the exercises.

3.1.2. Red Team and Blue Team. Red-Blue exercises are often team-based and therefore add another complexity level in LA application—measuring team learning vs. individual learning. Several authors discuss learning impact; however, not evidence-based analysis from digital footprints (e.g., [68] describes organizing team-based exercise, where teams were directly monitored and evaluation of skills improvement is observational). Typical evaluation methods are score-boards, verbal feedback and after-action reports highlighting conclusions from manual analysis of exercise data [34] that do not apply an analytical approach using digital datasets. Some papers do describe data collection and perform initial LA on digital dataset, e.g., [14] breaking time into intervals that can be meaningful for different learning objectives (e.g., incident responding, team communication) with walk-through. Ošlejšek et al. [33] show that visual analytics tools could provide automated statistical analysis and an in-depth insight into the learner's behaviour using observation software (Fowler's analysis) [32]. Vykopal et al. [34] describe an interactive timeline visualisation allowing learners to explore a scoring timeline and details about individual events.

3.1.3. Simulation. Simulations are a very common type of exercise and take various forms, e.g., online vs. live, gamified, etc. There are some emerging examples of LA performed already. Thompson and Irvine [35] present basic LA data, such as time played and discuss abstraction layers for data analysis. No formal effectiveness assessment was performed and conclusions are based on ad-hoc student interaction and logs review [35]. Legato and Mazza [36] assume a set of regeneration points that correspond to skill achievement through learning. However, the model was not validated and numerical results are reported for illustrative purposes only.

Nicholson et al. [37] system uses a dynamic tailoring system, which maintains a model of student proficiency and adapts training difficulty, while providing detailed feedback. Santos et al. [38] use a post-simulation analysis using a variety of graphics and reports to verify the network traffic, which teams were attacked, which services are still vulnerable, teams activity rate and strategy used, etc. This data enables statistical evaluation of what happened during the simulation, including how teams perform compared to previous exercises [38]. Furfaro et al. [39] used cloud based learning system including a dashboard (for managing scenarios, agents and VMs, displaying system usage and statistics, etc.); a report tool (provides statistical data from logging engine, and queries for business intelligence analysis displayed on charts), and a set of development tools.

Many simulations have a gamification element. Tioh et al. [69] performed a literature review (18 papers including some kind of empirical effect measurement) and concluded "the question as to the effectiveness of serious games dealing with the training of cyber security is a difficult one to answer conclusively at this point". We identified additional papers with games for security specialists or students (e.g., [69], [40], [41]). However, none used empirical learning analysis from digital datasets.

Several authors focus on cognitive levels of learning process in articles covering simulation-type exercises, as simulations allow experimentation. Some examples include a simulation-based approach to understanding cybersecurity threats when attempting multiple actions, the user is provided with an "awareness" measure [70]; a socio-technical systems approach to support the emerging role of systems thinking and using an agent-based simulation tool to change the students' thinking [71] [72]; a computational model based on Instance-Based Learning Theory that proposes a way to analyse the cyber analyst's awareness at both threat level and attack scenario level [43]. Such human dynamic decision making analysis

can help to determine various player models at individual and aggregated levels [73].

3.1.4. Table-top. A table-top exercise is typically a meeting to discuss a fictional cyber emergency situation increasing participants' engagement and strengthening their awareness and competences in strategic decision-making [74]. Although typically digital traces are typically limited in table-tops, some research on the system architecture for tracking learning process is emerging. Brilingaite et al. [45] presents a model of a web-based environment that enables playing table-top exercises in person and remotely. The environment includes the visual representation of decision-making during the game and provides the comparison to the correct solution.

However, in many cases the data analysis is not presented, offering an experience without learning process or empirical impact data. Cozine et al. [75] examine the pedagogical approach to incorporate game play, specifically probability-based tabletop exercises, into course curricula and collected survey data from students enrolled in the courses. Ottis [46] presents a light weight tabletop exercise format that has been successfully used in cybersecurity education to demonstrate these and many other concepts to master level students.

3.1.5. Drill, Seminar, Discussion Based Game and Workshop. The research builds upon experience but lacks evidence-based measures or uses mainly surveys/self-assessments as a tool to analyse learning. Some research results start to emerge, such as [47] describing a workshop using clickers for running a series of questions that allow easy data collection and analysis.

3.1.6. Exercises with no clear classification. In several papers CSXs are described in generic manner as "exercise" (often in classroom environment and part of a course) or include multiple exercise types (e.g., platform allowing both Red-Blue team and individual simulation game). However, several empirical learning data analyses have been completed. Weiss et al. [15] express that simply recording the number of correct answers is inferior to in-depth assessments and explores the use of command line history and visualization. Authors follow the "path" taken by a student in command-line when completing different tasks and levels (for skills level measurement some commands were identified as significant) [15]. The "path" is visualized in graph that can be decomposed into chains and cycles [15]. Similarly, Labushange et al. [49] assesses technical skill level based on indexed similarity (i.e., participants were ranked based on commands usage to achieve objectives) and classifies actions that can be automatically deducted using the clustering of commands (e.g., combination of "ifconfig", "sudo apt-get install nmap" and "sudo nmap -sT targetIP" together was classified as reconnaissance). However, the paper does not go into details of how such clustering can be achieved [49]. Caliskan et al. [50] use educational data mining, machine learning and identifies metrics for learning effectiveness (predicting final grade) in the university classroom course with efforts to validate their predictive model. Caliskan et al. also compare participant evaluation metrics and scoring systems in [76]. Moore et al. [55] focuses on the development of specific individual skill levels and state that competence in progressively harder levels of capabilities was observed over time in relation to the training components. [52] apply automated mechanism for parsing log entries into blocks of time during which participants are focused on specific high-level objectives, with instrumentation capturing students' computer-based transactions [53]. Several authors propose evaluation metrics or data that should be collected. However, the actual learning data analysis is lacking so far. For example, [54] suggest metrics such as time, participant numbers who succeed and feedback, and [8] suggests number of detected attacks from total attacks for learning task of monitoring systems' security, etc.

3.2. Analysis of LA Process Described in CSXs

The explicit use of learning analytics and relevant vocabulary in the cybersecurity education (incl. CSXs) is in its early stages. However, recently there is more focus in the academic community on applying LA methods to improve the cybersecurity education [77]. The discussion below is organized by LA process to analyse main steps in a CSX's life-cycle. The process steps span across the life-cycle and are re-iterative, however the whole process needs to be designed in planning phase and instrumented to the cyber ranges.

3.2.1. Collection and Acquisition. Existing research focuses on data collection—i.e., how to build an exercise platform, cyber range, etc. However, the literature lacks considerations for what purpose and what data is actually collected, and also contains discussion on learner consent and ethical aspects. As the tendency is for collecting simple technical measures, rather than more complex cognitive learning measures (Section 4), often there is no clear connection whether it was collected for evidencing learning, and what data is relevant for evidencing learning.

3.2.2. Storage. How and what data is stored (and how long period) is mostly not covered (only few examples such as [32]). The security of information stored and privacy concerns (anonymization processes) appear not to be a high priority. Only one paper [23] was identified that describes security measures for captured data.

3.2.3. Cleaning. The data cleaning process is not typically described, however network traffic (pcap) and other datasets are expected to include non-relevant data. Few papers start discussing what processes were used to clean the data, e.g., [50]. Also the principles of privacy and anonymization needs to considered here.

3.2.4. Integration. As the exercises generate multiple datasets, combining multiple datasets and different formats, e.g., technical and timing data with self-reported learner data, is an important process step. Some examples were found about aligning timestamps of datasets, such as [52], [53] [14], [22]—however, no detailed methods are provided on how data is processed and time-synced.

3.2.5. Analysis. Due to LA being a novel research area, the data analysis performed has been limited, with a predominance of studies undertaking low-level analysis relating to the readily accessible data such as the reporting of number/frequency login times, number of messages posted, time online, etc to academic performance as measured by grades [78]. Similarly in CSXs, we did not identify commonly used tools or methods for data analysis (statistical, machine learning model, etc.). Similarly to overall LA field, the analysis has genreally been conducted using easily obtainable metrics (such as time), see Table 2 and in typically linked to high-level learning objective(s).

3.2.6. Representation and Visualization. The challenge is determining the relationship between visualizations and learning. The feedback about low level user actions—such as number of log ins, videos watched, or documents submitted—does not illustrate progress in learning for students or educators [79]. Visualizations and dashboards usefulness and effectiveness is not widely covered in the exercises. Several papers, such as [32] analyse the use of visualizations in CSXs in addition to describing the system architecture.

3.2.7. Action. Actions, such as intervention, optimization, systematic improvements (including design) are not necessarily evidence-based. Rather learning design choices are based on the authors' experience or the learners' self-reporting/survey evaluations. Some relevant research is emerging how to improve feedback loop from using digital traces, e.g. [17].

4. Inferring Learning from Digital Datasets—What to Measure?

The various frameworks in Section 2.1 have been developed for learning in CSXs and cover different aspects but none directly incorporate or utilize learning analytics processes. When inferring learning from the granular digital dataset, the challenge is linking learning objectives and competencies to the granular raw data to, as the design of a CSX should follow a top-down pattern [18]. The cyber range should be designed to allow such learning design and measurement process.

4.1. What Metrics are Collected and Analysed to Evidence Learning?

Table 2 summarises the learning indicators from digital datasets that have been used in academic research, which could be used as a starting point to brainstorm when selecting the metrics to measure that the training objectives have been achieved. It should be noted, that any papers on learning in the CSXs are based on the experience and interpretation of the authors or based on the traditional learner evaluation (e.g., feedback surveys, evaluation forms).

4.2. What to Consider in Choosing Metrics?

We should focus on measuring what we value. The metrics used in CSXs often focus on easily measurable data (e.g., time spent, number of attacks mitigated, etc.) and individual actions. However, the students are "too easily satisfied that a system is secure after identifying only one possible source of security for a system rather than seeking to explore the adversarial space more thoroughly" [83]. Thus it is important to understand not only whether the students found the correct answer but how they found it [15]. There is some research that starts to look into "how" the learner completes tasks (i.e., use of tools, attempts, submission of wrong answers), such as [52], [53], [49]. However, validation is limited (e.g., 4 participants [49]). In regards to teamwork and communication, there is some research, such as [81], [36] that have started to explore the use of analytics as evidence for achieving learning in teams.

Also as learning is complex cognitive process, the further research should focus on cognitive metrics, such as Knox et al. [82]. From the LA research, a similar measure to "cognitive presence" can be applied in cybersecurity training (e.g., "Active Learning Squared (AL2)" paradigm, which emphasises metacognition and uses both active student learning and machine learning [84], [85]).

Metrics are valuable, however, "being able to report upon a metric does not mean that you should use it, either in the tool, or in reporting its worth [3]". The metrics will depend on the exercise goals that in turn are guided by different pedagogical principles (e.g., behaviorist, cognitivist or constructivist) [12] and the wider evaluation model chosen [10]. Therefore, we need to be mindful of learner and learning process, and measurement should move towards mapping of digital traces describing student activity onto interpretable constructs of interest (e.g., Knowledge Components, Q-matrix), which facilitate actionable analytics [86].

5. Challenges in Implementing LA approaches in CSXs

Scientifically-valid evidence that learning outcomes were achieved in CSXs is difficult to obtain, especially as the exercise design, objectives, technology and learner characteristics vary. These factors make inter-institutional and between exercises comparisons difficult. However, sharing the measurement results would enhance measuring that the learning was achieved and new skills obtained.

The data analysis until now has been limited, with a predominance of studies undertaking low-level analysis using easily obtainable metrics, such as login times, time to complete tasks, number of attack mitigated, see Table 2. The related work did not reveal commonly used data analysis tools or methods (statistical, machine learning, etc.) in CSXs, but developing and sharing methods used would enhance validity of the results.

Security challenges, such as intrusion detection, insider threats, malware detection, and phishing detection lack exact algorithmic solutions and the boundary between normal and anomalous behaviour isn't clear-cut as attackers are continuously improving their techniques and strategies [87]. This also impacts LA, as it needs to keep up with moving algorithms and learning patterns. In addition, the challenge relates to data volume—one large exercise can create terabytes of data including multiple

TABLE 2. METRICS FROM DIGITAL DATASETS TO CONSIDER WHEN MEASURING LEARNING IN CSXS

Metrics	Reference	Learning Objective/Competency	Validated	Validation Method/Results
Technical Metrics				
Time and Time Periods				
Total completion time	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Time taken to win the exercise	[54]	Effectiveness of the overall exercise	No	Recommendation
Time before nmap	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Time spent on scenario (incl replays)	[35]	Network filters	No	Interpretation, CyberSiege, 1 lab, 149 students
Time taken to recover from a successful attack	[8]	Incident handling / response	No	Recommendation
Downtime of attacked service compared to attack duration	[8]	Perform DDoS	No	Recommendation
Time period during the attack response (5 timestamps)	[14]	Incident response / handling	No	Log analysis vs. self-reporting, Locked Shields, 19 teams
Time played	[35]	Various cybersecurity skills	No	CyberCiege, online platform
Mean time per action	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 CSXs
Commands, including count of commands				
Time-to-Detect	[80]	Defending network against attack	No	Bivariate regression analysis, multivariate regression analysis, and principal component analysis
Time-to-apProval (by team controller)	[80]	Defending network against attack	No	Bivariate regression analysis, multivariate regression analysis, and principal component analysis
Time-to-End	[80]	Defending network against attack	No	Bivariate regression analysis, multivariate regression analysis, and principal component analysis
Category Correct (NIST category of inject correctly identified)	[80]	Defending network against attack	No	Bivariate regression analysis, multivariate regression analysis, and principal component analysis
Total Commands Entered	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Reconnaissance Similarity (index of most accurate command)	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Number of File Commands	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
File Server Identification Similarity (index of most accurate command)	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Number of Incident Commands	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Incident Commands Similarity (index of most accurate command)	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Number of Threat Commands	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Threat Commands Similarity (index of most accurate command)	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Number of System Administration Commands	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
System Administration Similarity (index of most accurate command)	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Command usage	[20]	Various cybersecurity knowledge	Partial	Interpretation, feedback survey, CTF with lectures, no details of flags or commands
Count of Events, Objects or Individuals				
Number of scenario replays	[35]	Network filters	No	Interpretation, CyberSiege, 1 lab, 149 students
Number of successful attacks	[8]	Implement security configurations	No	Recommendation
Number of detected attacks from total number of attacks	[8]	Monitor systems' security	No	Recommendation
Number of attacks correctly identified	[8]	Analyse logs and do forensics	No	Recommendation
Number of open ports/services detected compared to total number of open ports	[8]	Perform scanning and enumeration	No	Recommendations
Number of successful backdoors accesses to target systems kept until the exercise end	[8]	Cover tracks and place backdoors	No	Recommendations
Number Actions Per Block	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 CSXs
Number Actions Per Block	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 CSXs
Number of finalist participants	[54]	Effectiveness of exercise	No	Recommendation
Number of participants succeeding brute-force attack	[54]	Effectiveness of exercise	No	Recommendation
Number of participants successfully exploited Windows vulnerability	[54]	Effectiveness of exercise	No	Recommendations
Compromised services as reported by attacking and defending teams	[81]	Various cybersecurity skills	No	Statistical analysis (team performance)
Number of attack and vulnerability reports per defending team	[81]	Various cybersecurity skills	No	Statistical analysis (team performance)
Attempted and successful attacks on teams DMZ, calculated by NIDS analysis	[81]	Various cybersecurity skills	No	Statistical analysis (team performance)
Number Different Software Tools	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 CSXs
Number Transitions Between Software Tools	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 CSXs
Number Returns to a Previous Software Tool	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 CSXs
Number of valid flags submitted	[19]	Basic encryption, access control, protocol analysis, web security, RE	Yes	Correlations over 3 datasets to final grade
Total number of logins over two months per weekday	[20]	Various cybersecurity knowledge	Partial	CTF, 3 iterations, no details of flags provided Interpretation, feedback survey CTF with lectures, no details of flags provided
Total number of logins over two months per hour	[20]	Various cybersecurity knowledge	Partial	Interpretation, feedback survey CTF with lectures, no details of flags provided
Tools, Commands and Methods Used by Learner				
Commandline (nmap, Linux bash history)	[15]	Network reconnaissance	No	Interpretation, 24 teams of students, 2 classes 2 schools
Reconnaissance Commands	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Use of Internet browsers	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Frequency of software tools use	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Number of software tools used	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Type of software (general vs specialised) tools used	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Choice of tools used	[56]	Efficiency of student actions	No	Idea proposed, no measurements
Programming languages used	[56]	Efficiency of student actions	No	Idea proposed, no measurements
Input logs				
Direct input (logs)	[56]	Not specified	No	Idea proposed, no measurements
String similarity metrics (using e.g., Levenshtein distance)	[56]	Efficiency of student actions	No	Idea proposed, no measurements
Blocks of activity	[52], [53]	Forensics	No	Interpretation, TracerFire, 26 participants, 2 exercises
Log data: frequent itemsets to learning behaviors	[22]	Various cybersecurity knowledge	Yes	Data analysis: 7-hours competition collecting 8,257 logs from CTF Server and 407,623 logs from GRR Server
Log data: start/end, incorrect flags, hints used, skipping level, displaying solution, game ID	[24]	Various cybersecurity skills	No	No detailed metrics analysis
Network data: dst_ip, dst_port, ip_proto, ip_len, signature, signature_gen, priority, class, status	[50]	IDS alerts, network sessions, or top destination IP addresses	No	Nave Bayes and decision tree algorithms. For results validation, k-fold cross validation, 10 iterations, 17 students, 1 lab
Service status (active/non-active, vulnerable/not-vulnerable)	[38]	Various cybersecurity skills	No	Idea proposed, no measurements
Network traffic, teams attacked, vulnerable services, teams activity rate and strategy used, network traffic peaks, protocols usage, etc.	[38]	Various cybersecurity skills	No	Recommendation to look at logs to statistically evaluate what attacks were most efficient, possible damage caused, threats easily defended and a team's performance to prior CSX
Tags: OS, programming language, vulnerability, language, associated CVE, tools	[23]	Various cybersecurity knowledge	No	Discussion and experience
Logs: automatic scoring, pcap, chats/emails screen capture, video and audio	[81]	Various cybersecurity skills	Yes	Statistical analysis (team performance)
Joint Information Exchange Environment and Chat logs (hashtags)	[80]	Defending network against attack	No	Bivariate regression analysis, multivariate regression analysis, and principal component analysis
Other—rankings, indexes, indicators, manual				
Success rate (correct answer) from total challenges	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Abandonment rate of challenges	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Challenge submission accuracy	[16]	Forensics	No	Interpretation, TracerFire, 11 participants
Submission data and completion rate of challenges	[21]	Forensics, cryptography, RE, web and scripting exploitation, binary exploitation	No	Discussion and experience, PicoCTFs, 1588 participating teams, survey data
Total Similarity (based on several similarity indexes)	[49]	Defending network against attack	No	Interpretation, 4 participants, self-developed cyber range
Proxy indicators (e.g., observable indicators of adversarial thinking)	[20]	Various cybersecurity knowledge	No	Recommendation, overall CTF evaluation model
Scalar unit for each mitigation completed under cooperation, algorithm	[36]	Attack mitigation	No	Algorithm to measure skill improvement, SO tool
Soft Skills / Cognitive Metrics				
Teamwork: task ownership changes, dashboard, timecounter, option to mark challenge as difficult or solved	[23]	Various cybersecurity knowledge, teamwork	No	Discussion and experience
Awareness measure/critical thinking/decision making	[70], [43], [72], [73]	Understanding threats and systems, making desicions	No	Discussion and experience
Cognitive Agility Index	[82]	Individual cognitive performance	Yes	Regression analyses, science based, validation of 31 participants

and big datasets. The large amount of data generated by automatic logs and sensors necessitates efficient and automated data and LA techniques. There may not be enough traces to identify learning patterns (e.g., short time of detection, gaps in time-line) and data may be very diverse (e.g., different OS, applications). Therefore, identification of the relevant learning traces requires techniques that can deal with such imbalance and diversity. To combine multiple datasets and formats, e.g., technical and timing data with self-reported learner data no detailed descriptions or methods are provided how data is processed and time-synced. However, some examples were found about aligning timestamps of datasets, such as [53], [14], [22].

Also, the CSXs and related studies mostly work over a short period but it is known that short-term interventions are not particularly effective at affecting behavioral change [88]. Thus longitudinal studies are needed to evidence learning and behavior change as result of the exercises, and also to separate from other learning.

6. Conclusion

The opportunity to improve the learning in CSXs as part of educational effort is missed without considering the learners experience, different learning styles and pace, and the impact of the learning environment. The application of learning analytics and analysing digital datasets can provide a deeper understanding of learning behaviour and lead to evidence-based improvement. The consideration of LA aspects is also vital for the cyber range developers, as they design the technological foundation of instrumenting exercises that enable the application effective LA methods.

We proposed a LA reference model to assist in implementing LA into the CSXs life-cycle to achieve a more adaptive design and measurement using evidence-based data from the learning environment. As a practical starting point, we shared extensive related work overview of existing research describing some aspects of learning analytics process and the analysis of empirical evidence from the digital datasets to assist in implementing the model across all exercise types. We described the learning indicators (metrics) used for evidencing learning in CSXs, with focus on analytical evidence from digital dataset. Such metrics are mainly simple technical measures (time, number of attacks mitigated, availability of service, etc.) that are not necessarily validated and may not evidence effective learning (i.e., metacognition achieved). With LA and evidence-based measurement, we also need keep in mind and validate that what we measure (i.e., metrics used) actually help learners to learn. In turn, the validated metrics have the potential to provide more detailed and evidence-based input that form an integral part of the comprehensive training evaluations.

Further work should seek to identify and validate what learning metrics are evidencing the learning process and learning improvement in CSXs. Understanding the current use of learning analytics in CSXs is expected to help setting the baseline for further research and practical implementation by combining two evolving disciplines. By doing this, the cybersecurity community can establish more evidence-based and systematic approach for the evaluation of learning impact that will enable the design of more effective learning experiences.

Acknowledgments. This work was partially supported by the ECHO project which has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement no 830943.

References

[1] Nita G Brooks, Timothy H Greer, and Steven A Morris, "Information systems security job advertisement analysis: Skills review and implications for information systems curriculum," *Journal of Education for Business*, vol. 93, no. 5, pp. 213–221, 2018.

[2] George Siemens, "Learning analytics: The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013.

[3] Kirsty Kitto, Simon Buckingham Shum, and Andrew Gibson, "Embracing imperfection in learning analytics," in *Proceedings of the 8th International Conference on LAK*. ACM, 2018, pp. 451–460.

[4] Zacharoula Papamitsiou and Anastasios A Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence.," *Journal of Educational Technology & Society*, 2014.

[5] Krzysztof Cabaj, Dulce Domingos, Zbigniew Kotulski, and Ana Respício, "Cybersecurity education: Evolution of the discipline and analysis of master programs," *Computers & Security*, vol. 75, pp. 24–35, 2018.

[6] Adrien Ogee, Razvan Gavrila, Panagiotis Trimintzios, Vangelis Stavropoulos, and Alexandros Zacharis, "The 2015 report on national and international cyber security exercises," ENISA, https://www.enisa.europa.eu/publications/latest-report-on-national-and-international-cyber-security-exercises.

[7] Art Conklin, "Cyber defense competitions and information security education: An active learning solution for a capstone course," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*. IEEE, 2006, vol. 9, pp. 220b–220b.

[8] Victor-Valeriu Patriciu and Adrian Constantin Furtuna, "Guide for designing cyber security exercises," in *Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy*, 2009, pp. 172–177.

[9] Gordon Wells and Guy Claxton, *Learning for life in the 21st century: Sociocultural perspectives on the future of education*, John Wiley & Sons, 2008.

[10] Hanne Foss Hansen, "Choosing evaluation models: a discussion on evaluation design," *Evaluation*, vol. 11, no. 4, pp. 447–462, 2005.

[11] Jason Kick, "Cyber exercise playbook," 2014, MITRE Corporation, https://www.mitre.org/publications/technical-papers/cyber-exercise-playbook.

[12] Mika Karjalainen, Tero Kokkonen, and Samir Puuska, "Pedagogical aspects of cyber security exercises," in *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2019, pp. 103–108.

[13] Mohamed Amine Chatti, Anna Lea Dyckhoff, Ulrik Schroeder, and Hendrik Thüs, "A reference model for learning analytics," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5-6, pp. 318–331, 2013.

[14] Kaie Maennel, Rain Ottis, and Olaf Maennel, "Improving and measuring learning effectiveness at cyber defense exercises," in *Nordic Conference on Secure IT Systems*. Springer, 2017, pp. 123–138.

[15] Richard Weiss, Michael E Locasto, and Jens Mache, "A reflective approach to assessing student performance in cybersecurity exercises," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. ACM, 2016, pp. 597–602.

[16] Austin Silva, Jonathan McClain, Theodore Reed, Benjamin Anderson, Kevin Nauer, Robert Abbott, and Chris Forsythe, "Factors impacting performance in competitive cyber exercises," in *Proceedings of the Interservice/Interagency Training, Simulation and Education Conference, Orlando, FL*, 2014.

[17] Valdemar Švábenskỳ, Jan Vykopal, and Pavel Čeleda, "Toward an automated feedback system in educational cybersecurity games," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE19)*, 2019.

[18] Margus Ernits, Kaie Maennel, Sten Mäses, Toomas Lepik, and Olaf Maennel, "From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises," in *ICCWS 2020 15th International Conference on Cyber Warfare and Security*, 2020.

[19] Tom Chothia and Chris Novakovic, "An offline capture the flag-style virtual machine and an assessment of its value for cybersecurity education," *USENIX Summit on Gaming, Games, and Gamification in Security Education*, 2015.

[20] Ronald S Cheung, Joseph Paul Cohen, Henry Z Lo, Fabio Elia, and Veronica Carrillo-Marquez, "Effectiveness of cybersecurity competitions," in *Proceedings of the International Conference on Security and Management*, 2012, p. 1.

[21] Peter Chapman, Jonathan Burket, and David Brumley, "Picoctf: A game-based computer security competition for high school students," in *3GSE*, 2014.

[22] Shian-Shyong Tseng, Sung-Chiang Lin, Ching-Hao Mao, Tsung-Ju Lee, Guan-Wei Qiu, and Ming-Han Lin, "An ontology guiding assessment framework for hacking competition," in *10th International Conference on Ubi-media Computing and Workshops*. IEEE, 2017, pp. 1–4.

[23] Nicholas Capalbo, Theodore Reed, and Michael Arpaia, "Rtfn: enabling cybersecurity education through a mobile capture the flag client," in *Proceedings of the International Conference on Security and Management*, 2011.

[24] Jan Vykopal and Milos Barták, "On the design of security games: From frustrating to engaging learning," in *ASE@USENIX Security Symposium*, 2016.

[25] Jan Vykopal, Martin Vizváry, Radek Oslejsek, Pavel Celeda, and Daniel Tovarnak, "Lessons learned from complex hands-on defence exercises in a cyber range," in *Frontiers in Education Conference*. IEEE, 2017, pp. 1–8.

[26] Valdemar Švábenskỳ, "Analyzing user interactions with cybersecurity games," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, 2019, pp. 1295–1295.

[27] Valdemar Švábenskỳ, Jan Vykopal, and Pavel Celeda, "Towards learning analytics in cybersecurity capture the flag games," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, 2019, pp. 1255–1255.

[28] Radek Ošlejšek, Vít Rusňák, Karolína Burská, Valdemar Švábenskỳ, and Jan Vykopal, "Visual feedback for players of multi-level capture the flag games: Field usability study," *arXiv preprint arXiv:1912.10781*, 2019.

[29] Elena Sitnikova, Ernest Foo, and Rayford B Vaughn, "The power of hands-on exercises in scada cyber security education," in *IFIP World Conference on Information Security Education*. Springer, 2009, pp. 83–94.

[30] Ryan Richards, Abdullah Konak, Michael R Bartolacci, and Mahdi Nasereddin, "Collaborative learning in virtual computer laboratory exercises," *Network, Security*, vol. 155, pp. 9, 2015.

[31] Pavel Čeleda, Jakub Čegan, Jan Vykopal, and Daniel Tovarňák, "Kypo–a platform for cyber defence exercises," *M&S Support to Operational Tasks Including War Gaming, Logistics, Cyber Defence. NATO Science and Technology Organization*, 2015.

[32] Radek Ošlejšek, Dalibor Toth, Zdenek Eichler, and Karolína Burská, "Towards a unified data storage and generic visualizations in cyber ranges," in *ECCWS 2017 16th European Conference on Cyber Warfare and Security*, 2017, p. 298.

[33] Radek Ošlejšek, Jan Vykopal, Karolína Burská, and Vít Rusňák, "Evaluation of cyber defense exercises using visual analytics process," in *Frontiers in Education Conference*. IEEE, 2018, pp. 1–9.

[34] Jan Vykopal, Radek Ošlejšek, Karolína Burská, and Kristína Zákopčanová, "Timely feedback in unstructured cybersecurity exercises," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM, 2018, pp. 173–178.

[35] Michael Thompson and Cynthia Irvine, "Active learning with the cyberciege video game," in *Proceedings of the 4th conference on Cyber security experimentation and test*. USENIX Association, 2011, pp. 10–10.

[36] P. Legato and R. M. Mazza, "Modeling and simulation of cooperation and learning in cyber security defense teams," in *Proceedings - 31st European Conference on Modelling and Simulation, ECMS 2017*, 2017, pp. 502–509.

[37] Denise Nicholson, Lauren Massey, R O'Grady, and E Ortiz, "Tailored cybersecurity training in lvc environments," in *MODSIM World Conference, Virginia Beach, VA*, 2016.

[38] Andre LM Santos, Fred A Freitas, Leandro J Martins, Rodrigo L Magalhes, and Saulo RF Hachem, "Towards a cloud-based cyber war simulator," in *Proceedings of SBGames*, 2012.

[39] Angelo Furfaro, Antonio Piccolo, Andrea Parise, Luciano Argento, and Domenico Saccà, "A cloud-based platform for the emulation of complex cybersecurity scenarios," *Future Generation Computer Systems*, vol. 89, pp. 791–803, 2018.

[40] Affan Yasin, Lin Liu, Tong Li, Jianmin Wang, and Didar Zowghi, "Design and preliminary evaluation of a cyber security requirements education game (sreg)," *Information and Software Technology*, vol. 95, pp. 179–200, 2018.

[41] Natalie Coull, Iain Donald, Ian Ferguson, Eamonn Keane, Thomas Mitchell, Oliver V Smith, Erin Stevenson, and Paddy Tomkins, "The gamification of cybersecurity training," in *International Conference on Technologies for E-Learning and Digital Entertainment*. Springer, 2017, pp. 108–111.

[42] Kyle E Stewart, Jeffrey W Humphries, and Todd R Andel, "Developing a virtualization platform for courses in networking, systems administration and cyber security education," in *Proceedings of the 2009 Spring Simulation Multiconference*, 2009, p. 65.

[43] F. Wu and C. Gonzalez, "How could cyber analysts learn faster and make better decisions?," in *24th Conference on Behavior Representation in Modeling and Simulation, BRiMS 2015, co-located with the International Social Computing, Behavioral Modeling and Prediction Conference, SBP 2015*, 2015, pp. 35–42.

[44] Bin Zhang, Kamran Shafi, and Hussein A Abbass, "Robo-teacher: A computational simulation based educational system to improve cyber security," in *Robot Intelligence Technology and Applications 2012*. Springer, 2013.

[45] Agnė Brilingaitė, Linas Bukauskas, Virgilijus Krinickij, and Eduardas Kutka, "Environment for cybersecurity tabletop exercises," in *ECGBL 2017 11th European Conference on Game-Based Learning*, 2017, p. 47.

[46] Rain Ottis, "Light weight tabletop exercise for cybersecurity education," *Journal of Homeland Security and Emergency Management*, 2014.

[47] Irfan Ahmed and Vassil Roussev, "Peer instruction teaching methodology for cybersecurity education," *IEEE Security & Privacy*, vol. 16, no. 4, pp. 88–91, 2018.

[48] S. Morgan and B. Lagesse, "Dynamically generated virtual systems for cyber security education," in *Proceedings of the International Conference on Cloud Security Management*, 2015, vol. Jan, pp. 187–193.

[49] William Aubrey Labuschagne and Marthie Grobler, "Developing a capability to classify technical skill levels within a cyber range," in *ECCWS 2017 16th European Conference on Cyber Warfare and Security*, 2017, p. 224.

[50] Emin Caliskan, Unal Tatar, Hayretdin Bahsi, Rain Ottis, and Risto Vaarandi, "Capability detection and evaluation metrics for cyber security lab exercises," in *ICMLG2017 5th International Conference on Management Leadership and Governance*. Academic Conferences and publishing limited, 2017, p. 407.

[51] Kelly J Neville and Jeremiah T Folsom-Kovarik, "Recommendation across many learning systems to optimize teaching and training," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2018.

[52] Robert G Abbott, Jonathan McClain, Benjamin Anderson, Kevin Nauer, Austin Silva, and Chris Forsythe, "Log analysis of cyber security training exercises," *Procedia Manufacturing*, vol. 3, pp. 5088–5094, 2015.

[53] Robert G Abbott, Jonathan T McClain, Benjamin Robert Anderson, Kevin S Nauer, Austin Ray Silva, and James C Forsythe, "Automated performance assessment in cyber training exercises," Tech. Rep., Sandia National Laboratories, Albuquerque, NM, 2015.

[54] Adrian Furtună, Victor-Valeriu Patriciu, and Ion Bica, "A structured approach for implementing cyber security exercises," in *8th International Conference on Communications*. IEEE, 2010, pp. 415–418.

[55] Erik Moore, Steven Fulton, and Dan Likarish, "Evaluating a multi agency cyber security training program using pre-post event assessment and longitudinal analysis," in *IFIP World Conference on Information Security Education*. Springer, 2017, pp. 147–156.

[56] Sten Mäses, Bil Hallaq, and Olaf Maennel, "Obtaining better metrics for complex serious games within virtualised simulation environments," in *European Conference on Games Based Learning*, 2017, pp. 428–434.

[57] Nina Wilhelmson and Thomas Svensson, *Handbook for planning, running and evaluating information technology and cyber security exercises*, Försvarshögskolan (FHS), 2011.

[58] "ISO 22398:2013 Societal Security—Guidelines for Exercises," International Organization for Standardization, https://www.iso.org/standard/50294.html.

[59] Agnė Brilingaitė, Linas Bukauskas, and Aušrius Juozapavičius, "A framework for competence development and assessment in hybrid cybersecurity exercises," *Computers & Security*, p. 101607, 2019.

[60] Menelaos Katsantonis, Panayotis Fouliras, and Ioannis Mavridis, "Conceptual analysis of cyber security education based on live competitions," in *Global Engineering Education Conference*. IEEE, 2017, pp. 771–779.

[61] Melissa Dark and Jelena Mirkovic, "Evaluation theory and practice applied to cybersecurity education," *IEEE Security & Privacy*, vol. 13, no. 2, 2015.

[62] Jelena Mirkovic, Melissa Dark, Wenliang Du, Giovanni Vigna, and Tamara Denning, "Evaluating cybersecurity education interventions: Three case studies," *IEEE Security & Privacy*, vol. 13, no. 3, pp. 63–69, 2015.

[63] Julia Nevmerzhitskaya, Elisa Norvanto, and Csaba Virag, "High impact cybersecurity capacity building," in *The International Scientific Conference eLearning and Software for Education*. "Carol I" National Defence University, 2019, vol. 2, pp. 306–312.

[64] Erik L Moore, Steven P Fulton, Roberta A Mancuso, Tristen K Amador, and Daniel M Likarish, "A short-cycle framework approach to integrating psychometric feedback and data analytics to rapid cyber defense," in *IFIP World Conference on Information Security Education*. Springer, 2019, pp. 45–58.

[65] Petteri Taitto, Julia Nevmerzhitskaya, and Csaba Virag, "Using holistic approach to developing cybersecurity simulation environments," *eLearning & Software for Education*, vol. 4, 2018.

[66] Chris Hart, *Doing a Literature Review: Releasing the Research Imagination*, Sage, 2018.

[67] Chris Eagle, "Computer security competitions: Expanding educational outcomes," *IEEE Security & Privacy*, vol. 11, no. 4, pp. 69–71, 2013.

[68] Brandon Mauer, William Stackpole, and Daryl Johnson, "Developing small team-based cyber security exercises," in *The International Conference on Security and Management, Las Vegas*, 2012.

[69] Jin-Ning Tioh, Mani Mina, and Douglas W Jacobson, "Cyber security training a survey of serious games in cyber security," in *Frontiers in Education Conference*. IEEE, 2017, pp. 1–5.

[70] John Burris, Wesley Deneke, and Brandon Maulding, "Activity simulation for experiential learning in cybersecurity workforce development," in *International Conference on HCI in Business, Government, and Organizations*. Springer, 2018, pp. 17–25.

[71] Erjon Zoto, Stewart Kowalski, Christopher Frantz, Edgar Lopez-Rojas, and Basel Katt, "A pilot study in cyber security education using cyberaims: A simulation-based experiment," in *IFIP World Conference on Information Security Education*. Springer, 2018, pp. 40–54.

[72] Edgar A. Lopez-Rojas Mazaher Kianpour Erjon Zoto, Stewart Kowalski, "Using a socio-technical systems approach to design and support systems thinking in cyber security education," in *CEUR Workshop Proceedings*, 2018, vol. 2107, pp. 123–128.

[73] Johan de Heer and Paul Porskamp, "Human behavior analytics from microworlds: the cyber security game," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2017, pp. 173–184.

[74] Carlos Arturo Martinez Forero, "Tabletop exercise for cybersecurity educational training; theoretical grounding and development," M.S. thesis, University of Tartu, 2016.

[75] Keith Cozine, "Thinking interestingly: the use of game play to enhance learning and facilitate critical thinking within a homeland security curriculum," *British Journal of Educational Studies*, vol. 63, no. 3, 2015.

[76] Emin Caliskan, M Oguzhan Topgul, and Rain Ottis, "Cyber security exercises: A comparison of participant evaluation metrics and scoring systems," *Strategic Cyber Defense: A Multidisciplinary Perspective*, vol. 48, 2017.

[77] Valdemar Švábenský, Jan Vykopal, and Pavel Čeleda, "What are cybersecurity education papers about? a systematic literature review of sigcse and iticse conferences," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 2020, pp. 2–8.

[78] Shane Dawson and George Siemens, "Analytics to literacies: The development of a learning analytics framework for multiliteracies assessment," *The International Review of Research in Open and Distributed Learning*, 2014.

[79] Ioana Jivet, Maren Scheffel, Marcus Specht, and Hendrik Drachsler, "License to evaluate: Preparing learning analytics dashboards for educational practice," in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, 2018, pp. 31–40.

[80] Diane S Henshel, Gary M Deckard, Brad Lufkin, Norbou Buchler, Blaine Hoffman, Prashanth Rajivan, and Steve Collman, "Predicting proficiency in cyber defense team exercises," in *Military Communications Conference, MILCOM 2016-2016 IEEE*. IEEE, 2016, pp. 776–781.

[81] Dennis Andersson, Magdalena Granåsen, Thomas Sundmark, Hannes Holm, and Jonas Hallberg, "Exploratory sequential data analysis of a cyber defence exercise," in *Proceedings of the International Defense and Homeland Security Simulation Workshop*, 2011.

[82] Benjamin. Knox, Ricardo Lugo, Kirsi Helkala, Stefan Sütterlin, and Øyvind Jøsok, "Education for cognitive agility: Improved understanding and governance of cyberpower," in *European Conference on Information Warfare and Security, ECCWS*, 2018, vol. 2018-June, pp. 541–550.

[83] Travis Scheponik, Alan T Sherman, David DeLatte, Dhananjay Phatak, Linda Oliva, Julia Thompson, and Geoffrey L Herman, "How students reason about cybersecurity concepts," in *Frontiers in Education Conference*. IEEE, 2016, pp. 1–5.

[84] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens, "Towards automated content analysis of discussion transcripts: A cognitive presence case," in *Proceedings of the 6th International Conference on LAK*. ACM, 2016, pp. 15–24.

[85] Kirsty Kitto, Mandy Lupton, Kate Davis, and Zak Waters, "Designing for student-facing learning analytics," *Australasian Journal of Educational Technology*, vol. 33, no. 5, pp. 152–168, 2017.

[86] Ran Liu and Kenneth R Koedinger, "Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains.," *Journal of Educational Data Mining*, vol. 9, no. 1, 2017.

[87] Rakesh Verma, Murat Kantarcioglu, David Marchette, Ernst Leiss, and Thamar Solorio, "Security analytics: essential data analytics knowledge for cybersecurity professionals and students," *IEEE Security & Privacy*, vol. 6, pp. 60–65, 2015.

[88] Maurice Hendrix, Ali Al-Sherbaz, and Bloom Victoria, "Game based cyber security training: are serious games suitable for cyber security training?," *International Journal of Serious Games*, vol. 3, no. 1, pp. 53–61, 2016.

Appendix 8

VIII

K. Maennel, K. Kivimägi, S. Sütterlin, O. Maennel, and M. Ernits. Remote technical labs: an innovative and scalable component for university cybersecurity program. In *Educating Engineers for Future Industrial Revolutions—Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2021

Remote Technical Labs: an Innovative and Scalable Component for University Cybersecurity Program Admission

Kaie Maennel[1], Kristian Kivimägi[1], Olaf Maennel[1], Stefan Sütterlin[1,2], and Margus Ernits[3]

[1] Tallinn University of Technology, Tallinn, Estonia
first.lastname@taltech.ee
https://www.taltech.ee/institutes/centre-for-digital-forensics-cyber-security/
[2] Ostfold University College, Norway
[3] Rangeforce
margus.ernits@rangeforce.com

Abstract. In response to the existing and predicted skills gap in cybersecurity, educational institutions establish an increased number of studies. Admission boards need to screen large numbers of applicants to identify those with the highest probability of successful completion. To address the current lack of scalable and validated admission procedures with predictive value, we present a validation of an innovative university admission process for a master level program including technical skills assessment via cloud-based virtual labs. A regression model based on data collected during admission assessment procedures is applied to predict later study performance in technical courses. The virtual labs assessing technical skills but also interview component had comparably high predictive values for study performance, indicating a complementary relationship of two distinct skill-sets. The primary conclusion of this research is that cybersecurity technical labs can be used to significantly improve the predictive value of traditional interview-based admission processes for the candidates' later success in technical courses.

Keywords: cybersecurity, exercises, predictive analytics, technical skills

1 Introduction

Cybersecurity professionals need to be trained on a variety of skills including identifying cyber threats and vulnerabilities, protecting information and resources, detecting, responding and recovering from cybersecurity events, etc [5],[25]. To meet the labour market's demands, academic institutions have established dedicated programs to train more specialists [5]. A variety of admission processes are in place aiming to select candidates of high quality and low dropout probability. To ensure effectiveness, transparency, and the possibility of further development, any such selection process should be scalable and validated. To date,

systematic validations of admission processes are scarce. Traditional knowledge-based assessments for technical skills may fail to detect high potential candidates, who fail to provide declarative knowledge at the time of assessment due to their interdisciplinary background [21]. We argue that an accurate assessment of study potential needs to assess technical skills and knowledge using knowledge questions and hands-on tasks in addition to interviews to assess learning capacity in an ecologically valid assessment procedure.

In the admission process to the international Cybersecurity masters program (MSc), we have used online interviews and technical assessments in addition to traditional admission procedures [21]. Currently, the technical assessments—Intro, HTTPS Security, SQL Injection and Botnet labs—are optional. We aim to demonstrate that the admission's technical assessment component is an accurate predictor to rank the candidates for cybersecurity studies and that this can predict their success in the technical cybersecurity subjects. As a result, the practical technical assessment tapping into learning potentials rather than pure pre-existing knowledge, may then replace less reliable technical questions during the interview. We use Cybersecurity Technologies (CST), a mandatory course for first year MSc students, to evaluate the technical skills in admission and in later studies. As the lab exercises focus on assessing the technical competencies, we leave the assessment of non-technical skills, which are at least equally important, to other validation methods.

We address following research questions (RQs):

- RQ1: Can admission procedure including the technical online labs on selected cybersecurity topics predict the students' performance in the technical subjects in the university curriculum? Is it a more accurate predictor than the admission interview component?
- RQ2: Do more comprehensive and complex cybersecurity technical assessments used at the beginning of the course predict student performance? Is this assessment appropriate as the basis to assign the students to the courses with different difficulty levels?
- RQ3: Is more comprehensive assessment necessary at the beginning of the course, or can the results of selected technical labs used during admission procedure also predict the students performance?

We analyse admission process and CST course completion data to model the prediction of the students' success applying a linear regression model. The model's main purpose is to statistically validate whether the novel use of virtual technical labs on varied cybersecurity topics is a significant predictor to measure candidates' technical skill level during their later studies in this core subject of cybersecurity technology. The results indicate that such methods to predict student performance with limited set of input data from the labs in the cybersecurity domain can be indicative. We also describe the data and evaluation method in sufficient detail to replicate the research and CST course design to assist in developing cybersecurity curricula.

2 Research Design

The first-year course are used for performance prediction in order to minimize impact of other variables in a student's academic life. We also evaluate whether the complex and comprehensive skill assessment conducted during the first CST lecture predicts students' later study performance more or less accurately than the admission labs. This in turn validates that the admission labs are relevant and predict the students' success in the cybersecurity technical subjects at an early time point critical for admission selection. Pearson's bivariate correlations are used for correlational analysis of parametric variables [4]. Kendall's Tau is used to analyse non-parametric measures of relationships between columns of ranked variables (value of 0–no relationship or 1–relationship) [11].

2.1 Ethics, Privacy Data and Data Security

Aiming to scale the university admission process by incorporating the technical online labs raises a variety of ethical implications. An aim of this research is to reduce sole dependence on the decisions of human interviewer and course instructor by adding an additional component of evaluating the skills using the technical labs. Ethical considerations such as fostering trust, transparency, student control over data, right of access and accountability [24] are followed. The labs completion in the admission is voluntary, the applicant/student receives the results automatically. The role of the labs is described as being a part of the admission process and the course grading. Also, as online or digital interactions produce a data trail of a person's activity, privacy and data security aspects are important. We have pseudonymised the data with unique identifiers to ensure the privacy of individuals. The data was stored on the university' server, with access to the research material is restricted to only the some of the authors and selected university personnel directly associated with this study and admission process.

3 Related Work

One purpose of assessments during admission processes is to early distinguish between students who are likely to perform well or drop out [23]. Prediction modeling in the university admissions (including in Science, Technology, Engineering and Mathematics (STEM) disciplines) frequently used. A meta-analysis of academic literature on the prediction of student performance in computing courses is described in [14] summarising 357 articles. This review shows the relevance of predictors such as GPA, demographics, learning behavior data, etc [14]. However, there is a gap of knowledge regarding the use of technical online labs as part of admission assessment and performance prediction in cybersecurity programs. Most prediction models require knowledge of previous performance or are mainly based on demographic data (e.g., [7], [18], [10], [20], [19]). However, in global assessment proceudres such information may not necessarily be

available, reliable, or comparable and admission decisions thus include limited or incomplete data. Prediction models not relying on legacy data do usually not address gamified technical exercises as a relevant predictor for future learning success([16]). [22] categorizes methods used for predicting performance into four high level categories: Decision trees, Regression, Clustering, Dimensionality reduction / other. [17] suggests choosing multiple linear regression models when predicting the average academic performance. [18] uses past performance data and applies a decision tree algorithm to identify students who are likely to fail in advance. Other methods include Logistic Regression, Decision Tree, Random Forest, Naive Bayes and Adaboost models [10], [13], [1], [19]), [9]. We apply a multiple linear regression model and leave comparison with other modeling methods as further work.

In cybersecurity, [3] describes a recruiting tool that provides an 8-hour training and competition framework. This approach requires significant time commitment from faculty and students as it is designed as live learning event. [8] proposes a model for predicting cybersecurity aptitude beyond a general-intelligence approach, where the constructs of tasks, work roles, and people can be used to create assessments of applicants. However, a general intelligence assessment is very time-consuming, may not tap into the relevant skills related to the technical tasks in focus and do not require additional efforts over time (motivational aspects) that virtual technical labs offer. In cybersecurity, there are few examples of student modeling, such as [6] using log data to predict course grade, [15] predicting team proficiency, [2] assigning specific exercise in accordance to preparedness, [26] analysing learning activities (reading lab materials and working on lab tasks) association with students' learning performance in a course. However, these papers do not build prediction models in context of admissions with limited data.

This paper builds upon the previous work described in [21].

4 Admission Process and Study Program/Technical Courses

We provide an overview over the study program, the technical labs used in the admission process, the CST course and an overview of the data collected and used in the prediction model.

4.1 Cybersecurity Masters Program

The cybersecurity curriculum consists of general studies, core studies, special studies, free choice courses and graduation thesis. During the studies the students have to choose a specialty—Cybersecurity, Digital Forensics or Cryptography. All specialities require completion of general studies, core studies (Cybersecurity Defence Technology), a selection of free choice courses, specialty courses and a master's thesis.

4.2 Admission's Technical Labs and Interview

The detailed overview of the admission process and all its components is presented in [21]. To apply a feasible selection of the many possible cybersecurity skills in the admissions, the technical exercises represent topics from basic to advanced skill levels as follows:

- Introduction lab (25 minutes)—essential command line skills (Git, apt-get, Apache server);
- HTTPS Security (45 minutes)—basic level skills connected to command line, public key infrastructure, and server administration basics;
- SQL injection (90 minutes)—intermediate level skills connected to attacking SQL databases (SQL, SQL injection); and
- Botnet (45 minutes)—advanced level skills connected to network scanning skills, text parsing (programming skills are beneficial) and SQL injection skills.

The choice of these labs is based on typical attack vectors that the applicants are likely to encounter in their future cybersecurity jobs and require different skill levels (from essential to advanced). Each lab represents a pre-determined skill level from basic to advanced [21].

Interviews last usually for 10-15 minutes. The interviews includes few technical questions, which aim to measure the candidate's knowledge and logical thinking [21].

4.3 Cyber Security Technologies

The CST's learning objective is to provide a coherent understanding of technology (theory) and provide a hands-on learning (experiential learning). The main focus is on the tools and methods for securing networks, operating systems and web applications.

CST1 and CST2 CST is split into two sub-courses, Cyber Security Technologies 1 (CST1) and Cyber Security Technologies 2 (CST2). While both courses follow the same study plan, CST1 is aimed towards beginners and CST2 for advanced students who are already familiar with cybersecurity technologies. All students must attend either course version in their first study year.

Students in CST I are introduced the topics related to the fundamentals of networking, information security and cybersecurity. Students learn about the different types of technologies with the learning objective to understand when and where these tools should be utilised, and how to configure and deploy specific tools to their own environment. CST2 requires a programming, system administration or information security background and basic knowledge in networking, operating systems and web applications. If students from CST2 feel they do not have enough knowledge on a certain topic, then they are free to attend CST1.

Initial Assessment to Assign Students to CST1 or CST2 To determine whether a student is assigned to CST1 or CST2, a mandatory skill assessment exam is given in the first lecture. Students complete a "WASE Assessment", which is 180 minutes long complex virtual lab where an individual has to investigate a website and try to regain control. Students have to know various types of web application vulnerabilities and how to exploit them (web-servers side, incl. SQL authentication bypass, reconnaissance, privilege escalaton, command injection, path traversal, blind SQL injection, etc. and web-clide side, incl. reflected, stored and DOM based XSS, session hijacking, CSRF, etc.). The requirements of this test are high to avoid ceiling effects. The relevant outcome variable is the progress made within three hours.

Course Assignments The course is designed as a combination of theory and hands-on exercises aiming to challenge the students along the way with different assignments as milestones rather with one final exam. The assignments consist of group work, home labs, individual tasks, discussions and online hands-on technical assessments, see Table 1. Groups (consisting of 4-6 students) are self-formed and all group works are completed with the same group.

Table 1. Course assignments in CST1 and CST2

Assignment	CST 1	CST 2
Individual Assignment	Malware Lab	Malware Lab
Individual Assignment	Quiz	Vulnerability Testing
Group Work 1	Security Principles	Company X-Part 1
Group Work 2	Information Gathering and Vulnerability Testing	Company X-Part 2
Group Work 3	Authentication and Access Control	Company X-Part 3
Group Work 4	Logging and Log Analysis	Company X-Part 4
Online Technical Test	SOC-Security Compromised	SOC-Security Compromised
Group Work 5	Certificates and Public Key Cryptography	Company X-Part 5
Group Work 6	Risk Management	Company X-Part 6

5 Data Collection, Cleaning and Integration

The data originate from multiple sources, see Table 2. The unstructured data was converted into structured data by removing irrelevant information (i.e., duplicated data if student attends both courses, not enrolled at MSc program, registered but not taking CST, is not 1st year student), verified and pseudonymized. For correlation analysis, the data was integrated into one data-set, also newly aggregated/calculated fields were added named "Ranks". Descriptive statistics are presented in Table 3. The admission and course results of 60 students are included in this analysis.

Table 2. Data collected and used in prediction model as variable

Field Name	Value	Unique data values	Description
Student Name	Firstname. Lastname	60	An identifier to integrate data together from different sources. Once each student receives an ID number, this field is removed
Student ID	Student 0X	60	Each student is identified with an ID number from 001
Course	(1) Beginner (2) Advanced (3) Advanced/ Beginner	3	3 unique values: students eligible for CST2 are classified as advanced, students eligible for CST1 are classified as beginners, students eligible for CST2 but take both courses are classified as advanced/beginners
Admission Interview Score	50-100	60	Admission interview results, minimum threshold to accept candidates into the program is 50 points
Admission Online Technical Assessment Score	0-400	59	Admission Technical assignment results —4 different labs, each worth up to 100 points
Assignment 1 Lab Results	0-100	60	First assignment is identical for both courses; students are expected to accomplish this individually at home. Task is to analyse 3 malware samples, answer questions and write report. This is same task for both courses and can be used to measure student performance, regardless of course enrollment
CST1 Individual Assignments	0-400	52	Aggregated results of all individual assignments for CST1
CST1 Group Works	0-600	52	Aggregated results of all group assignments for CST1
CST1 Course Total	0-1000	52	Aggregated results of all assignments for CST1
CST2 Individual Assignments	0-200	11	Aggregated results of all individual assignments for CST2
CST2 Group Works	0-300	11	Aggregated results of all group assignments for CST2
CST2 Course Total	0-500	11	Aggregated results of all assignments for CST2
WASE Assessment Score	0-150 000	60	All students have to complete WASE Assessment in order to be assigned to CST1 or CST2. As all students complete this– data can be used to measure and compare student performance
SOC-System compromised Progress	0-100	60	Online assessment, that all students have to complete in class, regardless of course enrollment
SOC-System compromised Duration	0-5h	60	All students are encouraged to finish the lab, therefore measuring progress is redundant as most get 100% score. Time to complete assessment is more relevant metric—this can be used to compare students' understanding the problem and how fast they can solve it

Table 3. Descriptive statistics of data

Data	Mean	Median	Min	Max	Standard Deviation	Range
Admission Labs Results	280.2	293.0	0	400	117.8	400
Admission Interview Results	73.90	77.25	50	97.5	12.09	47.5
WASE Assessment	32 000	20 000	0	150 000	40 497	150 000
Lab 1	89.5	100	0	100	26.32	100
System Compromised Points	85.05	100	0	100	32.7	100
System Compromised Time	1.9	2.07	0	4,42	0.98	4,42
CST 1 Individual Assignments	249.1	268.3	0	310	63.15	310
CST 1 Group Assignments	354.13	380	0	400	91.76	400
CST 1 Course Total	687.9	739.05	81.25	845	160.74	763.75
CST 2 Individual Assignments	175	200	0	200	60.2	200
CST 2 Group Assignments	189.1	190	170	200	10.44	30
CST 2 Course Total	473.2	490	300	590	68.3	290

6 Results and Discussion

6.1 RQ1: Admission labs and later success

We used Pearson's one sided correlation test method to account for the directed hypotheses that admission and CST-performance are positively related. The variables of course performance, admission interview rank, admission technical assessment rank and admission rank were added to the correlation matrix, see Table 4.

There is a positive correlation between the admission assessments and the later course performance, while the strongest positive correlation is between

Table 4. RQ1: Pearson's one sided correlation test

Course Performance	Admission Interview Rank	Admission Lab Rank	Admission Rank
Pearson's r	0.378**	0.432**	0.492***
p - value	0.005	0.001	< .001

"Course performance" and the resulting"Admission Rank" combining both interview and virtual lab results.

Admission interview and technical results are added to a stepwise regression model with predictor variables interview and remote labs and course performance as the dependent variable, see Table 5. In Model 1, course performance is used as dependent variable and admission interview rank as only predictor. Model 2 adds the remote lab result as additional predictors.

Table 5. RQ1: Admissions and course success linear regression model

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
1	0.378	0.143	0.126	14.448	0.143	8.507	1	51	0.005
2	0.495	0.245	0.214	13.700	0.102	6.727	1	50	0.012

The assessment technique interview is in itself a statistically significant predictor for later student performance and explains .126=12.6% of Student performance variance. Adding the additional predictive effect of the virtual lab, the explained variance in study performance almost doubles from 12.6 to 21.5%. This significant increase (p=.012) means that interviews and labs are complementary methods that both predict different, but relevant, aspects of the later student performance results as measured in CST courses.

Table 6. RQ1: ANOVA test

Model	Model type	Sum of Squares	df	Mean Square	F	p
1	Regression	1775.822	1	1775.822	8.507	0.005
	Residual	10646.291	51	208.751		
	Total	12422.113	52			
2	Regression	3038.279	2	1519.139	8.094	< .001
	Residual	9383.834	50	187.677		
	Total	12422.113	52			

Relationships between the predictor variable and response is shown in Table 7. The standardized betas show the relative weight of the predictors. The

Table 7. RQ1: Relationships between predictor variable and response

Model		Unstandardized	Standard Error	Standardized	t	p
1	Admission Interview Rank	0.340	0.117	0.378	2.917	0.0005
2	Admission Interview Rank	0.231	0.118	0.257	1.955	0.056
	Admission Technical Assessment Rank	0.246	0.095	0.341	2.594	0.012

single standardized betas are low to moderate, but provide a highly significant predictive value when combined. The predictors' scores are medium correlated with each other, which means that the interview scores and technical skills assessment via virtual labs share both common factors and relevant unique variance. For example, attitude, eagerness and technical interest are checked in the interview process but are also relevant for the technical performance. An explained variance of 21.4% is considered to be large in behavioral science convention [12].

6.2 RQ2: Complex assessment at course start and later success

To evaluate whether WASE assessment predicts student performance and in turn assign students to appropriate CST course, we used Kendall's Tau B testing method. From the model output we see that Kendall's Tau B coefficient is 0.502 and p-value $\leq .001$. There is a strong positive correlation between WASE assessment and course performance. This indicates that WASE assessment is a valid predictor of the students' performance. This also indicates that the students have been assigned to the correct CST course.

6.3 RQ3: Admission labs vs. complex assessment and later success

To evaluate whether a WASE assessment is necessary or admission lab results could be used to assign students on CST1 and CST2, we used Pearson's one-sided correlation testing method because there is directed hypothesis that WASE assessment and CST-performance are positively related.

Table 8. RQ3: Pearson's one-sided coefficent

Course Performance	Without WASE Admission Interview Rank	WASE Assessment Rank
Pearson's r	0.329*	0.548***
p - value	0.016	.001

A significant positive correlation between the input and admission interview rank is evident, see Table 8. However, the correlation between course performance and WASE assessment rank is stronger with the correlation coefficient 0.548. When using stepwise regression to further explore the relationship between course performance, admission result and WASE assessment, we can see

that course performance without WASE assessment is dependent variable and "Admission technical rank" and "WASE assessment rank" are predictor variables.

Table 9. RQ3: WASE and technical assessment linear regression model

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
1	00.322	0.104	0.086	14.727	0.104	5.890	1	51	0.019
2	0.551	0.304	0.276	13.106	0.200	14.401	1	50	$< .001$

Looking at the linear regression model in Table 9, the admission technical lab results significantly (.019) predict the student performance and explain .086=8.6% of performance variance. This means that interview in its own is a valid predictor for assigning students into correct CST course. However, when adding WASE assessment then R square is 3 times higher from 8.6 to 27.6. This increase is significant (p=.001), which means that the two variables are complementary methods that both predict different aspects of the student performance.

Table 10. RQ3: Relationships between predictor variable and response/coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	Admission Technical Rank	0.231	0.095	0.322	2.427	0.019
2	Admission Technical Rank	0.047	0.098	0.066	0.484	0.630
	WASE Assessment Rank	0.543	0.143	0.516	3.795	$< .001$

Relationships between the predictor variable and response is shown in Table 10. The combined weight of the predictor scores are relatively highly correlated, which means that the admission technical assessment and WASE assessment have common factors. The correlation is relatively high, which means that both methods contribute also individually and unique parts that predict the student performance score. Overall, WASE assessment has a stronger correlation with student performance.

7 Future Work

As future work, validation of the hands-on technical exercise tasks by correlating it with general intelligence, other cognitive skills, and domain-specific knowledge is suggested. This will improve our understanding for what is tested in this task (construct validation). A longitudinal study with new students and larger sample size will be continued to validate and refine the algorithm. In addition, as we applied multiple regression model, the comparison with larger datasets and other prediction modelling methods would be beneficial.

8 Conclusion

We evaluated using technical labs as a novel part of graduate level admissions for cybersecurity program, to predict students' later success in studies. Such an approach can be a scalable evaluation of technical skills, but still incorporate human evaluation to enable balanced approach for the ethical and evidence-based decision making and assessment. While the labs used in this paper are specific for the technical skills for a cybersecurity program, incorporating this type of assessment may also spark interest in other STEM programs.

While we acknowledge that this analysis is an initial attempt with relatively small sample size to assess whether such technical skill labs can be used as a significant predictor to assess potential candidates' skill level and future study success in technical topics, is shows some promise based on the regression analysis. This analysis however also shows that the interview score is not redundant either—both admission methods are complementary to each other addressing different, but equally relevant aspects of the student performance.

In addition to the prediction model, we shared our experience and description of the admission process and Cyber Security Technologies course design.

References

1. Ahadi, A., Lister, R., Haapala, H., Vihavainen, A.: Exploring machine learning methods to automatically identify students in need of assistance. In: Proceedings of the 11th Annual International Conference on International Computing Education Research, pp. 121–130 (2015)
2. Aoyama, T., Nakano, T., Koshijima, I., Hashimoto, Y., Watanabe, K.: On the complexity of cybersecurity exercises proportional to preparedness. Journal of Disaster Research Vol **12**(5), 1081 (2017)
3. Augustine, T.A., DeLooze, L.L., Monroe, J.C., Wheeler, C.G.: Cyber competitions as a computer science recruiting tool. Journal of Computing Sciences in Colleges **26**(2), 14–21 (2010)
4. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
5. Cabaj, K., Domingos, D., Kotulski, Z., Respício, A.: Cybersecurity education: Evolution of the discipline and analysis of master programs. Computers & Security **75**, 24–35 (2018)
6. Caliskan, E., Tatar, U., Bahsi, H., Ottis, R., Vaarandi, R.: Capability detection and evaluation metrics for cyber security lab exercises. In: ICMLG2017 5th International Conference on Management Leadership and Governance (2017)
7. Campbell, J.P., DeBlois, P.B., Oblinger, D.G.: Academic analytics: A new tool for a new era. EDUCAUSE review **42**(4), 40 (2007)
8. Campbell, S.G., O'Rourke, P., Bunting, M.F.: Identifying dimensions of cyber aptitude: the design of the cyber aptitude and talent assessment. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 59, pp. 721–725. SAGE Publications Sage CA: Los Angeles, CA (2015)
9. Castro-Wunsch, K., Ahadi, A., Petersen, A.: Evaluating neural networks as a method for identifying students in need of assistance. In: Proceedings of the ACM SIGCSE Technical Symposium on Computer Science Education, pp. 111–116 (2017)

10. Chen, Y., Johri, A., Rangwala, H.: Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In: Proceedings of the 8th International Conference on LAK, pp. 270–279. ACM (2018)
11. Cliff, N., Charlin, V.: Variances and covariances of kendall's tau and their estimation. Multivariate Behavioral Research **26**(4), 693–707 (1991)
12. Cohen, J.: A power primer. Psychological Bulletin **112**, 155–159 (1992)
13. Gardner, J., Brooks, C.: Evaluating predictive models of student success: Closing the methodological gap. arXiv preprint arXiv:1801.08494 (2018)
14. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N.: Predicting academic performance: a systematic literature review. In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, pp. 175–199 (2018)
15. Henshel, D.S., Deckard, G.M., Lufkin, B., Buchler, N., Hoffman, B., Rajivan, P., Collman, S.: Predicting proficiency in cyber defense team exercises. In: Military Communications Conference, pp. 776–781. IEEE (2016)
16. Hlosta, M., Zdrahal, Z., Zendulka, J.: Ouroboros: early identification of at-risk students without models based on legacy data. In: Proceedings of the 7th International LAK, pp. 6–15. ACM (2017)
17. Huang, S., Fang, N.: Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education **61**, 133–145 (2013)
18. Kabra, R., Bichkar, R.: Performance prediction of engineering students using decision trees. International Journal of Computer Applications **36**(11), 8–12 (2011)
19. Liao, S.N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W.G., Porter, L.: A robust machine learning technique to predict low-performing students. ACM Transactions on Computing Education **19**(3), 1–19 (2019)
20. Loh, C.S., Sheng, Y., Ifenthaler, D.: Serious games analytics: Theoretical framework. In: Serious games analytics, pp. 3–29. Springer (2015)
21. Maennel, K., Mäses, S., Sütterlin, S., Ernits, M., Maennel, O.: Using technical cybersecurity exercises in university admissions and skill evaluation. Proceedings of the 14th IFAC/IFIP/IFORS/IEA Symposium on Analysis Design and Evaluation of Human Machine Systems, Tallinn, Estonia, IFAC-PapersOnLine (2019)
22. Muthukrishnan, S., Govindasamy, M., Mustapha, M.: Systematic mapping review on student's performance analysis using big data predictive model. Journal of Fundamental and Applied Sciences **9**(4S), 730–758 (2017)
23. Papamitsiou, Z., Economides, A.A., Pappas, I.O., Giannakos, M.N.: Explaining learning performance using response-time, self-regulation and satisfaction from content: an fsqca approach. In: Proceedings of the 8th International Conference on LAK, pp. 181–190. ACM (2018)
24. Pardo, A., Siemens, G.: Ethical and privacy principles for learning analytics. British Journal of Educational Technology **45**(3), 438–450 (2014)
25. Paulsen, C., McDuffie, E., Newhouse, W., Toth, P.: Nice: Creating a cybersecurity workforce and aware public. IEEE Security & Privacy **10**(3), 76–79 (2012)
26. Zeng, Z., Deng, Y., Hsiao, I., Huang, D., Chung, C.J.: Improving student learning performance in a virtual hands-on lab system in cybersecurity education. In: Frontiers in Education Conference, pp. 1–5. IEEE (2018)

Curriculum Vitae

1. Personal data

Name	Kaie Maennel
Date and place of birth	25 April 1976, Estonia
Nationality	Estonian

2. Contact information

Address	Tallinn University of Technology, School of Information Technology, Department of Software Sciences, Akadeemia tee 15A, 12168 Tallinn, Estonia
E-mail	kaie.maennel@taltech.ee

3. Education

2017–…	Tallinn University of Technology, School of Information Technology, Information and Communication Technology, Ph.D. studies
2015–2017	Tallinn University of Technology and Tartu University (joint degree), School of Information Technology, Cybersecurity (Digital Forensics), MSc, *cum laude*
2004–2008	Concordia International University Estonia, International Business, BSc *magna cum laude*

4. Professional Certifications

2018–…	Palo Alto Networks, Certified Trainer
2017–…	ICASA Estonian Chapter (CISA), student member (certification in progress)
2010–…	Certified Internal Auditor (CIA) by The Institute of Internal Auditors
2009–…	Estonian Certified Auditor with signing authority in Estonian Republic
2009–…	ACCA Professional Accounting Certification, member (Fellow status)

5. Language competence

Estonian	native
English	fluent
Russian	basic
German	basic

6. Professional employment

2017– …	Tallinn University of Technology, Early Stage Researcher
2015– …	Deloitte Audit Eesti AS, Deloitte CE Audit Learning Leader, Senior Manager
2009–2015	Deloitte LLP (Nottingham, UK), Global Audit Learning Team Leader, Manager
2008–2009	Deloitte Audit Eesti AS, Manager
2005–2008	Deloitte & Touche Tohmatsu (Adelaide, Australia), Senior Analyst
2001–2005	Deloitte Audit Eesti AS, Assistant to Senior
2001–2001	Estonian Ministry of Environment, Internal Auditor
1999–2001	Estonian Accounting Standards Board, Managing Secretary
1998–1999	Trailway Rental AS, Financial Analyst-Administrative Assistant
1994–1998	OU APEK Piimandus, Secretary-Accountant

7. Voluntary work

2016–...	Locked Shields Cyber Security Exercise, White Team
2020	Presenting at Serious Games for Cyber Security Workshop, Joint Cyber Security Centre, Adelaide, Australia
2019	Writing article about Digital Forensics to edasi.org (published in July)
2019	Evaluation committee member for ICT thesis contest, TalTech
2019	Writing article about Cyber Hygiene to Estonian Accounting Journal (published in April)
2019	Reviewing Cybersecurity Course study materials for Estonian High Schools
2019	CyCon (upcoming) student volunteer (organised by NATO CCD COE, mainly media support)
2018	CyCon student volunteer (organised by NATO CCD COE, mainly media support)
2018	Public lecture on 22 November in Mektory, Taltech about cyber awareness and hygiene
2018	Student team member at Cyber 9/12 Student Challenge in Geneva 5-6 April
2018	Mentor in CyberSecurity Summer School, Tallinn University of Technology
2017	Invited speaker at Women in Cybersecurity workshop at TalTech on (11 November)
2017–	House Union, Revision Committee member (financial audit)

8. Computer skills

- Operating systems: Windows, MacOS, Linux
- Document preparation: Latex, Word
- Programming languages: Python (learning)
- Scientific packages: R, NVivo

9. Honours and awards

- 2017, MSc Cyber Security with Cum Laude
- 1998, B.A. International Business with Magna Cum Laude

10. Defended theses

- 2017, Improving and Measuring Learning Effectiveness at Cyber Defence Exercises, MSc, supervisor Prof. Rain Ottis, Liina Randmann; Raimundas Matulevičius, University of Tartu, Faculty of Science and Technology, Institute of Computer Science
- 2008, Impact of Agricultural Policy in Estonia, supervisor Dr. James O'Neill, Concordia International University Estonia

11. Supervised theses

- 2018, Kristiina Renel, Master's Degree, (sup) Kaie Maennel; Kristjan Kikerpill, Compliance with EU Personal Data Protection Framework in the Context of Public Sector Logging, Tallinn University of Technology, School of Information Technologies, Department of Software Science
- 2020, Kristian Kivimägi, (juh) Kaie Maennel; Olaf Manuel Maennel, Predicting Students' Success Using Technical Labs as Part of University Admission to a Cyber Security Program, Tallinn University of Technology, School of Information Technologies, Department of Software Science

- 2021, Yadav, Kapil (juh) Kaie Maennel, Information Security Management for Teleworking in Small and Medium Enterprises during the COVID-19 Crisis, Tallinn University of Technology, School of Information Technologies, Department of Software Science

12. Field of research

- Cybersecurity
- Learning Analytics

13. Scientific work
Papers

1. K. Maennel, R. Ottis, and O. Maennel. Improving and measuring learning effectiveness at cyber defense exercises. In *Nordic Conference on Secure IT Systems*, pages 123–138. Springer, 2017
2. M. Kont, M. Pihelgas, K. Maennel, B. Blumbergs, and T. Lepik. Frankenstack: Toward real-time red team feedback. In *IEEE Military Communications Conference (MILCOM)*, pages 400–405. IEEE, 2017
3. T. Lepik, K. Maennel, M. Ernits, and O. Maennel. Art and automation of teaching malware reverse engineering. In *International Conference on Learning and Collaboration Technologies*, pages 461–472. Springer, 2018
4. K. Maennel, S. Mäses, and O. Maennel. Cyber hygiene: The big picture. In *Nordic Conference on Secure IT Systems*, pages 291–305. Springer, 2018
5. K. Maennel, S. Mäses, S. Sütterlin, M. Ernits, and O. Maennel. Using technical cybersecurity exercises in university admissions and skill evaluation. *IFAC-PapersOnLine*, 52(19):169–174, 2019
6. M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020: 15th International Conference on Cyber Warfare and Security*. Academic Conferences and Publishing Limited, 2020
7. K. Maennel. Learning analytics perspective: Evidencing learning from digital datasets in cybersecurity exercises. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 27–36. IEEE, 2020
8. K. Maennel, K. Kivimägi, S. Sütterlin, O. Maennel, and M. Ernits. Remote technical labs: an innovative and scalable component for university cybersecurity program. In *Educating Engineers for Future Industrial Revolutions—Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2021

Conference presentations

1. Maennel, Kaie *Improving and Measuring Learning Effectiveness at Cyber Defence Exercises*, The 22nd Nordic Conference on Secure IT Systems, 8 November 2017, Tartu, Estonia
2. Maennel, Kaie *Cyber Hygiene: The Big Picture*, The 23rd Nordic Conference on Secure IT Systems: 29 November 2018, Oslo, Norway

3. Maennel, Kaie *Team Learning in Cybersecurity Exercises*, The 5th Interdisciplinary Cyber Research conference, 29 June 2019, Tallinn

4. Maennel, Kaie *Using Technical Cybersecurity Exercises in University Admissions and Skill Evaluation*, The 14th IFAC/IFIP/IFORS/IEA symposium on Analysis Design and Evaluation of Human – Machine Systems (HMS 2019), 17 September 2019, Tallinn

5. Maennel, Kaie *Learning Analytics Perspective: Evidencing Learning from Digital Datasets in Cybersecurity Exercises*, Cyber Range Technologies and Applications (CACOE) Workshop, IEEE European Symposium on Security and Privacy, 7 September 2020, Cyberspace

6. Maennel, Kaie *Remote Technical Labs: an Innovative and Scalable Component for University Cybersecurity Program*, 23rd International Conference on Interactive Collaborative Learning (ICL), 23 September 2020, Cyberspace

Elulookirjeldus

1. Isikuandmed

Nimi	Kaie Maennel
Sünniaeg ja -koht	25.04.1976, Tallinn, Eesti
Kodakondsus	Eesti

2. Kontaktandmed

Aadress	Tallinna Tehnikaülikool, Infotehnoloogia Teaduskond, Tarkvarateaduste Instituut Akadeemia tee 15A, 12168 Tallinn, Estonia
E-post	kaie.maennel@taltech.ee

3. Haridus

2017–…	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Info- ja Kommunikatsioonitehnoloogia, doktoriõpe
2015–2017	Tallinna Tehnikaülikooli ja Tartu Ülikool (ühine õppekava), Infotehnoloogia teaduskond, Küberkaitse (Digitaalne eksertiis), MSc *cum laude*
2004–2008	Concordia Rahvusvaheline Ülikool Eestis, Rahvusvaheline majandus, BSc *magna cum laude*

4. Erialased sertifikaadid

2018–…	Palo Alto Networks, Sertifitseeritud koolitaja
2017–…	ICASA Eesti chapter (CISA), õpilasliige (sertifikaat omandamisel)
2010–…	Sertifitseeritud Sisaudiitor (CIA), Siseaudiitorite Instituut (The IIA)
2009–…	Vannutatud Audiitor allkirjaõigusega Eesti Vabariigis
2009–…	ACCA, liige (Fellow staatuses)

5. Keelteoskus

eesti keel	emakeel
inglise keel	kõrgtase
vene keel	algtase
saksa keel	algtase

6. Teenistuskäik

2017– …	Tallinna Tehnikaülikool, doktorant-nooremteadur
2015– …	Deloitte Audit Eesti AS, Deloitte CE Auditi Koolitusjuht, Seenior manager
2009–2015	Deloitte LLP (Nottingham, UK), Global Audit Learning Team Leader, Manager
2008–2009	Deloitte Audit Eesti AS, Manager
2005–2008	Deloitte & Touche Tohmatsu (Adelaide, Australia), Vanemanalüütik
2001–2005	Deloitte Audit Eesti AS, Assistent kuni Seenior
2001–2001	EV Keskkonnaministeerium, Siseaudiitor
1999–2001	EV Raamatupidamise Toimkond, Tegevsekretär
1998–1999	Trailway Rental AS, Finantsanalüütik–administrator
1994–1998	OU APEK Piimandus, Sekretär–raamatupidaja

7. Vabatahtlik töö

2016–...	Locked Shields küberkaitseõppus, valgete meeskond
2020	Koolitus Tõsimängud Küberkaitses esineja, Joint Cyber Security Centre, Adelaide
2019	IKT lõputööde konkursi hindamiskomisjoni liige, TalTech
2019	Artikkel digitaalse ekspertiisi kohta edasi.org (ilmunud juuli)
2019	Artikkel küberhügieeni kohta Raamatupidamisuudised ajakirjas (ilmunud aprillis)
2019	Eesti keskkoolidele suunated küberkaitse õpiku ülevaatus ja tagasiside andmine
2019	CyCon vabatahtlik (NATO Küberkaitsekeskuse konverents, meediatoetus)
2018	CyCon vabatahtlik (NATO Küberkaitsekeskuse konverents, meediatoetus)
2018	Avalik loeng Mektory, Taltech 22. novembril küberteadlikkusest ja -hügieenist
2018	Õpilasmeeskonna liige Cyber 9/12 Student Challenge võistlusel Genfis 5.-6. aprillil
2018	Mentor Küberkaitse Suvekoolis (C3S), TalTech
2017	Kutsutud kõneleja Women in Cybersecurity konverentsil TalTechis 11. novembril)
2017–	Korteriühistu, Revisjonikomisjoni liige (audit)

8. Arvutioskused

- Operatsioonisüsteemid: Windows, MacOS, Linux
- Kontoritarkvara: Latex, Word
- Programmeerimiskeeled: Python (omandamisel)
- Teadustarkvara paketid: R, NVivo

9. Autasud

- 2017, MSc Küberkaitses *cum laude*
- 1998, BSc Rahvusvahelises Majanduses *magna cum laude*

10. Kaitstud lõputööd

- 2017, Kaie Maennel, magistrikraad, Improving and Measuring Learning Effectiveness at Cyber Defence Exercises (Õpitulemuste parendamine ja mõõtmine küberkaitseõppustel), juhendajad Rain Ottis; Liina Randmann; Raimundas Matulevičius, Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Arvutiteaduse instituut

11. Juhendatud lõputööd

- 2018, Kristiina Renel, magistrikraad, (juh) Kaie Maennel; Kristjan Kikerpill, Compliance with EU Personal Data Protection Framework in the Context of Public Sector Logging, Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Tarkvarateaduse instituut
- 2020, Kristian Kivimägi, (juh) Kaie Maennel; Olaf Manuel Maennel, Predicting Students' Success Using Technical Labs as Part of University Admission to a Cyber Security Program, Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Tarkvarateaduse instituut
- 2021, Yadav, Kapil (juh) Kaie Maennel, Information Security Management for Teleworking in Small and Medium Enterprises during the COVID-19 Crisis, Euroopa Isikuandmete Kaitse Raamistikule),

12. Teadustöö põhisuunad

- küberkaitse
- õpianalüütika

13. Teadustegevus
Teadusartiklite, konverentsiteeside ja konverentsiettekannete loetelu on toodud ingliskeelse elulookirjelduse juures.

ISSN 2585-6901 (PDF)
ISBN 978-9949-83-702-1 (PDF)