The Enhanced Security in Process System - Evaluating Knowledge Assistance

Anna-Pia Lohfink* Vera M. Memmesheimer† Frederike Gartzky‡ Christoph Garth§
Technische Universität Kaiserslautern, Germany

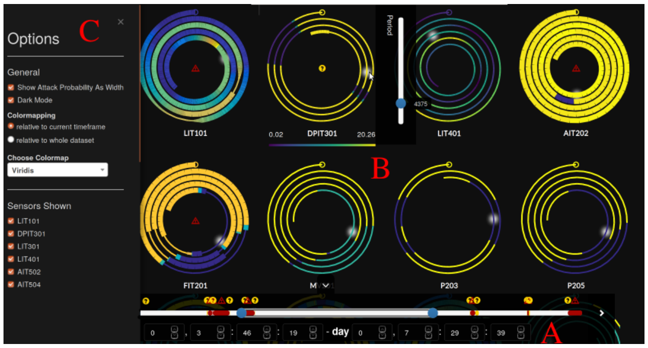
Abstract

We present evaluation results of our enhancements to the Security in Process System [14] developed by Lohfink et al. to support triage analysis in operational technology networks. To ensure fast and appropriate reactions to anomalies in device readings, this system communicates anomaly detection results and device readings to incorporate human expertise and experience. It exploits periodical behavior in the data combining spiral plots with results from anomaly detection. To support decisions, increase trust, and support cooperation in the system we enhanced it to be knowledge-assisted. A central knowledge base allows sharing knowledge between users and support during analysis. It consists of an ontology describing incidents, and a data base holding collections of exemplary sensor readings with annotations and visualization parameters. Related knowledge is proposed automatically and incorporated directly in the visualization to provide assistance that is closely coupled to the application, without additional hurdles. This integration is designed aiming on additional support for correct and fast detection of anomalies in the visualized device readings. We evaluate our enhancements to the Security in Process System in terms of effectiveness, efficiency, user satisfaction, and cognitive load with a detailed user study. Comparing the original and enhanced system, we are able to draw conclusions as to how our design narrows the knowledge gap between professional analysts and laymen. Furthermore, we present and discuss the results and impact on our future research.

Index Terms: Security and privacy Intrusion—anomaly detection and malware mitigation; Human-centered computing—Visualization systems and tools

1 Introduction

Modernization and Industry 4.0 lead to the connection of operational technology (OT) and information technology (IT) networks that were separated before. Without this physical separation, and equipped with modern, high level components, OT networks become vulnerable to attacks, especially since they are often less secured than deemed appropriate for home and office IT [8]. Efforts and difficulties that need to be overcome when securing OT networks lead to recent research in cyber security with the aim to detect attacks in available information, such as sensor and actuator readings during production. To ensure short reaction times in spite of the large amount of data that needs to be analyzed, automated anomaly detection algorithms that often incorporate machine learning approaches are applied. While being fast, anomaly detection algorithms are subject to uncertainty, requiring to incorporate human experience and expertise in the alarm chain. This approach is followed and visually supported by Lohfink et al. with the Security in Process System (SiP) [14].



Options ×

General
☑ Show Attack Probability As Width
☑ Dark Mode

Colormapping
● relative to current timeframe
○ relative to whole dataset

Choose Colormap
Viridis

Sensors Shown
☑ LIT101
☑ DPIT301
☑ LIT301
☑ LIT401
☑ AIT502
☑ AIT504

Figure 1: **The Security in Process System**: the three main parts are the time slider A, the spiral chart B and the options panel C. The time slider provides overview over the complete data set and indicates areas with warnings and alerts. In the spiral chart, readings and abnormality ratings of devices are shown by color and line thickness respectively. Reproduced from [14].

The SiP System was developed to be used by professional analysts and laymen. To increase trust in its machine learning components and diminish knowledge differences between different analysts, we incorporated knowledge assistance in the system. Furthermore, our enhancements render collaboration between analysts possible and facilitate knowledge transfer from professional analysts to laymen.

We realized these enhancements based on the Knowledge Rocks Framework [13] and integrated a knowledge base in SiP to provide easily accessible support for analysts. As part of this knowledge base, we defined an ontology with callback functions that is able to classify detected anomalies automatically using procedural and machine learning approaches. Based on this classification, stored instances of anomalies from the knowledge base are proposed to the analyst: stored readings are shown using the same visual encoding as used for the readings of the currently analyzed data set in the enhanced time slider. Selecting a proposed instance, its annotations from the knowledge base are shown and its readings are added to the spiral plots of all contained sensors, by combining the spiral plots with a stream graph.

To evaluate our design of the knowledge-assisted, enhanced SiP (eSiP), we performed a detailed user study, focusing on the comparison of SiP with, and without knowledge assistance. The results of our study show that using eSiP, more anomalies and false positives were identified correctly. We furthermore observed a learning effect in terms of task completion time for correct responses in both systems. Overall, SiP with assistance turned out to be the preferred system and to decrease discrepancies in effectiveness and satisfaction between users with and without experience in visual analytics. The improved satisfaction and success rates suggest that the system is successfully supporting laymen in cyber security. Considering these results, we believe that knowledge assistance will increase user acceptance as well as trust in the system.

*e-mail: lohfink@cs.uni-kl.de
†e-mail: memmesheimer@cs.uni-kl.de
‡e-mail: gartzky@cs.uni-kl.de
§e-mail: garth@cs.uni-kl.de

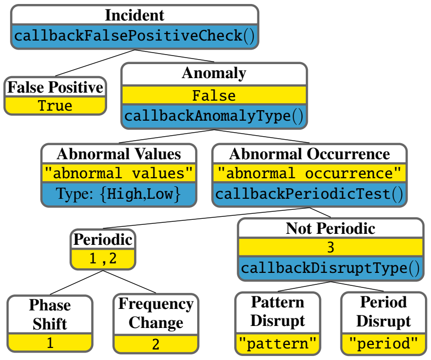


Figure 2: The **incident based ontology** in eSiP. A detailed description of classes and callback functions is given in the supplemental material. Reproduced from [13].

2 Related Work

Knowledge-assisted visualization aims to incorporate knowledge into the visualization process to support users [16]. Its utility is well understood: already in the late 90's, Fujishiro et al. developed the GADGET system that supports users in choosing suitable visualization systems for their goals under given constraints (e.g. data properties) [5]. Examples from cyber security visualization are KAMAS, developed by Wagner et al. [20], and the QCAT analytics system by Walton et al. [21]. KAMAS is a knowledge-assisted visualization system for behavior-based malware analysis. Its knowledge base consists of behavior patterns that are extracted from recorded data. The QCAT analytics system aims on anomaly detection using Queries with Conditional Attributes (QCATs), which are stored in the knowledge base. Both systems have a similar structure and aim on collaboration. Their knowledge bases can be browsed, the contents applied, changed and new defined by the user.

An analysis of further examples for knowledge-assisted cyber security was given by Böhm et al. [1]. Their full list of analyzed papers and results is available at `http://bit.ly/2LgrG3p`. Böhm et al. found, that collaboration tools are up to now rarely found in cyber security. With our system, we contribute to closing this gap.

For related work concerning SiP, that is concerning time series visualization, anomaly detection in time series and detection and visualization of periodic behavior, we refer to the Security in Process paper [14]. Backgrounds and related work on knowledge-assisted visualization, underlying models, its concrete implementation in systems, and the use of ontologies in these systems are given in the Knowledge Rocks paper [13].

2.1 The Security in Process System

SiP visually supports triage analysis in OT networks. Since in OT-networks, components and used protocols oftentimes don't allow the implementation of security measures, these need to be implemented based on existing information –that is readings of sensors and actuators (devices) in the system. Using these readings, SiP scans for abnormal behavior (incidents) using machine learning tools. Three approaches for anomaly detection based on the periodic nature of readings in OT-networks were evaluated by Duque Anton et al. [4], namely one-class support vector machines, isolation forests and matrix profiles. Matrix profiles turned out to be the most suitable approach due to their robustness to different kinds of data, their applicability without the requirement of extensive training and the obtained results. They have been presented by Yeh et al. [22].

The result of the automated anomaly detection is an anomaly score for the readings of every device. This score represents the conformity of the current readings with previous readings and is categorized as "normal", "warning" and "alert" based on its height and adaptable, application dependent boundaries. The anomaly score and the readings are then visualized to allow the analyst to revise them and react accordingly. To exploit the periodical behavior of industrial processes, spiral plots encode the readings as color. The abnormality rating is given as line thickness, where thin lines represent normal behavior (spiral plots in Figure 1B). Users can select the shown time frame in the time slider and change the period of the individual spiral plots (Figure 1A). The time slider represents the complete data set, and reflects increased anomaly ratings of intervals with warning ? and alert ⚠ signs.

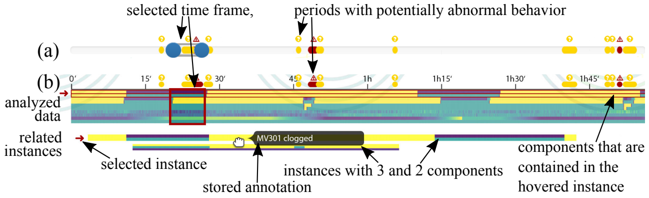


Figure 3: **Original** (a) **and enhanced time slider** (b): combining the time slider functionality with a linegraph, the time slider now gives an overview of the readings of all devices and realted instances from the database. Reproduced from [13].

The used data set contains readings from PLCs monitoring a modern water treatment process. Polluted water is treated in six stages and checked by different sensors until it is clean or re-enters the process. The readings of the system behavior under different attacks and in normal mode are provided by the iTrust, Centre for Research in Cyber Security, SUTD [9, 15].

3 eSiP - Enhancing SiP with Knowledge Assistance

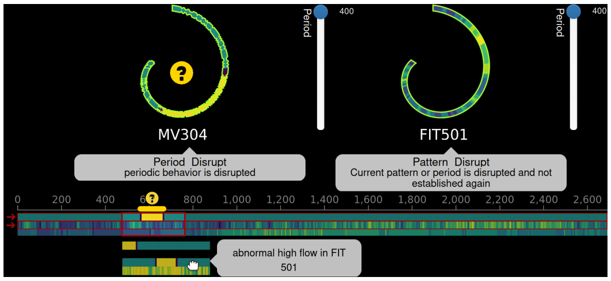
To increase trust in the SiP System, we enhanced it with a knowledge base to amplify the support for anomaly detection and render collaboration possible. To do so, we used the Knowledge Rocks framework and created a knowledge base that consists of an ontology and a connected database where events of interest are stored with their classification. eSiP was implemented using D3.js [2] and python 3, framed by bottle [7]. All details on the system design and enhancements are given in the Knowledge Rocks Paper [13].

3.1 Implementation of the Knowledge Rocks Framework

Since SiP deals with incidents, we defined the ontology part of the knowledge base accordingly (Figure 2). Coupling the ontology with callback functions, eSiP is able to classify input data automatically, and to suggest fitting instances from the database. This is done based on the classification of the instances and the distance between stored instances and the analyzed data calculated based (like the anomaly detection) on the matrix profile method. A detailed description of the ontology classes and callback functions, that allow the traversal of the ontology, are given in the supplemental material. To be able to classify incidents, procedural as well as machine learning approaches (namely isolation forests) are employed.

To integrate the knowledge base in SiP, we enhanced the time slider, the spiral plots, and added an ontology visualization. Possible workflows are illustrated in the supplemental material.

The Enhanced Time Slider. We combined the time slider functionality with a linegraph as described by Kincaid et al. [12]; a comparison of the original and enhanced time slider is given in Figure 3. The linegraph shows the current data set, with highlighted intervals of increased anomaly rating as in the original time slider. The additional information provides an overview of the complete data set and thus increases reproducibility of the anomaly detection results that rely on pattern matching in the data. Optionally, the devices' readings are clustered such that similar patterns are



Period Disrupt
periodic behavior is disrupted

Pattern Disrupt
Current pattern or period is disrupted and not established again

abnormal high flow in FIT 501

Figure 4: **Hovering an instance**: contained devices are highlighted in the time slider and the spiral chart. Annotations are shown for contained devices and the instance.

adjacent in the time slider, adapting the order in the spiral chart accordingly. Similar to the original version of the time slider, the selection frame –indicating the interval that is currently shown in the spiral plots– can be re-positioned via drag and drop and its width can be changed using handles.

Below the time slider, following its visual encoding and device order, related instances from the database are suggested, based on the classification of the currently analyzed data. These instances can be selected via mouse click and moved relative to the analyzed data via drag and drop. The initial position of every instance is the position resulting from the minimum distance between data and instance using matrix profiles. Hovering a suggested instance, contained devices are highlighted in the time slider and spiral charts, and stored annotations are shown (Figure 4). Holding *Alt* when selecting a related instance, visualization settings from the database are prescribed. The stored settings are period, color map, and color map reference frame.

The Enhanced Spiral Plot. We enhanced the spiral chart –showing the selected interval of the data set– to additionally contain selected instances. To do so, we combined the spiral plot with a stream graph (Figure 5a,b) similar to the approach by Jiang et al. [10]. The analyzed data is the innermost spiral and the line thickness of data and instances (indicating the anomaly score) accumulates. At the beginning of a selected instance, a handle is shown that allows re-positioning of the instance relative to the analyzed data, similar to dragging the instance below the time slider. Selected instances have the same period as the analyzed data to prevent misinterpretation.

Identifying abnormal high or low values was a challenging task in SiP: only professional analysts, knowing the normal range of sensors, could identify them without relying on an abnormality in the pattern. In eSiP, a classification as “Abnormal Values” triggers a visual cue; the interval containing the abnormal values is highlighted (Figure 5c).

To avoid overlapping between spiral twists if multiple instances are selected or the anomaly rating is high, we limited the thickness of each strand. If necessary, this limit is reduced automatically.

Ontology Visualization. Visualizing the ontology supports the user in revising the automated classification and (re-)classifying data if necessary. In addition, it provides a common vocabulary to talk about incidents, decreasing the knowledge gap between professional analysts and laymen. We implemented a basic ontology visualization in a browsable tree layout, similar to Figure 2.

3.2 Increasing Reproducibility and Trust

Both –the original SiP and the eSiP– use machine learning approaches for anomaly detection and incident classification respectively. The design of our enhancements to the SiP System aim on decision support and generating trust towards the results of these approaches. The enhanced time slider provides a holistic overview over the analyzed data and hence the base for the automated anomaly detection, making its results more comprehensible and justified for the user. The enhanced spiral plot communicates the results of both, the automated classification via visual cues and annotations, and the anomaly detection via alert and warning signs and line thickness. The detailed comparison of stored incidents with the current readings in the enhanced spiral plot aims on increasing trust in anomaly detection results (if they fit the stored results), or the overall system by providing decision support (opposing the anomaly detection result if required). Providing direct access to the acting ontology, its classes and callback functions and their documentation increases the reproducibility of classification results. Furthermore, our design aims on providing knowledge of professional analysts to laymen, for example via the terminology and structure of the ontology and example results from analysis performed by experts. Doing so, we hope to decrease the discrepancy in terms of effectiveness between professional analysts and laymen, resulting in a higher over all effectiveness and rendering effective collaboration possible.

3.3 Collaboration

Our system supports distributed and asynchronous collaboration, that is the analysts are neither required to be in the same location, nor to interact at the same time. With regard to the map of groupware options presented by Johansen [11], this corresponds to the *different time/different place* category. Using our system, all interaction takes place via stored incidents and annotations in the knowledge base. This enables laymen to benefit from expertise and experience of professional analysts, narrowing the knowledge gap between them. The knowledge base holds exemplary data for every stored incident. This allows everybody to comprehend the reasoning behind the classification of an incident and enables people to discover similar patterns independently. As a consequence, the knowledge base is a valuable collection of examples that can also be used for educational purposes. In our evaluation, we focus on the benefits individual users gain from the knowledge base. We chose this focus since the system is currently not in use, resulting in a limited knowledge base and a lack of operators that could take part in our user study. Together with an evaluation of the operational system, increasing the possibilities for direct interaction between users will be part of our future research.

4 Evaluating Knowledge Assistance

4.1 Experimental Design and Procedure

We performed an online user study with 15 participants (6 female, 9 male, age 24 - 44) to evaluate the usability of eSiP. Since the system aims to close the knowledge gap between professional analysts and laymen in cyber security, we compare the performance of laymen with and without knowledge assistance i.e., evaluating the impact of the ontology and knowledge base that was established with the help of professional analysts.

We conducted our study with a heterogeneous group of participants: 13 participants reported to have a technical background (IT/electrical engineering). We furthermore recorded the number of participants having experience in visual analytics (7), with spiral plots (5), and with SiP (3). We consider the effective, efficient, and satisfying individual interaction with the system to be a necessary requirement for successful collaboration in a heterogeneous group of system users. To address this first step, the present study investigates system usage on an individual level.

A within-subject design was used to evaluate how system usability is affected by the knowledge assistance incorporated in eSiP: Each participant completed the tasks T1 - T3 with (eSiP) and without (SiP) knowledge assistance. Tasks T4 - T5 require features that are new in the eSiP and were hence not performed in SiP. We provide details on the tasks in Section 4.2. To avoid learning effects, the order of the systems was assigned randomly. In total, 9 participants started with SiP and 6 with eSiP. Prior to each task, we provided an

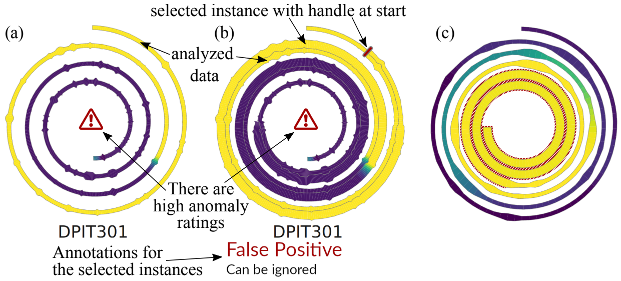


Figure 5: **Original** (a) **and enhanced spiral plot** (b): selecting an instance, it is added to the spiral plot - stream graph combination. Else, it remains as in (a). (c) **Abnormal high values** are highlighted.

explanatory video to introduce the system and establish a common knowledge basis. Each video had to be watched at least once and could be replayed throughout the completion of the task.

In line with ISO 9241 - 11, we assessed usability via effectiveness, efficiency, and satisfaction. To this end, we measured success rates and task completion times (TCTs) throughout the experiment. Furthermore, we posed questionnaires about satisfaction and cognitive load after the completion of the tasks with SiP and eSiP respectively. At the end of the experiment, we asked for the preferred system.

4.2 Tasks

🟦**T1 Risk evaluation.** In this task, we presented readings rated as warnings by the automated anomaly detection. The participants were asked to decide whether the readings are abnormal or normal in four sub-tasks, where three contained abnormal, and one normal sensor readings. Using eSiP, proposed incidents from the knowledge base support this decision by providing knowledge of professional analysts on comparable readings, aiming on an increase in trust in the system.

🟪**T2 Alert revision.** The participants were asked to check readings that were rated as attacks for abnormal behavior, and identify each example as true (two examples) or false positive (one example). Again, stored incidents from the database could be selected using eSiP, supporting the decision.

🟧**T3 Determining periods.** The participants were asked to use the period slider to find the period of the shown sensor readings. Three different examples were used in this task: one example with one peak per period and two examples with two peaks per period. Using eSiP, participants could apply stored periods to the spiral plots. Since stored periods have been determined by humans, they are likely to be more accurate than the automatically detected period that is proposed by the system. Hence, if a matching incident is stored in the knowledge base, this incident is on the one hand able to increase trust in the system by proposing appropriate period and visualization settings. On the other hand, the patterns that occur when the correct period is set in the spiral plot reassure the user in the assessment of the correspondence between the current readings and the selected incident.

🟩**T4 Spotting irregularities.** To evaluate the enhanced time slider, we presented a time slider containing an anomaly and asked the participants to enter the start time of a suspicious pattern. The required information to do so are are not contained in the original time slider, but only in the enhanced time slider. Hence, this task is only performed with eSiP.

🟨**T5 Incident classification.** In this task, incidents in the sensor readings had to be classified according to the ontology. A detailed knowledge of the ontology that is used for classification and proposing incidents, increases the reproducibility of automated choices by the system, and thus trust in its decisions. To this end, we provided four out of five enhanced spiral plots representing a pattern disrupt, period disrupt, phase shift, frequency shift, and abnormal high values along with a picture of the ontology and background information on frequency and phase of signals.

4.3 Implementation of the Online Evaluation Tool

Due to the COVID-19 pandemic, our study was carried out remotely; participants independently navigated through our web-based evaluation tool that guided them through different tasks that were solved using SiP or eSiP. In the beginning, they were asked to give personal information and watch a video explaining the purpose of the system. Each participant used both, SiP and eSiP whereby the first system was assigned randomly. After watching the explanatory video at least once, the participant could click a start button to see the task description and start the timer. Depending on the task, the participants had to provide their answer via single choice (🟦T1, 🟪T2, 🟨T5) or via entering a number (🟧T3, 🟩T4). Clicking a submit button stopped the timer for the respective example and either immediately started a new timer while showing the next example, or lead the participant to the next task. For each participant, the collected data was saved in json files on the server (see supplemental material).

4.4 The Knowledge Base

As described above, the knowledge base consists of the acting ontology and a database that holds instances of data that contain different incidents. The acting ontology was developed in cooperation with professional analysts and will hardly change in the operational phase. The content of the data base on the other hand is expected to change and grow. We faced the challenge, how to populate the data base with incidents that can be used in the user study, on a task-driven base. For every task that is executed with eSiP, we added incidents to the knowledge base that represent a spectrum of possible related incidents, specifically for the data used in the task. Like this, we avoided influences of the automated classification on the evaluation results (which also incorporates the influence of trained ML approaches). Further, this decreased the complexity of required operations and hence prevented waiting times during the evaluation, even if many people evaluated the system at once.

4.5 Results

Effectiveness. We assessed effectiveness in terms of success rates. To this end, we measured whether participants selected the correct response option in tasks 🟦T1, 🟪T2, and 🟨T5. For 🟧T3 we accepted answers in a range of 500s; for 🟩T4 the frame for correct responses was 960s wide. Both ranges were determined example based.

Comparing eSiP and SiP, higher success rates were found for eSiP in 🟦T1 and 🟪T2. As shown in Figure 6a, we found that eSiP especially supports the correct detection of anomalies and false positives, which is the main purpose of both systems. Using eSiP, the participants identified more presented anomalies (64%) than with SiP (47%). In 🟪T2, the false positive example was identified by 67% of the participants using eSiP but only by 13% using SiP. Concerning 🟧T3, both systems performed equally well in terms of success rates. In total, 89% of the periods were determined correctly, affirming the suitability of spiral plots for periodic behavior.

Decreasing the discrepancy in effectiveness between visual analytics experts and novices is an important requirement for effective collaboration between them. Comparing participants with (7) and without (8) experience in visual analytics, we found that the absolute difference between success rates of the two groups was lower for eSiP than for SiP (Figure 6c). Hence, our enhancements of the system helped aligning the effectiveness for participants with and without experience in visual analytics.

🟩T4 and 🟨T5 were only performed with eSiP. We found a particularly high success rate (93%) in 🟩T4, showing that anomalies can be detected effectively with the additional information provided by the enhanced time slider. Using our ontology, 31% of the pattern

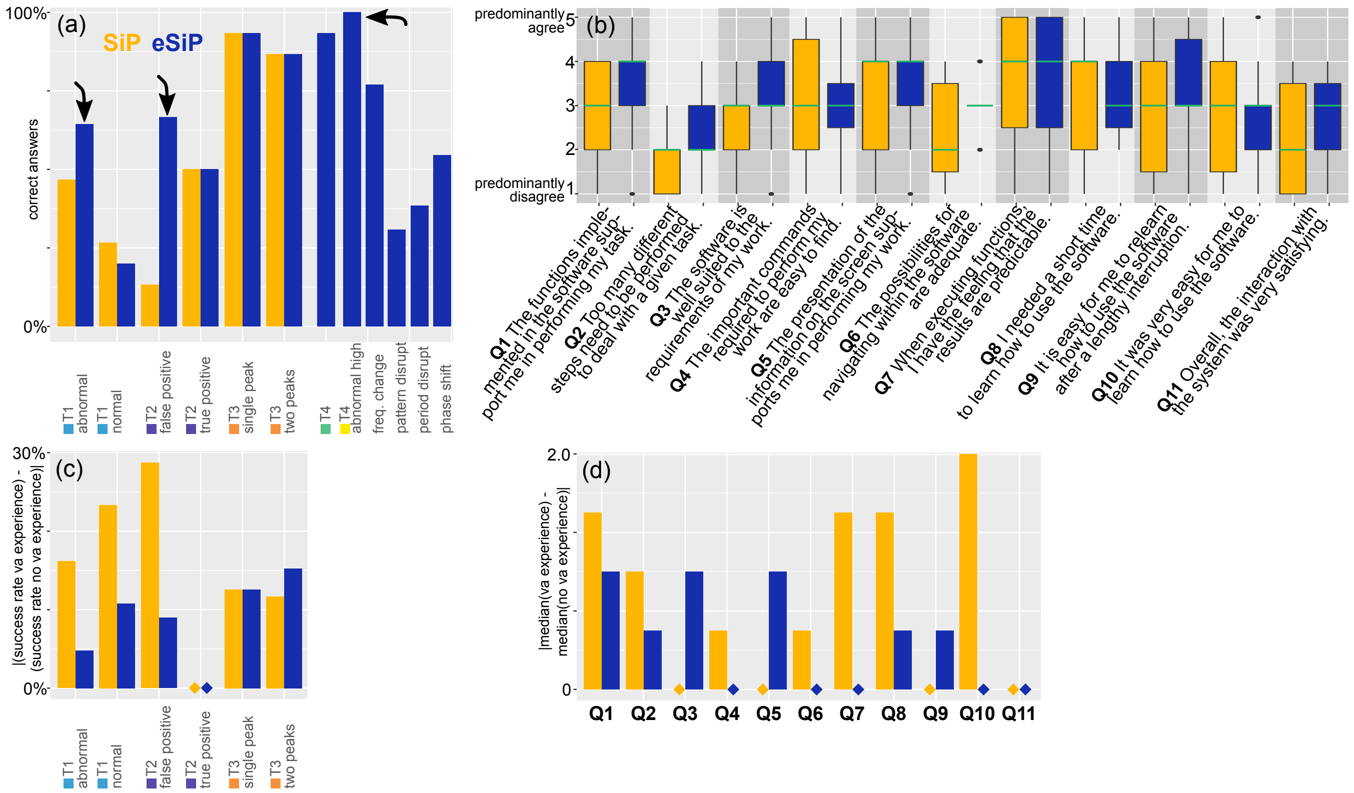


Figure 6: Evaluation results for **effectiveness** (a) assessed via success rates and **satisfaction** (b) assessed via questionnaire: Using eSiP more abnormalities and false positives were detected than with SiP; all examples containing abnormal high values were classified correctly with the visual cues in the enhanced spiral plot (a). The results in (b) (quartiles, median (green), data range and outliers) show that eSiP was rated to have more supportive functions and possibilities to navigate, to be better suited for the task, easier to relearn, and more satisfying to interact with. **Absolute discrepancy in effectiveness** (c) and **satisfaction** (d) between participants with and without experience in visual analytics. For the majority of tasks and questions the absolute difference between success rates (a) and median questionnaire ratings (b) of experienced and unexperienced users is lower when using eSiP compared to SiP.

and 38% of the period disruptions were classified correctly. Phase shifts (55%) and frequency changes (77%) were identified more often. 100% of the examples representing abnormal high values were classified correctly, demonstrating the high effectiveness of the visual cue in the enhanced spiral plot.

Efficiency. Throughout the experiment we recorded TCTs to assess efficiency. We measured a median TCT of 45s for 🟩T4 and of 173s for completing the four examples in 🟨T5. To compare SiP and eSiP, we considered the TCT of the fastest correct response for 🟦T1, 🟪T2, and 🟧T3. Comparing the respective TCTs for each participant, we did not find that one of the two systems generally leads to a faster completion of the tasks. Instead, we found that most participants were able to complete 🟦T1 (80%) and 🟧T3 (93%) faster with the second system (SiP or eSiP, depending on starting condition). For these participants, we measured an average TCT decrease of 52% in 🟦T1 and of 43% in 🟧T3 when comparing the fastest correct responses using the first to using the second system. We interpret this decrease in TCT as increasing trust and high learnability of the system.

Satisfaction and Cognitive Load. To assess and compare the user satisfaction of SiP and eSiP, the participants filled a questionnaire and selected their preferred system. Similar to [14], they should indicate their agreement with 11 statements on a five-level Likert scale (predominantly disagree (1) - predominantly agree (5)) about the systems suitability for the tasks, controllability, conformity with user expectations, learnability, and overall satisfaction.

As presented in Figure 6b, the functions offered by eSiP were rated to be more supportive (Q1) and the software to be more suited for task performance (Q3). Similarly, the navigation in eSiP was rated better than in SiP (Q6). Both systems received relatively low scores in Q2, indicating that the number of steps to perform a task is adequate. We believe that the slightly higher score assigned to eSiP reflects the available knowledge assistance functions. Both systems received especially high scores concerning the predictability of results (Q7). It is easier for the participants to relearn eSiP (Q9) and the overall interaction with eSiP was rated more satisfying (Q11). In total, 80% of the participants preferred eSiP over SiP, affirming that knowledge assistance is indeed helpful.

We observed that for most questions the absolute discrepancy between the median rating of participants with and without experience in visual analytics was lower for eSiP than for SiP (Figure 6d). Again, we interpret this decreasing discrepancy (i.e., a more equal level of satisfaction) as a result of the knowledge assistance features incorporated in eSiP.

Since cognitive overload could impede the fast and correct identification of attacks during extended periods of usage, we measured cognitive load via the NASA-task load index [6]. To compute the weighted rating (0 – 100), we followed the procedure described in [17]: the participants were asked to rate mental, physical, and temporal demand as well as performance, effort, and frustration after task completion with each system and to weight the sources of workload at the end of the study.

While the average weighted workload rating measured for eSiP (65) was slightly higher compared to SiP (59), we observed the overall weighted workload and ratings for the single workload scales to differ highly among participants and systems. In total, 6 out of 15 participants experienced lower workload with eSiP than with SiP.

4.6 Limitations

In our study the group of participants is limited to people with little experience in cyber security. Hence, our evaluation investigates how system usability and laymen's performance is affected by the ontology and knowledge base that was established with the help of professional analysts rather than on the explicit comparison between professional analysts and laymen. The evaluation of our goal to narrow or close the gap between professional analysts and laymen is thus performed implicitly: Our results show that laymen achieved higher success rates using eSiP for tasks that would otherwise require professional analysts. From this, we conclude that the knowledge of professional analysts was successfully transferred to and used by laymen.

We further found that the knowledge assistance features incorporated in eSiP helped decreasing the absolute discrepancy in effectiveness and satisfaction between users with and without experience in visual analytics. Effects concerning the experience with spiral plots, SiP, and a technical background could not be investigated as the respective groups of participants were unbalanced.

As the study had to be carried out remotely, the assessment of satisfaction and cognitive load was limited to subjective feedback. We expect observation techniques as well as the collection of physiological data to deliver additional insights concerning the interaction with the system and varying levels of cognitive load. However, these tools were not applicable while keeping social distance.

5 Conclusion and Future Work

We designed eSiP to enhance the support for triage analysis in OT networks, increase trust in its machine learning components and enable cooperation via the system. Based on the Knowledge Rocks Framework, we integrated a knowledge base in the system; automatically suggested incidents from this knowledge base are then incorporated in the visualization, providing direct support without additional hurdles.

The majority of participants selected eSiP as the preferred system and rated it to be more satisfying to interact with, easier to relearn, and its additional functions to be more supportive. Hence, we rate our design as successful. Especially the enhanced spiral plot turned out to be particularly supporting for the detection of anomalies and false positives – crucial tasks in triage analysis. Supporting visual cues resulted in an optimal recognition rate for abnormal high values. With the additional information provided in the enhanced time slider the participants were able to spot irregularities in the sensor readings directly from the time slider, and results of the automated anomaly detection algorithm became reproducible. Overall, discrepancies in effectiveness and satisfaction between participants with and without experience in visual analytics were reduced while using eSiP.

Due to the observed decrease in TCT from the first to the second system, we expect performance to further increase after additional training. We believe that in addition, appropriate training will support collaboration via the knowledge base and thus help leveraging the benefits of the knowledge assistance provided in eSiP.

In line with Sweller [19], we believe that extraneous load, that is the amount of cognitive load evoked by system design, depends strongly on the user's previously established knowledge. In order to adjust the system design respectively, further insights regarding the variation of cognitive load across different tasks are needed. This could also bring further insights regarding tasks that require "too many different steps" (Q2) in eSiP. Thus, we will take into account continuous cognitive load assessment in the future, e.g., via changes in pupil diameter as described in [3]. Furthermore, we will consider tracking gaze patterns during system usage, to analyze how eSiP's features influence and support the user's solution strategies. In particular, the impact of the currently implemented and further visual cues, and other guidance options are part of our future work. In this study, eye tracking was not applicable due to the COVID-19 pandemic.

Based on the evaluation of individual system usage, we plan to investigate how eSiP enhances collaboration during real-world usage. To this end, we will recruit analysts and ask them to annotate stored instances according to the ontology and use the annotations that were added by other collaborators. In this setting, a further evaluation of the enhancement's impact on trust will also be possible.

To further promote collaboration in the system, adding unknown incidents with a "request" for a classification and annotations by experts will be tested. Also, increasing the possibilities for direct interaction between users similar to messengers and comment sections for incidents will be evaluated. Since most of the decisions in the acting ontology are based on correlation (in particular the ones based on machine learning), but some decisions base on causality, a distinction of these two classification sources in the system might be of benefit for the user.

Misleading information in the knowledge base is a problem for all knowledge assisted systems. A corrupt knowledge base can cost trust and destroy the credibility of the system. While preventing the input of misleading information is only possible on user-side, eSiP has several mechanisms to make it more robust against such cases: the instances are proposed based on classification and pattern matching. A wrong classification in the knowledge base hence makes it unlikely for this incident to be proposed. In addition, multiple instances are proposed, giving users the opportunity to compare and filter out proposed instances that differ from all others. For ontologies, different evaluation methods are available [18]. Researching further possibilities to maintain gained trust is an interesting research topic for the future.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 252408385 – IRTG 2057

References

[1] F. Böhm, N. Rakotondravony, G. Pernul, and H. Reiser. Exploring the role of experts' knowledge in visualizations for cyber security. In *IEEE Symposium on Visualization for Cyber Security*, VizSec '18, October 2018. doi: 10.5283/epub.38044

[2] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185

[3] A. T. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, M. Raubal, and I. Giannopoulos. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173856

[4] S. D. Duque Anton, A. P. Lohfink, C. Garth, and H. D. Schotten. Security in process: Detecting attacks in industrial process data. In *Proceedings of the Central European Cybersecurity Conference 2019*, CECC 2019. ACM, New York, NY, USA, 2019.

[5] I. Fujishiro, Y. Takeshima, Y. Ichikawa, and K. Nakamura. Gadget: goal-oriented application design guidance for modular visualization environments. In *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, pp. 245–252, 1997. doi: 10.1109/VISUAL.1997.663889

[6] S. G. Hart. Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909

[7] M. Hellkamp et al. Bottle: Python web framework. `https://bottlepy.org/docs/dev/`, 2016. software, version 0.12.

[8] V. M. Igure, S. A. Laughter, and R. D. Williams. Security issues in SCADA networks. *Computers & Security*, 25(7):498–506, 2006. doi: 10.1016/j.cose.2006.03.001

[9] iTrust Centre for Research in Cyber Security. Secure water treatment (SWaT) testbed. Technical Report 4.2, Singapore University of Technology and Design, October 2018.

[10] S. Jiang, S. Fang, S. Bloomquist, J. Keiper, M. J. Palakal, Y. Xia, and S. J. Grannis. Healthcare data visualization: Geospatial and temporal integration. In *VISIGRAPP (2: IVAPP)*, pp. 212–219, 01 2016. doi: 10.5220/0005714002120219

[11] R. Johansen. Teams for tomorrow (groupware). In *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, vol. 3, pp. 521–534, 1991. doi: 10.1109/HICSS.1991.184183

[12] R. Kincaid and H. Lam. Line graph explorer: Scalable display of line graphs using focus+context. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '06, p. 404–411. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1133265.1133348

[13] A.-P. Lohfink, S. D. D. Anton, H. Leitte, and C. Garth. Knowledge rocks: Adding knowledge assistance to visualization systems. `arXiv:2107.11095`, 2021. Preprint, to appear at IEEE Vis 2021.

[14] A.-P. Lohfink, S. D. D. Anton, H. D. Schotten, H. Leitte, and C. Garth. Security in process: Visually supported triage analysis in industrial process data. *IEEE Transactions on Visualization and Computer Graphics*, 26(4):1638–1649, 2020. doi: 10.1109/TVCG.2020.2969007

[15] A. P. Mathur and N. O. Tippenhauer. SWaT: a water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pp. 31–36, April 2016. doi: 10.1109/CySWater.2016.7469060

[16] S. Miksch, H. Leitte, and M. Chen. *Knowledge-Assisted Visualization and Guidance*, pp. 61–85. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-34444-3_4

[17] National Aeronautics and Space Administration. NASA TLX paper and pencil version instruction manual. `https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php`. questionnaire, accessed 06/25/2021.

[18] J. Raad and C. Cruz. A survey on ontology evaluation methods. In A. L. N. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, and J. Filipe, eds., *KEOD 2015 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 2, Lisbon, Portugal, November 12-14, 2015*, pp. 179–186. SciTePress, 2015. doi: 10.5220/0005591001790186

[19] J. Sweller. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22(2):123–138, 2010. doi: 10.1007/s10648-010-9128-5

[20] M. Wagner, A. Rind, N. Thür, and W. Aigner. A knowledge-assisted visual malware analysis system: Design, validation, and reflection of kamas. *Computers & Security*, 67:1–15, 2017. doi: 10.1016/j.cose.2017.02.003

[21] S. Walton, E. Maguire, and M. Chen. Multiple queries with conditional attributes (qcats) for anomaly detection and visualization. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security*, VizSec '14, p. 17–24. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2671491.2671502

[22] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh. Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 1317–1322, 12 2016. doi: 10.1109/ICDM.2016.0179