Evaluation of Cyber Situation Awareness - Theory, Techniques and Applications

Georgi Nikolov
g.nikolov@cylab.be
Royal Military Academy
Brussels, Belgium

Axelle Perez
axelle.perez@ulb.be
Université libre de Bruxelles
Brussels, Belgium

Wim Mees
w.mees@cylab.be
Royal Military Academy
Brussels, Belgium

ABSTRACT

In recent years the technology field has grown exponentially, bringing with it new possibilities, but also new threats. This rapid advancement has created fertile grounds for new sophisticated cyber attacks, exhibiting a high degree of complexity. In an ever evolving cyber landscape, organizations need to dedicate valuable resources in enhancing their understanding of emergent threats for the purposes of identification, analysis and mitigation. To accomplish this task, they rely on Cyber Situation Awareness (CSA), a framework designed for the purposes of managing the virtual environment. This is achieved through the perception and comprehension of the behaviors therein, be that benign or malicious, followed by modeling the future state of the environment based on the gathered information. In this paper, we will discuss how exactly the theory of Situation Awareness has been applied to the cyber domain. Further on, we will present various techniques used for handling the large quantity of complex data and managing the dynamic nature of the environment by Cyber Situation Operation Centers (CSOC) and discuss in detail a number of methodologies that have been designed for the evaluation of the level of CSA. Finally, we will provide specific examples of simulated scenarios for the application of the CSA assessment techniques.

CCS CONCEPTS

• **Human-centered computing → Visualization design and evaluation methods**; • **Security and privacy → Network security**; **Intrusion detection systems**.

KEYWORDS

Cyber Situation Awareness, Visualization, Evaluation Methodologies, Assessment

ACM Reference Format:
Georgi Nikolov, Axelle Perez, and Wim Mees. 2024. Evaluation of Cyber Situation Awareness - Theory, Techniques and Applications. In *The 19th International Conference on Availability, Reliability and Security (ARES 2024), July 30-August 2, 2024, Vienna, Austria.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664476.3670921

1 INTRODUCTION

The rapid development and growth of new technologies has enabled organizations to establish complex network infrastructures, supporting a large quantity of users and generating enormous amounts of data. This presents a major challenge for the management and security of the network. Indeed, the complexity of current infrastructures can be a benefit as well as drawback- malicious actors can more easily infiltrate the network and stay undetected for long stretches of time. Cyber analysts have to fight an uphill battle to sift through the terabytes of daily-generated data and separate the benign from the malicious. Even with the availability of highly sophisticated tools, which can collect, process and even generate mitigation solutions through the use of Machine Learning, the human element remains a vital part of the analysis loop and needs to maintain a high degree of Situation Awareness (SA).

The concept of Situation Awareness is something very intrinsic to humans. Since the dawn of times we have had the need to observe our surroundings and extract valuable information, vital for our survival. It is a concept that is understandable for many on a subconscious level, but only relatively recently a codified definition of what it entails has been presented. The first attempt to define SA was in the military domain, defining the need for gathering information before the enemy, analyzing it and acting upon it [11]. Therefor, it is not a surprise that a concrete definition, and the most widely used nowadays, was first introduced in the field of aviation by Mica Endsley [7]. In recent years, a lot of importance has been put in applying these principles to the cyber domain by designing systems which can help enhance the three levels of SA and offer a higher degree of human-computer interaction through the implementation of Visual Analytics. In this paper, we will describe in Section 2 how Endsley defined Situation Awareness and how her definition has been extended to the cyber environment. In Section 3 we will address the metrics used to measure SA and present a review of multiple methodologies used for the evaluation of SA and measuring the performance of operators, when using systems designed with Situation Awareness in mind. Finally, in Section 4 we will present two use cases that can be used during the evaluation and talk about how an assessment methodology can be applied to the scenarios.

2 CYBER SITUATION AWARENESS

2.1 Definition of Situation Awareness

At its core, Situation Awareness describes how a person perceives and understands inputs, leading to the creation of specific models describing possible future changes to their surroundings and what decisions and actions can result from that. The specific definition

ARES, July 30 – August 2, 2024, Vienna, Austria

ACM ISBN 979-8-4007-1718-5/24/07
https://doi.org/10.1145/3664476.3670921

proposed by Endsley [7], that became the cornerstone of future SA research, goes as follows:

> "Situation awareness is the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future."

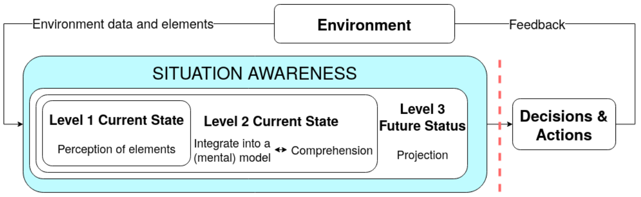

Figure 1: Situation Awareness Stages

The three stages, or levels, of SA defined by Endsley [7], shown in Figure 1, are as follows:

- **Perception (level 1)** - This phase refers to the capability of humans to monitor, detect cues in the environment and basic recognition. These capabilities are vital in the cyber domain for the surveying of the different network elements (events, people, systems, etc.) and their current state.
- **Comprehension (level 2)** - Comprehension consists of interpreting and understanding the significance of the previous disjointed elements. Through comprehension humans can attach meaning to the observed information and possible patterns emerge. In essence, this phase translates raw data into a contextual form that can support informed decision making.
- **Projection (level 3)** - Projection refers to the ability to employ the knowledge gathered during the Comprehension phase and construct a model of the near future. The projection phase thus encompasses predicting future states of the cyber environment based on the information gathered and comprehended in the previous two stages.

2.2 Situation Awareness in the Cyber domain

This definition can be applied to the cyber domain, but not without some adjustments. Indeed, applying the concept presents some difficulties, as described by Husák et al. [15]. Contrary to the real world, in the cyber environment the operator is a passive observer instead of an active participant. They need to rely on information collected by sensors and systems and can't use their own senses to perceive the environment and draw conclusions. Another problem is the "border-less" nature of the cyber environment- it is difficult to establish which part of the environment an analyst should focus their attention on. The issue of differentiation and limitation of the cyber environment stems from bad optics - if we regard the electronic components as the integral parts, then everything is interconnected and the environment has no borders. To remedy this, we need to take into account the logic behind how the electronic components are interconnected, which establishes specific borders.

Another important aspect of CSA is brought up by the authors in [15]- the aspect of the taxonomy of Cyber Situational Awareness. They base their taxonomy on the work done by Evesti et al. [10], but expand the top level to accommodate the 3-level model defined by Endsley [7]. The new taxonomy presented in [15] is shown in Figure 2. The Perception level has to be expanded to deal with the issues related to data. Because of the nature of any cyber environment, data is produced in large quantities and at a rapid pace. This can lead to an overload of information where important events may be hidden by the amount of noise created.

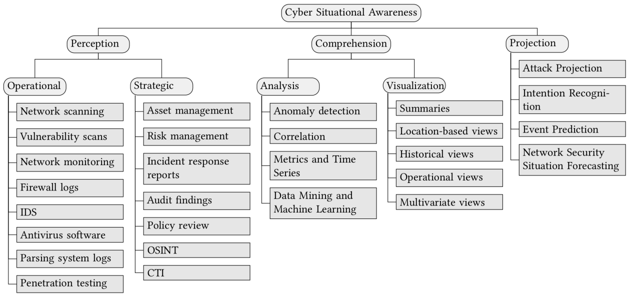

Figure 2: CSA taxonomy as defined by [15]

Furthermore, the capability of SA systems to correctly perceive and interpret data is largely based on the quality of data [1] as the data collected from a multitude of different sources can be very different. As described in [1], the data produced by sensors can be of a variable type:

- **dynamic** - network data has the tendency to change over time, describing the ever-evolving topology of the network and its specifications.
- **one-off** - data generated in the form of reports, most often by analysts, which are made after a data review.
- **alert-based** - data generated on the basis of alerts produced by systems such as Intrusion Detection Systems (IDS) or Intrusion Prevention Systems (IPS).
- **Intelligence sharing** - data relying on the exchange of information between different organisations. Cyber Threat Intelligence (CTI) is often used to share Indicators of Compromise (IoCs) with the public, strengthening the defences against recent malware and vulnerabilities. Platforms such as the Malware Information Sharing Platform (MISP) are highly used, but can also provide the drawback of large volumes of information, not always useful or usable, being produced daily.
- **Raw data** - all other types of data can be regarded as raw data. This encompasses data produced from log generating tools (firewalls, Operating System audit logs, etc), packet sniffers or tools that generate variety of dumps.

Precise management tactics and policies are needed to optimize the Perception level of CSA to be able to establish a complete picture of the environment. Further, the Comprehension level needs to be expanded to better deal with the large quantity of data through the use of Visual Analytics, bringing often purely text-based data to the visual medium. Finally, a set of methods for correctly and efficiently creating attack models for the correct Projection of the evolution of any threat need to be applied.

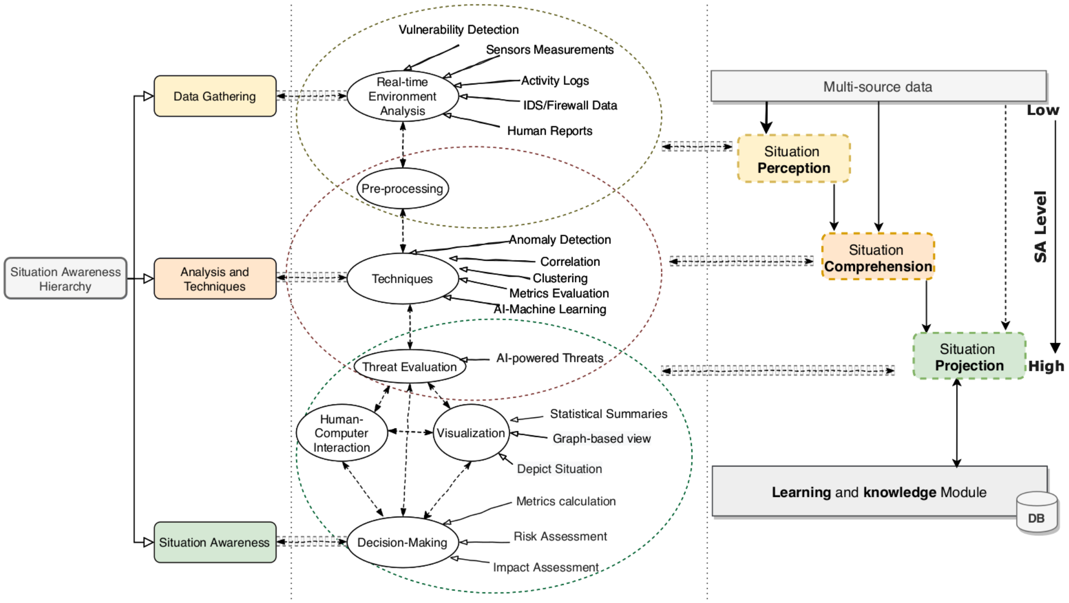

Figure 3: CSA hierarchy as described by [1]

2.3 Hierarchy of Cyber Situation Awareness

A lot of work has been done on defining specific techniques and methods used for the enhancement of each phase of the Situation Awareness. They have been collected and described by Alvizadeh et al. in [1], presented in a diagram shown in Figure 3. The SA hierarchy is based on the three layers of awareness introduced by He and Li [13], with each layer encompassing different methods and techniques necessary for enhancing the result produced in the corresponding layer. The authors have taken the time to review papers detailing different approaches used for the data gathering, analysis and gaining SA. Two points of interest that we can observe in their proposed hierarchy is the reliance on data pre-processing and the inclusion of Machine Learning techniques. Regarding the data challenges we discussed in Subsection 2.2, pre-processing is important to handle the large volume and variety of data present. Pre-processing can help transition between the Perception and Comprehension phases through the use of techniques to homogenize the data for easier analysis. Machine Learning, on the other hand, has become a very popular topic in the last couple of years, with the development of ML-powered IDS and IPS systems. This can greatly help speed up the analysis and partially alleviate the work of the operator. One aspect of ML that needs to be taken into consideration is the reliance on datasets to train the learning based detection tools. Alvizadeh et al. [1] discuss two prominent datasets used for the training and testing, specifically the KDD-Cup 1999 [31] and NSL-KDD [17]. Other research has been done in the hopes of producing valid datasets, such as the CSE-CIC-IDS2018 [29] and Unraveled [22] datasets. This has greatly advanced the development of ML-powered detection, but we need to be cognisant of the challenges when discussing the use of datasets in cyber security. Indeed, as described by the authors in [15], the problem lies in the duality of the datasets, those captured in a live environment versus synthetically produced ones. Artificial datasets offer ground-truth and extensive documentation on network topology, but it is not always evident generating background traffic, random noise and anomalies, which are present in a real network. Contrary to that, when using a dataset captured in a live environment, we have very little information available on the ground-truth and what types of threats to expect in the data, which is not very well suited for the ML approach. On top of that, anonymization measures need to be taken into account to protect the private information of the users.

3 EVALUATING CYBER SITUATION AWARENESS

3.1 Situation Awareness design

Understanding the principles of Cyber Situation Awareness makes it possible to design better network management and analysis tools. The CSA level of an operator is closely coupled with the way they can visualize and interpret the information. Applying the principles of Visual Analytics can enhance the representation of the information and help the users get a better understanding of the situation at hand. The paper written by Varga et al. [34] considers the importance of visualization as well as the human factor to propose a case study of two human machine interface design approaches: the user centered and the system based. As presented in [34], the relationship between visualization and CSA is of great importance. The authors refer to the work done by D'Amico [4] presenting a nine-way taxonomy to illustrate the relationship between different cyber analysis practices, the visualizations used and the stages of SA. The three stages of CSA are strongly linked and visualization is crucial to achieve and maintain CSA. As shown in Figure 4, through the use of specialized visualizations the attention of the operator can be oriented to the information of importance during the Perception Phase, propose meaningful means to explore and filter data during the Comprehension phase and to correctly predict the future state of the environment during the Projection phase. Through all three stages a continuous reporting is done to explain what has been observed, essential for the decision-making process.

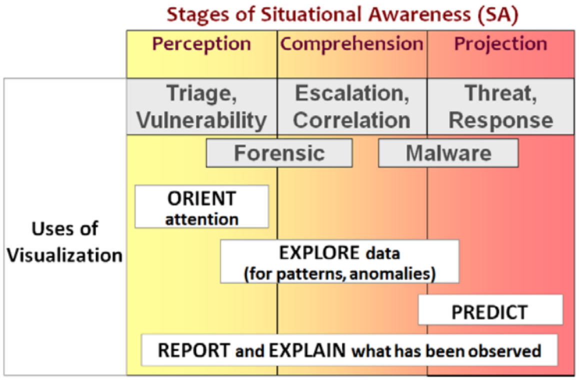

Figure 4: Relationship between the stages of situational awareness, the use of visualization and the types of analysis performed [4]

Depending on the intended user and the objectives, the visualization needs to be adapted to meet the expected needs. One way to design adequate visualizations is to decide on which factor to focus-the user or the system. As proposed in [34], visualization can be separated into user centric and system centric approaches. The human centric approach takes into account the experience and knowledge of the operator, offering a modular visual representation that can be used to apply their skills and know-how. Incorporating the user's domain knowledge is vital for the enhancement of the CSA of the operator. Through this, the operator can quickly identify data of importance (Perception), analyse it to detect abnormal patterns (Comprehend) and determine how this will influence the future state of the analysed machines (Projection). To facilitate this, best practise would be to follow the taxonomy of interactive dynamics for visual analysis [14]. Contrary to the user centric approach, the system based approach has greater focus on the optimal way to represent the complex structure of the system. By reproducing the system in a detailed visual way, the user can better understand the system they are working with, quickly identify problems and deploy solutions or mitigation techniques. It is important to note that because of the visual scale, it is difficult to determine the specific cause of a problem and more in-depth analysis will be needed. In both cases, specialized techniques need to be developed to bolster the CSA of the operator. Much research has been done to propose and develop various ways of representing data, as shown in the systematic literature review on the topic of CSA visualizations, done <sby Jiang [16]. The paper presents interesting examples of various visualizations, but also states the obvious lack of research into certain aspects, such as integration of human and organization input, limited interaction techniques and lesser focus on enhancing the Projection phase of CSA.

The human-machine interaction is of major importance when discussing CSA. To correctly design an interface that facilitates the users to perceive, comprehend and project information in a cyber environment, one must consider the different challenges analysts face daily dealing with large quantities of various data types. As shown in Section 2.2 in Figure 2, visualization is a major aspect in the Comprehension phase of CSA. Before we can correctly evaluate the level of SA of a user, first we need to assess how intuitively and effectively the visualization tools aid the user experience. This is vital to ensure that the visualization helps to enhance the decision-making capabilities and alleviates the workload. Through our co-operation with the NATO Visual Analytics for Complex Systems Research Task Group [23], we have gathered a selection of elements that need consideration when evaluating a cyber visualizations and separated them into two groups- elements pertaining to the user experience and those defining the visualization tool used. The characteristics shown in Figure 5 can be evaluated using methods aimed at a formal assessment of a visualization design [25]. It is important to keep in mind that the two groups are interdependent on each other. The user specific characteristics encompass how the operator interacts with the visualization tools and how the experience of the user is enhanced by them. Further, the user experience is influenced by the visualization tools design and implementation. A usable and useful visualization tool aids the analyst to more effectively accomplish tasks, reducing their cognitive load and enhancing their decision-making process.

3.2 Measuring Situation Awareness

As mentioned in Section 2, there has been much work done in the field of CSA, specifically in defining how the concepts of Situation Awareness can be applied to the cyber domain, how the taxonomy needs to be expanded and the underlying hierarchy. This is all good and well, but an aspect that is still lacking is how exactly can we score the CSA of an operator. Indeed, as Situation Awareness is not a tangible concept and is highly dependant on the environment/tools/operator, it is not evident to present a concrete methodology for its evaluation. Nevertheless, effort has been invested into the design and implementation of different procedures to estimate the degree of SA that a user exhibits. When examining the evaluation of CSA, the focus is often on the human-machine interaction and more precisely on how the users benefit from the specific tools at their disposal. Scholtz et al. [28] have previously discussed the intricacies in the evaluation of Visual Analytics environments and proposed a practical approach to doing so. More often than not, evaluating visualization tools needs a specialized environment and appropriate data to create specific scenarios to be used during the evaluation. These environments can be physical [28] or virtual [26], through the use of a Cyber Range [3]. In either case, after the decision on the type of environment, datasets and scenarios to be used during the evaluation, a concrete metric needs to be selected for the evaluation of a CSA tool. Munir [21] proposes multiple possible metrics that can be used for the assessment, such as: *timeliness* (of the data present and amount of noise), *accuracy*

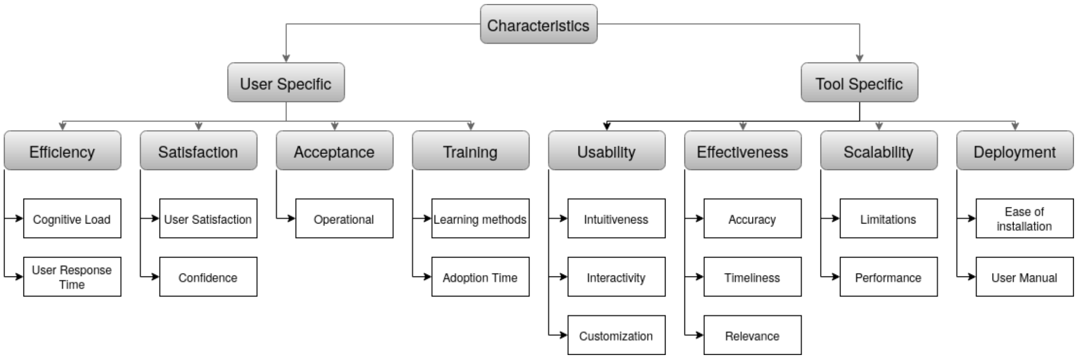

Figure 5: Characteristics used for the evaluation of the design and implementation of visualizations

(objective accuracy of the user's SA) , *trust*, *credibility* (probability of detection vs false alarms), *availability* (of information and systems), *workload* (to be handled by the operator), *cost*, *attention*, *performance* (assesses successful completion of the mission and decision making) and *scope* (single- or multi-level scenario). Each of these metrics evaluates a specific characteristic that is of vital importance to assess the design of a CSA system. Furthermore, a specific assessment technique needs to be chosen, one that is most appropriate for the given situation. There are different approaches, described by Nguyen et al. [24] :

- **Freeze-Probe Techniques** - the evaluation scenario is paused at specific moments and questionnaires are presented to the operator, pertaining to the current state of the environment. The questionnaires are used to evaluate and score the different stages of CSA. A disadvantage of this technique is the significant time needed for the assessment preparation, but offer a subjective way to score the CSA.
- **Real-Time Probe Techniques** - as with freeze-probe techniques, questionnaires are provided to the operator to assess their CSA, but this happens in real-time, without freezing the environment and blanking out the screen.
- **Post-Trial Self-Rating Techniques** - the operator is asked to assess their own level of CSA at the end of the scenario. It is easier to execute, but the highly subjective manner of evaluation can skew the results and is highly dependant on the operator's own subjective measure of their performance.
- **Observer-Rating Techniques** - an expert is tasked to observe the operator's behavior and actions during the scenario and provide a CSA rating. The evaluation is limited as the expert can't observe the internal process of CSA.
- **Performance-based Rating Techniques** - a set of characteristics of performance are determined for specific events during the scenario, recorded during the operator's run and later used to attribute a CSA score. It is important to note that it is assumed that a user's performance corresponds to a good SA, but that not might be true as the performance is linked to the experience and knowledge of the operator.
- **Process Indices-based Rating Techniques** - during the assessment, specific indices are taken into account and recorded. Such indices could be eye or mouse movement and keyboard strokes. These indices are used to calculate the CSA score, but they might not indicate the true level as, for example, looking or clicking on an element on the screen does not indicate it was correctly identified as significant during the scenario.

3.3 Cyber Situation Awareness evaluation methodologies

Evaluating Cyber Situation Awareness involves assessing how intuitively and effectively the visual tools and interfaces facilitate the users to perceive, comprehend, and project information in a cyber environment. Evaluation is vital to ensure that the visualizations do indeed help to enhance the decision-making capabilities of those monitoring and responding to cyber threats. To correctly evaluate not only the CSA level of an operator, but also the validity of a CSA design, a well defined methodology needs to be applied and a formal way to calculate the final result must be employed. The CSA evaluation techniques are presented in Table 1, an expanded version of the one provided by [21], together with a short description of each methodology and how they can be applied for the assessment of CSA.

3.3.1 *Situation Awareness Global Assessment Technique.*

The Situation Awareness global assessment technique (SAGAT) was introduced by Endsley [5][6] and is defined as a Freeze-Probe technique, as explained in Subsection 3.2. A specific scenario is designed beforehand and appropriate queries are prepared to be presented to the operator at specific points. The queries cover elements of the Perception, Comprehension and Projection levels, in order to assess the knowledge and understanding of the current state of the environment. The answers are attributed a score, which reflects whether the operator's level at that specific moment. The SAGAT methodology was first used in the aviation field, but since then has easily been adapted for use in many different domains. The main advantage is the objective and impartial nature of the assessment of all three levels of CSA. The main disadvantages are the time needed to prepare a custom scenario, the inability of operators to prepare for the queries as the freezes are designed to happen at random moments, as well as the fact that the answers rely on memory (the screens are blanked out during the freezes).

Table 1: Extension of the comparison of CSA evaluation methodologies presented by [21]

Metric	SAGAT	SART	SPAM	SABARS	NASA TLX	CDM
Timeliness		✓			✓	✓
Accuracy	✓			✓		✓
Trust		✓		✓		✓
Credibility	✓	✓	✓	✓		✓
Availability				✓		✓
Workload		✓			✓	
Cost	✓			✓		
Attention		✓				
Performance	✓	✓	✓	✓	✓	✓
Scope	✓	✓	✓	✓	✓	✓

3.3.2 Situation Awareness Rating Technique. The Situation Awareness Rating Technique (SART) [32] is a Post-Trial Self-Rating Technique. This implies it is up to the operator to estimate their level of CSA at the end of the scenario. The final score is calculated based on a variety of bipolar scales, reflecting the degree to which operators perceive a demand on their cognitive resources, the supply of resources available and the their understanding of the issues at hand. A comparison between SART and SAGAT [9] shows that contrary to SAGAT, SART is easier and straightforward to implement, as the queries do not need customization depending on the scenario or domain. However, the major disadvantages of SART are related to the inability of operators to correctly assess their own SA as often they cannot know whether the situation was correctly understood. The subjective nature of the scoring often reflects the operator's own belief in their knowledge and capabilities, which may not correctly reflect the reality, basing their ratings on their perceived performance instead on their SA.

3.3.3 Situation Present Assessment Technique. the Situation Present Assessment Technique (SPAM)[8] is a Real-Time Probe Technique. SPAM is comparable to SAGAT, as queries are prepared beforehand relating to a specific events at various stages of a scenario. Contrary to SAGAT, SPAM provides the queries in real time, without interrupting the operational task of the operators. The questions may pertain to the different phases of CSA and also to the past, present and future states of the environment. An extra dimension of scoring is introduced in SPAM, taking into account the time necessary for the operators to answer the queries. The real-time nature of the assessment permits the users to continue with their work and answer when they see fit, but prompt answers would signify better understanding of the information and score better. Another aspect of SPAM that needs to be taken into account is that it might be considered intrusive and hinder the ability of operators to focus on their task, lowering their performance. Moreover, this methodology might, instead of evaluating the actual SA level, measure the ability of operators to look up information as the displays are not blanked when answering the queries.

3.3.4 Situation Awareness Behavioral Rating Scale. The Situation Awareness Behavioral Rating Scale (SABARS) is another methodology originating from the military domain [19][20]. SABARS is an Observer-Rating Technique, relying on expert's observation to evaluate and score the level of CSA. The methodology consists of rating behaviors and actions relevant to developing SA. Such evaluations can effectively distinguish between experienced and inexperienced platoon leaders. The Observer-Rating Technique presents couple of disadvantages- the expert can observe only a limited amount of participants and must adapt to the scale of the evaluation. Second, the expert-observers need to be experienced professionals in the field and be impartial in their scoring. As stated previously in Subsection 3.2 , the expert can't observe the internal process of CSA in the operator and can only make assumptions, bringing an element of subjectivity to the assessment.

3.3.5 NASA Task Load Index. The mental workload of the operator has great impact on their capability of assessing a situation and forming the opinions about the state of the environment. An accepted methodology of calculating the human mental workload is the NASA Task Load Index (NASA TLX) [12], an objective assessment done by the operator, rating their experience with a task on a multi-dimensional scale. The questions answered by the user relate to the mental, physical and temporal demand of the tasks. As with SART, NASA TLX is a Post-Trial Self-Rating Technique, focusing on three very specific aspects of the operator's SA. There have been studies to determine if the methodology is mathematically meaningful [2], providing proof that the methods used for the combination of dimension can be considered meaningless and each aspect should be analyzed separately to get a better understanding of the dimension scores.

3.3.6 CDM. The CDM methodology is based on Cognitive Task Analysis (CTA) [27], a Post-Trial Self-Rating Technique, used for determining the decision points used by the operator during a scenario. This is done by looking at a set of metrics such as goal specification, cue identification, decision expectation, decision confidence, information reliability, information integration, information availability, information completeness, decision alternatives, decision blocking, decision rules and decision analogy [21]. Each of these metrics evaluates a specific characteristic of the decision chain and how that relates to the SA of the operator. The various goals of the operator at different decision points are assessed, together with the information used when formulating the specific decision. The questionnaire also asks the user to rate the confidence they have in the taken decision and if it can be considered as an expected course of action at that time. Alongside those criteria, the user is also prompted to rate the different information dimensions- such as how reliable it was, how well the information was integrated in the undertaken task, the availability of information and its completeness. All these criteria shape the decision of the operator and his performance during the three stages of SA.

4 PRACTICAL APPLICATION OF EVALUATION TECHNIQUES

The application of evaluation methodologies is no trivial task. A lot of effort needs to be spend on the technical side- creating a realistic testing environment which can be explored by the analysts, and on the logistic side where sufficient amount of participants need to take part to gather significant results. Luckily, the nature of the cyber domain facilitates the technical set-up through the use of various technologies such as cyber ranges [3] and online voice/video communications. Indeed, creating a simulated network can be done quickly, populating it with realistic looking background data through the use of frameworks such as GHOSTS [33]. The difficulty comes in the creation of scenarios that mimic real-world attacks, deciding how the CSA level will be evaluated and establishing the correct structure for the assessment. Each technique demands specific preparations be set in place:

- **Probe Techniques** - either Freeze or Real-Time Probe Techniques demand that specific points of time are selected, where the operator can be questioned about the state of the scenario. In the case of Freeze-Probe Techniques the screens need to be blanked and sufficient time given to complete the questionnaire. For the Real-Time Probe Technique, this is not an issue, as the user can fill in the answers when they want. Furthermore, the scenario needs to be designed in such a way that data the operator is looking for can be found through the use of the available tools. The data can have different forms (logs, network packets, executable, etc.) and needs to be generated in such a way to simulate real-world anomalous behavior. Another aspect that needs to be taken into account is if the questionnaire will be presented in text form or a digital form, each having its own set of requirements.
- **Post-Trial Self-Rating Techniques** - a questionnaire needs to be prepared and presented to the operators at the end of the scenario. Contrary to probing techniques, the preparation of these questionnaires takes less time as they are not scenario specific and focus more on the user's personal assessment of specific aspects, be that their level of perception, comprehension and projection, or the level of mental and physical workload.
- **Observer-Rating Techniques** - Experts in the field need to be contacted and introduced to the technologies used, the scenario prepared and the specific states of the environment that need to be detected and analyzed by the operators. The observer needs to have intimate knowledge of the prepared scenario to correctly assess the operator's level of CSA. Indeed, a sufficient amount of time needs to be dedicated to the preparation of the scenario and the familiarization of the expert with it.

4.1 Evaluation Scenarios

In this subsection we will present two example scenarios which can serve for the evaluation of CSA. We choose to present a scenario incorporating an external Advanced Persistent Threat (APT) attack and one using an insider threat due to relevance, complexity and alignment with the roles and responsibilities of a cyber defense analyst.

In the case of the outsider threat scenario, an APT attack involves a sophisticated adversary with a specific target and an extended intrusion period. Public and private organizations often face similar attacks on a daily basis and analysts need to recognize subtle indicator of compromise, such as slow, stealthy intrusions that traditional security measures can miss. Moreover, APT attacks involve multiple phases- reconnaissance, initial compromise, establishing persistence, lateral movement, data collection and exfiltration [35]. By assessing the ability of the operator to identify and understand attack techniques, comprehend the evolving attack landscape and anticipate adversary actions, we can cover the three levels of CSA. On the other hand, an insider threat scenario involves individuals within an organization who exploit their position for malicious purposes. This is pertinent for the evaluation of CSA because insiders often have legitimate access, making them harder to detect compared to external threats. Evaluating CSA in this context assesses whether participants possess the ability to differentiate between normal and anomalous behavior, as well as identify indicators of insider threat. Moreover, an insider threat scenario emphasizes the importance of behavioral analysis. Cyber defense analysts need to understand basic user behaviors and recognize deviations that might indicate malicious intent.

4.2 Technical setup

The technical setup for the proposed scenarios is presented in Figure 6. Both scenarios share similar infrastructure:

- Simulate background traffic through the use of the GHOSTS Framework [33]. This is done to mimic as close as possible a real-world environment and obfuscate the suspicious behavior so it is not easily discernible by the analysts.
- We use a centralised Security Information and Event Management (SIEM) system in the form of the Elasticsearch Kibana Logstash (ELK) stack [30]. This will serve as a centralised repository for the various collected logs and Kibana can be used as a visual tool to analyze the information. Alongside other Analysis tools, the operators will need to go through the data generated in the network to detect any abnormal or malicious activity.
- The state of the environment will be represented by logs generated by a proxy, SNORT [18] and netflow.
- A separate sub-network will host the various analysts, they will have access to the ELK stack and the collection of analysis tools.

4.3 Scenario description

In both scenarios the analysts will focus on the data generated in the simulated network. The major difference is the type of data that will be generated during the scenarios and the specific indicators that the analysts need to detect and analyse for their CSA

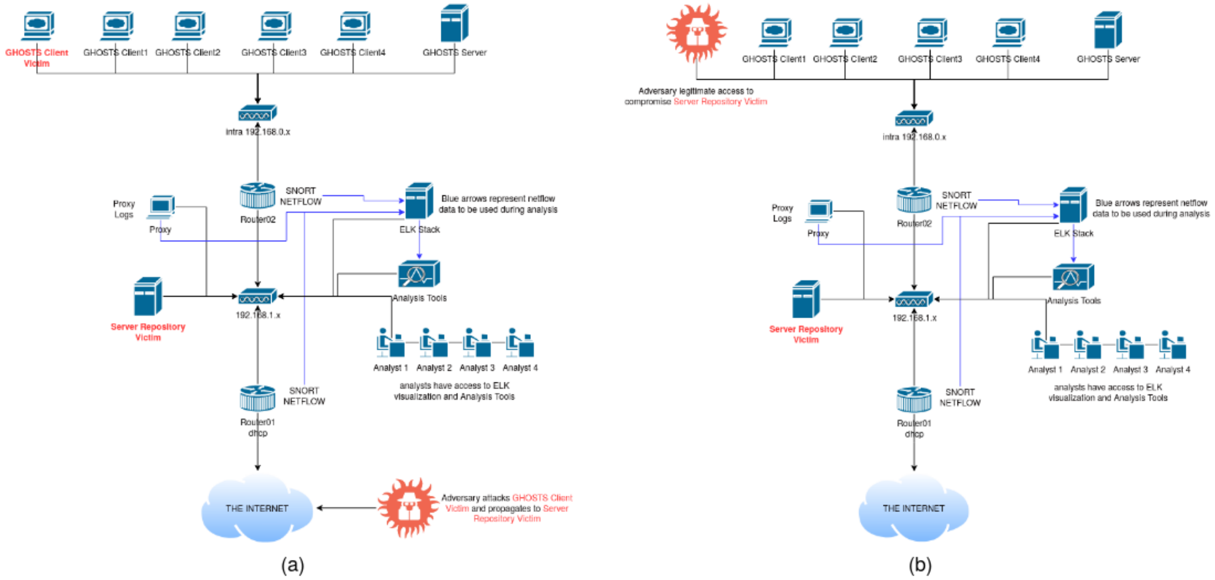

Figure 6: Network topology for (a) external threat scenario and (b) insider threat scenario

assessment. In the exterior threat scenario, the attack will use a phishing attack to get an initial foothold in the network, using a simple malware that will serve as a beacon for the establishment of initial connection to the compromised computer. From there, the attacker will do reconnaissance through the network and look to set up a command & control channel to further solidify their foothold. This can be done by gathering poweruser passwords or compromising the Active Directory and escalating their privileges. Once the permanent connection is set up, the attacker will target important information on the repository server and begin exfiltration.

For the interior threat scenario, suspicious activity by a disgruntled employee will be simulated in the form of chat/email correspondence and access to specific websites. The employee will access data, which they normally should not, in the form of sensitive information about the company and its clients, and start copying the information to external websites such as Dropbox and WeTransfer. Finally, the employee will use an online anonymous platform to blackmail the company, threatening to release the information if they don't pay a sum of money.

4.4 Evaluation Process

To showcase how the evaluation of the CSA level will be accomplished using the two scenarios presented earlier, we will use the SAGAT methodology as an example.

The first step of the evaluation consists of a general introduction of the assessment, explaining the purpose and the objectives that need to be met. Participants will be asked to complete a short form, gathering information about their years of experience, technical skills and occupied position. As SA is highly dependant on the experience and knowledge of the participants, it is essential to establish a baseline, which can be used later in the scoring process. Once the evaluation is underway, each participant will have a personal machine with connection to the data repository and various analysis tools. During the scenario, 5 freezes will occur at randomly determined moments, excluding the first 4 minutes of the trial and keeping enough time between each freeze for the operators to be able to explore the data. The questionnaire for the outsider threat scenario will consist of between 15 and 20 questions, the questionnaire for the insider threat scenario will consist of between 10 and 15 questions. The queries will cover the three levels of CSA, based on what the participants have observed during the time before the freeze. The questions will be presented in a Google Form, occupying the screen of the participant. The queries are dependant on the scenario, which is used for the evaluation. As mentioned previously, the three levels of SA will be evaluated, with questions pertaining to the Perception, Comprehension and Projection phases. To provide concrete examples of the questionnaires that can be presented to the operator, we will go over the three CSA levels for the external threat scenario and for each, present possible questions that can be used:

- Perception
  (1) Multiple questions regarding specific timestamps such as: specific time of received phishing email, execution times of malware, access timestamps for the data repository
  (2) Questions regarding information relating to the phishing email and its attachments
  (3) Questions about discovery of network events (network scanning by the attacker) and activity on the data repository
- Comprehension
  (1) Possible Indicators of Compromise that can be extracted from the attack for future use
  (2) What specific techniques were used for establishing persistence

(3) What type of data was targeted and what exfiltration techniques were used

- Projection

(1) What possible backdoors are still present and how that impacts future security of the environment
(2) Possible proactive measures to mitigate future attacks of this nature
(3) Questions regarding the information that needs to be included in the incident documentation to support future investigations

At the end of the scenario, the questionnaires will be reviewed and each query will be assigned a score depending on its validity- 0 if false and 1 if true. The queries will be grouped depending on which level of CSA they evaluate and will be combined to obtain a final score corresponding to that specific level. Having three separate scores per level will allow the participants to get a clearer idea of the level of CSA they have attained, and where they need to improve.

Each assessment is concluded by a debriefing session consisting of a questions and answers session, discussion about possible events that have been overlooked and collecting feedback to improve the scenario and evaluation in the future.

5 FUTURE WORK

In this paper we have shown the evolution of the Situation Awareness methodology and how its concepts haven been applied to the cyber domain. Much effort has been dedicated to defining a robust taxonomy and hierarchy for the CSA, as shown in Section 2.2 and 2.3. Further, we have shown that there is a strong connection between the visualization tools used and the level of CSA of the analyst in Section 3.1. The field of Situation Awareness design offers many possibilities in the research and development of new visualization tools with strong emphasis on the enhancement of the different levels of SA. The topic of human-machine interaction is more relevant than ever with the introduction of new AI-powered tools and visualization capabilities. We have presented a set of characteristics that can be used for the assessment of the validity and effectiveness of visualizations and how they enhance this interaction. The graphical representation of the complex nature of the cyber environment is a quickly developing field of Visual Analytics and the need for a robust framework for the evaluation of these graphical representations is needed.

Alongside the need for the evaluation of visualizations, we have shown that the theory of SA evaluation has been well established, but sorely lacks transition from the theoretical field to practical application. We have shown theoretical comparison of various evaluation techniques and intend in the future to organise practical evaluations as described in Section 4. The field of CSA will greatly benefit by concrete examples of the application of the various CSA evaluation techniques to real world scenarios and the evaluation and scoring techniques used throughout those assessments.

6 CONCLUSION

Situation awareness has been a popular topic since its introduction in the aviation sector and is a principle that can be applied to a variety of domains. Specifically for the cyber domain, it has become more and more prevalent as attacks on organizations have multiplied and become more sophisticated. Quickly identifying and analyzing malicious activities in the network is paramount for the continuous functionality of any system. In this paper we have presented an overview of the various ways the concept of Situation Awareness has been extended to apply to the cyber domain. Nevertheless, the application of SA comes with new problems- the large quantities of data generated daily by the infrastructures, the dynamic nature of the cyber environment and the difficulty of designing tools for analysis. Analysts have to rely on information gathered by sensors instead of their own senses and the quantity of information to be analyzed can quickly overwhelm them if they don’t use appropriately powerful visualization tools. To enhance the human-machine interaction and aid the analyst to better understand and manage their environment, new visualizations need to be designed with CSA in mind. The validity of these new visualizations needs to be proven by assessing their effectiveness, usefulness and usability.

In this paper we presented a selection of techniques for the assessment of the Cyber Situation Awareness, together with a proposed practical applications on how to use these techniques. Numerous approaches are available to evaluate an operator’s level of CSA. Nevertheless, there is a dearth of comparisons between these approaches and no established means to ascertain the most suitable methodology for a particular situation. The state of the art mainly focuses on the theory or the application of SA for the military domain, but rarely on how to apply these methodologies for the cyber domain. We are confident that the proposed evaluation framework can serve to evaluate not only the validity of CSA assessment techniques, but also correctly score the level of CSA of the participants. In the future, we aim to deploy the proposed scenarios in a cyber range environment and apply the assessment methodologies we presented to gather practical results from the participants and continue our comparison with real-world data.

REFERENCES

[1] Hooman Alavizadeh, Julian Jang-Jaccard, Simon Yusuf Enoch, Harith Al-Sahaf, Ian Welch, Seyit A Camtepe, and Dong Seong Kim. 2021. A Survey on Threat Situation Awareness Systems: Framework, Techniques, and Insights. *arXiv preprint arXiv:2110.15747* (2021).

[2] Matthew L. Bolton, Elliot Biltekoff, and Laura Humphrey. 2023. The Mathematical Meaninglessness of the NASA Task Load Index: A Level of Measurement Analysis. *IEEE Transactions on Human-Machine Systems* 53, 3 (2023), 590–599. https://doi.org/10.1109/THMS.2023.3263482

[3] Thibault Debatty and Wim Mees. 2019. Building a cyber range for training cyberdefense situation awareness. In *2019 International Conference on Military Communications and Information Systems (ICMCIS)*. IEEE, 1–6.

[4] A D'Amico. 2011. Visual Analytics for Cyber Defense Decision-Making. *Washington, USA., VAC* (2011).

[5] Mica R Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*. IEEE, 789–795.

[6] Mica R Endsley. 2017. Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational awareness*. Routledge, 129–156.

[7] Mica R Endsley. 2017. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*. Routledge, 9–42.

[8] Mica R Endsley. 2021. A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. *Human factors* 63, 1 (2021), 124–150.

[9] Mica R Endsley, Stephen J Selcon, Thomas D Hardiman, and Darryl G Croft. 1998. A comparative analysis of SAGAT and SART for evaluations of situation awareness. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 42. Sage Publications Sage CA: Los Angeles, CA, 82–86.

[10] Antti Evesti, Teemu Kanstrén, and Tapio Frantti. 2017. Cybersecurity situational awareness taxonomy. In *2017 international conference on cyber situational awareness, data analytics and assessment (Cyber SA)*. IEEE, 1–8.

[11] Richard D Gilson. 1995. Special issue preface. *Human factors* 37, 1 (1995), 3–4.

[12] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[13] Changlin He and Yufen Li. 2017. Survey of network security situation awareness. In *2017 International Conference on Computational Science and Engineering (ICCSE 2017)*. Atlantis Press, 136–141.

[14] Jeffrey Heer and Ben Shneiderman. 2012. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue* 10, 2 (2012), 30–55.

[15] Martin Husák, Tomáš Jirsík, and Shanchieh Jay Yang. 2020. SoK: Contemporary issues and challenges to enable cyber situational awareness for network security. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*. 1–10.

[16] Liuyue Jiang, Asangi Jayatilaka, Mehwish Nasim, Marthie Grobler, Mansooreh Zahedi, and M Ali Babar. 2022. Systematic literature review on cyber situational awareness visualizations. *IEEE Access* 10 (2022), 57525–57554.

[17] H Günes Kayacik, A Nur Zincir-Heywood, and Malcolm I Heywood. 2005. Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets. In *Proceedings of the third annual conference on privacy, security and trust*, Vol. 94. Citeseer, 1723–1722.

[18] Jack Koziol. 2003. *Intrusion detection with Snort*. Sams Publishing.

[19] Michael D Matthews and Scott A Beal. 2002. Assessing situation awareness in field training exercises. *US Army Research Institute for the Behavioral and Social Sciences* 31 (2002).

[20] Michael D Matthews, Jarle Eid, Bjorn Helge Johnsen, and Ole Christian Boe. 2011. A comparison of expert ratings and self-assessments of situation awareness during a combat fatigue course. *Military Psychology* 23, 2 (2011), 125–136.

[21] Arslan Munir, Alexander Aved, and Erik Blasch. 2022. Situational awareness: techniques, challenges, and prospects. *AI* 3, 1 (2022), 55–77.

[22] Sowmya Myneni, Kritshekhar Jha, Abdulhakim Sabur, Garima Agrawal, Yuli Deng, Ankur Chowdhary, and Dijiang Huang. 2023. Unraveled—A semi-synthetic dataset for Advanced Persistent Threats. *Computer Networks* 227 (2023), 109688.

[23] NATO 2020. *IST Research Group on "Visual Analytics for Complex Systems"*. Retrieved June 10, 2024 from https://www.sto.nato.int/search/Pages/activities_results.aspx?k=IST-184&s=Search%20Activities

[24] Thanh Nguyen, Chee Peng Lim, Ngoc Duy Nguyen, Lee Gordon-Brown, and Saeid Nahavandi. 2019. A review of situation awareness assessment approaches in aviation environments. *IEEE Systems Journal* 13, 3 (2019), 3590–3603.

[25] Jakob Nielsen. 1994. Usability inspection methods. In *Conference companion on Human factors in computing systems*. 413–414.

[26] Georgi Nikolo and Wim Mees. [n. d.]. Evaluation through deployment of the Multi-agent System For Advanced Persistent Threat Detection framework in a Cyber Range environment. ([n. d.]).

[27] DAVID O'HARE, Mark Wiggins, Anthony Williams, and William Wong. 1998. Cognitive task analyses for decision centred design and training. *Ergonomics* 41, 11 (1998), 1698–1718.

[28] Jean Scholtz, Catherine Plaisant, Mark Whiting, and Georges Grinstein. 2014. Evaluation of visual analytics environments: The road to the Visual Analytics Science and Technology challenge evaluation methodology. *Information Visualization* 13, 4 (2014), 326–335.

[29] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. 2018. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* 1 (2018), 108–116.

[30] Sung Jun Son and Youngmi Kwon. 2017. Performance of ELK stack and commercial system in security log analysis. In *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*. IEEE, 187–190.

[31] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. 2009. A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee, 1–6.

[32] Richard M Taylor. 2017. Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational awareness*. Routledge, 111–128.

[33] Dustin D Updyke, Geoffrey B Dobson, Thomas G Podnar, Luke J Osterritter, Benjamin L Earl, and Adam D Cerini. 2018. *GHOSTS in the Machine: A Framework for Cyber-Warfare Exercise NPC Simulation*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA.

[34] Margaret Varga, Carsten Winkelholz, and Susan Traber-Burdin. 2016. Cyber situation awareness. *NATO/OTAN (STO-MP-IST-148)* (2016).

[35] Tarun Yadav and Arvind Mallari Rao. 2015. Technical aspects of cyber kill chain. In *Security in Computing and Communications: Third International Symposium, SSCC 2015, Kochi, India, August 10-13, 2015. Proceedings 3*. Springer, 438–452.