Joonas Forsberg

MEASURING THE TECHNICAL PERFORMANCE OF A SECURITY OPERATIONS CENTER

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2022

ABSTRACT

Forsberg, Joonas
Measuring the Technical Performance of a Security Operations Center
Jyväskylä: University of Jyväskylä, 2022, 81 pp.
Cyber Security, Master's Thesis
Supervisor(s): Frantti, Tapio

This research examines the current state of the performance indicators and other metrics used to measure the technical performance of a Security Operations Center (SOC), as based on empirical experience, the current methods for measuring the technical performance of different types of SOCs are inadequate. Without properly constructed performance indicators or metrics, it is difficult to evaluate the actual performance of a SOC, which makes it difficult to assess the concrete impact a SOC has in terms of overall cyber defence capabilities.

Design Science methodology is used as the research methodology in this research. The outcome of the research is a design science artifact, a novelty metric selection framework, that can be used to construct metrics to measure the technical and non-technical performance of a SOC. The design science artifact was successfully demonstrated by constructing five metrics that can be, as such, adopted by different types of SOCs to improve the technical performance measurement capabilities of their threat detection capabilities.

The original hypothesis is supported by the literature reviewed within the research, as the commonly mentioned metrics revolved mostly around operational activities. Furthermore, the research concluded that the current methodologies to construct metrics and the commonly deployed metrics are inadequate to measure the technical performance of a SOC. The literature outlined a limited amount of technical performance metrics, but the ones evaluated, were considered to be invalid according to the metric selection framework.

The design science artifact and the metrics utilized to demonstrate the metric provide means for SOCs to construct metrics and measure their technical performance, but further research around the subject is required to enable comprehensive industry-standard measurement capabilities to emerge.

Keywords: security operations center, soc, csoc, cyber security operations center, metric, measurement, technical, performance, cyber defence, performance indicator

TIIVISTELMÄ

Forsberg, Joonas
Tietoturvavalvomon teknisen suorituskyvyn mittaaminen
Jyväskylä: Jyväskylän yliopisto, 2022, 81 s.
Kyberturvallisuus, pro gradu -tutkielma
Ohjaaja(t): Frantti, Tapio

Tutkimuksessa selvitetään tietoturvavalvomon (engl. Security Operations Center, SOC) suorituskykyindikaattoreiden tämän hetkistä kyvykkyyttä mitata tietoturvavalvomon teknistä suorituskykyä. Empiirisen kokemuksen perusteella voidaan todeta, että tällä hetkellä yleisesti käytössä olevat menetelmät eivät ole riittäviä erilaisien tietoturvavalvomoiden teknisen suorituskyvyn mittaamiseen. Teknisten suorituskykyindikaattoreiden puute aiheuttaa sen, että tietoturvavalvomoiden teknistä suorituskykyä on hankala mitata, jonka seurauksena tietoturvavalvomon käytännön vaikutusta organisaation kyberpuolustuskyvykkyydelle on hankala määrittää.

Tutkimuksessa käytetty tutkimusmenetelmä on suunnittelutiede, joka tuottaa iteratiivisen prosessin lopputuloksena artefaktin. Työn tuloksena syntynyt artefakti on uudenlainen menetelmä, suorituskykyindikaattoreiden valintakehys, jonka avulla voidaan luoda teknisiä sekä epäteknisiä suorituskykyindikaattoreita. Luotuja suorituskykyindikaattoreita voidaan käyttää hyväksi tietoturvavalvomon suorituskyvyn mittaamisessa. Artefakti esiteltiin onnistuneesti luomalla viisi metriikkaa, joita voi sellaisenaan käyttää tietoturvavalvomoiden teknisen suorituskyvyn mittaamisen parantamiseen uhkien havainnointikyvykkyyden saralla.

Tutkimuksen aikana suoritettu kirjallisuuskatsaus tukee alkuperäistä hypoteesiä, sillä kirjallisuudessa useimmiten mainitut metriikat mittaavat pääasiallisesti tietoturvavalvomon operatiivisia toimia. Tämän lisäksi tutkimuksessa päädyttiin johtopäätökseen, jonka perusteella nykyiset menetelmät suorituskykyindikaattoreiden luomiseen ja olevassa olevat suorituskykyindikaattorit eivät ole riittäviä teknisen suorituskyvyn mittaamiseen. Kirjallisuudessa mainitut tekniset suorituskykyindikaattorit osoittautuivat epäpäteviksi valintakehyksellä arvioitaessa. Artefakti ja sen esittelyyn luodut mittarit mahdollistavat tietoturvavalvomoille suorituskykyindikaattoreiden luomisen sekä teknisen suorituskyvyn mittaamisen parantamisen artefaktin esittelyssä käytetyillä metriikoilla. Tästä huolimatta, aihepiiri vaatii tarkempaa tieteellistä tarkastelua, jonka pohjalta voidaan luoda kattava alan standardi tietoturvavalvomoiden teknisen suorituskyvyn mittaamiseen.

Avainsanat: tietoturvavalvomo, tekninen suorituskyky, mittaaminen, kyberpuolustus, mittari, suorituskykyindikaattori, soc, csoc

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

GLOSSARY

Benign true-positive	A detection or a security incident that is technically a true-positive incident but the activity causing the detection is benign and thus not malicious.
Detection	A detection originating from a positive hit on a monitoring rule, which acts as an impulse to begin the security analysis process.
Event	Processed log event or telemetry data without significant or identified security value.
False-positive	A detection or a security incident that ends up being created as a result of an incorrectly functioning monitoring rule and the activity associated with the detection is not malicious.
Monitoring rule	Concrete manifestation of a monitoring scenario constructed in a detection technology, for example, SIEM or IDS, which generates detections upon a positive hit.
Monitoring scenario	A formalized description of a monitoring rule that attempts to detect the manifestation of a threat scenario.
Raw event	Unprocessed log event or telemetry data originating directly from the source entity.
Security event	Processed log event or telemetry data with significant security value.
Security incident	A confirmed or suspected true-positive detection results in a security incident that triggers a security incident response process.
Threat scenario	A formalized description of a threat that could potentially affect the security of the organization.
True-positive	A detection or a security incident that is both technically correct and ends up being a result of malicious activity.

1 INTRODUCTION

Cyber threats have evolved dramatically within the past couple of years, as they have become more sophisticated and complex, and as a result, have a bigger impact on the operational activities of the affected organization. The usage of general-purpose malware has declined recently, but simultaneously more advanced threats such as supply chain compromises, ransomware and other extortion activities, attacks against critical infrastructure, disinformation campaigns, and targeted business e-mail compromises are trends among others that are continuously increasing in terms of volume (European Union Agency for Cybersecurity [ENISA], 2021). The risk of threats related to cybersecurity is a major concern in several recently published corporate risk surveys (Allianz, 2022; Caldwell, 2021; PwC, 2022) and the global risk report published by the World Economic Forum (2022) depicts that cybersecurity failure could become a significant risk to the world, being on par on short term with the debt crisis, and on medium term with biodiversity loss. The increased sophistication and the associated risks require organizations to step up their defensive capabilities to combat the evolving threat landscape, as traditional malware and network defences are not enough to protect the organization from modern-day cyber threats.

One strategy for organizations to increase their cyber defence capabilities is to build an in-house Security Operations Center (SOC) or outsource the operations to a dedicated managed security services vendor. According to Nathans (2014), SOC is typically responsible for the detection of security incidents and the related incident response activities resulting from a true-positive security incident. Nathans also determined that depending on the size and the needs of the organization, the SOC can consist of a single person responsible for the security or a larger team of operative personnel working in 24/7 shifts, developers, architects, and managers that together form a coherent collection of different competences put together to prevent, detect and respond to cyber threats targeting the organization. To quantify the operative efficiency and capabilities of a SOC, the performance should be measured through a set of commonly agreed metrics and other performance indicators. A systematic literature review on SOCs performed by Vielberth, Böhm, Fichtinger, and Pernul (2020) concluded that the

currently established metrics are insufficient for measuring the performance of a SOC, which means there is currently no industry standard framework that can be used to measure the performance of a SOC.

Multiple publications on SOCs focus mostly on quantitative metrics related to vulnerabilities and operational metrics, especially volume- and time-based metrics, to measure the performance of SOCs (Ahlm, 2021; Kokulu et al., 2019; Nathans, 2014). Based on empirical experience from real-world implementations, most SOCs have not successfully implemented valid metrics for other categories, such as threat detection capabilities, threat hunting, or other technical areas within the SOC. Quantitative metrics are useful for managerial level staff members to measure and optimize the usage of human resources performing the analysis work, and other quantitative metrics, such as vulnerability-related data, can provide an overview of the overall exposure to known threats within the environment (Nathans, 2014). In practice, they are inefficient to measure the capabilities of the SOC, as they are unable to measure the effectiveness of the detection capabilities and other protective controls. Measuring the false-positive rate of the monitoring rules is a common solution to combat this issue (Ahlm, 2021; Nathans, 2014), but looking at the false-positive rate alone can be misleading in a practical sense since the outcome of such metrics can be manipulated either subconsciously or consciously by the detection engineering team, which typically produces the threat detection capabilities for a SOC. Several studies have been performed on general security metrics (Böhme, 2010; Pendleton, Garcia-Lebron, Cho, & Xu, 2016; Salmi, 2018) and in addition, there are also commonly referenced industry standards such as ISO/IEC 27004:2016 (2016) and NIST SP 800-55 (Chew et al., 2008) that organization can use to measure the effectiveness of their information security, but as determined by Vielberth et al. (2020), there is a lack of literature on how to specifically measure SOC as a function or entity.

The lack of commonly accepted methods to measure the technical performance of a SOC is especially prominent when the SOC is being outsourced to a third-party vendor, as it makes it difficult to measure the quality of the service provider during the tendering and production phases. In practice, this often leads to a situation where the tendering process produces a suboptimal result, either by selecting the vendor that has the lowest total cost of service or the most convincing sales material to provide a false sense of quality to the procurement team. A trend has been observed in recent years in Finland, where the select vendors are chosen for a proof of concept phase during the tendering process of outsourcing the SOC, in which an attack simulation is performed in an environment temporarily monitored by the vendor and afterwards, their threat detection capabilities are evaluated. The trend is essentially a manifestation of the difficulty of evaluating the technical capabilities of different vendors during the tendering process and while it can provide meaningful insights when comparing the different vendors, the results are not truly comparable between the vendors, as the number of resources allocated to the proof of concept phase can significantly impact the final results and thus skew the vendor selection based on false conclusions.

1.1 Research methodology

Design science research methodology has been selected to be used as the primary research methodology of this research. The initial hypothesis of the research is, based on empirical experience, that there are no suitable frameworks out there that could be efficiently used to measure the technical performance of a SOC. As a result, the expected outcome of the research is a novelty framework that would provide an industry-standard way to measure the technical performance of a SOC.

Design science as a research methodology has been traditionally utilized within the field of engineering (Hevner, March, Park, & Ram, 2004; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) but it has also been incorporated as a commonly used research method into the field of information systems research (Peffers et al., 2007). As such, it is suitable for solving what is fundamentally an engineering problem within the field of information security, which is closely related to the field of information systems. The research method can be used to create an information technology artifact, which is a solution to an information technology problem observed by an organization (Hevner et al., 2004). As per the definition by Peffers et al. (2007), the methodology used in this research is an iterative process consisting of six activities, which are:

1. Identify the problem and motivation
2. Define the objectives of a solution
3. Design and development
4. Demonstration
5. Evaluation
6. Communication

Peffers et al. (2007) determined the first activity to consist of identifying and demonstrating the problem by providing the fundamental principles behind the research problem, effectively justifying the need for the artifact to provide a solution to the problem at hand. The second activity consists of defining the solution based on the problem statement and the knowledge of the related topics. The third activity is about creating an artifact to provide a solution to the problem. In the fourth activity, the artifact is demonstrated in association with the problem. The fifth activity evaluates the effectiveness of the artifact as a solution to the problem statement. The sixth and final activity involves the communication of the problem, artifact and the results of the study to a broader audience, which is commonly achieved by writing a research paper about it (Peffers et al., 2007).

The first activity of the design science research methodology is covered in the chapter 1, the second activity is the chapter 6 backed by a literature review in chapters 2, 3, 4 and 5. The third and fourth activities are covered in the chapter 7 and the fifth activity in the chapter 8. The sixth activity is effectively the entirety of this thesis and as such, none of the chapters is explicitly associated with the sixth activity.

The purpose of the literature review is to provide sufficient theoretical background on the subject and enable a way to construct and utilize the design science artifact in a way that relevant metrics for SOCs can be produced with it. In addition to topics related to SOCs (chapter 2) and the way the SOCs are currently measured (chapter 5), the literature review focuses on metrics and measurements on a general level (chapter 4), and adversary behaviour (chapter 3), which is expected to produce information that can be utilized during the creation of the metrics.

The primary search engines used for the literature review to discover academic journals and other publications are JYKDOK[1], IEEE Xplore[2] and Google Scholar[3]. Google Search[4] is used for discovering commercial and other non-scientific publications used as supportive material for the published material. A keyword search is used as the search method for the discovery of literature related to SOCs and metrics. The search query for SOC-related literature is *'(security OR cyber security) (operations OR operation) (center OR centre)'* and the metric-related search query is *'(metric OR measurement OR "performance indicator")'*. Additionally, when searching for SOC-related metrics, the two queries are combined with an *AND* operator. There are no predefined keywords for other topics covered within this thesis and as a result, the remaining topics are discovered with free-text search based on the specific topic expected to be discovered.

Peer-reviewed research published in academic journals is considered the preferred source for information, but commercially produced material, industry standards, and other non-academic literature is used as supporting material if the preferred source for information does not exist or the information contained within the source is inadequate. The systematic literature review about SOC-related publications by Vielberth et al. (2020) is used as the primary source for information within this research, as the information contained in the research is comprehensive and it manages to stitch together the information scattered throughout multiple publications to a coherent description on how SOCs are seen from the academic perspective.

1.2 Research question and scope limitations

The objective of this research is to determine which metrics and other performance indicators are relevant for measuring the technical performance of a SOC. A technical performance metric within the context of this thesis is defined as: "a qualitative or quantitative indicator derived from one or more measures resulting from the activities performed by a SOC, which describes how well the SOC can utilize technologies to prevent, identify, detect and respond to cyber threats affecting the organization". The objective of the research leads to the following

[1] https://jyu.finna.fi/
[2] https://ieeexplore.ieee.org/
[3] https://scholar.google.com/
[4] https://www.google.com/

research questions:

1. What frameworks are available to measure the performance of a SOC?
2. What are the commonly mentioned key metrics used to measure a SOC?
3. Can the common metrics be used to measure technical performance?
4. How can the metrics be improved to enhance the reporting capabilities of technical performance?

The focus of this thesis is on the technical performance of a SOC, which leads to process- and people-related metrics being out of the scope of this research, unless they have a direct relation to the technical performance. Additionally, to ensure the resulting artifact is neutral in terms of technologies used by the SOCs, the metrics related purely to capabilities provided by technology are out of the scope of this research. In other words, within the context of this research, it is irrelevant whether product A is unable to detect a threat while product B can.

2 SECURITY OPERATIONS CENTER

The SOC is considered to be a pivotal function in the overall cyber defence capabilities of modern enterprise organizations and it is being referred to in literature and by practitioners by multiple different names, such as Cyber Security Operations Center (CSOC), Computer Security Incident Response Team (CSIRT), Computer Incident Response Team (CIRT), Computer Security Incident Response Capability (CSIRC) and Network Operations and Security Center (NOSC) (Knerler, Parker, & Zimmerman, 2022). However, it might be worth noting that function is likely to focus on a slightly different aspect of the overall cyber defence depending on its name, for example, CSIRT and CIRT focus more on post-attack incident response activities (Vielberth et al., 2020) while CSOC takes a more holistic approach covering multiple different cyber defence functions (Knerler et al., 2022).

2.1 What is a Security Operations Center?

One common way to describe a SOC is through a People, Processes and Technologies (PPT) framework (Knerler et al., 2022; Vielberth et al., 2020) where each subcategory is seen as a separate building block for running a successful SOC. Vielberth et al. (2020) summarized in their study that the people block describes the people associated with the SOC and their required competencies, the process block describes how the people are interacting with each other and how security incidents are handled, and the technology block describes the tools the work is done with. The study also argued that the PPT framework can be expanded to include Governance and Compliance, which enables organizations to utilize the SOC as a function to ensure compliance with various standards, such as ISO/IEC 27001, GDPR, or PCI-DSS. Furthermore, including governance and compliance as a part of the PPT framework will bring a more structured approach to managing the SOC via maturity assessments and metrics, which can be used to better determine the current state of the SOC (Vielberth et al., 2020).

Vielberth et al. (2020) concluded that the people aspect of the framework

consists of the people who are working in the SOC or are otherwise working in close collaboration with the SOC. Based on the study, the responsibilities in a SOC are typically structured by different tiers of analysts and a manager, who is responsible for managing the operations. Following the definitions within the study, tier 1 is typically responsible for reviewing the detections and performing triage on the alerts, which in practice means, they confirm the detection and thus determine whether there is a security incident, determine the severity of the security incident and enrich it with additional data. If the security incident cannot be mitigated at tier 1, it is typically escalated further to tier 2 (Vielberth et al., 2020).

As determined by Vielberth et al. (2020), tier 2 reviews the security incident and performs an in-depth analysis utilizing more advanced techniques with the primary purpose of determining the scope and the impact of the security incidents, and tier 3 handles the major security incidents and performs digital forensics, which are activities the tier 2 is usually unable to perform. The study also concluded, that in addition to the response activities, tier 1 is responsible for configuring and managing tools, tier 2 is responsible for designing and implementing strategies to contain and recover from security incidents and tier 3 is responsible for proactively identifying and mitigating threats affecting the organization. In addition to the responsibilities, there are typically several roles within the SOC, which can be managerial roles, such as incident response coordinators or security managers, technical roles, such as security analysts or security engineers, consulting roles, such as security architects or consultants, and other external personnel working closely with the SOC (Vielberth et al., 2020). The typical roles and responsibilities within SOC and their interactions are depicted in figure 1.

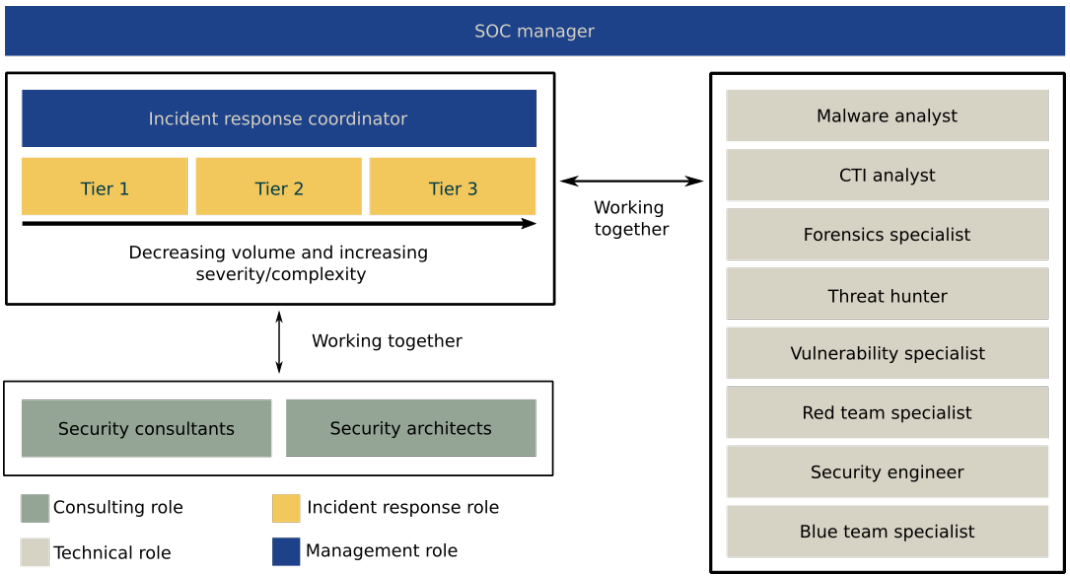

FIGURE 1 Interactions between various roles and responsibilities in SOC according to Vielberth et al. (2020, p. 8)

The work performed at a SOC is heavily driven by processes, as the work is structured around the prevention, detection, and response of security incidents, which means the work can often be quite hectic, and as such, without properly defined

processes there is a risk of SOC not being able to properly eradicate an active threat from the environment. Usually, some form of security incident management process is used as a basis for the entire operation, for example, the Computer Security Incident Handling Guide (NIST SP 800-61), which consists of four phases "preparation", "detection and analysis", "containment, eradication and recovery" and "post-incident activity", as depicted in figure 2 (Vielberth et al., 2020). The purpose of the preparation phase as defined by Cichonski, Millar, Grance, and Scarfone (2012) is to ensure the SOC has the necessary visibility and potential to detect and respond to security incidents and that they are ready to perform such activities in the later stages of the process. They also determine that the preparation stage also includes activities that are aiming the prevent security incidents altogether, such as malware prevention or user awareness training. They further define that the detection and analysis consist of responding to security incidents by analyzing, documenting, and prioritizing them, and finally notifying the necessary people of the detected security incidents. The third phase of the process is about containing the security incident by limiting the potential damage the incident can cause, followed by evidence gathering and identifying the attackers' host. Once the impact of the incident has been limited and the source of the attack has been identified, the threat can be eradicated and recovery actions can be started (Cichonski et al., 2012). The final stage of the incident response lifecycle, post-incident activity, does not contain activities that the SOC would typically be responsible for.

Other processes utilized in a SOC can include processes such as data collection process, automation of response activities via a Security Orchestration, Automation and Response (SOAR) tooling (Vielberth et al., 2020) or a detection engineering process, which is used to create, validate and tune monitoring rules (Knerler et al., 2022). SOCs can also have supplementary processes that are not directly tied to security or incident management, such as problem management or change enablement processes.

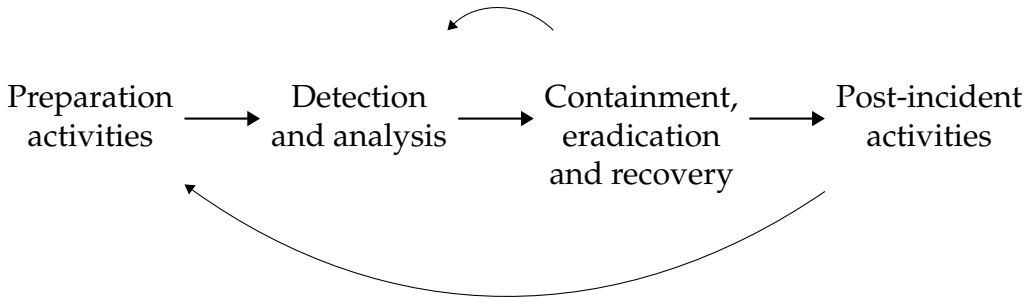

FIGURE 2 Incident response life-cycle as depicted by Cichonski et al. (2012, p. 21)

The technology describes the tooling and associated supplementary technical solutions used by the SOC personnel to run the SOC processes (Vielberth et al., 2020). According to the book by Knerler et al. (2022), the tooling typically includes security information and event management (SIEM) software to collect, process, and correlate events collected from log sources to generate detections for

the analysts to investigate, endpoint detection and response (EDR) tooling running on endpoints that utilize machine learning and analytical rules to generate detections, a network sensor to collect network data and detect intrusions, different platforms, such as incident management, big data analysis, or a SOAR platform. EDR software is considered to be an essential tool in a modern SOC, as it also allows the analysts to perform response activities on the endpoints (Knerler et al., 2022), such as investigating the endpoint via a remote command prompt or isolating the endpoint to prevent communication to destinations other than the EDR server. A SOAR tool can be used to automate workflows and certain response activities to enable faster containment of potential security incidents, which can be a valuable tool to both decrease the workload of the security analysts and reduce the impact of the security incidents (Knerler et al., 2022).

Vielberth et al. (2020) have conducted a systematic literature review of SOC-related literature and have determined, that the definition of SOC is not universal across the literature, which makes it difficult to determine what SOC is. As a result of the systematic literature review, Vielberth et al. (2020) defined SOC as follows:

> The Security Operations Center (SOC) represents an organizational aspect of an enterprise's security strategy. It combines processes, technologies, and people to manage and enhance an organization's overall security posture. This goal can usually not be accomplished by a single entity or system but rather by a complex structure. It creates situational awareness, mitigates the exposed risks, and helps to fulfill regulatory requirements. Additionally, a SOC provides governance and compliance as a framework in which people operate and to which processes and technologies are tailored. (Vielberth et al., 2020, p. 4).

Within the context of this thesis, the above definition of SOC is being used. Furthermore, the SOC as a function can be further split into multiple collections of tightly interlinked functional areas, which according to Knerler et al. (2022) are the following:

1. Incident triage, analysis, and response
2. CTI, hunting, and analytics
3. Expanded SOC operations
4. Vulnerability management
5. SOC tools, architecture, and engineering
6. Situational awareness, communications, and training
7. Leadership and management

Some more mature or larger SOCs could have elements from all of the functional areas above, but a single operations center does not necessarily need to cover all of the areas to be a functional part of the overall cyber defence capabilities (Knerler et al., 2022). The SOC services that were determined to be the most relevant within the context of this thesis, are "Incident triage, analysis, and response", "CTI, hunting, and analytics" and "Expanded SOC operations".

2.1.1 Incident triage, analysis, and response

Based on the definition by Knerler et al. (2022), incident triage, analysis, and response form the backbone of a SOC, and without these functions, it is not possible to run a viable operation, as the security incident management process is operated within this function. According to their definition, the incidents are detected, analyzed, contained and recovered from, based on the work conducted in this function. The tier 1, tier 2 and tier 3 analysts are typically performing real-time security monitoring by utilizing security technologies, such as SIEM or EDR, to react and analyze detections to determine whether they are true-positive or not, and thus evaluate whether they become a security incident or not (Knerler et al., 2022). Incidents are traditionally categorized either as false-positive, meaning the detection logic is incorrect, benign true-positive, meaning the detection logic is correct but the detection was proven to be legitimate, and true-positive, meaning the detection is considered to be a legitimate security incident.

According to Knerler et al. (2022), hundreds of millions of raw log data are typically collected every day to either a SIEM, log management, or a big data analysis platform. The raw data is collected from various log sources, networks, servers, endpoints, clouds, applications, and many others, which forms the basis for threat detection. Unnecessary noise is filtered out from the raw data, the filtered data is reduced to interesting security events out of which detections are created based on the monitoring rules resulting from the detection engineering process (Knerler et al., 2022). Detections can be referenced as alerts, alarms, offenses, or incidents depending on the source, but the fundamental definition is, that they are not yet considered to be security incidents. Detections are a result of monitoring rules that are created as a part of the analytics process based upon monitoring scenarios that aim to detect adversary behaviour and as a result, be able to gain actionable detections for the SOC analysts to investigate further. Figure 3 depicts how the raw data is transferred to actionable detections by the SOC analysts and the related personnel.

According to Knerler et al. (2022), operating without a SIEM or a data analysis platform can also be a viable option to consider for smaller and more focused SOCs. They mentioned that in such situations, the detections are mostly based on an EDR and other security products, which can be used purely based on their native capabilities, or the SOC can augment the capabilities based on a detection engineering process, and in such setups, the data used for incident response purposes is typically stored in log management or a cloud-native storage platform and is retrieved when necessary. Some more mature SOCs can also opt-in to build their in-house SIEM-like capabilities based on a big data platform and augment their detection capabilities with machine learning and other statistical methodologies to discover anomalies within the massive amounts of data collected to the platform (Knerler et al., 2022).

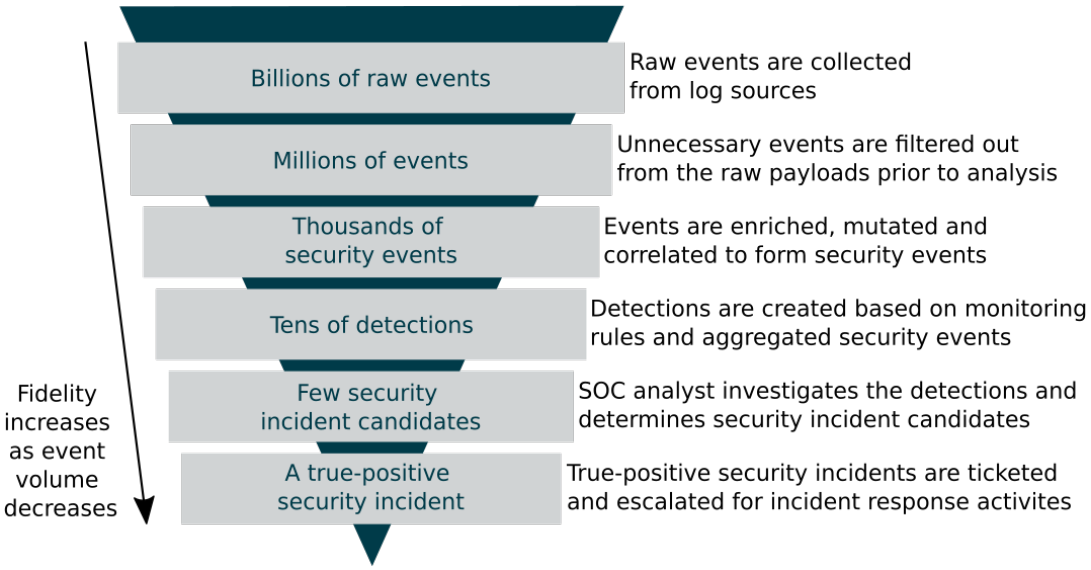

FIGURE 3 The SOC funnel: from raw data to actionable detections

2.1.2 Cyber threat intelligence, hunting, and analytics

Knerler et al. (2022) introduce cyber threat intelligence (CTI), threat hunting, and analytics as functions that are complimenting the incident triage, analysis, and response functions by providing them with actionable threat intelligence in the form of tactics, techniques, and procedures (TTPs), that are used to quantify the behaviour of adversaries. They also discuss Indicators of Compromise (IoC) that are forensic artifacts attributed to attackers, such as IP addresses of Command and Control (C2) servers and hashes of malware used to gain an initial foothold within the environment. The TTPs can be used as a basis for the detection engineering process, which contributes to the analytics function by providing actionable direction to the creation, tuning, and optimization of monitoring rules and thus enabling the SOC to detect threats associated with real-world adversaries (Knerler et al., 2022). The threat scenarios are modelled as monitoring scenarios that often are describing the threat, the detection logic, and the response activities to mitigate the impact of the threat. According to Knerler et al. (2022), monitoring rules can either be conditional expressions modelling a suspicious chain of events based on the TTPs or they can also be mathematical models or machine learning algorithms that aim to discover anomalies from the events stored in a data analytics platform. Without the analytics function, SOC would have to rely on the native detections produced by the technologies used at the SOC, which could considerably decrease the quality of the detections SOC is dealing with (Knerler et al., 2022).

As defined by Knerler et al. (2022), threat hunting is a proactive activity performed either by SOC analysts or dedicated threat hunters, and it goes beyond the threat detection capabilities provided by the monitoring rules to detect adversaries that are either previously unknown or are using techniques that are hard to model as monitoring rules or the capabilities are otherwise not yet there.

Threat hunting is typically based on a hypothesis, meaning the hunting activities are structured in a way that they have a specific objective the hunting activity is aiming to accomplish, which for example means, a hypothesis can be that an unknown threat actor is performing activity X to achieve objective Y and hunting is about looking at indicators of activity X from within the environment (Knerler et al., 2022).

2.1.3 Expanded security operations center operations

In addition to the traditional SOC functions, Knerler et al. (2022) establishes the concept of expanded SOC operations that includes proactive methods within the SOC in the form of attack simulations, adversary deception, insider threat prevention, and other active defence-related methodologies. According to them, the expanded operations are typically a part of more established and larger SOCs. They also state that attack simulation can be done as a part of the detection engineering process to validate the monitoring scenarios, or it can be a separate function organized in the form of red or purple teaming exercises, adversary emulation, or any other kind of testing activities similar to a penetration test. As Knerler et al. summarized, in all of the previous situations, the fundamental idea is to validate the detection capabilities of the SOC by utilizing offensive techniques in the live environment to see how the detection and incident response capabilities hold against the techniques used by real-world adversaries. The detection of insider threats can also be a separate operation from the day-to-day operations of the SOC in situations where an insider threat is a substantial concern for the organization and needs to be taken care of with higher priority, and thus, it could be justified to run insider threat detection as a separate function from the SOC with own people, processes and technologies rather than relying on the same elements utilized by the SOC (Knerler et al., 2022).

2.2 Maturity models

A maturity model can be used to assess the maturity of an organization or a process. The maturity is usually expressed as tiers of maturity where each tier contains multiple requirements the organization or process must fulfil to reach a certain level of maturity. There are several cybersecurity-related maturity models currently being used, such as Cybersecurity Maturity Model Certification (CMMC) by the U.S. Department of Defense (2021), Framework for Improving Critical Infrastructure Cybersecurity by the National Institute of Standards and Technology (NIST 2018) and Cybersecurity Capability Maturity Model (C2M2) by the U.S. Department of Energy (2022). While the general-purpose security maturity models are covering topics relating to the SOC, they are not directly applicable to measuring the maturity of a SOC. For this purpose, some frameworks are available, such as the SOC-CMM by Van Os (2016), CTI-SOC2M2 by

Schlette, Vielberth, and Pernul (2021) and the ENISA CSIRT Maturity Framework, which is more aimed towards assessing the maturity of national CSIRT functions (Dufkova et al., 2022), such as The National Cyber Security Centre Finland (NCSC-FI), but can to some degree be applied to SOCs.

The SOC-CMM maturity model by Van Os (2016) introduces a capability maturity model for SOCs as a design science artifact, which is based on items discovered in a systematic literature review and a survey on SOCs. Within the model, the maturity aspect is measuring the maturity of the business, people, and process domains, and is based on the five levels of maturity: initial, managed, defined, quantitatively managed, and optimizing, which are defined according to the Capability Maturity Model Integration (CMMI) model. Based on the SOC-CMM model, the initial level means there is no practical progress on an area that has been identified, managed level states that the activity is formalized so that it can be reliably repeated. The model also determined defined activities as something that are documented and formalized, quantitatively managed activities as something that are being measured for optimization on a process level, and the optimizing being a level where the activities are optimized on an organizational level. The capabilities are measured on a score between 0% and 100% and are not associated directly with CMMI levels, but instead, the results are normalized on a five-point scale upon which capability targets can be defined (Van Os, 2016). The usage of the SOC-CMM model is further clarified in a whitepaper published at a later date (Van Os, 2018), which summarized the usage of the SOC-CMM compactly but does not provide additional insights on the tool itself.

Another example is a CTI-centric capability and maturity framework CTI-SOC2M2 by Schlette et al. (2021), which is an interesting model for measuring the SOCs, as it does not measure the overall capabilities or the maturity of the SOC similarly as SOC-CMM does, but it rather assesses SOCs based on how well the usage of CTI is incorporated in the core services. The fundamental idea behind the framework is that without the usage of CTI, a SOC is unable to act swiftly on emerging threats and thus is unable to identify and mitigate post-exploitation activities resulting from the initial breach, increasing the impact of the breach. Similar to the other frameworks, the framework has maturity levels that are based on maturity tiers, which in the framework are none, initial, core, extended, and visionary, and to move from one maturity level to another, a specific capability level must be reached for specified services. For example, to reach the initial maturity level of the CTI-SOC2M2 framework, capability level two must be reached within the services "Log & event management", "Security monitoring, analysis & threat detection" and "Vulnerability management". Each service has a source and format for CTI data and the scoring is done based on how well the sources are utilized within the domain, for example, the initial maturity level being the source for the CTI is assessed and the visionary level requires a mechanism to keep track of the changes within the CTI and the related formats (Schlette et al., 2021).

3 CYBER ADVERSARY BEHAVIOUR

As the purpose of a SOC is to protect the environment against security threats and prevent detections from becoming high-impact security incidents, understanding how cyber adversaries are operating is an essential part of planning the overall defensive capabilities as well as the operational activities. A threat landscape report published by ENISA (2021) and similar whitepapers published by several commercial organizations (CrowdStrike, 2022; Firstbrook et al., 2022; Sophos, 2022) mention an increase in the sophistication of cyber attacks conducted by adversaries leading into the decline of general purpose malware and other run-of-the-mill techniques that traditional cyber defences are effective against to. As adversaries are constantly evolving and circumventing traditional cyber defences, cyber defenders need to understand what they are against, which is why understanding adversary behaviour is essential to construct a viable cyber defence operation.

3.1 Attribution of cyber adversaries

The types of adversaries can be generalized into three different groups: cybercrime actors, state-sponsored actors, and hacktivists (CrowdStrike, 2022; ENISA, 2021). The definition of the adversary types varies from one source to another, but ENISA (2021) determines that cybercrime actors are in it for personal gain, which means they are likely to utilize cost-effective techniques and target organizations that could potentially hand out a hefty payment. Based on the report, this is usually done either through the extortion of data or selling tools, such as ransomware or C2 frameworks, or other information stolen from the victim, such as credentials or confidential information, on a dark web marketplace. State-sponsored actors usually have different motives, and although some state-sponsored actors have conducted intrusions purely for monetary gain, actors attributed as state-sponsored have focused mostly on espionage and sabotage driven by geopolitical tensions (ENISA, 2021). As an example, it is considered that a state-sponsored

threat actor originating from Russia was behind the SolarWinds supply-chain breach in 2020, as the focus of the campaign was placed on espionage and the post-compromise activities targeted critical organizations in the United States (Willett, 2021). A recent report by Microsoft (2022) demonstrates that threat actors associated with the government of Russia have utilized cyber attacks to sabotage Ukrainian infrastructure before a kinetic strike conducted as a part of the Russo-Ukrainian War, which fits well into the definition of the activities performed by state-sponsored adversaries. The last adversary group, hacktivists, are considered to be actors that typically are performing relatively traditional unsophisticated attacks, such as denial of service and defacements, as a protest against organizations, which can minorly impact the target organization, but ultimately, the threat is nowhere near the other adversary groups in terms of potential impact (ENISA, 2021).

As stated by Rid and Buchanan (2015), attribution of cyber adversaries is fundamentally about identifying the entity behind the cyber attack, but the attribution as a process and the methodologies associated with it have been a source of debate among practitioners. Furthermore, they also mention that attribution of the adversary is an important part of assessing the potential impact of security incidents, but the attribution of adversaries can sometimes be difficult to perform accurately. Organizations can have difficulties performing the attribution in-house and as a result, they have to rely on public resources to do so. However, public attribution can sometimes be driven by geopolitics and it is not common for intelligence agencies to disagree on the attribution on a national level, which can lead to inaccurate attribution especially when it comes to state-sponsored threat actors (Egloff, 2020).

The term Advanced Persistent Threat (APT) is used to describe threat actors that are well organized, properly funded, and are aiming to establish long-term persistent access to the target environment (Cole, 2012). One way to perform the attribution is by analyzing the TTPs used by the adversary when traversing through the target organization attempting to meet their ultimate objective (Bahrami et al., 2019; Cole, 2012) but as concluded by Rid and Buchanan (2015), a high-quality attribution cannot be done with purely a technical routine, but instead requires a multi-layered approach to complete the analysis work, such as figuring out the reasoning for the selection of victims or performing extensive analysis on how the payloads delivered to the victim organization are programmed.

Although the attribution of threat actors is a relatively complicated problem, some frameworks attempt to map out how adversaries are behaving. Such frameworks include the Diamond Model (Caltagirone, Pendergast, & Betz, 2013), MITRE ATT&CK® Framework (Strom et al., 2018), Lockheed Martin Cyber Kill Chain® (CKC) (Hutchins, Cloppert, & Amin, 2011) or the Unified Kill Chain (UKC), that attempts to bring best parts of MITRE ATT&CK and the CKC together in a unified model (Pols, 2017). Out of these frameworks, the most commonly observed frameworks are the MITRE ATT&CK Framework and the CKC. However, as the CKC is a linear model describing the end-to-end process of an entire attack, that focuses on preparation (reconnaissance, weaponization), ex-

ploitation (delivery, exploitation, installation), and post-exploitation (command and control, action on objectives) activities (Hutchins et al., 2011), it has been a subject to some criticism among practitioners as it is unable to accurately model the adversary behaviour between stages. For example, the phases "Command and Control" and "Action on Objectives" can contain lateral movement and privilege escalation between them, but when using CKC, these activities cannot be distinguished, making the usage of CKC relatively ineffective in the practical applications, even though the fundamental idea behind CKC is still standing on solid ground.

3.2 MITRE ATT&CK Framework

The MITRE ATT&CK Framework is a collection of adversary TTPs with the purpose of mapping out adversary behaviour throughout the attack lifecycle (Strom et al., 2018) based on evidence observed in the real world. Version 11 of the framework consists of fourteen tactics (Mitre Corporation, 2022) that represent the reasoning behind the actions of the adversary, and several hundreds of techniques that model out how the objective is met. For example, within the paper published by Strom et al. (2018), the tactic "Credential Access" consists of techniques describing methods for stealing credentials, such as "OS Credential Dumping" or "Brute Force". The tactics and techniques as defined in the paper are still relatively abstract, but the procedures attached to the tactics describe the specific way of performing the technique. For example, a procedure can be a description of a method that utilizes PowerShell to dump credentials from Local Security Authority Server Service (LSASS) or a specific way to perform password spraying. The way the adversaries utilize specific procedures is what the framework is basing the attribution on; although credential dumping as a technique is the same, different threat actors might have opposite procedures from one another to accomplish this objective (Strom et al., 2018).

The framework can be utilized by SOCs to design and build their monitoring scenarios around real-world threat scenarios rather than relying on vendor-produced content or attempting to build the capabilities completely in-house (Ahlm, 2021; Strom et al., 2018). Based on the description by Strom et al. (2018), one of the key features of the framework is the possibility to utilize the TTPs to perform adversary emulation. They state, that when performing an adversary emulation, a threat actor relevant to your organization is chosen and the effectiveness of the procedures used by the threat actor group is tested in a live environment. They argue that this activity measures the effectiveness of the security controls in terms of preventing the execution and being able to detect both successful execution and the execution attempts blocked by the security controls. Fundamentally, it is about providing actionable TTPs, that can subsequently be prioritized by the defenders to improve the organizational defences against cyber threats (Strom et al., 2018).

There is some criticism among the cyber security practitioners of using the framework to assess the performance of SOCs and for a good reason. The research on the subject is relatively limited and it is currently unknown whether it is a good or a bad thing for the industry to base threat detection on a single framework. Additionally, several commercial organizations are using the framework to advertise their product capabilities in terms of coverage, which is in direct contradiction with the design philosophy of the framework as per the publication by Strom et al. (2018), who are the original authors of the framework. Targeting 100% coverage is not reasonable, as the solution would likely end up producing a large amount of false-positive or benign true-positive detections, making the SOC inefficient at completing its primary objective. Effective threat detection is a balance of several factors and utilizing the framework as a basis for creating actionable threat detection capabilities is a vital part of the overall strategy, meaning it should not be the sole source of information, but it could be argued that it is the best we have available at the moment.

3.3 The Unified Kill Chain

A whitepaper by Pols (2017) introduces the UKC, which was designed to improve the linear approach taken by the CKC and extend it with elements from the MITRE ATT&CK framework and multiple different variations of the CKC. They demonstrate that the UKC can be used to describe the end-to-end process of an attack and instead of being completely linear, it contains three separate iterative phases describing the journey of the adversary within the compromised environment, with an ultimate target of achieving the final objective they have defined for themselves. If an adversary fails to meet the objectives within a single iteration of the phase they are in, they can adjust their approach and re-iterate the phase for as long as required to meet the objective of the phase and move on to the next phase (Pols, 2017). The comparison of tactics between CKC, MITRE ATT&CK, and the UKC is shown in table 1.

The first stage of the UKC by Pols (2017) is "Initial Foothold" in which the attacker aims to gain a persistent presence within the targeted environment. The first stage contains adversary tactics that when used together, aim to breach the initial defences of the organizations and provide the adversary with the necessary level of access to pivot further into the environment, which is the objective of the first phase. The second stage within the UKC is "Network Propagation", which depicts the actions required by the adversary to gain a sufficient level of access within the target environment to achieve their ultimate objective, usually meaning access to a critical asset within the environment. The third phase is "Action on Objectives", which contains phases to achieve the objectives, usually in the form of data collection, exfiltration, and impacting the systems, by for example distributing ransomware malware in the environment (Pols, 2017). The phases of the second stage of the UKC are a good example of how the CKC falls short, as

TABLE 1 The Unified Kill Chain stages with Cyber Kill Chain and MITRE ATT&CK

	Cyber Kill Chain	MITRE ATT&CK	Unified Kill Chain
Initial Foothold			
Reconnaissance	*	*	*
Weaponization	*	*	*
Delivery	*	*	*
Social Engineering			*
Exploitation	*	*	*
Persistence	*	*	*
Defence Evasion		*	*
Command and Control	*	*	*
Pivoting			*
Network Propagation			
Discovery		*	*
Privilege Escalation		*	*
Execution		*	*
Credential Access		*	*
Lateral Movement		*	*
Access			*
Action on Objectives			
Collection		*	*
Exfiltration		*	*
Impact		*	*
Objectives	*		*

none of the stages of the CKC are contained within the stage. The second stage of the UKC highlights the fact that in many cases adversaries have a long way between Command and Control and their objectives, which offers the defenders multiple chances to detect the adversary and evict the threat.

Pols (2017) mentions that one way to use the UKC is to plan the defensive strategies in a way that an adversary has difficulties progressing from one stage to another, which limits the potential impact of the adversary's actions within the environment. It is difficult to completely prevent the compromise of an internet-connected device so accepting the fact can be a viable way to re-prioritize the defensive efforts, which can potentially lead to a more effective security posture against advanced adversaries (Pols, 2017).

4 METRICS AND MEASUREMENTS

Following the definition by Savola (2007), the primary purpose of a metric is to measure how well a business process, product, or resource is behaving, and upon which a business decision can be made. Typically, a metric consists of one or more discrete single point-in-time measurements, from which metrics are then derived from (Savola, 2007). For example, a measurement could be a temperature measurement taken every hour and a metric could be the daily average temperature calculated from the hourly measurements. The metric could then be used to track the changes in the average temperature over years to determine the impact of climate change on the average temperatures at a specific location.

Within the context of cyber security, Black, Scarfone, and Souppaya (2008) state that organizations can utilize metrics to verify the effectiveness of their cyber security program by making observations about the measures behind the metrics. This can, for example, help with the verification of security controls, identify areas of improvement or conclude the value of investments by observing long-term trends (Black et al., 2008). According to Savola (2013), to produce quality metrics for security, the resulting metrics must conform to three fundamental quality criteria: correctness, measurability, and meaningfulness. According to his definition, correctness is about the metric being implemented correctly and providing error-free results, measurability requires the metric to have defined dimensions, quantities, or other qualities, and meaningfulness is fundamentally about the metric being relevant and fit for purpose. The usability of the metrics should also be considered, as it has a direct impact on the practical meaningfulness of the metric (Savola, 2013).

4.1 Metrics and key performance indicators

The literature about security metrics uses the terms metrics and key performance indicator (KPI) almost interchangeably, although their fundamental meaning is slightly different from one another. A KPI is aligned with the strategy of the com-

pany, it has a significant impact on the company, it is non-financial, the period for the measurements is less than a week and it is tied specifically to single or multiple teams working in close collaboration (Parmenter, 2019). For example, a KPI in the SOC could be the number of high-priority security incidents this week that have not been handled according to the service level agreement. Following the definitions by Parmenter (2019), measuring the total number of incidents not handled as per service level agreement for a longer period of such, such as a month or a quarter, would be considered a performance indicator, as the scope is not limited to high-priority incidents and the period is longer. He also expands the definition of indicator slightly further and defines that there there are two groups of performance measures, result indicators and performance indicators, and in these groups, there are select key indicators that have a more significant impact on the business and thus, are called KPIs and key result indicators. Result indicators are used to measure the activities spanning multiple teams and can be financial or non-financial, while the performance indicators are non-financial, tied to a specific team, and measure a limited group of activities (Parmenter, 2019). Metrics on the other hand are general purpose standards of measurements as per the definition by Merriam-Webster[5], so it could be stated that all KPIs are metrics, but not all metrics are KPIs.

Most of the metrics mentioned in the literature are either performance or result indicators and it is up to the organization to decide whether an indicator can be considered a key indicator or not. For an organization that provides managed security services and running a SOC is a major source of their revenue, a performance indicator displaying the number of incidents that were not handled according to a service level agreement could be a KPI. For a larger organization for which the revenue from the SOC business is only a fraction of the total revenue, the same indicator could simply be a simple performance indicator as the impact of the metric is considerably lower.

4.2 Constructing metrics

An example structure of a metric is depicted in figure 4, in which the metric "number of security incidents with reaction SLA violations" is derived from the measurements of the reaction time for detections and tickets originating from multiple different sources. The configuration of the metric only selects the measurements where the reaction time exceeds a pre-determined value used to define the maximum time allowed to pass until the analysts react to detection or a ticket. The value of the metric could be zero even though there would be several measurements within the two categories if none of the reaction times pass the pre-determined upper limit for the reaction time and thus be in breach of the service level agreement (SLA).

There are several ways to construct or select metrics that are for measur-

[5] https://www.merriam-webster.com/dictionary/metric

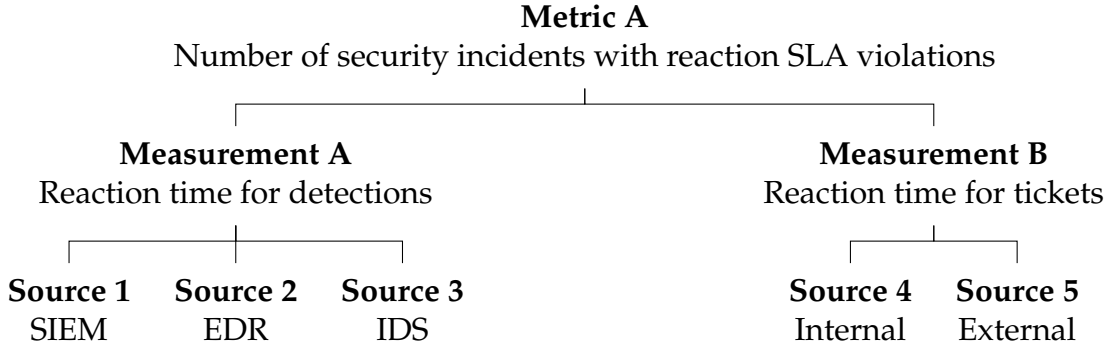

FIGURE 4 Structure of a metric

ing the outcome of an activity. A paper by Doran (1981) describes a method that utilizes the S.M.A.R.T. method, which is an acronym for specific, measurable, assignable, realistic, and time-related, to select a suitable metric. Furthermore, he argues that S.M.A.R.T. framework defines objectives that should be met when selecting metrics and it is not necessary to completely fulfil all of the objectives, but the closer the metric matches the criteria, the smarter the metric is. Within the framework, specific means that the metric should target a specific area or an entity, measurable means the metric needs to have a quantifiable indicator of progress, assignable should determine who takes responsibility for the metric, realistic means achieving realistic results and time-related specifies when the results are available (Doran, 1981).

Brotby and Hinson (2013) introduces another acronym-based methodology, which is the PRAGMATIC method that consists of nine metametrics (predictive, relevant, actionable, genuine, meaningful, accurate, timely, independent, and cheap) depicted in table 2, which are essentially scoring criteria for the metric itself. As per the definition in the book, the PRAGMATIC method has scoring criteria in a range between 0% and 100%, and the overall score for a metric can either be the average score of the metametrics or the average score after the individual metametrics are weighted based on selected criteria. Based on the method, the resulting score defines how well-structured the metric is, which in practice means, the higher the percentage, the better the metric is. The method contains a guideline that has a verbal definition for four different scores upon which the score for the given metametric can be determined (Brotby & Hinson, 2013). The definitions for the metametrics are also depicted in table 2.

In summary, the S.M.A.R.T. and PRAGMATIC methods are more about how metrics can be selected and which principles should be embraced when a metric is constructed. There are also other methods to construct metrics, for example, the Annex A of the ISO/IEC 27004:2016 (2016) standard summarizes a measurement information model contained in the ISO/IEC 15939 standard and describes how specific attributes related to an entity can be converted into an information product that can be used for conducting business decisions. Within the standard, the process starts by selecting which attributes from an entity should be measured, after which the measurements are converted to derived measures,

TABLE 2 Metametrics as defined by Brotby and Hinson (2013)

Metametric	Definition
Predictive	Metric can be used to predict what will be happening in the future.
Relevancy	Metric must produce relevant information for the intended targets of the metric.
Actionable	Metric must be influenced by the organization the metric is expected to measure.
Genuine	The metric should be objective, and provide credible and unambiguous information based on real-world information.
Meaningful	The metric should provide information that can be consumed by the intended audience of the metric.
Accuracy	The metric should be precise and provide correct information.
Timeliness	The metric should be timely in a way that any actions performed can be swiftly observed.
Independence	The metric should be objective and it should not be possible to manipulate the metric.
Cheap	The metric should be cost-effective and have a high net value.

out of which an indicator, or a metric, is ultimately constructed. The indicator is then interpreted to make an information product that can be used for conducting a business decision (ISO/IEC 27004:2016, 2016).

The study conducted by Savola (2013) focused on determining the quality of a security metric, rather than attempting to provide guidelines on how metrics can be scored or constructed. The outcome of the study was that a security metric should be correct, measurable, meaningful, and to some degree usable. The model is built in a way that related criteria are linked to the primary criteria, for example, unbiasedness and representativeness are linked to correctness, and objectiveness and applicability are linked to meaningfulness (Savola, 2013). Similar to S.M.A.R.T. and PRAGMATIC methods, the model can be used as a guideline when metrics are being constructed or a metric is selected from a group of metrics, by for example ranking the metric based on correctness, measurability, meaningfulness, and usability similar to the PRAGMATIC model.

4.3 Problems and pitfalls of metrics

Metrics can provide valuable insights into the performance of an organization but when incorrectly constructed, they can become counterproductive and as a result, decrease the performance of an organization rather than improve it. There is a possibility the metric can be incorrect either due to a problem with the raw measurements, a programming error has happened in the algorithm used to derive the metric, or someone is either maliciously or non-maliciously interpreting the metric incorrectly to steer their agenda forward in the organization (Brotby & Hinson, 2013). A study by Hauser and Katz (1998) argued that it might also be possible that certain metrics, which are hard to influence by the activities of team members, might lead to a situation where short-term decisions are favoured over long-term decisions. They argued that short-term decisions could have a swift

positive impact on the metric value and thus are subconsciously seen by the team as a better decision to make. There is also a possibility that a metric is precisely wrong, meaning that something is measured with high accuracy but the metric does not improve the business process it is supposed to improve, and as a result, it leads to negative consequences (Hauser & Katz, 1998).

We can also approach the problems with metrics from the quality perspective. As pointed out by Savola (2013), the top three criteria for a metric should be correctness, measurability, and meaningfulness. An attribute closely related to correctness is unbiasedness and objectivity, meaning the interpretation of the metric should not be influenced by the beliefs or biases a person looking at the metric might have, as otherwise, it could have unforeseen consequences for the interpretation of the metric. A similar conclusion could be reached in the other categories, for example, reproducibility is closely related to measurability. If a result from a metric cannot be reliably reproduced, the metric could produce an incorrect value and thus be subject to unconsciously interpreting it incorrectly.

The problem can also be with the quantity of the metrics. If the security metrics are not complete and they have significant gaps in some areas, it might not be possible to correlate multiple metrics together to gain an understanding of the entire situation, which then could lead to incorrect decisions. On the other hand, having too many metrics can also be a source of incorrect decisions, so it is crucial to balance the quantity and the quality of the metrics to create a coherent collection of metrics to base the decisions (Brotby & Hinson, 2013). A book by Parmenter (2019) presents an idea of a 10/80/10 rule as a way to structure the collections of metrics, in which there are ten key result indicators, a sum of eighty performance and result indicators, and ten KPIs covering the entire organization. He also argues that smaller organizations can manage with fewer indicators while larger organizations should figure out whether they can reduce the number of indicators or combine some cross-business unit metrics into one generic metric. An exception to the rule can be made if an organization is running multiple distinct businesses with different business models and as such, it is also entirely possible to deploy multiple sets of indicators following the 10/80/10 rule within a single organization (Parmenter, 2019).

5 SECURITY OPERATIONS CENTER METRICS

This chapter reviews a select group of SOC-related literature, both academic and commercial, to determine what kind of metrics can be discovered from the literature. In addition to literature specific to SOC, some of the key sources for generic security metrics are also reviewed to determine how well the generic security metrics can be applied in the SOC context. The objective of the SOC-related literature review is to gain a sufficient understanding of the metrics that could be considered to be common or typical for SOC and therefore, be considered to be established metrics for SOCs.

5.1 Published literature on Security Operations Center metrics

Vielberth et al. (2020) performed a structured literature review of the current state of SOC-related academic research published between 1990 and 2019. One of the conclusions they reached in the research was that the general level of governance and compliance-related aspects of SOC-related research are immature and while there is a significant amount of research about security metrics, the same cannot be said about SOC metrics-related research. Based on the research, they have identified four major groups for metrics: general SOC, people, technical, and governance and compliance metrics. The general metrics mentioned in the research consisted of coverage and general performance metrics, such as average analysis time or mean time to detect. They also introduced a couple of people-related metrics, such as metrics to measure the efficiency of the people working at the SOC, such as the number of incidents closed per shift. Finally, they outlined several technical metrics, such as threat, vulnerability, risk, alert, incident, and resilience-related metrics, such as threat actor attribution, cost per incident, or the number of incidents. Governance and compliance metrics include maturity and other compliance-related metrics, such as the number of policy violations or the percentage of systems with tested security controls (Vielberth et al., 2020). The research provides an excellent summary of the common metrics described

in the academic literature and it succeeds in highlighting the lack of a common group of metrics used to measure the technical performance of a SOC. It contains some relevant metrics for measuring the technical performance of a SOC, such as false-positive rate, mean time to detect, threat actor attribution, and defensive efficiency (Vielberth et al., 2020) but fails to demonstrate meaningful metrics on several areas of the SOC-CMM framework such as automation & orchestration, threat hunting, and detection engineering & validation.

Within a book by Nathans (2014), the SOC-related qualitative and quantitative metrics are mostly discussed within the context of two primary domains: metrics that are utilized by the management of the SOC and metrics that are related to vulnerabilities. The management-related metrics contained within the book consist of qualitative metrics such as the top 10 oldest tickets, and quantitative metrics such as tickets per incident type, number of tickets, number of tickets solved within agreed limits, number of incidents per department, mean time to resolution, mean time to first response, or average analyst downtime between tickets (Nathans, 2014). On an overall level, the metrics associated with the SOC management revolve mostly around tickets, their categories, and the way they are handled by the analysts, and the primary purpose of these metrics is to manage the staff and demonstrate that time-based contractual obligations such as the service level agreement are met within the SOC and as such, do not provide a concrete way to measure the technical performance of a SOC.

Keeping track of vulnerabilities within the environment is a vital part of ensuring the security of the monitored environment as per the definition by Nathans (2014). He defines that vulnerability-related information is crucial for SOCs to be able to meet their primary goal of detecting potential security incidents. Nathans also argues that SOCs are usually not the ones who are applying the fixes to the vulnerabilities, but instead, they work in close collaboration with system administrations responsible for applying the fixes. As per the definition by Nathans, SOC can provide the system administrators meaningful information about the impact of the vulnerabilities, instructions on how to properly mitigate the vulnerabilities or information if they have already been exploited and as a result, incident response activities are required. The metrics related to the vulnerabilities are categorized similarly to the management-related metrics and there are qualitative metrics, such as the top 10 vulnerable endpoints, quantitative metrics such as the number of vulnerable endpoints, number of vulnerabilities per severity, number of unknown assets, and the time it took to apply a patch that fixed the vulnerability (Nathans, 2014). While metrics related to vulnerabilities are more relevant from the technical performance point of view compared to supervisor-themed metrics, the proposed metrics are not sufficient to be used to demonstrate the technical capabilities of a SOC. However, as the book by Nathans focused mostly on high-level concepts related to SOC it was expected to not contain concrete lists of metrics to be implemented by SOCs.

Agyepong, Cherdantseva, Reinecke, and Burnap (2020) have constructed a framework in which the SOC is split into multiple functions and each function is measured separately to determine the actual performance of the analysts within the context of a function, which are monitoring and detection, analysis, response

and reporting, intelligence, baseline and vulnerability, and policies and signature management. Furthermore, the framework proposes that each function should monitor its performance in terms of quantitative (absolute numbers derived from analysts' tasks, time-based measures), and qualitative (quality of analysts' analysis and report) performance metrics, out of which the qualitative metrics are seen as subjective and as such, hard to measure in practice (Agyepong et al., 2020). The framework does not provide concrete measurement mechanisms upon which organizations can implement the metrics defined in the framework, which leads any implementation based on the framework to be different from one another, making it difficult to compare the results between multiple organizations.

In his master's thesis, Keltanen (2019) proposed to utilize results from a customer survey as a way to measure the performance of an outsourced SOC. The focus of the thesis is placed on the way the customer survey should be structured and how to build meaningful metrics to be measured with the questionnaire sent out to the customers. The method chosen for developing the metrics within the thesis is the Goal-Question-Metric (CQM) method, where a goal is first defined, then a question is asked on how the goal can be achieved, and finally, a metric provides an answer to the question is created. Keltanen then ranks the metrics based on the PRAGMATIC method, where the metric is ranked based on submetrics referenced as metametrics, such as how genuine or accurate the metric is. The resulting score will help to evaluate different metrics between one another, making it possible to determine which metric is considered to be the most important (Keltanen, 2019). The study does not present concrete metrics that could be used by SOCs to measure their performance but rather focuses on how the metrics can be constructed.

Kokulu et al. (2019) published a research in which they performed a qualitative study on the issues observed by SOC practitioners. The study was based on interviewing eighteen persons working in a SOC, ten of them being managers and eight being security analysts. They concluded, that the key issues for SOCs are: lack of visibility on networks and endpoints, protections against phishing are inefficient, false-positives appear to have no significant impact on the operative activities of the SOCs, current performance metrics are ineffective and the analysts and managers disagree on several high-impacting topics, such as level of automation, tool functionality, and evaluation metrics. The research implied that quantitative metrics, such as mean time to response, mean time to detection, and the total number of incidents, are more common within SOCs and that the common metrics include mostly metrics that are seen as beneficial by most of the managers but useless by the analysts. The analysts mentioned that the metrics selected for measuring the performance are used to measure completely irrelevant things and are used for demonstrating a false improvement to the upper management rather than demonstrating the actual performance of a SOC (Kokulu et al., 2019). While the research did not contain any meaningful metrics for the technical evaluation of SOC performance, the evaluation metrics being a subject of disagreement between the security analysts and the managers is a good example of the mismatch between the practitioners and managers in general, which could be a contributing factor on the lack of common metrics for measuring the

technical performance of SOCs.

A SOC framework proposed by Onwubiko (2015) consists of log collection, analysis, incident response, reporting, personnel, and continuous monitoring. He defines that the reporting part of the framework consists of metrics used to evaluate the performance of the SOC and to determine the return on investment. The framework does not contain a concrete set of metrics a SOC should take into use, but instead, it provides a top five examples that should be taken into account, which are the number of incidents, the performance of the cyber operations (true-positive, false-positive, false-negative, true-negative), top ten cyber attacks, a summary of policy violations and a summary of privileged use misuse detections (Onwubiko, 2015). Rather interestingly, the research is among the few that discusses false-negative detections, but unfortunately, it does not deliver any actionable information about this or any other metrics for that matter.

5.2 Commercial whitepapers and publications

In addition to scientific research and other publications on the subject, there is a plethora of commercial material available that discusses the metrics to measure SOCs. As scientific research and other published works are fairly limited in terms of content and availability, looking into unpublished works could provide additional insights into the available metrics for the SOC, as seen by the industry and the SOC practitioners.

Zimmerman and Crowley (2019) held a presentation at the annual FireEye Cyber Defense Summit of 2019 about practical SOC metrics, which laid out seven focus groups for the metrics. The first group they introduced was about the health of the data feed, presenting the idea that the status of the data ingestion should be monitored to detect if there are any large-scale gaps within the visibility due to sensor issues. The second group was about the coverage of the monitoring in terms of the percentage of environments covered, linked to different computing layers, and the number of device groups covered. The third group within the presentation consisted of vulnerability relating metrics, for example, the percentage of assets covered by the vulnerability assessment, and the fourth group focused on the monitoring rules and highlighted some generic metadata about the monitoring rules themselves, such as the kill chain or MITRE ATT&CK mapping or which data sources the monitoring rule depends on.

The fifth group within the presentation by Zimmerman and Crowley (2019) was all about the analysts' performance and provided metrics such as the true-positive rate per analyst or the true-positive rate of escalations made by the analyst. The sixth group within the presentation was about incident handling and mentioned the usual operational metrics, such as time to detect, time to react, and time to containment, and it also further presented more advanced types of metrics, such as the relation of proactive vs reaction work, incorrect conclusions, and insufficient threat eradication. The final and seventh area was about risk priori-

ties and general hygiene and contained metrics about discovered and mitigated vulnerabilities (Zimmerman & Crowley, 2019). The presentation contained several metrics that are useful for the demonstration of the technical performance of a SOC but lack the justification behind the metrics and as such, the authority of the author is the only basis for the metrics.

Gartner published a guide (Ahlm, 2021) discussing the industry best practices for building and operating a modern SOC, among which a set of fundamental metrics were defined. The metrics contained in the guide focus mostly on quantitative metrics and they were grouped into four categories, which are incident volume, incident detection, incident response, and incident impact. The incident volume group, as described in the guide, consisted of metrics focusing on the metadata of the incidents, such as the total number of incidents or incident severities. The incident detection group within the guide paid more attention to the detection capabilities and included metrics such as the total number of use cases or false-positive rate for the said use cases. The incident response group focused on how long it takes for the incident to be handled and had metrics such as time to detect and time to contain, and the final group of incident impact contains metrics such as financial loss or brand impact and attempts to quantify the actual impact of the security incidents in the form of relatively simple metrics (Ahlm, 2021). The metrics for most parts are not relevant for the technical performance of a SOC, except for the false-positive rate and the time to contain. It could even be argued that some of the metrics presented in the paper, such as the total number of use cases, could be considered harmful, as focusing purely on the quantity of the use cases with no regard to the quality could end up contributing to a false sense of security, as a large number of monitoring rules does not directly correlate with a better detection capability.

A survey by SANS Institute (Crowley & Pescatore, 2019) surveyed 355 organizations about common and best practices for SOCs. Among the topics they surveyed, there were questions about which metrics are used by the SOC or measure its performance. The results of the survey conclude, that quantitative metrics such as the number of incidents handled and time from detection to containment to eradication are the most common ones used by the surveyed organizations. According to the survey, the three least commonly deployed metrics were "Losses accrued vs. losses prevented", "Monetary cost per incident" and "Avoidability of the incident" all of which are relatively difficult to implement, and as such, it was concluded that the results are not surprising. Out of the metrics mentioned, there were a few metrics that are relevant for measuring the technical performance of the SOC, which are "Threat actor attribution", "Time from detection to containment to eradication", "Thoroughness of eradication" and "Thoroughness and accuracy of enterprise sweeping" (Crowley & Pescatore, 2019).

A guide for measuring SOC published by Logsign (n.d.) groups the recommended metrics into two categories, metrics for security operations and metrics related to business requirements. Within the guide, the number of security incidents was raised as the most important metric for security operations, as the metric can be used to derive additional information, such as whether the overall number of security incidents is increasing or decreasing. The guide also depicted

additional metrics for the security operations, such as the number of alerts per analyst, the number of alerts closed by automation, the number of false-positive alerts, and the average time to detect a security incident. Metrics related to business requirements were also mentioned, which are essentially about measuring how the SOC is running from the business perspective, and the metrics include the productivity of security analysts and the number of incidents impacting the business (Logsign, n.d.). The metrics described in the guide are somewhat relevant for measuring the technical performance of a SOC, but the justifications for the metrics are practically non-existing and as such, the whitepaper is similar to the other commercial material and is relying purely on authority without any scientific justification behind the metrics.

A blog post by Simos and Dellinger (2019) presented some of the key metrics used at the Microsoft SOC. Time to acknowledge (TTA) was the first metric mentioned within the blog, which is about the responsiveness of the SOC. According to the definition in the blog, the TTA measures the time between the alert being raised and the time an analyst begins the investigation process. The second metric outlined in the blog is time to remediate (TTR), which measures the time it takes to contain an incident from the time of detection. The lower the TTR the less time the adversary has within the environment, and the third metric measures the number of incidents remediated grouped per response type, meaning either manually by an analyst or automatically by automation technologies. The fourth and last metric is escalation between tiers, which is used to track the workload between SOC tiers (Simos & Dellinger, 2019). The metrics mentioned in the blog are mostly about measuring the response capabilities of the SOC and are highly relevant to the technical performance of the SOC.

5.3 Generic security metrics

In addition to SOC-specific frameworks and methodologies, several publications describe ways of measuring information security on a general level rather than focusing on a specific topic within the field of information security. One such document is the Performance Measurement Guide for Information Security (NIST SP 800-55) published by the National Institute of Standards and Technology (NIST), which describes a way for the organization to create, select and implement metrics for monitoring the state of security program on an overall level (Chew et al., 2008). Another publication is the ISO/IEC 27004:2016 standard, which describes guidelines that can be utilized by organizations for measuring the effectiveness of the information security management system implemented as per requirements defined in the ISO/IEC 27001:2013 standard (ISO/IEC 27004:2016, 2016).

The NIST SP 800-55 definition by Chew et al. (2008) contains some examples of metrics that organizations can deploy to meet a portion of the requirements defined by the Federal Information Processing Standard (FIPS) 200, which is used to describe the minimum requirements for federal information systems within the

United States. The metrics mentioned in the SP 800-55 are not directly targeted to be deployed to measure a SOC, but some of the metrics are similar to the metrics commonly observed in the SOC-related literature, such as the number of incidents or the number of incidents reported within an agreed timeframe. Similar to the ISO/IEC 27000 family of standards, there are some relations between the NIST publications, as the metrics described in the NIST SP 800-55 are referring to security controls in the NIST SP 800-53. However, the fundamental idea of NIST SP 800-55 is not about providing a list of metrics for organizations to deploy but instead providing detailed guidelines on how organizations can implement a security measurement program themselves, describing a process of developing and implementing metrics (Chew et al., 2008).

The ISO/IEC 27004:2016 (2016) standard describes the rationale, characteristics, types of metrics, and processes relating to measuring the effectiveness of the information security management system (ISMS). Based on the standard, the rationale behind the measuring is that although the controls described in the ISO/IEC 27001 standard are in place, there is no guarantee that they would remain effective for eternity and as a result, several of the controls also enforce the implementation of a metric to evaluate the effectiveness of the control. In addition, within the standard, the metrics validate the results of the implementation process and provide additional benefits for the organization, such as increased accountability and support for the decision-making process. The standard also discusses the characteristics of the security metrics and it provides information on the general properties of metrics, suggestions on what business processes to monitor, what to measure, and when and by whom the monitoring of the metrics is performed.

The ISO/IEC 27004:2016 (2016) standard groups the types of metrics into two categories: performance and effectiveness metrics. Within the context of the standard, the performance metrics are used for measuring the progress and the effectiveness of the implementation of the ISMS processes, and the effectiveness metrics are used to measure the impact of the implemented processes, for example, cost savings or a degree of customer trust gained or maintained by the ISMS program. To support the creation of effective metrics, the standard introduces six processes: identity information needs, create and maintain measures, establish procedures, monitoring and measure, analyse results, and evaluate information security performance and ISMS effectiveness (ISO/IEC 27004:2016, 2016). Similar to NIST SP 800-53, the metrics proposed by the ISO/IEC 27004 standard are examples of generic security metrics that are not directly related to SOCs. However, a functional SOC can have an impact on the effectiveness of the metrics, such as a decrease in the cost of security incidents, which is one of the sample metrics contained in the document.

In his master's thesis, Salmi (2018) conducted a survey of information security metrics implemented in large Finnish corporations, in which he identified 28 security metrics categorized into either management, operational or technical metrics. The study contained no metrics that would be directly related to the technical performance of a SOC, but several metrics are closely related to the typical activities of a SOC, such as the business impact of security incidents,

characteristics of security incidents and system vulnerabilities (Salmi, 2018). The characteristics of the observed security metrics mentioned in the thesis are matching closely the quantitative metrics mentioned in other literature, especially the ISO/IEC 27004. It could be concluded that in Finland, organizations are building their security metrics on the requirements of the ISO/IEC 27004 standard, and as a result, are not effectively measuring the technical performance of their SOCs in a way that would be strongly incorporated in the metrics related to the information security management system related processes.

Pendleton et al. (2016) performed a literature survey on the system security metrics. The research classified systems based on two groups of systems: enterprise systems and computer systems. The study defined, that the term enterprise system refers to a group of individual interconnected computer systems that together form an enterprise system, and that the computer system refers to an individual entity, consisting of a single self-contained node, device, or computer. The research resulted in a framework, which categorizes security metrics based on four sub-metrics: metrics of system vulnerabilities, metrics of defence power, metrics of attack or threat severity, and metrics of situations. All of the sub-metric groups defined within the research contain metrics that can be used to measure the performance of a SOC, and for example, the metrics of situations group discusses metrics that are also commonly referred to in the SOC-related literature, such as incident rate, which is used to measure how often computer systems are infected with malware, or cost of incidents, which measures the monetary losses resulting from incidents. The remaining groups are telling a similar story, the metrics of the attack group contain a metric for measuring the success of detection of malware associated with advanced persistence threat actors, and the defence power group discusses intrusion detection metrics, such as false- or true-positive rates, and the vulnerability group contains metrics to measure the lifetime of a vulnerability (Pendleton et al., 2016). Despite introducing multiple useful concepts, the metrics described in the research were quite abstract on a practical level, but regardless, they contained enough information to enable a relatively straightforward process to derive metrics dedicated to a SOC based on the metrics demonstrated in the research.

5.4 Summary of Security Operations Center metrics

As a summary of chapter 5, table 3 describes the top thirty most commonly seen metrics in the literature. Some of the metrics were combined into a generalized term instead of having two separate rows for what is essentially the same thing, such as "time to resolution" and "time to incident closure". Another general level observation from the metrics was that the terminology does not appear to be consistent among the literature. For example, the metric "mean time to detect" is used to describe at least two different behaviours, the time it takes to react and perform the analysis of the alert (Agyepong et al., 2020; Ahlm, 2021; Crowley &

TABLE 3 Top 30 metrics in the literature

	Chew et al. (2008)	ISO/IEC 27004:2016 (2016)	Salmi (2018)	Vielberth et al. (2020)	Nathans (2014)	Keltanen (2019)	Agyepong et al. (2020)	Kokulu et al. (2019)	Onwubiko (2015)	Crowley and Pescatore (2019)	Ahlm (2021)	Logsign (n.d.)	Simos and Dellinger (2019)	Zimmerman and Crowley (2019)
Number of security incidents	*	*	*	*	*		*	*	*	*	*	*		
Mean time to reaction						*	*	*			*	*	*	*
Number of vulnerabilities	*	*	*	*	*			*						*
False-positive rate				*	*	*			*		*	*		
Mean time to detect				*			*	*			*	*		*
Mean time to resolution			*			*		*		*	*			*
Cost of security incidents		*		*		*				*	*			
Detections per category		*			*						*	*		
Mean time to vuln. remedy				*	*			*					*	
Number of vulnerable devices			*	*	*									*
% of employees trained	*	*	*											
% of standard systems	*	*	*											
Analyst productivity				*								*		*
Coverage of vuln. scanning		*	*											*
Downtime due to sec. incidents			*							*	*			
Incident avoidability					*					*				*
Incidents with business-impact				*							*	*		
Mean time to containment								*		*				*
Mean time to triage				*	*									*
Number of incidents per shift				*			*			*				
Number of monitored assets			*	*										*
Number of patched vulns.	*		*											*
Number of risk per severity		*	*	*										
Resolution SLA breaches		*	*		*									
Severity of sec. Incidents				*				*			*			
Threat actor attribution				*						*				*
# of automated incidents												*	*	
Mean time to escalation								*						*
Quality of eradication										*				*
Reaction SLA breaches	*													*
Sources of detection												*		*

Pescatore, 2019) and time it takes for the SOC to become aware of the incident (Logsign, n.d.; Vielberth et al., 2020). Some of the publications did not provide enough information to make a clear distinction between the two (Kokulu et al., 2019; Zimmerman & Crowley, 2019). However, as the systematic literature review on the SOC metrics performed by Vielberth et al. (2020) provided a separate metric for the average analysis time in addition to the time to detect, a conclusion could be reached that the correct definition for time to detect metric would be the time between the initial activity of the adversary and the first detection caused by the activities. This viewpoint is also supported by the Computer Security Incident Handling Guide (NIST SP 800-61), which separates the detection and analysis as a separate activity within a single phase of the incident response life cycle (Cichonski et al., 2012). Table 3 follows the terminology used in the lit-

erature and thus is partially inconsistent. As a result, it succeeds in highlighting a common problem with SOC-related metrics in the literature.

The most common metric was the total number of security incidents, which was mentioned in eleven of the fourteen publications included in table 3. This is not a surprise, given the number of incidents is mentioned as a specific metric in both NIST SP 800-53 and ISO/IEC 27004, and it is a metric that can easily be collected and used outside of the scope of SOC. Vulnerability-related metrics are also relatively common, with the count of vulnerabilities being mentioned in seven publications, and mean time to vulnerability remediation and the count of vulnerable devices were mentioned in four publications. Operative metrics, such as mean time to reaction, detection, resolution, containment, triage, and escalation were also commonly mentioned, but the publications also contained a few technical performance metrics. The technical performance metrics mentioned in the source literature were false-positive rate, threat actor attribution, the number of security incidents closed with automation, and the quality of eradication. Many of the publications did not explain the metrics with enough detail to draw a definitive conclusion of the actual meaning of the metric, which means some of the items summarized in table 3 are subjective to some degree, as the terminologies used were slightly different between the publications, which is likely to skew the results slightly.

The literature review further emphasizes the need for a common framework that could be used for the performance evaluation of SOCs globally, as the metrics presented in table 3 are scattered broadly and as mentioned earlier, there appear to be no common definitions for the so-called key metrics. The published literature is mostly focusing on operational SOC-related metrics or general security metrics. Although the commercial whitepapers provide slightly better technical performance metrics, they fall short in several ways, for example, the lack of proper justification is seen throughout the commercial whitepapers, which means they are attempting to push their message through based purely on authority and are not attempting to justify their views. Although the scientific research around the subject is limited, several studies have arrived at a similar conclusion (Agyepong et al., 2020; Keltanen, 2019; Kokulu et al., 2019; Vielberth et al., 2020), which supports the original hypothesis of lack of commonly available technical performance metrics for measuring the performance of a SOC. The lack of standard technical performance metrics could be attributed to the lack of a sufficiently mature governance model for SOCs as pointed out by Vielberth et al. (2020) in their structured literature review about SOCs.

6 SOLUTION OBJECTIVES

The objective of the solution is the documentation of an approach for the creation of SOC-related metrics. Additionally, the solution should be used to create three to seven metrics that can either augment existing technical performance metrics or introduce completely new metrics and provide capabilities to measure the technical performance in an area where previous metrics are either incomplete or entirely missing. Based on the literature review, the requirements for the solution are the following:

1. A selection criteria for the creation of metrics should be well-defined.
2. A separate quality criteria for the metrics should be defined and the metrics created with the framework should conform with it.
3. The metrics should be directly associated with a specific SOC function.
4. The metrics should be universal and not tied to a specific technology or an organizational structure.
5. The metrics should be justified either by scientific research, industry standards, or through other means that considerably decrease the subjectivity of the metrics.

The justification for the item 1 is that there does not appear to be an industry-standard framework for the creation of metrics, especially for a SOC. There are selection criteria such as S.M.A.R.T. (Doran, 1981) and PRAGMATIC (Brotby & Hinson, 2013) that are used for the creation of security metrics, but the selection criteria remain subjective at best, especially when the context is shifted outside from general security metrics. The creation of a comprehensive framework for measuring the performance of a SOC is out of the scope of this research, and the focus is placed on enabling the creation of selection criteria that can be used as a part of the comprehensive framework. The resulting selection criteria will be based on a combination of existing methods and interpretations within the context of a SOC.

Although the creation of a comprehensive framework is out of the scope of this research, the quality criteria for the SOC metrics should be defined. The justification for the item 2 is that based on the empirical experience, many of the

metrics used for measuring SOC are of low quality especially when it comes to bias and objectivity. The literature review did not prove the hypothesis to be false nor did it confirm it to be true either. As a result, to ensure the resulting metrics are of high quality, quality criteria for SOC metrics should be defined.

The justification for the item 3 and item 4 is the same, which is the universal applicability of the resulting metrics. In practice, this means the metrics should be usable by most SOCs out there, given they include the function to which the metric is tied. Finally, the justification for the item 5 is that the metrics in some situations do not appear to be backed up by scientific research or by any other method, especially when it comes to commercial sources, which means the metrics are purely based on the authority of the author and not necessarily on anything meaningful.

7 MODEL CREATION AND TESTING

This chapter consists of the creation of the design science artifact in the form of a metric selection framework. The metric selection framework can be used to construct metrics used to measure the performance of a SOC and consists of the requirements for the metric to be valid and the characteristics of the metric, which should be documented for each metric created with the selection framework. The design science artifact is evaluated by utilizing it to create metrics that are used to measure the technical performance of a SOC. Scoring and weighting of the metrics have been left out of the framework, as they have too many variables to result in objective scoring of the metrics.

The model is being constructed based on requirements defined in two separate stages, as depicted in figure 5. The first stage consists of the solution objectives that are defined as a part of the design science research methodology in the chapter 6. The second stage consists of the metric selection framework, as seen in figure 6, that consists of the requirements the metric should conform to and the characteristics that must be describable in the metric documentation.

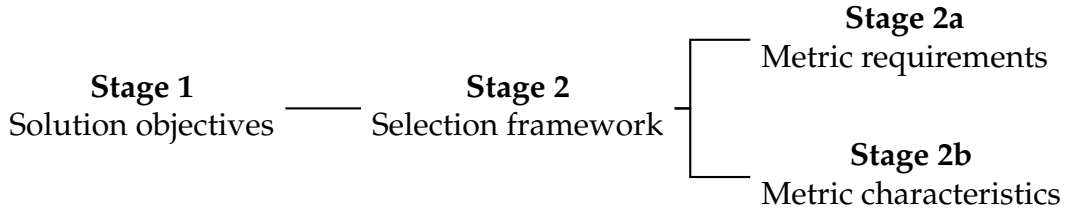

FIGURE 5 The metric construction stages

7.1 Quality criteria for the metrics

As the quality of the metrics is a critical factor when it comes to metric adaptation, the metrics should conform to formal quality criteria to ensure the metrics can

contribute something back to the organization. Lack of quality can lead to a wide variety of issues, such as SOC personnel seeing metrics as useless (Kokulu et al., 2019) or the focus being placed on measuring topics that are irrelevant for the organizational performance (Hauser & Katz, 1998).

The quality criteria for the SOC metrics is based on the model presented by Savola (2013), which concludes that correctness, measurability, and meaningfulness are the primary quality criteria for security metrics. The study also concluded, that on a practical level, usability is also a factor to take into consideration, and as such, it is added as the fourth criterion within the quality criteria. The quality criteria are shown in table 4, which describes the characteristics of the quality criteria within the context of a SOC.

TABLE 4 The quality criteria for the SOC metrics

Criteria	Characteristic	Description
Correctness	Granularity	The metric should provide the necessary granularity to tie the metric to a specific function or team within the SOC.
	Completeness	The metric should completely fulfil the goal defined in the metric documentation. If the metric cannot alone fill the goal, the metric should be coupled together with another metric.
	Objective and unbiased	The results should not be influenced by the activities performed by the person setting up the metric and the bias should be minimized to acceptable levels defined in the metric description.
Measurability	Availability	Measurements used to construct the metric must be automatically available in a reliable and consistent format.
	Reproducibility	The metric must be reproducible by different persons across multiple organizations given access to the same measurements.
Meaningfulness	Impactful	The metric must have an impact on the daily activities and it must be capable of showing the progression of development efforts.
	Clarity	The interpretation of the metric must be unambiguous and consistent across the entire lifecycle of the metric.
	Comparability	The result of the metric must be comparable between multiple SOCs even between organizations.
Usability	Portability	The metric must be usable by multiple different SOCs and not be dependent on their size, structure, service model, or parent organization.
	Controllability	The team the metric is used to measure must be capable of keeping the metric value between the expected values.
	Scalability	The metric must be able to behave consistently with low and high volumes of measurements.
	Presentable	It must be possible to visually present the information the metric is expected to provide.

7.2 The metric selection framework

The metric selection framework used for the creation of the metrics used for measuring a SOC is depicted in figure 6. It consists of two parts, the metric requirements, and the metric characteristics. To qualify to be a valid metric, all the criteria defined in both the metric requirements and the characteristics must be fulfiled. The metric requirements are:

1. The metric has a clear and well-defined goal.
2. The owner of the metric is clear.
3. The results are not dependent on third parties.
4. The metric can be justified.
5. The metric is tied to a success factor.
6. The metric is aligned with the quality criteria.

The metric should have a well-defined goal, which means the metric must be meaningful as per the quality criteria by Savola (2013) and the PRAGMATIC methodology by Brotby and Hinson (2013). On a practical level, the goal can be defined arbitrarily, but as there is a distinctive disconnection between the general security metrics and the SOC-related metrics, it would best to have a connection to the information security management program deployed within the organization, for example, the ISO/IEC 27001, which subsequently creates a link to the metrics defined in the ISO/IEC 27004. For example, the metric "mean time to reaction" could have a goal to increase the probability of the impact of a security incident being limited only to a single entity and thus decrease the costs associated with the security incidents. The metric could be linked to the measurement construct "B.8 Security incidents cost" in ISO/IEC 27004, which has a direct relation with clause ten in the ISO/IEC 27001 standard (ISO/IEC 27004:2016, 2016).

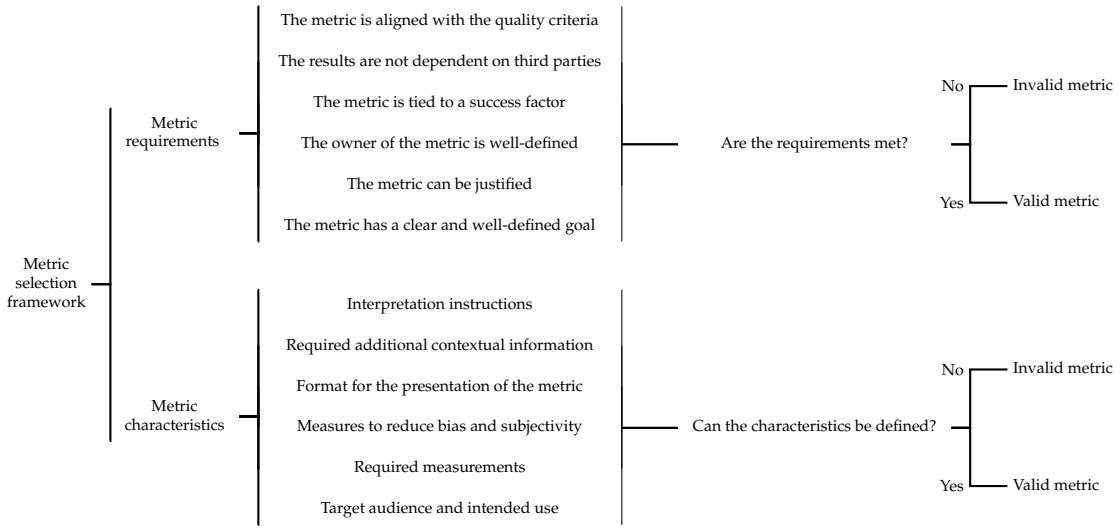

FIGURE 6 The metric selection framework

The metric should also be assignable directly to a function within the SOC. The

function definitions can differ between SOCs and some of the functions mentioned in the section 2.1 could be in practice produced by the same team, for example, threat hunting and incident response can be performed by the same team, although they are considered to be different SOC functions as per the definitions within the section 2.1. By assigning the metric to a specific function, it is possible to establish clear ownership of the metric and target the metric to a specific area as per the S.M.A.R.T criteria (Doran, 1981), and define a relationship with the SOC-CMM by Van Os (2016), or any other capability maturity model organizations are using to measure the maturity of their SOC.

If the metric is influenced by a third party, the metric does not measure only the performance of the SOC, but rather the entire chain related to threat detection and incident response. In the big picture, measuring the end-to-end capabilities is an important factor to consider, but as the objective of the metrics within the context of this thesis is to measure the performance of the SOC, the metrics cannot be influenced by third parties. For example, if the SOC is outsourced to a third-party vendor and the incident response is done internally by the customer organization, metrics such as "mean time to resolution" or "mean time to containment" are dependent on the activities performed by the customer and as such, are not viable metrics for measuring the performance of the outsourced SOC.

Establishing a connection between the indicator and the critical success factors (CSFs) is a fundamental requirement for KPIs as described in the book by Parmenter (2019). On a practical level, establishing a connection between the CSFs and the KPIs enforces organizations to identify the factors that are critical contributors to organizational performance and thus qualify to be measured with the KPIs. The same logic can also be applied to performance indicators and other metrics in a way that if they are tied to critical or non-critical success factors, the metrics are tied to something concrete that contributes to the performance of the organization, and as such, ensures that the metrics are meaningful and relevant as per the PRAGMATIC methodology by Brotby and Hinson (2013).

The metrics should also be justifiable by scientific research or industry standards, or through other means documented in the metric description. If the metrics are justifiable, there is no need to rely on the authority of the source as it is possible to evaluate the credibility of the justification and thus decide whether the metric is reputable or not, which in turn decreases the bias and the subjectivity of the metric, and as a result, increases the correctness of the metric as per the quality criteria defined by Savola (2013).

And finally, the metric should for the most part conform to the quality criteria as defined in section 7.1, as based on empirical experience, the quality of the existing metrics is not sufficient to measure the performance of a SOC and as such, any metrics created with the selection framework should be of high quality. It is expected that no metric can completely fulfil the quality criteria in all situations and within different types of SOCs. For example, some metrics are unlikely to be applicable for both outsourced and in-house SOCs as with outsourced SOCs the scope of the monitoring could be restricted or the service does not include custom monitoring rule development.

If the metric has passed the requirements defined for the metric, the metric

can be constructed. The metric can have different characteristics depending on multiple variables, such as which SOC function it relates to or who are the stakeholders the metric is targeted to. The fundamental characteristics that should always be defined are the following:

1. Target audience and intended use
2. Measures to reduce bias and subjectivity
3. Required additional contextual information
4. Requirement measurements
5. Format for the presentation of the metric
6. Interpretation instructions

The target audience of the metrics and the intended use must be well defined because as pointed out by Kokulu et al. (2019), there is a mismatch between the evaluation metrics when it comes to SOC managers, the SOC analysts, and other technical personnel. In practice, this means the expected audience for the metrics must be carefully considered to prevent having non-actionable metrics to measure the performance of a specific team, which is against the PRAGMATIC methodology as defined by Brotby and Hinson (2013) and as pointed out by Hauser and Katz (1998), it could cause the teams to prioritize unfavourable short-term decisions to improve the metrics.

Unbiasedness and objectiveness are a major part of correctness and as such, if we can reduce bias to a minimum and ensure the metric is objective, we can fulfil the most important quality criteria as described by Savola (2013), which is correctness. A metric that would be completely objective and unbiased would be an unrealistic target, but by considering ways to reduce bias and subjectivity, and documenting the results, it is possible to improve the quality of the metric. It is also possible to identify if the metric is too biased or objective to not meet the quality criteria even though bias and subjectivity have been reduced to a minimum and thus the metric would not be genuine, as per the PRAGMATIC methodology by Brotby and Hinson (2013).

There can also be situations in which a counter-metric or other additional contextual information is needed to provide a better picture of the overall situation. For example, a SOC could have a metric to measure the number of distinct monitoring rules and the metric could then be used to measure the detection potential of the SOC. However, the high number of monitoring rules does not directly correlate with the performance of a SOC, since if a majority of the security incidents resulting from the monitoring rules are false-positive, the SOC is unlikely to be able to handle them effectively. So to measure the effective detection potential, the metric for the number of monitoring rules should be coupled with the false-positive rate to form a better picture of the expected potential.

As the metrics consist of measurements as depicted in figure 4, the source for the measurement data must be defined along with the format of measurement, the measurement interval, and any other information that affects the measurements or metric in any way. For example, a metric measuring the mean time to resolution would require each security incident to have two measurements, one

to measure the time when the incident was opened and another one to measure the time when the incident has been resolved.

Metrics must also be presented in a way that they clearly and consistently depict the information the metric is supposed to deliver. The metric can, for example, be presented either numerically, by visualization methods, such as various charts or time series graphs, or in a text format within a table. The way the metric is presented should provide the person interpreting the metric with the necessary information to make decisions based on the data seen.

Interpreting the metrics is an important factor to ensure the metrics provide valuable insights for the stakeholders. Although the fundamental idea is that the metrics themselves should be presented in a way they are self-explanatory, in the practical sense some metrics can be hard to interpret regardless. The metric documentation should include the expected way to interpret the results to ensure the metrics are not misinterpreted by the expected audience.

7.3 Model evaluation

According to the guidelines defined by Hevner et al. (2004), the evaluation of the artifact can either be observational, analytical, experimental, testing, or descriptive. As per the definition within the study, the observational evaluation method consists of either a case or field study, where the artifact is studied or monitored in a business environment. They also determined that the analytical evaluation contains static analysis, architecture analysis, optimization, and dynamic analysis, and aims to evaluate the artifact through the examination of qualities or properties. Hevner et al. further stated that the experimental methodology consists of controlled experiments and simulation, in which the artifact is evaluated in terms of qualities within a controlled environment or executed with simulated data. They also defined the testing as an activity that consists of either black- or white-box testing, where the artifact is evaluated either by inspecting its interfaces or testing internal functions within the artifact. The final method is descriptive, which can either be an informed argument or a scenario, in which a convincing argument is provided to justify the artifact or the artifact is demonstrated within a detailed scenario (Hevner et al., 2004).

The method for the evaluation of the model is a combination of descriptive and experimental methodologies. The descriptive methodology is used to create a scenario that utilizes the artifact documented in the section 7.2 to create three to seven metrics, which are documented within the section 7.4. As a result of the creation of the metrics, the utility of the artifact will be demonstrated. The scenario that demonstrated the artifact is defined as: "It must be possible to create metrics based on the metric selection framework in a way that the metric can be visually demonstrated by utilizing simulated data constructed from measurements, that could be realistically collected as a part of the day-to-day activities of a SOC following the concepts defined in the chapter 2". The scenario is consid-

ered to be valid if the metrics can be produced as per the scenario description, but in addition to the capability to create the metrics, the metrics must also be validated. The validation is to ensure the artifact can produce meaningful metrics that can be utilized in a real-world situation. The validation of the metrics is further discussed in the section 7.5, in which the experimental methodology is utilized to conduct a controlled experiment to validate the metrics.

7.4 Metrics for Security Operations Center

This section contains the metrics created by the metric selection framework described in the section 7.2. The content in the following sections describes how the metrics are meeting the objectives defined in the metric selection framework, as displayed in figure 6. The section forms a coherent description of the metric with a default assumption that the quality criteria are met by all of the metrics presented in the sections, and as a result, the exact descriptions of the quality criteria are omitted from the descriptions to avoid unnecessary repetition of the content in table 4. However, if there is any potential for non-conformities of the quality criteria under certain circumstances, those will be mentioned separately within the corresponding section.

The measurements used to construct the metrics in this chapter have been programmatically generated and presented by utilizing open-source tools and python modules, most prominently Jupyter Notebooks[6] and a python graphing library Plotly[7]. Additionally, scikit-learn, which is a python module used for data analysis and machine learning algorithm development (Pedregosa et al., 2011), is used for data manipulation and other transformations. The parameters for the creation of the data points are adjusted every 100 steps to create variation in the results over time, which aims for a better representation of the value evolution over time and thus, simulates how the metric behaves with changing conditions.

7.4.1 Distribution of detections among the Unified Kill Chain

A metric that depicts the distribution of detections among the UKC measures how effective is the SOC in detecting threats in the early stages of the UKC, and thus decreases the impact of a security incident. The goal of the metric is to focus the efforts of the development of threat detection capabilities on adversary techniques used within the initial foothold stage of the UKC, and as a result, prevent network propagation within the environment. The metric is tied to a function responsible for custom analytics and detection creation as per the definition by Knerler et al. (2022). If the SOC does not have a team or a group of people who are responsible for the custom analytics and detection creation function, the metric

[6] https://jupyter.org/about
[7] https://github.com/plotly/plotly.py

is not measuring the performance of a SOC but rather the tools the SOC is using. Therefore, without custom analytics and detection creation function, the metric is dependent on third parties and thus is not a valid metric for a SOC without the function present.

While the critical success factors as described by Parmenter (2019) are almost always specific to an organization, a broader approach can be taken with the definition of success factors as they can be derived directly from the expectations of the activities performed by the SOC. One of the mission statements of a SOC as depicted by Knerler et al. (2022) is to utilize proactive measures to prevent security incidents from materializing or becoming widespread and thus the capability to do so can be seen as a success factor for a SOC. Furthermore, the metric can be justified by the decreased cost of recovering from a security incident if the advancements of the adversary can be stopped in the earlier stages of the UKC, meaning the probability of the adversary being able to exfiltrate sensitive information or perform destructive actions has decreased as a result. The scientific research on the impact of early detections on the cost of security incidents is insufficient to draw a scientific conclusion, but a commercially produced cost of data breach report by IBM (2022) concluded that between March 2021 and March 2022, the average cost of a security incident exceeding two hundred days before identification was approximately 30% higher than a security incident with identification under two hundred days. However, adversaries can traverse across the UKC in mere minutes if the circumstances are correct and as such, an adversary that has been present within the environment for tens of days is likely to have a significant foothold already in place.

In terms of the quality criteria, the correctness criteria are achieved if all true-positive detections across different detection technologies and other non-technical detection sources, for example, reports from end-users, are augmented with a piece of information about which step of the UKC the detection is related to. If certain detections are left out of the measurements used to construct the metric, the metric is no longer complete and as a result, it can also be heavily biased. For example, leaving out the previously mentioned reports from end-users, which typically are related to the final stages of the UKC, will increase the bias and thus makes the metric difficult to evaluate. In terms of measurability, the criteria are fulfiled if the SOC processes enforce the definition of the UKC stage for each detection and security incident handled and the information automatically contributes to real-time service reporting.

When it comes to meaningfulness, the metric conforms to all of the characteristics. There could be some concerns about the comparability between different SOCs, as depending on the scope of the monitoring, the detection focus could be placed on different parts of the kill chain. For example, a SOC that focuses on monitoring endpoints and identities is likely to have a different distribution of detections, compared to a SOC that focuses on infrastructure and networks. While fundamentally different, both detection strategies could be effective in stopping adversaries but in terms of the metric, the one that catches the adversaries earlier could be seen as the better strategy, since the goal of the metric is to push the development of detection capabilities in the earlier stages of the UKC. Us-

ability could be slightly problematic for the metric, as the behaviour is likely to be inconsistent with an extremely low volume of detections but once the amount of detection increases the scalability should start to normalize. In practice, this means the time frame of the metric must be long enough to normalize the results. The target audience of the metric is the detection engineers and the SOC management, as the metric can be used to steer the development efforts in addition to being able to measure the effectiveness of the SOC, and as such, is a relevant metric for management as well.

The metric consists of the categorization of the UKC phase stored in the statistics collected from true-positive security incidents and detections. The data is not recorded for the reconnaissance or weaponization phases, since they can be difficult or impossible to prevent, and as such, stopping an attack within the weaponization phase is unlikely to happen. Before the presentation, the data must be normalized by utilizing the MinMaxScaler[8] utility class, which normalizes the dataset and ensures the data values are between zero and one. Normalization makes the metric easier to be compared between SOCs as the absolute values of detections do not influence the metric values. Additionally, to measure the evolution of the metric over time, the coefficient must be calculated regularly.

Once the dataset has been normalized, the metric can be visualized as two charts. A bar chart containing the distribution of normalized detections across the UKC and a line chart depicting the evolution of the metric over time. Within the bar chart, there is an additional visual representation that contains a linear regression trend line, which depicts the actual value of the metric in a form of a linear regression coefficient, which is based on the count and the distribution of detections among the UKC. If the coefficient of the linear regression is negative, the metric is considered to show a positive result, as it indicates that the number of detections is decreasing as the UKC is traversed forward, meaning the SOC is more effective in detecting the security incidents in the earlier stages of the UKC. The value of the coefficient can be followed over time to determine whether SOC is improving or not. If the coefficient trend is decreasing, SOC is improving in terms of the metric, meaning the new detections are starting to shift more towards the left and thus detections occur in the earlier stages of the UKC. As all of the necessary data is presented in the metric, the metric can survive on its own and does not require any counter-metric to properly interpret the results.

Figure 7 depicts a visual representation of the metric with a detection strategy focused on the initial foothold stage, meaning that a larger portion of the detections is within the initial foothold stage, compared to network propagation or the action on objectives stages. Figure 8 depicts the metric with a detection distribution focused on the network propagation. Both strategies have statistically a similar amount of detections in the final phase of the kill chain and thus both are as effective when it comes to preventing the adversaries from reaching their objectives, but the strategy with a focus on the initial foothold stage is better in terms of the metric, as it is more effective at preventing the incidents from traversing forward within the UKC.

[8] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

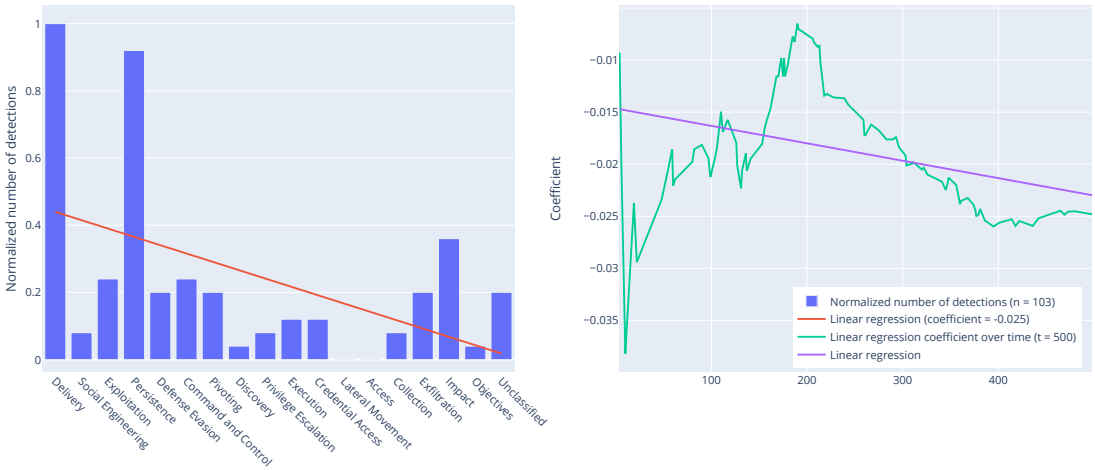

FIGURE 7 Initial Foothold focused detection strategy

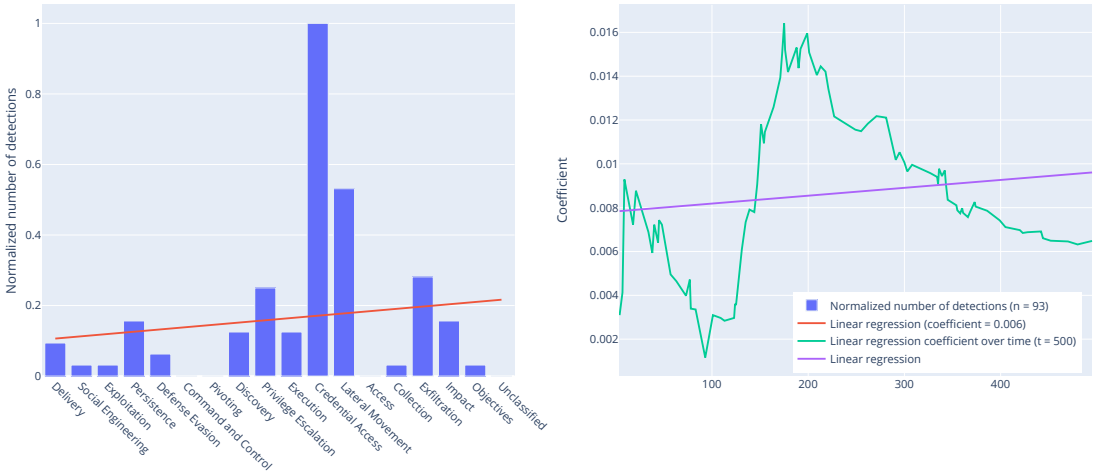

FIGURE 8 Network Propagation focused detection strategy

Figure 7 also shows that the value of the linear regression coefficient is decreasing over time, meaning the detection strategy is showing signs of improvement, whereas in figure 8 the trend is increasing, which shows signs of deterioration. Both figures are displaying a large variation at the beginning of the metric, which means the metric becomes accurate with a sufficiently large quantity of true-positive detections and as a result, the linear regression trend line alone could be ineffective at displaying the efforts of recent development activities. As the trend of the coefficient would primarily be used to measure the impact of the development efforts, the graph could be displayed only after a sufficient amount of true-positive detections have been recorded.

7.4.2 Number of verifiable monitoring rules

The metric "Number of verifiable monitoring rules" measures the portion of monitoring rules that can be verified either automatically or manually by executing actions that trigger the monitoring rules. The goal of the metric is to improve the

rate of verifiability of the monitoring rules, encourage detection engineers to better utilize threat intelligence as a part of their daily routines, and provide a mean for the SOC to demonstrate what attacks they can detect. Similar to the metric in section 7.4.1, the metric is tied to a function responsible for custom analytics and detection creation as per the definition by Knerler et al. (2022) and is valid only if the SOC has a team responsible for the creation of custom analytics.

As the purpose of a SOC is to mitigate risks and create situational awareness by utilizing technology to detect security threats affecting the organization (Vielberth et al., 2020), a fundamental success factor could be the possibility to detect relevant threats targeting the organization and as a result, reduce the risk level of the organization. The problem with an approach where the detections are not tested is the lack of visibility on the capabilities and the information on what risks can be reduced with high confidence. For example, a SOC could utilize a SIEM rule to detect password spraying against the organization and as such, the risk of password spraying is considerably reduced, as the expectation is that a SOC would be able to catch successful password spraying activities. However, without testing the rule, how can the SOC be sure the rule is working? By testing the rules by performing various methods of password spraying and observing the results of the activities, the SOC can increase their confidence in the capabilities to detect such activities in case all methods trigger an alert as expected. Furthermore, the metric can be justified by the capability requirements defined by the SOC-CMM framework version 2.2 (Van Os, 2022), which devotes an entire section to automated detection testing and adversary emulation, and to reach higher levels of maturity in the detection engineering & validation section, the SOC must be able to test the validity of their monitoring rules. Adversary emulation and the utilization of offensive techniques in the form of a red or purple team exercise are also seen as a vital part of the expanded SOC functionalities, described as a top strategy for world-class SOCs by Knerler et al. (2022).

In terms of the quality criteria, the metric is slightly problematic, as unbiasedness and comparability cannot be completely achieved, as the value of the metric is highly dependent on the strategy the SOC has taken for building up its detection capabilities. If for example, the SOC relies almost exclusively on native capabilities provided by a technology vendor, the metric will either show an abnormally low value due to not having any in-house monitoring rules or an abnormally high number due to having only a few rules that can be easily covered by a handful of testing scenarios. One possible solution to this could be the addition of testing capabilities for the native capabilities in addition to in-house capabilities, but as the capabilities of the technologies are usually not properly documented, the extent of the testing capabilities cannot be reliably measured. Another solution would be to measure the absolute number of testable detections, but the results would also be dependent on the detection strategy, as some SOCs can have a high number of relatively simple monitoring rules while some may have a low number of more complex monitoring rules.

Otherwise, the metric meets the quality criteria. The metric is granular enough to be tied to a specific SOC function and completely fulfils the goal defined for the metric. If the data is created and collected automatically as a part

of the rule development process, the measurability criteria are completely met. The metric is also impactful, as the metric can be used to direct the threat development efforts in a way that encourages the threat detection engineers to create testing cases for their monitoring rules. The output of the metric is consistent across the lifecycle as it is measuring the ratio between the monitoring rules and the testing scenarios, meaning it is not affected by external conditions and as a result, it meets the clarify and scalability characteristics of the quality criteria. The metric is usable by multiple SOCs, although the interpretation can slightly vary between the SOCs, due to the reasons mentioned earlier. The team responsible for threat detection can control the outcome of the metric, and it can be presented in terms of the current stage and historical progress, as demonstrated by figure 9.

To construct the metric, the creation of the measurements from the monitoring rule documentation or a similar source is required, as the data is not generated from the operational activities of a SOC. The creation of the measurement should happen when a new rule or test is added. The measurements required to be generated are, per MITRE ATT&CK tactic, the number of monitoring rules, the number of automated and manual tests, and the test coverage as a percentage compared against the total number of monitoring rules. Additionally, the mean value of the percentages should be calculated, which depicts the value of the metric at the given time. The tactic dimension is used to create a logical grouping capability for the monitoring rules, and at the same time, enable the metric to be combined with other metrics using the tactics as a dimension. Since not all monitoring rules are necessarily tied to any particular tactic, an unclassified category must be added among the tactics to take such situations into account in the metric.

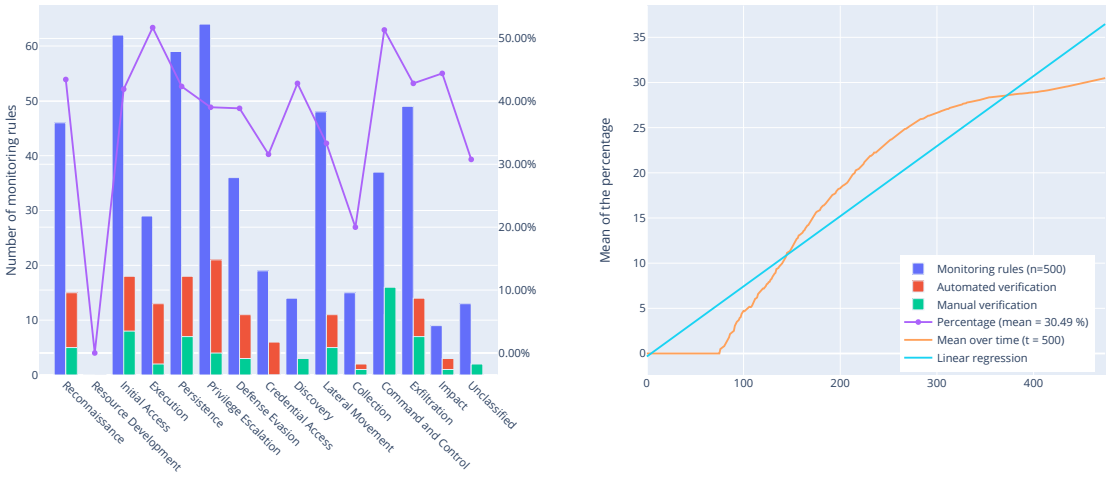

FIGURE 9 Number of verifiable monitoring rules

Figure 9 depicts two ways to present the metric. One as a combined bar and line chart and another one as a line chart, depicting the evolution of the metric over time. The bar chart displays, per MITRE ATT&CK tactic, the number of distinct monitoring rules and the sum of verifiable monitoring rules. Next to the bars, there is a trend line that displays the percentage of verifiable monitoring rules. The value of the metric is the mean of the percentage of the monitoring rules

covered per MITRE ATT&CK tactic at a given time. The intended way to interpret the metric is to follow the trend of the mean to determine whether the SOC is improving over time or not, by observing the trend line seen in the secondary chart.

For the metric to provide meaningful information, it needs a counter-metric that displays the number of distinct monitoring rules, seen as the blue bars in figure 9. If the SOC has a low number of monitoring rules, the metric value can be abnormally high, which will increase bias and decrease the comparability of results between different SOCs. Although, as previously mentioned, the number of monitoring rules can also be misleading, as there is a possibility to create a large number of low-fidelity monitoring rules that can be tested automatically, for example, the launch of a specific process, like powershell.exe, could generate a low severity detection that would subsequently be automatically closed as a benign true-positive detection. Without a unified approach for the creation of monitoring rules between SOCs, a way to reduce the bias of the metric to an acceptable level was not discovered.

Additionally, as the metric is measuring only the number of verifiable monitoring rules resulting from the SOCs development efforts, it fails to demonstrate the verification capabilities of native vendor detections. This makes the metric subjective, as some SOCs are likely to rely more on native capabilities for threat detection than others, making the metric inaccurate when attempting to compare SOCs with a different approach to the threat detection. A dimension that displays the number of vendor-native detections that can be tested could be added to the metric, but in doing so, the bias of the metric would continue to increase. A better solution would be to create a separate metric for measuring the number of verifiable vendor-native scenarios and construct another metric that takes both metrics into account.

To summarize, the metric alone is situational but when used as a part of a larger collection of metrics, it can provide additional insights to measure the overall quality of the monitoring rules. However, bias cannot be reduced enough to make this metric valid as per the metric selection framework. This does not mean the metric cannot be used by SOCs to measure and report their performance and progress, but rather that the metric is not valid between multiple SOCs, and as such, it cannot be used to mitigate the problems mentioned in chapter 1 as a motivation for this research. The audience of the metric can be both technical SOC personnel as well as the management of a SOC, as the value of the metric can be used to measure the progress of development in a way that supports both the technical teams and the management.

7.4.3 Distribution of detections by source

The goal of the metric is to determine to what extent the development efforts of the SOC can contribute to the detection of security incidents. On a practical level, if a large portion of detections originates from the native capabilities of the technologies in use, the detection engineering function may not be able to provide

additional value in the form of new monitoring rules. The metric can be tied to a specific function, which is the function responsible for the creation of custom analytics within the SOC. The results of the metric are somewhat dependent on third parties, as the technologies selected to protect the environment have an impact on the metric results. However, as the fundamental purpose is to compare the custom capabilities against the native capabilities, the metric is not dependent on third parties but rather influenced by them, which makes the metric pass the requirement.

As per the definition by Knerler et al. (2022), custom analytics and custom capability development are the functional areas of a SOC and as such, it could be stated that the capability to augment the detection capabilities provided natively by the technologies is a success factor for a SOC. On an overall level, the SOC-related literature does not succeed well in the definition of why custom analytics should be created in the first place. For example, Ahlm (2021), Knerler et al. (2022), Van Os (2022), and Vielberth et al. (2020) among others mention the creation of monitoring rules as a fundamental part of the SOC capabilities, but none are discussing in detail whether the creation of monitoring rules is something the SOC should focus on or not. However, as the literature appears to agree that the creation of monitoring rules is something SOCs should be doing, it acts as a justification for the metric. Additionally, the metric can also be justified by the fact that it can display whether it makes sense to invest in the development of custom capabilities or not.

Looking at the quality criteria in terms of correctness, the metric can be slightly problematic, as the metric can become biased if a SOC starts to replace native capabilities with a custom detection logic, and thus decreases the native detection capabilities. For example, the SOC can create a monitoring rule that raises an alert from native alerts produced by an anti-virus program only if a certain threshold exceeds, which blurs the line between native and custom capabilities. Technically the actual source of the detection is the anti-virus program, but the detection is raised as a result of custom capabilities. Additionally, whether the malware detection by an anti-virus program counts as a detection within the context of SOC capabilities or not, can also be subjective. The measurability criteria are achieved for as long as the data is collected from all true-positive and benign true-positive detections.

In terms of meaningfulness, the metric has the potential to impact the development efforts, but without additional measures to decrease the subjectivity and bias, the metric value can be difficult to compare between multiple SOCs, especially when differences in the tooling and the detection strategy can also have an impact on the metric. However, given the goal of the metric is to measure how well the SOC can augment the native capabilities, the problems related to the comparability could be considered to be a minor issue and as such, the small amount of bias and subjectivity does not make the metric invalid. There could also be slight problems in terms of clarity, as the metric can be hard to interpret, as it requires additional context about the detection strategy to properly interpret. In terms of usability, there are no difficulties to achieve any of the characteristics defined in the quality criteria, as the metric can be used as such by multiple SOCs,

the progression of the metric can be controlled by the threat detection function, the metric behaves consistently as it is measuring the ratio between native and custom detections capabilities of the SOC.

With MITRE ATT&CK tactic as the primary dimension for the metric, as seen in figure 10, the metric provides a method for the detection engineering function to align and focus their development efforts on specific tactics, which makes the metric relevant for them to help with the planning of the development activities. In addition to the detection engineering function, the SOC management can also benefit from the metric, as the value of the metric over time can be used to determine the direction and the impact of changes made within the detection engineering team and subsequently demonstrate the value the detection engineering function provides.

To decrease the subjectivity and bias of the metric, it is necessary to set certain limitations to the metric. Within the context of this metric, detections originating from native capabilities are something that technology, for example, SIEM, EDR, or an IDS, has provided the first indication of compromise with out-of-the-box capabilities and without correlation to other data sources. Meaning, if the source of the detection is an alert from an IDS or anti-virus program, it constitutes a native detection. However, if the anti-virus alert is correlated with other endpoint-related logs before generating a detection, it is constituted as a custom capability. Additionally, all detections that have been contained automatically and do not require further actions from the SOC, for example, an anti-virus program or an IDS preventing an infection or the delivery of a malicious payload, are out-of-scope for this metric. Furthermore, all false-positive detections and detections not originating either from a technology or a monitoring rule, such as reports from end-users, are also removed from the measurements of the metric, as they are unnecessary for achieving the goal of the metric. Some SOCs could also have a detection strategy in which they focus their development efforts on the detection of complex and seldomly occurring scenarios and otherwise rely on the native capabilities. In such a situation, the metric could be considered to not apply to them.

To construct the metric, the source for the detection in addition to MITRE ATT&CK categorization must be recorded for all true-positive and benign true-positive detections. One way to present the metric is seen in figure 10, which displays two charts. A bar chart displaying, for each MITRE ATT&CK tactic and an unclassified category for detections not mapped to MITRE ATT&CK, the current state of the metric in terms of the number of detections per category, and the ratio between native and custom detections. Additionally, the metric displays a line graph on a secondary y-axis depicting the score for each tactic and the overall value of the metric, which is the mean of the scores for each category. The line graph displays the evolution of the value of the metric over time along with a linear regression depicting the trend of the metric. The metric can be presented as such, as it does not require a counter-metric or any other metric to be properly interpreted.

The value of the metric is above one if native detections constitute a higher proportion of overall detections compared to custom detections, which means

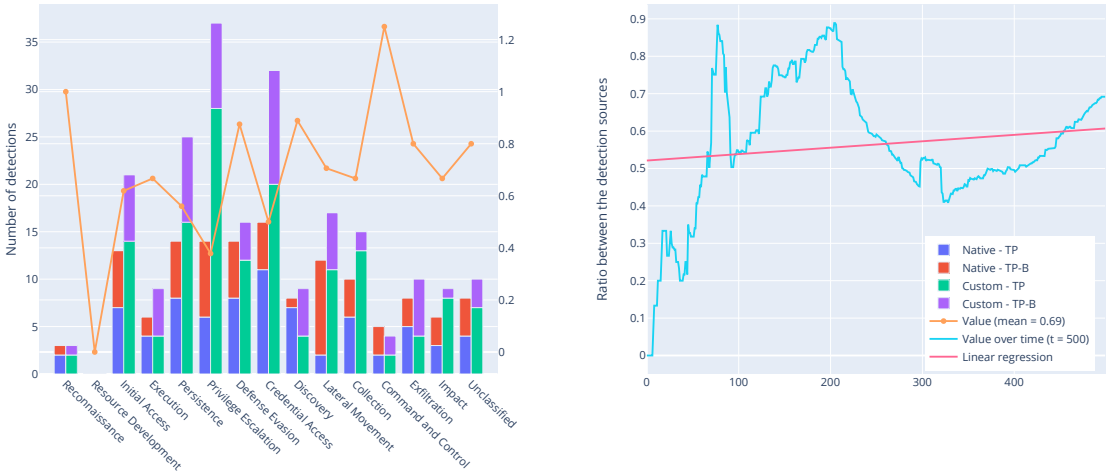

FIGURE 10 Distribution of detections per source

the closer the value is to zero, the better the value can be considered to be. However, in a practical sense, due to variations in detection strategies between SOCs, a value within the range of 0.80-0.20 can be considered to be a good value. If the value drops too close to zero, there is a possibility the SOC has not deployed the technologies properly, they are inefficient for threat detection, or are otherwise utilizing the technologies incorrectly. Furthermore, if the value closes to or exceeds one, the SOC is not capable to outperform the capabilities provided natively by the technologies.

7.4.4 Technical accuracy of the analysis

Measuring the quality of the analysis work (Agyepong et al., 2020; Zimmerman & Crowley, 2019) or eradication (Crowley & Pescatore, 2019; Zimmerman & Crowley, 2019) has been mentioned in the literature but the ways to measure the quality of the work has not been covered in detail. Measuring the quality of the analysis work is subjective and it is unlikely to be possible with a general-purpose metric, but measuring the accuracy of the analysis could provide some hints about the quality of the analysis work. A metric that measures the accuracy of the analysis has a goal to enable the capability to measure the technical quality of the analysis work and assign a quantifiable value on how well the analysts are performing on a technical level.

The metric is a collection of several factors that contribute to the quality of the analysis work. To calculate a value for the metric, an approach similar to the calculation of the net-promoter score (NPS) (Reichheld, 2004), in which customers are scoring a service on a range from 0 to 10, can be adopted. Based on the approach, scores 0-6 are detractors, 7-8 are passive and 9-10 are promoters, and the NPS is calculated by subtracting the percentage of detractors from the percentage of promoters, providing a score between -100 and 100. The formula for the calculation of the NPS is $NPS = (\frac{P_1 - D}{P_1 + P_2 + D}) * 100$ where P_1 are number of promoters, P_2 are the number of passives and D are the number of detractors. Although

in the academic sense, the NPS methodology has some issues for what is being used for (Bendle, Bagga, & Nastasoiu, 2019), but regardless of its limitations, the model succeeds in producing a value of the relationship of discouraged (detractors), neutral (passives), and encouraged (promoters) activities, and as such, it is a valid and relatively simple approach to take. Additionally, the modular approach enables organizations to extend the metric to include additional activities within the metric if required, based on the unique requirements of the organization.

Table 5 summarizes the activities included in the scope of the metric. Promoters are activities that should be encouraged to be performed continuously and are signs of a well-performing SOC, passives are activities that are expected from the SOC under normal operations and detractors are activities that the SOC should attempt to avoid, as they can harm the overall situation. For the sake of simplicity and to stay within the defined limits of this research, the number of activities is limited only to a few technical activities, but a practical implication of this metric can include additional technical and operational activities in the list of activities, such as whether detection was investigated according to time constraints (passive) or not (detractor), MITRE ATT&CK categorization was correct (passive) or incorrect (detractor) or an adversary behind a true-positive incident was attributed (promoter) or it remained unknown (passive).

TABLE 5 Grouping of activities per NPS category

Category	**Activity**
Promoters	True-positive incident was escalated to third-party. Escalated incident was not returned to SOC for further investigation. Original priority was correct throughout the incident lifecycle. No unknown entities before escalation.
Passives	Benign true-positive incident was escalated. Escalated incident was returned to SOC for further investigation. Priority of the security incident was adjusted after the initial analysis. Unknown entities before escalation. The initial conclusion on the returned incident was correct.
Detractors	False-positive incident was escalated. False-negative detection. The initial conclusion on a returned incident was incorrect.

In terms of the metric requirements, the owner of the metric is the management of the SOC, as the metric is constructed from several factors resulting from the analysis work. The activities themselves are something that can be tied to a specific function within SOC. As the data is collected from the activities performed by the SOC, the results are not dependent on third parties. The metric can be justified by the idea that if the analysis is of low quality or the quality keeps decreasing over time, the SOC might not be able to combat the challenges produced by a modern-day adversary or they might not have sufficient knowledge of the monitored environment. The same idea could be considered to be a success factor for a SOC, meaning that a SOC has to operate at a high quality to succeed in the modern threat landscape.

Looking at the quality criteria, as long as the activities selected for the metric are the same between SOCs, there should not be any major problems to con-

form with the quality criteria. As the metric could be considered to be a composite metric, the activities are the actual metric items that should conform to the quality criteria, and if the chosen activities are aligned with the quality criteria, the metric itself will also conform to the quality criteria. Portability and comparability could be slightly problematic if additional activities are added to the metric, but as it has been defined in table 5, the metric is portable and comparable between SOCs, as the activities are generic enough to not be influenced by the variations between different SOCs. The distribution of the activities can be slightly subjective, as there could, for example, be a difference between SOCs on how benign true-positive incidents are seen. Some organizations could consider them as a normal day-to-day activity and thus be classified within the passive category and some organizations could see them as something that should not happen often and as such, could be seen as a detractor.

The metric is more intended for the management of the SOC as they are likely to be more interested in the overall situation rather than focus on specific metrics. However, if the individual metric items are also constructed as a part of this composite metric, they are likely to be something that can be directly tied to a specific team and thus, be relevant for the individual teams as well. For example, the team performing the analysis work could be more interested in the capability to resolve all entities and the correctness of initial conclusions, whereas the detection engineering team could be interested to keep track of the true-positive rate and the correctness of the original incident priority. Reducing the bias and subjectivity can be difficult if the activities chosen as a basis for the metric are biased and subjective, to begin with. Adding classification criteria for the activities could help with the subjectivity of the activity grouping. However, to stay within the scope of the thesis, the creation of such criteria will not be discussed in detail. If the metric is deployed to production, there should be some degree of criteria used to classify the activities.

The measurements required to construct the metric vary depending on the activities chosen for the metric. When utilizing the activities shown in table 5, information about the security incident classification (true-positive, benign true-positive or false-positive), original priority, final priority, escalation status, number of unknown entities, the outcome of the initial conclusion and whether the security incident had to be re-investigated by SOC has to be recorded for each security incident. Additionally, any incidents not detected by the SOC, for example, the ones reported by end-users, must also be recorded, as they are known false-negative detections. A single security incident can contain one or more activities contributing to the metric.

The metric can be presented similarly to the other metrics, as seen in figure 11. Within the figure, there is a bar chart depicting the count of activity occurrences for each category (promoters, passives, detractors) per MITRE ATT&CK tactic. If there are zero items within the MITRE ATT&CK tactic, the tactic is omitted from the visualization. In addition to the count of activity occurrences, the NPS is displayed individually for each MITRE ATT&CK tactic to demonstrate the difference in the technical accuracy of the analysis between tactics. Due to the way the NPS is calculated, the actual value of the metric must be calculated from

all occurrences of the activities, rather than taking the average of the individual NPS. The secondary graph in figure 11 depicts the evolution of the NPS over time and a linear regression that demonstrates the trend of the evolution of the metric. If the NPS is above 0, it means there are more promoters than detractors, and as such, the higher the score, the better the value of the metric is.

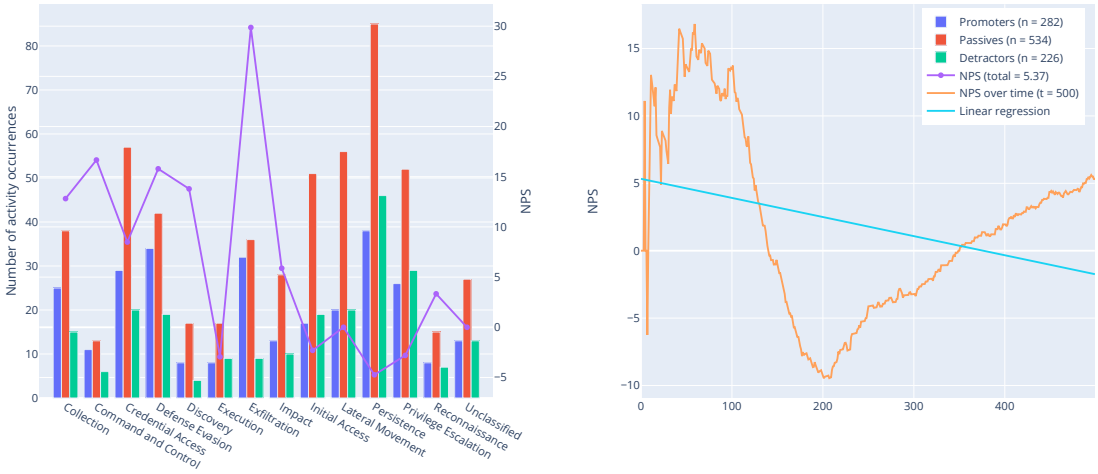

FIGURE 11 Detection accuracy NPS

The metric works as a standalone metric and does not require additional contextual information to be utilized. The value of the metric alone cannot be used to compare different SOCs as the value is affected by the selection of the activities. Additionally, even though SOC A and SOC B would have the same number of promoters and detractors, due to the way the NPS is calculated, the number of passives will impact the results. Despite the minor limitations and slight subjectivity, it is sufficient to meet the solution objectives and as a result, the metric can be considered to be a valid metric. An alternative form of the metric could be to ignore the passives altogether and only calculate the ratio between promoters and detractors, but by doing so, the activities that the SOC is expected to perform would have no impact on the metric value and as a result, the value of the metric would be heavily influenced by the extremes of either category, as the smoothing factor of the passive would be missing from the calculation.

7.4.5 Accuracy of automated containment

Automated containment could be used as one way to decrease the time it takes to contain a security incident and thus positively impact the metric that measures the mean time to contain a security incident, mentioned by Crowley and Pescatore (2019), Kokulu et al. (2019), and Zimmerman and Crowley (2019). If the mean time to containment could be pushed down to zero, the practical implication would be that none of the true-positive security incidents would be able to traverse down the UKC and reach the action on objectives stage, assuming that the detection would happen in the earlier stages of the UKC. Pushing the mean time to containment to zero is unlikely to be possible, but pushing the met-

ric value as low as possible would be a realistic objective for a SOC to pursue, and thus, the goal of the metric is to push down the value of the mean time to containment metric.

The metric can be justified by the fact that containment activity is the first post-detection activity within the incident response life cycle as described in the Computer Security Incident Handling Guide (NIST SP 800-61) by Cichonski et al. (2012). The purpose of the containment is to limit the impact of the incident, stop it from spreading further or prevent it from causing any further damage (Cichonski et al., 2012). Containment in the traditional sense can be disconnecting the device from the network or shutting it down (Cichonski et al., 2012) but modern tooling can offer more advanced capabilities, such as isolating an infected endpoint, sinkholing command and control traffic or revocation of compromised credentials. For example, by utilizing a SOAR platform, it is possible to automate such a response to security incidents (Knerler et al., 2022) and as a result, to automate the activities, the SOC needs to have a sufficient level of maturity to do so. Furthermore, automation capabilities are listed as aspects within the technology domain of the SOC-CMM maturity-capability model (Van Os, 2022), which further justifies the metric.

As has been previously mentioned, the capability to prevent security incidents from materializing or becoming widespread by utilizing proactive methodologies is one of the mission statements of SOC, as described by Knerler et al. (2022). As such, the capability to prevent the escalation of incidents can be defined to be a success factor for a SOC. As the metric is about an automated containment when a detection contributing to a true-positive security incident is triggered, the metric can be tied to the team responsible for the detection engineering and other automation activities. Furthermore, the metric is not directly dependent on third parties, but certain types of SOCs could have difficulties performing automated containment, for example, outsourced SOCs without jurisdiction over the incident response process within the customer environment.

The correctness and measurability criteria are achieved if the outcomes of all the security incidents independent of the source are recorded and used as a source when constructing the metric. By utilizing a confusion matrix to construct the metric and calculating the F-score as the value of the metric, the objectivity of the metric can be significantly reduced, as the F-score can be used to quantify the ratio between correct and incorrect activities. The correct activity is a containment when an incident is true-positive and not containing when it is anything else, and incorrect activity is a failure to contain when an incident is true-positive and containing when it is anything else. A false-negative detection is always incorrect, as, without detection, there is no possibility to perform automated containment.

However, since SOCs could have different response capabilities, the metric can be slightly biased if the theoretical response capabilities are not identical between the SOCs. In terms of meaningfulness, the metric is impactful as it has an impact on the daily activities of the detection engineering team by providing them with data about the effectiveness of the response automation. Additionally, the metric is expected to behave consistently and be clear to interpret after a few detections have been recorded, but as previously mentioned, there could be

some issues with comparability as the level of automation capabilities can vary between SOCs. Looking at the usability, the metric can be adapted by all SOCs, and the detection engineering team has control over the metric, for as long as they can perform automated responses.

To avoid confusion between the classification for the confusion matrix and the typical classification schema for security incidents, the security incident classification must be mapped to the confusion matrix classifications, based on the information on whether a containment was done or not, since it is not a one-to-one mapping between the classifications. The categorization has been summarized in table 6. Security incidents cannot be true-negative, since detections are only raised based on monitoring rules that can lead to either true-positive, false-positive, or benign true-positive detection. Within the confusion matrix, benign true-positive and false-positive incidents are classified as true-negative if containment has not been performed, meaning it has been correctly identified to not be a true-positive incident and a correct containment activity (no containment) was selected.

TABLE 6 Confusion matrix compared to security incidents and state of containment

Confusion matrix classification	**Security incident classification**	**Containment state**
TP	True-positive	Yes
FP	Benign true-positive	Yes
FP	False-positive	Yes
FN	True-positive	No
FN	False-negative	No
TN	Benign true-positive	No
TN	False-positive	No

Figure 12 depicts one possible way to visualize the metric with a confusion matrix on the left side and on the right side, a line graph depicting the evolution of the normalized F-Score over time along with a linear regression depicting the trend of the metric. The values used to construct the confusion matrix and calculate the F-score have been normalized with the MinMaxScaler[9] utility class to remove the detection volumes from the metric, which are not necessary for the interpretation of the metric, and thus making the metric easier to compare between different SOCs.

A study by Goutte and Gaussier (2005) defines the formula for calculating the F-score as $F_1 = \frac{2PR}{P+R}$. They also define that the value produced is the harmonic mean of precision (P) and recall (R), which means the value of the F-score is between zero and one. Precision is the ratio between true-positives and all identified elements ($P = \frac{TP}{TP+FP}$) and recall is the ratio of true-positives and all relevant elements ($R = \frac{TP}{TP+FN}$) (Goutte & Gaussier, 2005). A score of one indicates that both precision and recall are perfect and thus, the higher the F-score is, the better the value of the metric is.

To construct the metric, a measurement containing the security incident classification and a boolean value whether an automated containment was done or not, must be recorded for every security incident independent of the source of

[9] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

the detection. The metric works as such and does not require additional contextual information to be properly interpreted, although if a SOC cannot automate the containment, the metric value will always be zero and as such, is not applicable for such SOCs. Interpretation of the metric is relatively simple, the higher the F-score is, the better the accuracy of the automated containment within the SOC is. This means there are fewer unnecessary endpoint isolations and password resets resulting from the automated containment activities while being able to perform containment when the activity is relevant. Depending on the situation and the monitored environment, it might be viable to try to reach a high recall value, which means that a higher number of relevant incidents are contained, but among the contained incidents, there are also a lot of incidents that should have not been contained.

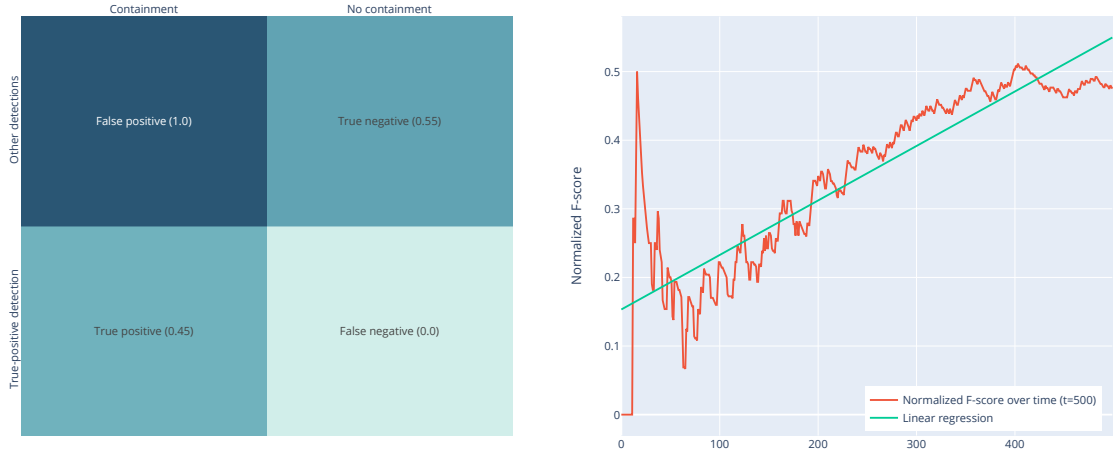

FIGURE 12 Automated containment of true-positive security incidents

Although the metric remains subjective to some degree, it can be considered to be valid from the selection framework point of view. If the metric is used in production, the metric could be coupled with additional visualization to display the F-Score per MITRE ATT&CK tactic and containment activity, for example, endpoint isolation or password reset, to be able to better understand which containment activities upon which MITRE ATT&CK tactics are performing better or worse than the others.

7.4.6 Other considered metrics

In addition to the metrics described above, this section briefly discusses some of the more significant metrics that have been considered to be selected but ended up being left out of the scope due to multiple different reasons. The metrics described below are valid on a theoretical level and could be utilized by SOCs, even though they were not included in the previous sections of this research.

Time to detect an incident is one of the more commonly mentioned metrics in the literature, and as was previously mentioned, there are also several interpretations for the metric. A metric that monitors the time between the initial

detection and the first activity by the adversary could have a been good metric to be added to the list of metrics. For example, if an adversary first opens a remote desktop connection at three o'clock, but is first detected at half past three as the adversary performs another activity, it would mean there would be a 30-minute time window between the first activity and the detection, meaning the time to detect an incident is 30 minutes. On a practical level, this metric would have become too biased based on the selection framework, as it would require SOCs to perform an extensive investigation of all true-positive incidents to determine the true first activity rather than just containing and eradicating the adversary during the investigation response process. Although it would be a good practice to perform a deep analysis on all true-positive incidents, in practice, this may not be performed if there is no evidence to indicate the original entry point for the adversary was other than the infected device.

Another metric that was considered was the percentage of actionable data to measure how well the data collected is utilized by the SOC. The definition would have been that actionable data is something that directly contributes to a monitoring rule, or is used for threat hunting or incident response purposes, and any other data would be considered to be non-actionable. The problem with the metric is that defining which data is used for incident response activities is relatively difficult to determine, and as such, the metric would be highly subjective. Additionally, some organizations can utilize the same platform used for threat detection for long-term log archival for compliance and other purposes, which would also skew the metric significantly. Being able to understand the collected data and optimizing it based on how the data is used, is a fundamental component when it comes to building an effective SOC, and as such, the metric could probably be made valid within a single SOC by measuring the portion of classified data compared to unclassified data.

The relevancy of the detection capabilities compared to relevant APT actors was also considered as one metric. The fundamental idea behind the metric was that SOCs should focus the development efforts of their detection capabilities on relevant MITRE ATT&CK techniques used by adversaries that are active within the industry of the organization the SOC is monitoring, or the organization is through some other way at an elevated risk of being targeted by a select group of APT actors. The problem with the metric was that the definition of relevant APT actors ended up being difficult to define in a way that would be neutral across multiple SOCs. Additionally, it was difficult to determine the value to be measured, as looking purely at the coverage of the techniques of MITRE ATT&CK framework can be misleading, and as such, the measurements would have to be brought down to the level of procedures, which ended up being too complex to construct to stay within the intended scope of this thesis. Understanding the relevancy of the monitoring is something that would be an important factor for SOCs and as such, the metric would be relevant, but probably quite complex to implement.

Coverage of the environment was also considered as a metric, where the value of the metric would be indicating how well the environment is covered by the security monitoring capabilities. The idea was dismissed relatively quickly as

organizations can have completely different ways to architect their environments and as such, the creation of a generalized model of the environment required to construct the metric was determined to be too complex for this thesis. However, it is an important aspect for SOCs to understand the breadth of the monitoring and as such, the metric is valid but can be difficult to implement.

7.5 Model validation

To validate the metrics produced by the artifact, a controlled experiment is conducted, which falls within the experimental design science evaluation method as per the definitions by Hevner et al. (2004). Mettler, Eurich, and Winter (2014) have proposed an evaluation framework for design experiments, which introduces three layers that can be used for the evaluation process of a design science experiment. The first layer proposed within the study is the user layer, which is based on how the artifact is used or misused by the study subject. They also argued that the design science experiment should describe the user characteristics, sampling procedure, and a description of the experimental setting to evaluate the impact of the user on the results of the experiment. The second layer of the framework is the use of the artifact, which should describe how the artifact is used within the experiment. The key areas within the use layer are the usage situation, the usage scenario, and information on whether the users are manipulating the artifact during the tests (Mettler et al., 2014).

The final layer of the framework by Mettler et al. (2014) is utility, which describes the evaluation metrics and the outcome of the experiment resulting from the use of the artifact by the user. They also state that during the experiment, defining the evaluation criteria and other metrics is something that should be properly documented to ensure the results can be replicated. Furthermore, additional metrics must be defined that can mediate and moderate the effects of user-specific attributes on the results of the experiment (Mettler et al., 2014).

In terms of the user layer, the metrics are evaluated within a SOC that provides managed SOC services to large and medium enterprises and as such, the user of the artifact is not any particular user but rather the entire SOC. Since the SOC is not interacting directly with the artifact but simply acts as a source for the measurements, the user layer is not that important when it comes to the validation of the model. The user itself has an impact on the utility layer, as was described earlier, and thus it is vital to highlight the user of the artifact, as the results could vary if the test would be executed in a SOC that provides security monitoring for in-house purposes. As the validation is about the utility provided by the metric, the use layer is not applicable. Within the utility layer, the metrics are judged by an evaluation metric: "Can the measurements required to construct the metric be collected within the SOC used as the test subject?", meaning is it possible for the SOC on a theoretical level to take the metrics constructed by the metric selection framework into production use.

The demonstration of metrics with live data could also have been a valid evaluation metric for the metrics, but due to the metrics requiring historical data to provide information other than a snapshot of the current state, the evaluation would require a longer period to be properly evaluated. Therefore, within the timeframe allocated to this thesis, a long-term evaluation is not possible to be conducted for metrics that do not have a sufficient level of historical measurements to be available. The metrics have been tested with simulated data during the creation process and the expectation is that the metrics behave in a similar way when real-world measurements are used as a source for the metrics, and thus, it was determined to be sufficient to demonstrate the capabilities of the metrics. Furthermore, simulation is considered to be a viable testing methodology as determined by Hevner et al. (2004), which means the metrics were already tested with a viable testing methodology during the creation process and as such, the testing conducted focuses on validating the measurements and the feasibility of the collection of the measurements required to construct the metric.

The validation plan therefore can be defined as: "The valid metrics created with the design science artifact are tested in a SOC that provides managed SOC services to large and medium enterprises. For each metric, the required measurements are attempted to be collected from the existing SOC functions, and if the measurements can be collected or can be made available, the testing is considered to be successful."

7.5.1 Required measurements

Before the data collection can be started, it is necessary to define the measurements that are required to construct the metrics described in the section 7.4. Out of the metrics generated, the metrics that were determined to be valid were the distribution of detections among the unified kill chain, distribution of detections by source, technical accuracy of the analysis, and the number of true-positive incidents with automated containment. The metric Number of verifiable monitoring rules was considered to be invalid by the criteria defined by the metric selection framework. The required measurements are summarized in table 7.

TABLE 7 Required measurements to construct the metrics

Measurement	Type
Security incident classification	String
Detection classification	String
State of automated containment	Boolean
Original priority of the security incident	String
Final priority of the security incident	String
Information about whether a security incident has been escalated	Boolean
Information about who the security incident has been escalated to	String
Number of unknown entities related to a security incident	Integer
Information about whether the initial conclusion was correct or not	Boolean
Information about whether the security incident had to be re-investigated by SOC	Boolean
Original detection source	String
MITRE ATT&CK tactic categorization	String
Unified Kill Chain stage	String

7.5.2 Collecting measurements

The classification for detections and security incidents is available in multiple locations, as several products produce detections for the SOC to investigate. The SOC in which the validity of the measurements is being validated is utilizing a SOAR platform to collect all detections into a single pane of glass, which unifies the analyst workflows across products and as a result, it is producing a unified classification of the detections and the security incidents alike. Furthermore, the SOAR platform is responsible for performing automated activities as a response to detections, and as a result, it can produce information on whether a security incident was automatically contained or not. The original priority, which is a result of the initial analysis performed in tier 1, and any changes to the priority as the analysis is progressing are also recorded by the SOAR system and made available for reporting purposes once the detections have been closed.

As the investigation workflow is managed by utilizing the SOAR platform, information about the escalation status and destination can also be produced by the SOAR platform. The escalation status is also available in the information technology service management (ITSM) system, which is being used to manage the security incidents between the SOC and third-party stakeholders, such as the customers or their service providers.

The SOC is resolving unknown entities, such as IP addresses without a hostname association, during the tier 1 workflow, but the information about the unknown entities that remain after the analysis has been concluded is not currently stored anywhere. The known entities are extracted to SOAR and linked to the ongoing investigation, but unknown entities remain for the analyst to manually identify as a part of the analysis process. As a result, reliably being able to provide a numerical value for the count of unknown entities for every detection handled by the SOC is unlikely to be realistically achieved even if the data would be currently recorded. It could be possible to construct a workflow that reminds analysts to record the unknown entities in the case, but as there is a human reliance on the activity, the data cannot be reliably and automatically collected, which violates the quality criteria as defined in table 4.

Information about the accuracy of the initial conclusion and whether the SOC had to re-investigate the security incident or not, can be produced as a part of the workflows used to manage the security incidents, as reassigning the case back to a security analyst after escalation to third-party means the case had to be investigated again by the SOC. During the closure workflow, the accuracy of the initial analysis must be confirmed by a human, which can cause imperfect measurements if automated closure workflows are utilized or the confirmation of the accuracy of the analysis is not done properly.

Detection source and MITRE ATT&CK tactic categorization are also available in the SOAR platform, and the UKC stage can be partially derived from the tactics, but the transition between phases is not recorded as the monitoring rules are tied to the MITRE ATT&CK framework. Additionally, the mapping of the detections is only partial as some of the sources for detections do not provide

MITRE ATT&CK categorization along the detection, which results in a partially manual categorization process that leads to an unnecessarily high number of unclassified detections, that skew the metrics that heavily rely on the accuracy of MITRE ATT&CK categorization, such as the distribution of detections among the UKC.

To summarize, as the SOC is utilizing a SOAR platform to manage the security incident management process, the information about the activities associated with the various activities within SOC is already available or can be made available with minor adjustments. If a SOAR system is not used to handle the incident response process, the information would have to be collected separately from various systems utilized by the SOC, such as SIEM, EDR, or an ITSM system, and stored in an external system to take care of the reporting, as it is necessary to combine information from multiple systems to construct the metrics. As the security incident management process can involve a lot of manual activities, the reporting requirements must be considered thoroughly to avoid having the SOC analysts manually record information that is not used for a specific purpose, and to ensure that recording the necessary information is enforced by a technical solution to avoid the metrics becoming biased or imperfect.

The results of the measurement validation are summarized in table 8. The validation of the measurement is considered to be a success if the measurement is already recorded, a partial success if it is not recorded at the moment but can be made available, and a failure if the measurement can not be collected without significant changes to the ways of working or the technical solutions in use.

TABLE 8 Results of the measurement collection

Measurement	**Success**	**Success (Partial)**	**Failure**
Security incident classification	*		
Detection classification	*		
State of automated containment		*	
Original priority of the security incident	*		
Final priority of the security incident	*		
Information about whether a security incident has been escalated		*	
Information about who the security incident has been escalated to	*		
Number of unknown entities related to a security incident			*
Information about whether the initial conclusion was correct or not		*	
Information about whether the security incident had to be re-investigated by SOC		*	
Original detection source		*	
MITRE ATT&CK tactic categorization	*		
Unified Kill Chain stage		*	

8 RESULTS AND DISCUSSION

The metric selection framework, which is the design science artifact constructed in this research following the design science methodology as described by Peffers et al. (2007), can be successfully used to construct metrics that can be used to measure the technical performance of a SOC, as was demonstrated by the metrics outlined in the section 7.4. Out of the five metrics that were constructed, four metrics were considered to be valid in terms of the metric selection framework. The invalid metric outlined in the section 7.4.2 had a problem with too much bias and thus was considered to be invalid, as ways to reduce the bias to acceptable levels were not discovered as a part of the metric construction phase. On an overall level, the remaining metrics suffered from similar issues, as reducing the bias and subjectivity was difficult to perform, and the entire definition of what is an acceptable level of bias or subjectivity remained relatively subjective as well.

The valid metrics created with the framework can help SOCs to push their detection capabilities more towards the earlier stages of the unified cyber kill chain to decrease the potential impact of the security incidents (section 7.4.1) and quantify the value of their detection engineering function (section 7.4.3). The metrics can also be used to provide insights on the activities performed as a part of the analysis process (section 7.4.4) and measure how precise and sensitive the workflows for automated containment are (section 7.4.5). While the metrics are by no means comprehensive, they can be used to measure the technical performance of a SOC within the respective areas, and as such, can be used to enhance the reporting capabilities related to the technical performance of a SOC, for as long as the required measurements can be made available.

The collection of the measurements required to construct the metrics was verified in a SOC that is offering managed SOC service to medium and large enterprises. Out of the thirteen measurements required to construct the metrics, six were already available, another six were not available but could have been made available with minor adjustments, and one was considered to not be something that could be made available without significant changes to the ways of working or the technical solutions utilized by the SOC. The SOC utilizes a SOAR platform to manage the detections and security incident management workflows,

which made the collection of the measurements relatively straightforward. A SOC that utilizes more than one system to manage the security incident management workflows is likely to require a third system to collect and combine the measurements from multiple systems, as the metrics require measurements from multiple sources, such as a SIEM and an EDR, to be considered complete and thus conform to the quality criteria associated with the metric selection framework. Despite the success of the validation, the validation of the measurements should be handled better during the metric creation process to ensure the measurements can be made available as expected.

During the creation of the presentation for the metrics, the objective was to provide a way for both the team and the SOC management to benefit from the outcome of the metric, which resulted in a two-part visualization strategy; one graph for the team the metric is intended for and another one for the management of the SOC. The teams responsible for the metrics may be more interested to gain additional insights from the low-level information outlined in the first graph while the SOC management could be more interested to follow the metric value evolution over time, as seen in the second graph. One of the mismatches between technical staff and managers outlined by Kokulu et al. (2019) was the interpretation of the values and meaning of the metrics deployed, and as a result, it was seen as a key factor to consider during the metric creation process but was not considered with sufficient detail in the metric selection framework. As the visualizations for the metrics were created by utilizing Jupyter notebooks, it remains unclear whether the metrics can be visualized as expected with tools commonly used by SOCs to publish and manage reporting dashboards. However, as the measurements required for the metrics need to be collected from several sources to create the visualizations for the metrics, SOCs should ensure their tooling and other reporting capabilities can support the complex visualizations required to present the information necessary to properly interpret the metric.

One of the limitations of this research is that the design methodology is not followed rigorously, as the feedback loop between the "Evaluation" and the "Design and development" activities are not completely enforced as per the design science methodology as outlined by Peffers et al. (2007). The decision to limit the number of iterations was done due to constraints related to the research schedule and as a result, after the first metric was successfully created, the iteration back to the "Design and development" stage was not performed. On a practical level, this means that any improvements discovered after the successful creation of the first metric are documented in this chapter and left out to be completed as a part of future research.

Although the artifact was able to produce metrics that can be used to enhance the reporting capabilities of the technical performance of a SOC, the link between the technical performance and the metric selection framework could have been slightly more concrete. The framework itself does not directly enforce the relationship between the metrics and technical performance. This does not necessarily make the selection criteria to be less useful, but it leaves the determination of whether the resulting metric measures technical performance or not up to the user of the artifact to decide. As an upside, the selection framework can be

additionally used to create non-technical metrics as well.

The metrics created with the metric selection framework revolved quite heavily around the detection capabilities provided by the detection engineering process as per the definitions by Knerler et al. (2022) and as a result, MITRE ATT&CK tactics are used as a key dimension for the presentation of the metrics. MITRE ATT&CK tactic was determined to be able to provide additional technical context for the visualization of the current state of the metric, as it can provide at a glance how the tactics are performing relative to others, which can be valuable information for the detection engineering team itself but might not be that relevant for the SOC management or other teams. Other functions that could be considered to be mostly about technical capabilities, such as threat hunting or cyber threat intelligence, were not covered during the evaluation of the artifact. As a result, additional research is required to validate the framework with functions other than detection engineering, as it remains unknown whether the metric selection framework can produce viable results when used in a context other than detection engineering.

One of the fundamental principles when it comes to the definition of KPIs according to Parmenter (2019) is to tie the indicators into critical success factors. As a result, a relation to a success factor was included as a requirement within the metric selection framework. However, the SOC-related literature does not discuss in detail the SOC success factors, which means the success factors are not properly taken into account when the metrics were constructed. There are some publications (Abd Majid & Zainol Ariffin, 2021; Crowley & Filkins, 2022; Vielberth et al., 2020) that discuss the success factors, but the success factors mentioned in the publications are too high-level to be able to contribute to the metric selection framework as was originally intended. Additional research around the low-level success factors for SOC, or more widely low-level success factors for cyber defence capabilities on an overall level, would be needed to utilize the framework as was originally expected. In addition to the success factors, the metrics should on an overall level be better justified by scientific research. Despite the lack of research on the area, organizations can and should identify their unique success factors and justifications for the metrics, and design the performance metrics around the parameters they have identified.

The literature review did not establish a clear pattern when it comes to the availability of SOC-related metrics, and as such, a conclusion was reached that no such framework currently exists. Many of the more commonly used metrics as summarized in table 3 are operative, and as a result, they cannot be used to measure the technical performance of a SOC. Some of the metrics can be used to partially measure the technical performance, such as the false-positive rate or number of incidents handled automatically, but they are at best imperfect when evaluated with the metric selection framework, due to lack of additional context that causes the metric to become overly biased. For example, a high true-positive rate, and subsequently a low false-positive rate, can be an indication of a well-performing SOC but without understanding the detection strategy or level of automation, the metric alone can be misleading, since the SOC could automatically close a majority of the false-positive and benign true-positive detections, and in-

stead use them for threat hunting or to provide additional context in the form of potentially related low-fidelity detections when investigating high-severity detections. The metric selection framework can be used to enhance the reporting capabilities provided by these technical metrics, as it can be used to reduce both the bias and subjectivity as well as take into account the additional contextual information required to properly interpret the metric.

It could be argued that most of the metrics are not inherently bad or insufficient for measuring technical performance, but rather they lack the necessary characteristics and definitions that are needed to properly interpret the value of the metric and thus reduce the objectivity and bias associated with the metric. Although the number of SOC-related publications has steadily increased since 2014, there are still several challenges that are needed to be solved to advance the level of SOC-related research, among them being the ineffective capabilities to measure SOC performance (Vielberth et al., 2020). Based on the conclusion reached within this research, the situation still appears to largely be the same, and thus, the field would need a considerable amount of novelty research to increase the maturity in terms of academic publications, which could be used as a basis to construct proper metrics to measure the performance of a SOC.

Additionally, the lack of maturity in the SOC-related literature could also be explained by the uncertain direction in which the SOC as a function should be heading into. Most publications focus on the idea that SOC by design is mostly a reactive function (Agyepong et al., 2020; Ahlm, 2021; Vielberth et al., 2020), and although some are bringing up more proactive measures to enhance the SOC operations (Knerler et al., 2022), the paradigm appears to revolve mostly around reactive capabilities combined with some proactive elements, such as threat hunting and cyber threat intelligence augmented monitoring rules.

Is the current SOC paradigm enough to protect organizations against modern threats? Should SOCs be better integrated with the overall cyber defence capabilities and be more involved with the prevention of security incidents, rather than just reacting and responding to them? How to determine the actual impact the SOC has on the overall cyber defence capabilities? It would appear that none of these questions is answered in the current SOC-related academic literature. Therefore, additional research about the future of SOCs is required, as the answers to these questions could cause a paradigm shift and change the role of the SOC to focus more on the prevention of security incidents and thus, potentially decrease the overall cost of cyber defence capabilities while simultaneously decreasing the costs associated with the practice. To accurately measure the technical performance of a reactive and a proactive SOC would likely require different metrics as the incentives of the two SOCs are different from one another. The lack of certainty in the direction where the SOC as an industry is heading and the large variety of different SOC approaches, could be one explanation for the lack of proper metrics to measure the performance of the SOC.

9 CONCLUSION

The outcome of this research further emphasizes the need for better capabilities to measure the technical performance of different types of SOCs, as the commonly used metrics focus on operational activities, and as such, are inadequate to measure the technical performance of a SOC. Furthermore, the metrics observed in the literature do not appear to be a result of a systematic development but rather be loosely based on generic security metrics or otherwise based on industry best practices without significant scientific justification.

This research resulted in a design science artifact, a novelty metric selection framework, that can be used to construct relevant metrics to measure both the technical and non-technical performance of a SOC. The literature reviewed during this research contained some metrics that could be partially used to measure the technical performance of a SOC, such as the false-positive rate or the number of incidents handled by automation. However, these metrics were determined to be invalid based on the metric selection framework, due to a lack of additional context to decrease the bias to an acceptable level.

As a part of the demonstration of the artifact, five unique metrics were created that can be as such used to improve the technical reporting capabilities related to the detection engineering and the incident analysis functions. The metrics created with the artifact were the following:

1. Distribution of detections among the Unified Kill Chain
2. Number of verifiable monitoring rules
3. Distribution of detections by source
4. Technical accuracy of the analysis
5. Accuracy of automated containment

Out of the five metrics, four were considered to be valid in terms of the metric selection framework. The invalid metric (number of verifiable monitoring rules) ended up being overly biased, which means it cannot be applied as a general-purpose metric to measure and compare different SOCs. Bias and subjectivity were also a concern for other metrics, but with the addition of contextual information, the bias and objectivity were reduced to an acceptable level, and as a

result, the metrics were classified as valid in terms of the metric selection framework. The metrics were validated by verifying that the measurements can be collected within a SOC service provider that produces managed SOC services for medium and large enterprises. The outcome of the validation of the measurements was that all but one measurement was either already recorded or could be made available with minor adjustments to the solutions used by the SOC. The outcome of the measurement validation was influenced by the utilization of a SOAR system to combine information from multiple sources and manage the workflow related to the security incident management process, thus a SOC that uses more than one system to manage its operation can have difficulties collecting the required measurements and combine them to construct the metrics.

One of the key limitations of this study is the relatively narrow selection of metrics that were used to demonstrate the framework, as the metrics created with the framework are mostly related to the detection capabilities of the SOC, and other functions, such as threat hunting or cyber threat intelligence, have not been included in the metrics chosen for the demonstration. As such, it remains unknown whether the framework is suitable for the creation of technical metrics for functions other than the ones that work closely with the security incident management process. Furthermore, the design science principle was not rigorously followed, and as a result, some of the iterative improvements for the metric selection framework, in addition to the wider demonstration, were left to be researched in the future.

Despite the minor limitations of the research, the framework and the metrics used for demonstration can be adopted by different types of SOCs to construct metrics they can use to measure and demonstrate their technical capabilities. Due to the lack of industry-standard reporting schema for the technical performance of SOCs, the SOC industry as a whole is encouraged, to enable industry-driven development of the measurement capabilities, be open and share the metrics they use to measure their technical capabilities, with the wider community. In addition to industry-backed development of the measurement of technical performance, additional academic research is needed on the subject.

Additional research is especially needed to better understand what makes SOC successful, what is the actual impact of a well-functioning SOC within the context of wider cyber defence capabilities of an organization, and if SOC as a function should focus more on the prevention of security incidents rather than responding to them, which appears to be the current standard approach for SOCs. With the addition of such scientific research, better metrics could be constructed, as the scientific research could be used to better justify the metrics constructed with the metric selection framework.

BIBLIOGRAPHY

Abd Majid, M., & Zainol Ariffin, K. A. (2021). Model for successful development and implementation of cyber security operations centre (SOC). *PLOS ONE*, *16*(11), 1–24

Agyepong, E., Cherdantseva, Y., Reinecke, P., & Burnap, P. (2020). Towards a framework for measuring the performance of a security operations center analyst. In *2020 international conference on cyber security and protection of digital services (cyber security)* (pp. 1–8).

Ahlm, E. (2021). *How to build and operate a modern security operations center*. Gartner Inc.

Allianz. (2022). *Risk barometer 2022*. Retrieved September 17, 2022, from https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/Allianz-Risk-Barometer-2022.pdf

Bahrami, P. N., Dehghantanha, A., Dargahi, T., Parizi, R. M., Choo, K.-K. R., & Javadi, H. H. S. (2019). Cyber kill chain-based taxonomy of advanced persistent threat actors: Analogy of tactics, techniques, and procedures. *Journal of Information Processing Systems*, *15*(4), 865–889.

Bendle, N. T., Bagga, C. K., & Nastasoiu, A. (2019). Forging a stronger academic-practitioner partnership–the case of net promoter score (NPS). *Journal of Marketing Theory and Practice*, *27*(2), 210–226.

Black, P. E., Scarfone, K., & Souppaya, M. (2008). Cyber security metrics and measures. In *Wiley handbook of science and technology for homeland security* (pp. 1–15).

Böhme, R. (2010). Security metrics and security investment models. In I. Echizen, N. Kunihiro, & R. Sasaki (Eds.), *Advances in information and computer security* (pp. 10–24). Berlin, Heidelberg: Springer Berlin Heidelberg.

Brotby, W. K., & Hinson, G. (2013). *Pragmatic security metrics*. Auerbach Publishers, Incorporated.

Caldwell, J. H. (2021). *Global risk management survey, 12th edition*. Retrieved September 17, 2022, from https://www2.deloitte.com/content/dam/insights/articles/US103959_Global-risk-management-survey-12ed/DI_Global-risk-management-survey-12ed.pdf

Caltagirone, S., Pendergast, A., & Betz, C. (2013). *The diamond model of intrusion analysis*. US Department of Defense.

Chew, E., Swanson, M., Stine, K., Bartol, N., Brown, A., & Robinson, W. (2008). *Performance measurement guide for information security* (tech. rep. No. NIST Special Publication (SP) 800-55, Rev. 1). National Institute of Standards and Technology.

Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). *Computer security incident handling guide* (tech. rep. No. NIST Special Publication (SP) 800-61, Rev. 2). National Institute of Standards and Technology.

Cole, E. (2012). *Advanced persistent threat : Understanding the danger and how to protect your organization*.

CrowdStrike. (2022). *2022 global threat report*. Retrieved August 26, 2022, from https://go.crowdstrike.com/rs/281-OBQ-266/images/Report2022GTR.pdf

Crowley, C., & Filkins, B. (2022). *Sans 2022 SOC survey*. SANS Institute. Retrieved November 17, 2022, from https://www.sans.org/white-papers/sans-2022-soc-survey/

Crowley, C., & Pescatore, J. (2019). *Common and best practices for security operations centers: Results of the 2019 SOC survey*. SANS Institute. Retrieved August 1, 2022, from https://www.sans.org/media/analyst-program/common-practices-security-operations-centers-results-2019-soc-survey-39060.pdf

Doran, G. T. (1981). There's a S.M.A.R.T way to write management's goals and objectives. *Management Review*, *70*(11), 35–36.

Dufkova, A., Stikvoort, D., Kossakowski, K. P., Maj, M., Benetis, V., & Gapinski, K. (2022). *ENISA csirt maturity framework*. Retrieved September 20, 2022, from https://www.enisa.europa.eu/publications/enisa-csirt-maturity-framework

Egloff, F. J. (2020). Public attribution of cyber intrusions. *Journal of Cybersecurity*, *6*(1).

European Union Agency for Cybersecurity. (2021). *ENISA threat landscape 2021: April 2020 to mid-july 2021* (I. Lella, M. Theocharidou, E. Tsekmezoglou, & A. Malatras, Eds.).

Firstbrook, P., Olyaei, S., Shoard, P., Thielemann, K., Ruddy, M., Gaehtgens, F., … Candrick, W. (2022). *Top trends in cybersecurity 2022*. Gartner Inc.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. (Vol. 3408, pp. 345–359).

Hauser, J., & Katz, G. (1998). Metrics: You are what you measure! *European Management Journal*, *16*(5), 517–528.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 77–105.

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

IBM. (2022). *Cost of a data breach report 2022*. Retrieved October 7, 2022, from https://www.ibm.com/downloads/cas/3R8N1DZJ

ISO/IEC 27004:2016. (2016). *Information technology — security techniques — information security management — monitoring, measurement, analysis and evaluation*. International Organization for Standardization. Retrieved from https://www.iso.org/standard/64120.html

Keltanen, P. (2019). *Measuring outsourced cyber security operations center* (Master's thesis, South-Eastern Finland University of Applied Sciences). Retrieved from https://urn.fi/URN:NBN:fi:amk-2019120925525

Knerler, K., Parker, I., & Zimmerman, C. (2022). *11 strategies of a world-class cybersecurity operations center*. The MITRE Corporation.

Kokulu, F. B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., & Ahn, G.-J. (2019). Matched and mismatched SOCs: A qualitative study on security operations center issues. In *Proceedings of the 2019 acm sigsac conference on computer and communications security* (pp. 1955–1970).

Logsign. (n.d.). *Guide for security operations metrics*. Retrieved August 2, 2022, from https://www.logsign.com/uploads/Guide_for_Security_Operations_Metrics_Whitepaper_2f999f27cc.pdf

Mettler, T., Eurich, M., & Winter, R. (2014). On the use of experiments in design science research: A proposition of an evaluation framework. *Communications of the Association for Information Systems*, *34*, 223–240.

Microsoft. (2022). *Defending Ukraine: Early lessons from the cyber war*. Retrieved August 26, 2022, from https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE50KOK

Mitre Corporation. (2022). *Enterprise matrix*. Retrieved August 29, 2022, from https://attack.mitre.org/versions/v11/matrices/enterprise/

Nathans, D. (2014). *Designing and building a security operations center* (S. Elliot, Ed.). Elsevier Science & Technology Books.

National Institute of Standards and Technology. (2018). *Framework for improving critical infrastructure cybersecurity*. Retrieved September 20, 2022, from https://doi.org/10.6028/NIST.CSWP.04162018

Onwubiko, C. (2015). Cyber security operations centre: Security monitoring for protecting business and supporting cyber defense strategy. In *2015 international conference on cyber situational awareness, data analytics and assessment (cybersa)* (pp. 1–10).

Parmenter, D. (2019). *Key performance indicators* (4th ed.). John Wiley & Sons, Incorporated.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pendleton, M., Garcia-Lebron, R., Cho, J.-H., & Xu, S. (2016). A survey on systems security metrics. *49*(4).

Pols, P. (2017). *The unified kill chain*. Retrieved August 29, 2022, from https://www.unifiedkillchain.com/assets/The-Unified-Kill-Chain.pdf

PwC. (2022). *2022 global risk survey*. Retrieved September 17, 2022, from https://www.pwc.com/us/en/services/consulting/cybersecurity-risk-regulatory/assets/pwc-global-risk-survey-report-2022-main.pdf

Reichheld, F. (2004). The one number you need to grow. *Harvard business review*, *81*, 46–54, 124.

Rid, T., & Buchanan, B. (2015). Attributing cyber attacks. *Journal of Strategic Studies*, *38*(1-2), 4–37.

Salmi, N. (2018). *The present state of information security metrics* (Master's thesis, University of Jyväskylä). Retrieved from http://www.urn.fi/URN:NBN:fi:jyu-201811234850

Savola, R. M. (2007). Towards a taxonomy for information security metrics. In *Proceedings of the 2007 acm workshop on quality of protection* (pp. 28–30).

Savola, R. M. (2013). Quality of security metrics and measurements. *Computers & Security*, *37*, 78–90.

Schlette, D., Vielberth, M., & Pernul, G. (2021). CTI-SOC2M2 - the quest for mature, intelligence-driven security operations and incident response capabilities. *Computers & Security*, *111*, 102482.

Simos, M., & Dellinger, J. (2019). *CISO series: Lessons learned from the Microsoft SOC—part 1: Organization*. Retrieved August 2, 2022, from https://www.microsoft.com/security/blog/2019/02/21/lessons-learned-from-the-microsoft-soc-part-1-organization/

Sophos. (2022). *Sophos 2022 threat report: Interrelated threats target an interdependent world*. Retrieved August 26, 2022, from https://assets.sophos.com/X24WTUEQ/at/b739xqx5jg5w9w7p2bpzxg/sophos-2022-threat-report.pdf

Strom, B. E., Applebaum, A., Miller, D. P., Nickels, K. C., Pennington, A. G., & Thomas, C. B. (2018). *MITRE ATT&CK®: Design and philosophy* (Revised). The MITRE Corporation.

U.S. Department of Defense. (2021). *Cybersecurity maturity model certification (CMMC) model overview*. Retrieved September 20, 2022, from https://www.acq.osd.mil/cmmc/docs/ModelOverview_V2.0_FINAL2_20211202_508.pdf

U.S. Department of Energy. (2022). *Cybersecurity capability maturity model (C2M2)*. Retrieved September 20, 2022, from https://www.energy.gov/sites/default/files/2022-06/C2M2%20Version%202.1%20June%202022.pdf

Van Os, R. (2016). *SOC-CMM: Designing and evaluating a tool for measurement of capability maturity in security operations centers* (Master's thesis, Luleå University of Technology). Retrieved from http://ltu.diva-portal.org/smash/get/diva2:1033727/FULLTEXT02.pdf

Van Os, R. (2018). *SOC-CMM measuring capability maturity in security operations centers*. Retrieved September 20, 2022, from https://www.soc-cmm.com/downloads/soc-cmm%20whitepaper.pdf

Van Os, R. (2022). *SOC-CMM basic assessment tool, version 2.2*. Retrieved October 19, 2022, from https://www.soc-cmm.com/downloads/latest/soc-cmm%202.2%20-%20basic.xlsx

Vielberth, M., Böhm, F., Fichtinger, I., & Pernul, G. (2020). Security operations center: A systematic study and open challenges. *IEEE Access*, *8*, 227756–227779.

Willett, M. (2021). Lessons of the solarwinds hack. *Survival*, *63*(2), 7–26.

World Economic Forum. (2022). *The global risks report 2022* (17th ed.). World Economic Forum.

Zimmerman, C., & Crowley, C. (2019). *Practical SOC metrics*. FireEye Cyber Defense Summit 2019. Retrieved July 25, 2022, from https://www.fireeye.com/content/dam/fireeye-www/summit/cds-2019/presentations/cds19-executive-s03b-practical-soc-metrics.pdf