Evaluation of Cyber Defense Exercises Using Visual Analytics Process

Radek Ošlejšek*, Jan Vykopal†, Karolína Burská* and Vít Rusňák†
*Masaryk University, Faculty of Informatics, Brno, Czech Republic
Emails: {oslejsek|burska}@fi.muni.cz
†Masaryk University, Institute of Computer Science, Brno, Czech Republic
Emails: {vykopal|rusnak}@ics.muni.cz

***Abstract*—This Innovative Practice Full Paper addresses modern cyber ranges which represent unified platforms that offer efficient organization of complex hands-on exercises where participants can train their cybersecurity skills. However, the functionality targets mostly learners who are the primary users. Support of organizers performing analytic and evaluation tasks is weak and ad-hoc. It makes harder to improve the quality of an exercise, particularly its impact on learners. In this paper, we present an application of a well-structured visual analytics process to the organization of cyber exercises. We illustrate that the classification derived from the adoption of the visual analytics process helps to clarify and formalize analytical tasks of educators and enables their systematic support in cyber ranges. We demonstrate an application of our approach on a particular series of eight exercises we have organized in last three years. We believe the presented approach is beneficial for anyone involved in preparation and execution of any complex exercise.**

***Index Terms*—visual analytics, cyber defense exercise, cyber range**

I. Introduction

Visual analytics (VA) is the science of analytical reasoning supported by interactive visual interfaces [1]. It is applied in various fields from biology or weather forecast [2]–[6] to education [7], [8]. As a specific case, we can consider applied cybersecurity training which is of our focus.

Various hands-on training programs which aim at improving attacking or defending skills of learners often augment theoretical cybersecurity education. While Capture the Flag (CTF) games focus on the attacking skills and learners solve one task at a time, Cyber Defense Exercises (CDX) are more complex events [9]. They mimic real-world operations of an organization under the attack of an unknown offender, and their participants work in teams on several issues at a time.

CDXs usually run in virtual environments called cyber ranges [10], [11]. Cyber ranges provide access to virtual computer networks where learners exercise their skills and abilities to protect the infrastructure against the attackers. The development in cyber ranges focuses mostly on tooling for learners who are the primary users and the instant assessment. Less attention is paid to analytical tools for organizers, which makes an in-depth evaluation and analysis tasks laborious and time-consuming. Nevertheless, these tasks are crucial in the process of continuous improvement of CDX events.

Related analytical tasks and visualizations discussed in this paper clarify analytical interests of organizers and provide a mapping to the general visual analytics model. This systematization is crucial for building awareness of organizational aspects so that the organizational process can be automated in cyber ranges. To demonstrate practical applicability, we present experience gained from the organization of a particular CDX series.

In the rest of this section, we describe key features of principles of visual analytics framework and cyber defense exercises. Section II discusses an adaptation and interpretation of the visual analytics framework in the context of CDX organization and its evaluation. We demonstrate a practical application of our approach on a case study described in Section III. Lessons learned from our experience follow in Section IV. Section V concludes the paper with the outline of follow-up work and research opportunities.

A. Visual Analytics

Keim et al. proposed a formal description of the visual analytics (VA) process in [12], [13]. They defined basic terms like data, models, visualization, and knowledge together with their modeling and analytical processes. However, this model is primarily system-driven. It focuses on automated data analyses and does not consider details of user-driven analytical tasks forming the knowledge via human reasoning.

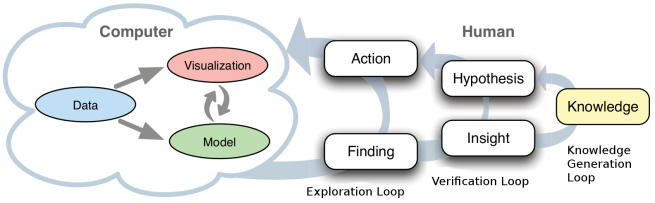

Figure 1. An overview of the knowledge generation model for visual analytics [14].

Sacha et al. in [14] provide a solution that extends the computer part of Keim's model with hierarchically connected human loops, as shown in Figure 1. In this model, we define *knowledge* as a "justified belief" that our understandings in a problem domain are correct [15]. A central role in building knowledge play *hypotheses*. In the beginning, they can be

vaguely defined using many unknown factors; then they can be gradually refined to produce deeper *insight* into a problem domain. Sufficiently approved insights and hypotheses can be accepted as new knowledge which can affect or initialize further hypotheses. During the *exploration loop*, analysts either verify or disprove hypotheses via *actions* that manipulate data and models utilizing interactive visualizations. Gained *findings* have no interpretation. A finding would be an unusual peak in a graph, for instance, that attracts analyst's attention. To understand the peak and then to gain insight, the peak has to be interpreted by the analyst. It often requires further actions to be performed. Meaningful findings can lead to gaining insight into the problem domain.

B. Cyber Defense Exercises

The CDX is an exhausting event which usually spans one to several days of a very intensive engagement of learners. It includes familiarization with the infrastructure, hands-on experience, and the evaluation phases. Hands-on part runs by a prescribed game scenario. However, the whole CDX life cycle is even more demanding for organizers. It spans several months and involves dozens of highly skilled people in multiple domains (cybersecurity, education, law).

Persons involved in the CDX usually form four teams. Learners, mostly ICT professionals, are organized in several Blue teams consisting of at least three people. *Red team* members represent attackers who run attacks against the Blue teams. Scoring and controlling game scenario and rules are tasks of the *White team* members. They also represent several avatar characters (company users, management, lawyers), or journalists who interrupt the game with various inquiries on the Blue teams. Members of the *Green team* maintain the underlying infrastructure of the exercise.

As we can deduce, CDX examines both technical and soft skills of learners. Every Blue team tries to protect a dedicated IT infrastructure while facing multiple issues at a time. They are forced to prioritize and assign tasks ranging from technical issues (e. g., hacked server) to interaction with avatar characters (e. g., creating press news). The learners can only presume whether their actions were correct or not based on the minimal feedback in the form of a total score of the team. Example techniques for automated assessment of learners performance are in [16], [17]. The organizers can usually access detailed overview based on game scoring rules.

The actual exercise (gameplay) is only one of the four phases of the CDX life-cycle. A *preparation phase* spans several months before the event. Its outputs are a detailed game scenario, infrastructure deployed in the cyber range, scoring system, and game rules. A *dry run* phase involving testers in Blue teams helps to find flaws in the rules and the game scenario. Learners play the CDX game in an *execution phase*. An *evaluation phase* concludes the life-cycle. As a result, organizers use the outcomes in the next run. Data sources for the evaluation span from learners' feedback to automatically acquired data from the cyber range (e. g., computer logs, configuration changes, users' actions). More details about an exercise life cycle can be found in [18].

Knowledge of organizers of CDX is collective and continuous. Collective means that there are many organizers involved who share their experience to build and reuse the knowledge. Continuousness comes from the fact that methods of exploratory and confirmatory analyses used in the exploration loops of CDX usually produce approximate results leading to uncertain insight. Only repeating the analysis through multiple exercises can improve the insight by making it gradually more and more credible to be finally accepted as a piece of knowledge. Our goal is to adapt Sacha's VA framework for CDX so that we can build and share the knowledge systematically and efficiently via functionality provided by cyber ranges.

II. Visual Analytics in Cyber Exercises

Application of the VA framework on the organization of CDX requires clarifying the type of data, available models, visualizations, and also human-driven analytical processes depending on these computer-related elements. In what follows, we discuss the VA model from these individual perspectives and define a classification scheme that helps us to understand how the VA model fits requirements of CDX organizers.

A. Hypothesis-driven Analytical Goals

Hypotheses actively drive the unified VA model. Moreover, a vast amount of various teams and user roles involved in the organization can introduce a considerable amount of different objectives. These aspects could make the adoption of the unified VA framework and its systematic support in cyber ranges impossible.

To cope with these doubts, we divide common analytical goals to three distinct categories which enable us to (a) clarify the hypotheses, (b) classify them according to the goals, and (c) map verification loops of the hypotheses to individual phases of CDX life cycle. Figure 2 overviews goals and their mapping to CDX life-cycle phases.

Goal 1 – Evaluation of exercise content and parameters: One of the most challenging tasks in the organization of cyber defense exercises is to make an exercise useful and to keep learners motivated to finish it. Therefore, hypotheses related to scenario difficulty, learners' confidence and satisfaction, learners' skills, and many other qualitative aspects, are often formulated.

During the *preparation phase*, organizers estimate and prepare key exercise parameters, such as a storyboard, a task schedule, penalty types and their values or types of attacks. Their improper values can make the exercise too complicated, too dull or unrealistic, which can quickly make learners frustrated. Therefore, organizers usually utilize a prior insight or knowledge gained from previous runs (results of their evaluation phase). Moreover, skills and experience of prospective learners are often ascertained employing self-evaluation questionnaires gathered in this phase. Results of

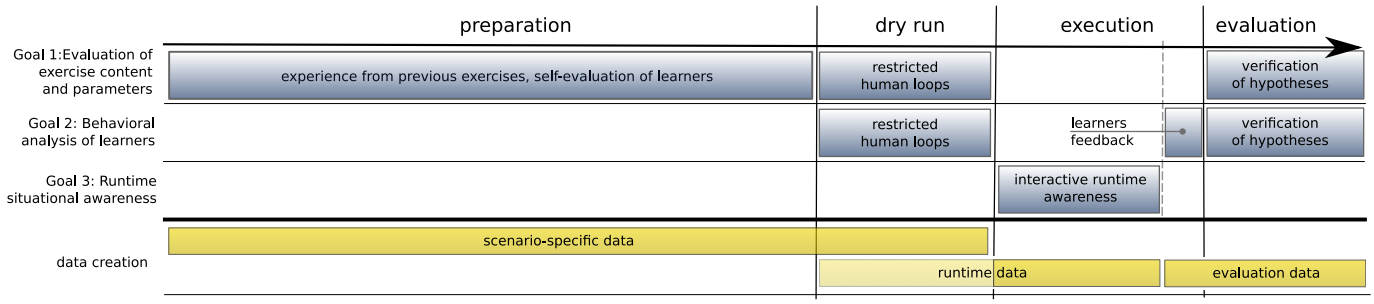

Figure 2. Mapping of goals and data to phases of CDX life cycle.

this analytic loop are used to create well-balanced teams and to adapt exercise parameters to them.

Exercise parameters are tested and adjusted during the *dry run*. Note, that the verification loops are limited because participants involved in the dry run differ from learners and then also the results are approximate.

Hypotheses are verified during the *evaluation phase* when statistical models, knowledge-discovery models, exploratory visualizations, and other tactics of verification loops applied to exercise parameters and the data gathered during the exercise are brought into action. Gained insight and knowledge are used by organizers to prepare even better and more attractive exercises in the future.

Goal 2 – Behavioral analysis of learners: Study of the behavior of learners during an exercise can reveal relevant facts about their motivation, learning impact or level of knowledge. Gained information is useful for (a) learners as they can learn about themselves, their strengths, weaknesses, and mistakes; (b) exercise contractors, usually learners' employers, who can learn about the skills of their employees; (c) security experts and researchers who can reveal and compare atypical defense strategies, collaboration strategies, and other behavioral patterns. Therefore, organizers should be supported in these types of behavioral analyses so that they can verify behavior-related hypothesis and provide reasonable feedback to participating parties. Let us note that this goal is also partially related to the previous one because behavioral analysis of learners can also reveal problems caused by exercise parameters. For example, if organizers detect that several teams gave the exercise up in a particular phase of attacks plan, then they can infer insufficient difficulty of these attacks or inadequate readiness of learners.

The *dry run* is used to verify infrastructure, required data, and analytical loops. However, Blue teams involved in the testing are different from target learners. Neither the data nor possible analytical results are usually valid for gaining general knowledge, and they are erased before the execution phase.

At the end of the *execution phase*, it is convenient to provide feedback to learners so that they can analyze their behavior and learn from their mistakes immediately after the exercise. However, such feedback requires automatic data gathering and mediation to learners through intuitive and interactive analytical tools integrated into the cyber range. Adoption of the VA model by cyber ranges would help to achieve this valuable functionality.

The main effort related to the behavioral analysis is dedicated to the in-depth verification of corresponding hypotheses during the *evaluation phase*. The organizers present these results to learners during post-exercise workshops few weeks after the exercise.

Goal 3 – Runtime situational awareness: During an exercise, organizers monitor and analyze the situation on the "battlefield" and actively intervene if necessary. They have to analyze the situation from their perspective and interact with the system continuously. However, it is important to realize that runtime situational awareness provided to learners is intentionally very limited because in CDX the realism is of high importance. Therefore, giving an insight, which is not available in the real world, is undesirable. On the contrary, the goal of CDX is often to train learners in gaining the insight by themselves.

We can consider situational awareness as a process of making simple runtime hypotheses in the users' mind. The hypotheses are evaluated via interactive visual tools mediating access to the infrastructure and proving insight into its internal processes and developments. Interactions of learners produce data for the verification of organizers' hypotheses.

The situational awareness plays an essential role during the *execution phase*. It is not a passive process when a user is only notified of important events. On the contrary, actions and visualizations of situational awareness have to enable learners and organizers to interact with the system actively and then to affect its state. These interactions and changes in state are often monitored and used for following analytic tasks of the previous two analytical goals during the CDX evaluation phase.

B. Data

As hypotheses defined for cyber exercises are changing then also requirements on data are frequently changing. Moreover, CDX is often unstructured from the learner's perspective. For instance, a cyber range can generate network flows (data transmitted through networks), hosts and network characteristics (e. g., network throughput, memory size), system logs, questionnaires or scenario penalties. Variability and heterogeneity of data put high demands on the adaptability of the monitoring and storage infrastructure and make the design intriguing, as discussed in [11], [17], [19].

To clarify data required for visual analytics of CDX, we classify them according to the phases of exercise life cycle, as shown in Figure 2. As classification criteria, we use data creation. However, it is worth to point out that data created in particular phase can be used to solve any analytical goal at any phase of the exercise life cycle.

Scenario-specific data: Configuration data defined by organizers usually in the preparation phase and possibly adjusted during the dry run. These data include, for example, a division of learners to teams, network topology, and network properties, the definition of required exercise services running on defended networks, types of penalties and their values or attack schedule. This category also includes answers to questions of various learners surveys.

Exercise runtime data: A system-generated data gathered and stored during the execution phase of an exercise. They represent quantitative operational data providing digital evidence of the behavior of users and applications during the exercise. Exercise runtime data is often based on the scenario-specific data and include, for example, particular penalty points assigned to teams, information about (un)availability of exercise services at given time or logs from hosts. Exercise runtime data is also collected in the dry run for testing purposes and then deleted.

Evaluation data: User-generated data provide qualitative information. This data is gathered either from learners at the end of exercise via post-exercise surveys, specialized feedback visualizations, or from organizers during the evaluation phase where additional data can be inserted to verify hypotheses. For example, structured informal notes about the behavior of learners noticed by organizers during the exercise can be added to the dataset.

C. Models

Models derived from data can be as simple as descriptive statistics or as complex as a data mining algorithms. Their usage is also reasonable in the context of cyber exercises. For example, complex networks [20] could be used to capture relationships between learners to simulate and analyze their behavioral patterns like collaboration or defense strategies. Knowledge discovery approaches to anomaly detection [21]–[24] can reveal significant exercise parameters or learners with remarkable skills.

Nowadays, standard statistical models are used extensively for the evaluation of exercises [25]–[29]. On the contrary, the utilization of advanced models is exceptional and ad-hoc just because of missing conceptual solution to repeated analytical tasks in the CDX domain.

D. Visualizations

As for the visualizations, some classifications and perspectives allow us to cover different targets of existing models and available data. Our classification divides visualizations into three basic categories providing insight into data according to analytical goals.

Exercise infrastructure overview: Interactive visualizations that give us a complete overview of the structure and state of network infrastructure and help us to monitor running services. These visualizations are beneficial for runtime situational awareness (analytical Goal 3). However, the network topology overviews usually represent primary access points used by learners to interact with the infrastructure. Therefore, visualizations equipped with functions monitoring interactions of learners can help us to gather the data related to learners' behavior and then to verify hypotheses of the analytical Goal 2.

Visual insight into the exercise progression: Visualizations that aim at providing insight into the state and development of an exercise. Some insight can be gained from the discussed views on exercise infrastructure. For example, inaccessibility of services dues to a successful Red team's attack. However, it is not usually enough, and both learners and organizers need specialized views covering the exercise state. For example, Blue teams should be informed about the development of their score, while Red, Green, and White teams should have a detailed overview of planned and performed attacks and their successfulness so that they can distinguish between expected behavior and failures in the infrastructure, for instance, and then intervene properly.

This category of visualizations is useful primarily for the analytical Goal 3 because it provides situational awareness to both parties. At the same time, it can be helpful in verifying exercise parameters (Goal 1). For example, if the schedule of the exercise is not rich enough, or the Blue team is too busy or bored, the exercise parameters could be considered wrong and adjusted for future runs.

Feedback visualizations: The goal of visualizations providing interactive visual feedback is to gain insight into the exercise as well, but not from the perspective of current exercise progression. Instead, this insight is retrospective, aiming at learning from runtime mistakes, wrong decisions, or improperly estimated exercise parameters. Providing timely intuitive feedback to learners is crucial for improving the impact of the exercise. However, if the feedback is extended with the possibility to comment or rank events by the learners actively, then it can be even more useful. This kind of learners' reflection can help to reveal inappropriate exercise parameters (Goal 1) and to gather a data related to the behavior of individual learners (Goal 2).

III. Case Study

In this section, we illustrate the application of visual analytics process in CDX which we distilled from eight runs of Cyber Czech exercise series held in 2015–2017. Each run lasted two days and involved about 20 learners located in one physical place. We follow the exercise life cycle and put tasks and components of the visual analytics model into the context of individual phases so that their continuity is better recognizable. The iterative principle of visual analytics process results in multiple iterations over the refined and/or redefined hypotheses. Table I summarizes results for primary hypotheses, as discussed in what follows.

Table I
OVERVIEW OF THE VISUAL ANALYTICS PROCESS MAPPED ON THE INITIAL HYPOTHESES.

Human	**Hypothesis**	*The participants improve their skills*	*Every single participant is involved*	*Exercise infrastructure is stable and responsive enough to resemble realistic settings*
	Insight	Fairly confirmed. Individual learners would be affected by their skills and skills of teammates ⇒ hypotheses H1a and H1b. A novel ways of prerequisite testing are desired.	Fairly confirmed. Detailed per-user data required in the future. The level of involvement would stand as an indicator of cost-efficiency of the exercise ⇒ hypothesis H2a.	Uncertain results. Current views on monitoring data provide situational awareness but no statistics for evaluation of stability and responsiveness of the infrastructure.
	Actions	Organizers: Data definition, configuration of data sources (sub-systems) and visualizations, evaluation. Learners: Filling questionnaires, interaction with the cyber range and feedback visualizations.		
	Findings	Majority of the learners confirmed they learned new skills or re-shaped existing ones. Some learners did not learn anything new. Some others admitted the lack of necessary skills.	Majority of the learners declared they were involved. The data was, however, collected on a per-team basis. We were not able to objectively measure the level of involvement of every single participant.	A considerable amount of issues reported by learners. The Green team was aware of most of them. Several issues remained unnoticed.
Computer	**Data**	Data from scoring and auditing systems, pre- and post-exercise questionnaires.	Post-exercise questionnaires.	Data from the monitoring system, notes taken by organizers.
	Models	Descriptive statistics	Descriptive statistics	N/A
	Visualization	Feedback visualization	Feedback visualization	Nagios, network topology

A. Hypotheses

Hypotheses are either formulated during the preparation phase of a CDX or reused from previous exercises. We formulated several hypotheses, from which we selected the following we consider as the most important:

H1 – Participants improve their skills: First and foremost, the exercise should be useful for learners. It should deliver any educational value: either in technical, organizational or communication level. In particular, learners should develop or exercise skills required for incident handling and resolution, including reporting and communication with other parties outside their team, as well as working under stressful conditions. This hypothesis is related to the analytic Goal 1.

H2 – Every single participant is involved: Costs and effort invested in preparation and execution of the complex exercise should be utilized efficiently. Each learner should benefit from participating in the exercise. The content of the exercise should be rich enough to engage each participant. This hypothesis is related to Goals 1 and 2.

H3 – Exercise infrastructure is stable and responsive enough to resemble realistic settings: The CDX infrastructure is complicated. Multiple instances of separate environments of individual teams are deployed and transparently emulated on a restricted and complex infrastructure of the cyber range. However, any virtualization issue should not affect the user experience of real-life infrastructure regarding performance, response or failures. This hypothesis is related to the analytic Goals 1 and 3.

B. Preparation Phase

During the preparation phase, it is necessary to define data to be gathered for further analysis, configure data-related components of the cyber range, and prepare the graphical user environment. In particular, we perform the following preparation actions.

Preparation of surveys, formulation of questions: To verify our hypothesis, we use pre- and post-exercise questionnaires. This *evaluation data* are related to qualitative aspects of learners and exercise, e. g., participants' skills, their exercise experience, their opinion on difficulty or usability. We currently use the external Google Forms system to define and process questionnaires, which complicate the evaluation and integration of gained answers with internal data measured and stored in the cyber range.

Scoring subsystem settings: A scoring subsystem is used for penalization of Blue teams. Concrete penalties assigned to learners represent an *exercise runtime data* which are collected during the later phases of CDX life cycle. During this preparation phase, a *scenario-specific data* is used to define scoring rules. Attack plans, objectives, and their penalty values are set according to expected goals of the exercise and learners' skills.

Infrastructure monitoring settings: Green team members configure the infrastructure monitoring subsystem to keep track of the health of the virtualized networks and their underlying infrastructure. This step requires to specify a *scenario-specific data* like topology details, IP addresses of monitored hosts, network ports of watched services, or required timeouts. *Exercise runtime data* produced by the monitoring subsystem during the execution phase in the form of events is used for situational awareness of the Green team and for the automatized penalization of Blue teams for inaccessibility of network services (e. g., web, mail) that are under their management in the scenario. Currently, we use the Nagios monitoring system running in the cyber range as a standalone application. Its tighter "out of the box" integration into the cyber range would

bring better connection to other internal data and then more effective situational awareness and analysis.

Configuration of auditing capabilities: While the infrastructure monitoring subsystem monitors infrastructure, the auditing subsystem monitors events that are related to the behavior of users and applications. This step includes the configuration of probes and internal auditing capabilities of the cyber range so that we can monitor required events like access to hosts, e-mail delivery, host reboots, or a history of commands run on a host by learners.

Configuration of runtime visualizations: Visualizations used in the cyber range are generic and highly configurable to cope with a wide variety of user goals. For example, the interactive network topology shows network-related *exercise runtime data* like current utilization of links or the state of nodes. However, this kind of situational awareness is undesirable for CDX, and the organizers have to adjust provided visualizations so that they satisfy specific requirements of the exercise.

C. Dry Run and Execution Phase

During the execution phase, interactions of various participants mingle. Moreover, the interactions reflect different levels and details of human loops of the visual analytics process. For instance, learners' actions produce data for exploration and verification loops of organizers. To discuss relevant activities meaningfully, we describe them from the viewpoints of the individual teams involved in the CDX.

Blue teams: In our exercise, the Blue teams produce data for verification loops of hypotheses *H1 – Participants improve their skills* and *H2 – Every single participant is involved*. Observations made by learners during their interactions with the network infrastructure lead to further interactions motivated to fulfill exercise tasks. Interactions are monitored and stored for verification loops of hypotheses of organizers. Moreover, learners also fill pre- and post-exercise questionnaires to evaluate their input knowledge and exercise experience respectively.

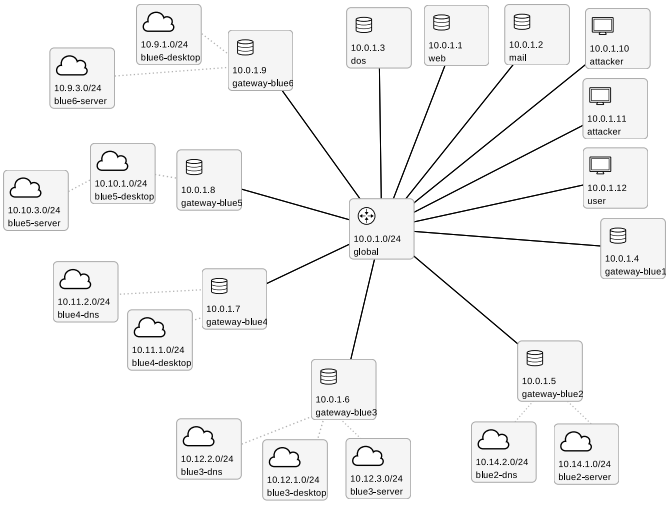

Figure 3. Interactive network topology visualization.

During the exercise, learners use two runtime visualizations for situational awareness: network topology (see Figure 3) and scoreboard. The former provides an overview of the network they administer and enables them to access individual hosts. The latter provides a score overview of all teams. The score includes both automatically collected data from the cyber range (e. g., availability of the web service or database) and inputs from other teams (answered questions of their users or journalists).

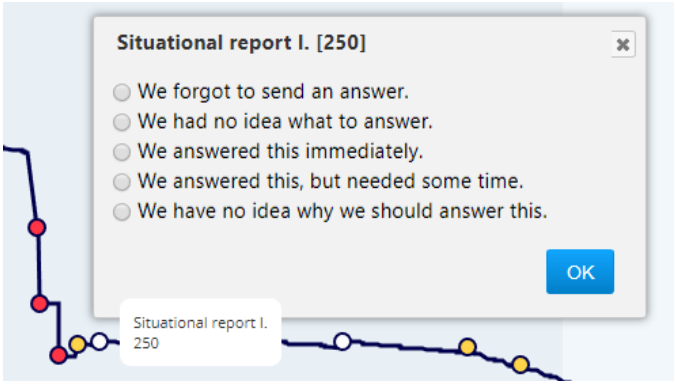

Figure 4. An example of learners' evaluation of their actions

At the end of the exercise, learners get access to a specific visual-analytics tool for personalized feedback [30]. It displays the score development throughout the time enhanced with data points containing brief descriptions of reasons why score changed. These are coupled with analytical questions related to a retrospective evaluation of learners' actions by organizers, as shown in Figure 4.

Green team: The Green team provides technical support during the exercise and produces data for *H3 – Exercise infrastructure is stable and responsive enough to resemble realistic settings*. The exercise execution is time-demanding since every issue needs to be solved as quickly as possible. Therefore, the more in-depth analysis of the issues and their solutions are mostly summarized in the evaluation phase.

The principal visual analytics tools of the Green team are network topology and Nagios dashboard. The network topology has the same capabilities as for the Blue teams, but it is rarely used. Most of the operations are done through the Nagios, which provides an overview of all Blue teams' networks. Also, it also displays service nodes that are not accessible by Blue teams, e. g., attackers' hosts.

Red team: Red team members perform the set of attacks according to the plan and enter penalty points based on the Blue teams' counter-actions. The gathered data about attacks and their successfulness is used to test H1 and H3 hypotheses.

The primary modus operandi is using a command line interface. The visualization tools are limited to a simple static schedule of the attacks made beforehand. For gaining better game situational awareness, the Red team needs to cooperate closely with the green one. For instance, to assess the success of their attacks, they need to know whether the attack was

unsuccessful due to an adequate counter-action of a Blue team or because of an outage of the cyber range.

White team: While the Red team focuses on *hard-skills*, White team members enter penalties based on *soft-skills* findings. They provide data for testing H1 and H2 hypotheses. They need to know the state of the game since some of their injects are time-related to attacks of the Red team. Therefore, they need to cooperate with both green and Red teams. Unfortunately, we have no tool for this type of orchestration and situational awareness available nowadays. The teams have to synchronize their activities via external communication channels.

Besides other activities, the White team also play a role of ordinary users. To do that, members of the team use the topology visualization to access hosts, simulate frequent utilization of the network, and interact with Blue teams.

D. Evaluation phase

Main outcomes of the evaluation phase are hypotheses for the second iteration of the visual analytics process. These are based on the findings and insight gained from the verification loops of the original set of hypotheses:

H1 – The participants improve their skills: We analyzed learners' pre- and post-exercise surveys and exercise scores using standard statistical models and exploration loops. We found out that the majority of learners confirmed they learned new skills or re-shaped existing ones. However, some learners reported that they did not learn anything new and some others admitted they lack some necessary prerequisite skills. Both extremes may indicate flaws in the selection of individual learners, their grouping to a team, or structure and content of exercise tasks.

H2 – Every single participant is involved: The vast majority of collected exercise runtime data captures actions of teams, not individual learners. The only data sources we analyzed to verify this hypothesis were post-exercise surveys and reflections provided by individual learners using an application providing automatically generated feedback visualizations. Although the results indicate that almost all participants felt involved during the exercise, the level of their involvement is unknown and may vary widely due to the absence of objective and rich data tracking all available modes of interactions.

H3 – Exercise infrastructure is stable and responsive enough to resemble realistic settings: Exercise runtime data serve for determining the exercise score and monitoring the exercise infrastructure. The current view of the data does not provide any statistics for evaluation of stability and responsiveness of exercise infrastructure. The only data sources are learners' post-exercise surveys and notes of issues taken by members of teams of organizers, in particular by Green team. A considerable amount of learners reported various issues. The organizers were fully aware of some of them, but there were several issues unnoticed. Again, the objective data source would help to clarify this bias.

E. Derived Hypotheses

Consequent hypotheses derived during the iterative VA process usually emerge. Due to the space restrictions of the paper, we end up with the outline of the hypotheses for the second iteration to illustrate the process continuation.

H1a – The difficulty of the exercise was adequate for learners: One of our observations related to H1 is that cohorts of dry-run and execution participants bias the perception of the difficulty level. The input knowledge of actual learners (not testers) needs to be considered. However, the pre-exercise survey relies only on self-assessment of learners' skills that could introduce an unwanted bias. We advocate complementing the self-assessment survey by a quiz or a practical task that would test the required skills objectively. As a result, prospective participants would have skills adequate to the exercise difficulty. The hypothesis relates to the analytic Goals 1 and 2.

H1b – Learners form well-balanced teams: While the CDX is based firmly on teamwork, grouping people of different skills into well-performing team covering as much as prerequisite skills is crucial. Weaknesses of one shall be balanced by strengths of another member of the team and vice versa. Since the teams are formed before the exercise, the pre-exercise survey questions should be refined to acquire more accurate data for optimal team balancing. The hypothesis relates to the analytic Goal 2.

H2a – Participants' involvement stands as an indicator for cost-efficiency of the exercise: CDX is a costly event. While person-months spent and costs of computational resources can be calculated relatively easily, answering the question whether the costs are adequate is tough. Participants' involvement could be a good indicator. The organizers should be able to determine the involvement ratio for each participant as well as the overall involvement of the whole group. The methodology for gaining the involvement ratio should combine outcomes from post-exercise questionnaires with analysis of participants' behavior and actions (e. g., from the automatically collected logs and commands they entered). As a side effect, the organizers should be able to provide learners with personalized feedback on their strengths and weaknesses. The hypothesis relates to the analytic Goals 1 and 2.

H2b: The set of exercise tasks covers relevant security issues: Thousands of threats exist, but only a subset of them is relevant these days. Attacks selected for the exercise should exploit recent and relevant threats rather than out-of-date and insignificant ones. While the obsolete threats can be suitable for the educational purpose, the organizers should carefully consider and select those, that are relevant nowadays (i. e., participants can experience them in their work). Strongly outdated threats (e. g., those that focus on no more used version of an operating system or a web server) are inappropriate. Common Vulnerability Scoring System (CVSS) by FIRST[1] can be used to assess the relevance of the threats. The hypothesis relates to the analytic Goal 1.

[1] https://www.first.org/cvss/

F. Output Knowledge

Verification of hypotheses H1–H3 brought a valuable insight regarding the usability of our cyber range, attractiveness for learners and the level of impact on them. However, the total number of teams that have been involved in the exercise series and provided data for the verification is relatively small. For this reason, we perceive conclusions formulated in this section as insight only. To be able to declare them as a justified knowledge, we need to verify them on more runs.

IV. Lessons Learned

Organization of complex CDX is usually ad-hoc hence inefficient: Modern cyber ranges support the organization of complex CDX. However, the organization discussed in Section I-B requires a vast amount of manual work and interventions in the infrastructure. Data useful for the optimization of CDX organization and improvement of exercise experience is often not gathered at all, or the data processing is not systematical. The data is usually exported manually from internal data sources and then processed in external tools ex-post. Our goal is to organize CDX efficiently, to use data as soon as possible (often at runtime), and to evaluate the impact on learners and the overall quality regularly. The classification of analytical tasks and their visual analytics elements discussed in this paper in the context of CDX life cycle would help us to solve this goal by identifying and clarifying processes that can be systematically supported by the cyber range.

Hypothesis-centered approach to CDX is suitable: Hypotheses actively drive the visual analytics model used as a unified framework for our approach. Although we were not using this hypothesis-centered way of thinking during the realization of previous exercises intentionally, we have come to realize that in fact, we were thinking in this way intuitively in many cases. Moreover, we found this kind of mindset handy for the definition of required data and the design of supporting interactive visual tools during the preparation of new exercises.

Positive impact on learners and organizers: Integration of even a few preliminary features of the visual analytics process into our cyber range brought positive outcomes from both learners and organizers, as shown in [30]. The application of the VA process to the organization of CDX also encouraged us to formalize attack plans, objectives, and other scenario-related events. Consequently, they are used for systematic analysis and runtime coordination of Green, Blue, White, and Red teams.

Structure of CDX-related knowledge was clarified: Nowadays, organizers have defined several processes prescribing how cyber defense exercises and their validation results should be documented and shared among team members so that exercises can be continuously adjusted and improved. However, the documentation is informal. It was not clear so far what the CDX-related knowledge exactly means and how to structure the pieces of information. Classification of CDX processes and elements discussed in this paper brings clear terminology and semantics which are suitable for formal knowledge modeling, e. g., using formal ontologies.

V. Conclusions and Future Work

Cyber defense exercises are complex education events requiring a significant amount of efforts of interdisciplinary teams. Application of the visual analytics process proves beneficial to CDX organization and evaluation. As we demonstrated on our case study, the iterative approach of human loop helps us in the identification of issues and leads us towards concrete suggestions for improvements in the organization of CDX. Simultaneously, application of the visual analytics process clarified the structure of the CDX-related knowledge enabling its better management.

However, we need to explore and revise the VA components further. While having raw data from the exercises, we have an unclear notion about the valuable data models applicable in this domain. The useful visualizations are alike. The lack of VA tools integrated into cyber ranges is a severe weakness of nowadays. These tools could provide automated statistical analysis as well as more in-depth insight into the learner's behavior during the game. They could also help in improving the process of CDX organization.

In this paper, we focus mainly on the organizers' viewpoint. Learners could also apply the VA process even though they have entirely different experience than organizers. They are focused on particular tasks related to the exercise content rather than the overall process. For this reason, we leave this topic for our future work.

Acknowledgments

This research was supported by the Security Research Programme of the Czech Republic 2015–2020 (BV III/1 – VS) granted by the Ministry of the Interior of the Czech Republic under No. VI20162019014 – Simulation, detection, and mitigation of cyber threats endangering critical infrastructure.

Access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144, is greatly appreciated.

References

[1] P. C. Wong and J. Thomas, "Guest editors' introduction–visual analytics," *IEEE Computer Graphics and Applications, 24 (5): 20-21*, vol. 24, no. PNNL-SA-41935, 2004.

[2] M. Krone, B. Kozlikova, N. Lindow, M. Baaden, D. Baum, J. Parulek, H.-C. Hege, and I. Viola, "Visual analysis of biomolecular cavities: State of the art," *Computer Graphics Forum*, 2016.

[3] B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege, "Visualization of biomolecular structures: State of the art revisited," *Computer Graphics Forum*, vol. n/a, no. n/a, pp. n/a–n/a, 2016. [Online]. Available: http://dx.doi.org/10.1111/cgf.13072

[4] K. Lawonn, N. Smit, K. Bühler, and B. Preim, "A survey on multimodal medical data visualization," *Computer Graphics Forum*, 2017. [Online]. Available: http://dx.doi.org/10.1111/cgf.13306

[5] A. Diehl, L. Pelorosso, C. Delrieux, C. Saulo, J. Ruiz, M. E. Gröller, and S. Bruckner, "Visual analysis of spatio-temporal data: Applications in weather forecasting," *Computer Graphics Forum*, vol. 34, no. 3, pp. 381–390, May 2015.

[6] A. Diehl, L. Pelorosso, K. Matkovic, J. Ruiz, M. E. Gröller, and S. Bruckner, "Albero: A visual analytics approach for probabilistic weather forecasting," *Computer Graphics Forum*, vol. 36, no. 7, pp. 135–144, Oct. 2017.

[7] R. Vatrapu, C. Teplovs, N. Fujita, and S. Bull, "Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. ACM, 2011, pp. 93–98.

[8] C. Vaitsis, G. Nilsson, and N. Zary, "Big data in medical informatics: improving education through visual analytics." in *MIE*, 2014, pp. 1163–1167.

[9] E. G. Díez, D. F. Pereira, M. A. L. Merino, H. R. Suárez, and D. B. Juan, "Cyber exercises taxonomy," INCIBE, Tech. Rep., 2015. [Online]. Available: https://www.incibe.es/extfrontinteco/img/File/intecocert/EstudiosInformes/incibe_cyberexercises_taxonomy.pdf

[10] "Cyber ranges," May 2017, https://www.nist.gov/sites/default/files/documents/2017/05/23/cyber_ranges_2017.pdf.

[11] J. Vykopal, R. Ošlejšek, P. Čeleda, M. Vizváry, and D. Tovarňák, "Kypo cyber range: Design and use cases," in *Proceedings of the 12th International Conference on Software Technologies - Volume 1: ICSOFT*. Madrid, Spain: SciTePress, 2017, pp. 310–321. [Online]. Available: http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=xwvv5mGUkNM=&t=1

[12] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual data mining*. Springer, 2008, pp. 76–90.

[13] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, "Mastering the information age solving problems with visual analytics," in *Eurographics*, vol. 2, 2010, p. 5.

[14] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, E. G., and D. A. Keim, "Knowledge Generation Model for Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, Dec 2014.

[15] E. Bertini and D. Lalanne, "Surveying the Complementary Role of Automatic Data Analysis and Visualization in Knowledge Discovery," in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, ser. VAKD '09. New York, NY, USA: ACM, 2009, pp. 12–20. [Online]. Available: http://doi.acm.org/10.1145/1562849.1562851

[16] A. Doupé, M. Egele, B. Caillat, G. Stringhini, G. Yakin, A. Zand, L. Cavedon, and G. Vigna, "Hit 'em where it hurts: A live security exercise on cyber situational awareness," in *Proceedings of the 27th Annual Computer Security Applications Conference*, ser. ACSAC '11. New York, NY, USA: ACM, 2011, pp. 51–61. [Online]. Available: http://doi.acm.org/10.1145/2076732.2076740

[17] R. G. Abbott, J. McClain, B. Anderson, K. Nauer, A. Silva, and C. Forsythe, "Log Analysis of Cyber Security Training Exercises," *Procedia Manufacturing*, vol. 3, pp. 5088–5094, 2015, 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2351978915005247

[18] J. Vykopal, M. Vizváry, R. Ošlejšek, P. Čeleda, and D. Tovarňák, "Lessons learned from complex hands-on defence exercises in a cyber range," in *2017 IEEE Frontiers in Education Conference*. Indianapolis, IN, USA: IEEE, 2017, pp. 1–8.

[19] R. Ošlejšek, D. Toth, Z. Eichler, and K. Burská, "Towards a unified data storage and generic visualizations in cyber ranges," in *Proceedings of the 16th European Conference on Cyber Warfare and Security ECCWS 2017*, N.-A. L.-K. Mark Scanlon, Ed. UK: Academic Conferences and Publishing International Limited, 2017, pp. 298–306.

[20] G. Chen, X. Wang, and X. Li, *Fundamentals of complex networks: models, structures and dynamics*. John Wiley & Sons, 2014.

[21] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowledge Eng. Review*, vol. 25, no. 2, pp. 137–166, 2010. [Online]. Available: https://doi.org/10.1017/S0269888910000032

[22] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 303–336, First 2014.

[23] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, May 2015. [Online]. Available: https://doi.org/10.1007/s10618-014-0365-y

[24] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19 – 31, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804515002891

[25] W. Schepens, D. Ragsdale, J. R. Surdu, J. Schafer, and R. New Port, "The cyber defense exercise: An evaluation of the effectiveness of information assurance education," *The Journal of Information Security*, vol. 1, no. 2, 2002.

[26] D. M. Nicol, W. H. Sanders, and K. S. Trivedi, "Model-based evaluation: from dependability to security," *IEEE Transactions on dependable and secure computing*, vol. 1, no. 1, pp. 48–65, 2004.

[27] A. Endert, C. Han, D. Maiti, L. House, and C. North, "Observation-level interaction with statistical models for visual analytics," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011, pp. 121–130.

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

[29] M. Granåsen and D. Andersson, "Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study," *Cognition, Technology & Work*, vol. 18, no. 1, pp. 121–143, Feb 2016. [Online]. Available: https://doi.org/10.1007/s10111-015-0350-2

[30] J. Vykopal, R. Ošlejšek, K. Burská, and K. Zákopčanová, "Timely feedback in unstructured cybersecurity exercises," in *Proceedings of Special Interest Group on Computer Science Education, Baltimore, Maryland, USA, February 21–24, 2018(SIGCSE'18)*. Baltimore, Maryland, USA: ACM, 2018, pp. 173–178.