Projektnummer/Project no
E716589
Uppdragsgivare/Client
FM
FoT-område
Operationer i cyberdomänen

Författare/Author
Jenni Reuben,
Hannes Holm

Datum/Date
2022-11-08
Memo nummer/Number
FOI Memo 8217

Cyber Defence Exercise - Cybon 2022

1 Introduction

A Cyber Defence Exercise (CDX) is an active learning environment that employs networked systems and orchestrated cyber security threats with the purpose to enable experiences similar to real-life cyber threat scenarios.

In a typical CDX, security analysts attempt to identify and prevent live-fire cyber-attacks carried out by one or more threat agents. Here, a group of security analysts that protect the same system is commonly referred to as a "blue team", and the threat agents as the "red team". While the focus of a CDX generally is to train blue teams (see e.g., [1]–[4]), creating and carrying out engaging threat agent campaigns is time consuming and costly. Consequently, there has been a number of research efforts to automatize red team activities. One such research effort called Lore has been developed within the project *Verktyg och Experiment för Computer Network Operations* (VECNO). Lore is able to automatically select and execute red team actions in a CDX [5].

This memo describes the planning, execution and follow-up of the CDX Cybon that was carried out in the scope of VECNO during 2022. In the CDX, Lore was employed as the red team and conscripts enrolled in a cyber-security training program at the Swedish Armed Forces (SAF) as the blue team. While the purpose of the CDX was to provide the participants a learning environment to exercise their cyber security defence knowledge acquired from their training program, it was also well suited to study the following research question:

- *How were the red team campaigns designed by Lore perceived by the blue team during Cybon 2022?*

2 Method

The aim of the study was to study how the blue team experienced the red team campaigns executed by Lore. In order to answer the research question stated in Chapter 1, data was acquired through three different collection measures:

- cyber incident reports submitted by each group
- a survey questionnaire that each participant of the CDX answered after the completion of each session
- a survey questionnaire that each participant answered at the end of the CDX.

The CDX Cybon was carried out over five days during May 2022 and involved ~40 security analysts (the blue team) employed by the Swedish Armed Forces (SAF), that were tasked to defend ~80 machines in a fictive organization called Cybon. The analysts worked in shifts during both day and night (i.e., around the clock) engaging in the CDX that ran for 5 days.

Each day presented the analysts with a different Lore scenario:

- **Day 1**: *Server exploits*: Very "loud" and "aggressive" behaviour that was perceived as being easy for the security analysts to manage. For example, heavy use of nmap and pivoting to internal machines through "public" servers exploited by remote software vulnerabilities.
- **Day 2**: *Bad USB*: Moderate difficulty for the security analysts. Very limited use of nmap and pivoting to internal machine through a payload on a USB drive mounted on an internal machine. Limited use of software exploits.
- **Day 3**: *Phishing*: Moderate difficulty for the security analysts. Similar to "Bad USB", but using a phishing email with a malicious document instead of a USB drive, and with a bit more focus on lateral movement through PSEXEC and similar methods.
- **Day 4**: *Responder*: High difficulty for the security analysts. A machine running the tool Responder was put through DHCP on an internal network to abuse victim broadcast requests such as mDNS.
- **Day 5**: *Bad employee*: Low difficulty for the security analysts. An angry employee with access to internal networks and user credentials to several machines, but a poor understanding of IT. Very "loud", e.g., major use of nmap and server software exploits, and actions were in general poorly chosen.

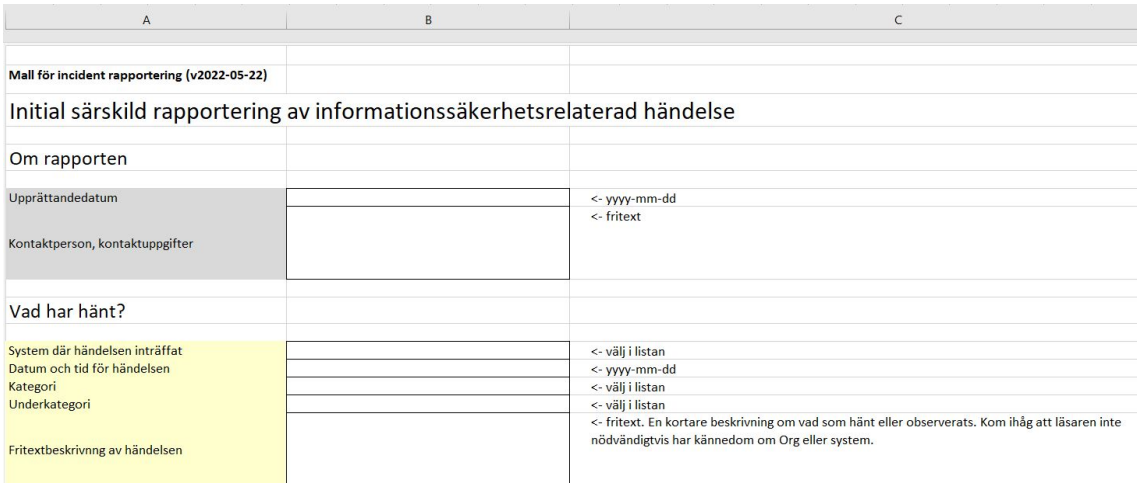
A script was used to reset the environment (everything apart from the blue team's own systems) after each threat agent campaign had been completed. In addition to the Lore scenarios, there were also simulated autonomous end-users interacting with different office computers in the fictive organization. These end-users conducted various tasks, such as browsing the web, interacting with an email client, and interacting with files on disk. Finally, there were also simulated external end-users that interacted with machines in Cybon that could be reached from the fictive Internet (e.g., a web server and a mail server).

Approximately 35 respondents completed the "participants' experience of Lore" questionnaire (see Section 2.2 for details). In the case of the final evaluation questionnaire though, we received responses from 33 respondents. As there is no constraint on the number of incident reports that each group should or shouldn't send in, we didn't keep a record of the number of incident reports filed by each group.

2.1 Incident Report Template

Figure 1 shows a screenshot of the incident report template designed for the purpose of gathering information regarding participants' understanding of Lore's actions in their networks.

Figure 1. Screenshot of incident report template designed for Cybon 2022 CDX.



A	B	C
Mall för incident rapportering (v2022-05-22)		
Initial särskild rapportering av informationssäkerhetsrelaterad händelse		
Om rapporten		
Upprättandedatum		<- yyyy-mm-dd
Kontaktperson, kontaktuppgifter		<- fritext
Vad har hänt?		
System där händelsen inträffat		<- välj i listan
Datum och tid för händelsen		<- yyyy-mm-dd
Kategori		<- välj i listan
Underkategori		<- välj i listan
Fritextbeskrivnng av händelsen		<- fritext. En kortare beskrivning om vad som hänt eller observerats. Kom ihåg att läsaren inte nödvändigtvis har kännedom om Org eller system.

2.2 Daily Survey Questionnaire on the participants' experience of Lore

The following questions were posted to the participants in order to capture their reactions to Lore's actions on their networks.

1. How do you rate the skill level of the threat agent?
2. How challenging was it to understand
    a. The threat agent's information gathering activities?
    b. The threat agent's attacks?
    c. The threat agent's command and control activities?
    d. The threat agent's goals?
    e. Which resources were compromised?
    f. Which countermeasures were required?
3. How realistic did you find the cyber-attacks?
4. How much did you learn today?

A Likert [1] scale from 1-7 was employed to get a quantified view of all the answers.

2.3 Survey Questionnaire to Evaluate CDX as a whole

The following questions were posted to the participants on the final day of the CDX, in order to capture their experience relating to CDX as a whole.

1) How rewarding did you find the CDX as a whole?
2) How instructive did you find the tasks in the CDX as a whole?
3) How much did you learnt during the CDX?

[1] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22 140,* 55.

3 Results

In this chapter we only report the descriptive statistics of subjective responses of participants' experiences of Lore (through the questionnaire presented in Section 2.2) and objective assessment of participants' understanding of Lore (through the incident reports filed by the participants). These are the most relevant results to the research question stated in Chapter 1.

3.1 Results from the questionnaire on participants' experience of Lore designed tasks

This section presents a statistical summary of responses to the following questions in the participants' perception of Lore questionnaire,

Q1: Hur upplevde du hotaktörens kunskapsnivå?

Q2: Hur utmanade upplevde du att det var att förstå: [Hotaktörens kartläggningsarbete?]

Q3: Hur utmanande upplevde du att det var att förstå: [Vilka angrepp som utfördes?]

Q4: Hur utmanande upplevde du att det var att förstå: [Hotaktörens tekniska kommunikationslösningar (eng: command and control)?]

Q5: Hur utmanande upplevde du att det var att förstå: [Hotaktörens mål?]

Q6: Hur utmanande upplevde du att det var att förstå: [Vilka egna resurser som kompromettterats?]

Q7: Hur utmanande upplevde du att det var att förstå: [Vilka motmedel som behövde utföras?]

Q8: Hur realistiska upplevde du att angreppen var

Q9: Hur lärorikt upplevde du denna del av övningen var (sista dygnen)?

From Table 1, it can be noted that the mean value of the responses for all the studied aspects (formulated as questions) lie more or less in the middle of the Likert scale, which means that Lore designed tasks are perceived positively by the CDX participants. Furthermore, the variances of the responses across the studied days did not vary much, which means the difficulty level of Lore scenarios do not elicit a negative experience among the participants.

Table 1: Descriptive statistics of the participants' experience of Lore, described from participants' responses to 9 different questions.

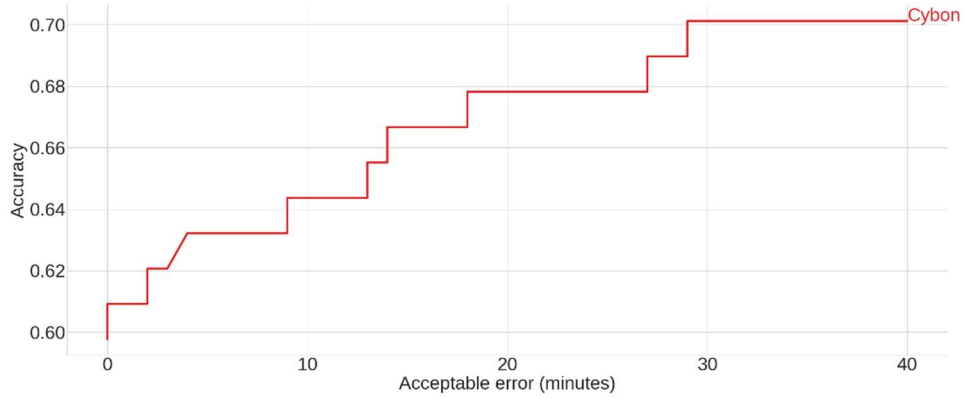
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Day 1	**Mean**	3.8	4.9	4.94	4.5	3.56	4.24	4.1	4.12	4.56
	STD	0.86	1.56	0.98	1.05	1.28	1.12	1.45	1.27	1.06
	Min	2	1	3	2	1	2	1	2	2
	Median	4	5	5	4	4	4	4	4	4
	Max	5	7	7	7	6	6	6	6	7
	MAD	0.71	1.29	0.73	0.84	1.09	0.95	1.17	1.05	0.85
	Variance	0.75	2.45	0.97	1.10	1.63	1.25	2.09	1.61	1.16
	Sample size	35	34	34	32	33	33	30	33	32
Day 2	**Mean**	4.03	4.56	4.63	4.76	3.76	4.01	4.15	4.13	4.66
	STD	0.89	1.4	1.11	1.00	1.14	0.92	1.22	1.14	1.24
	Min	2	2	2	2	2	2	1	2	2
	Median	4	5	5	5	4	4	4	4	5
	Max	5	7	7	7	6	6	6	6	7
	MAD	0.64	1.16	0.86	0.80	0.95	0.65	0.96	0.89	1.02
	Variance	0.79	1.98	1.24	1.01	1.29	0.85	1.50	1.29	1.54
	Sample size	30	30	31	30	30	29	26	30	30
Day 3	**Mean**	4.00	4.86	5.03	4.78	4.06	4.3	4.33	4.42	4.81
	STD	1.00	1.25	0.87	1.03	1.13	1.11	1.5	1.39	1.33
	Min	2	3	4	3	2	2	1	2	2
	Median	4	5	5	5	4	4	4.5	5	5
	Max	5	7	7	7	6	7	7	7	7
	MAD	0.77	0.97	0.63	0.82	0.89	0.92	1.20	1.18	1.02
	Variance	1.00	1.57	0.77	1.08	1.29	1.25	2.30	1.92	1.76
	Sample size	31	30	31	32	32	30	30	31	31

3.2 Results from the Incident Reports

During the five days, Lore conducted ~12 600 actions (9910 exploits, 1972 shell commands, 262 network scans, and 478 online password guessing attempts) and compromised 10 machines out of ~80 machines.

The ground truth was obtained from the network topology information of the hypothetical organization that the participants are tasked to defend and logs from the attack execution tool called SVED. Given the ground truth, we looked at if the reported victim machines (hostnames or FQDN or IPv4 addresses) and attackers (hostnames or FQDN or IPv4 addresses) machine are related to the time of the compromise reported in the incident report with some acceptable error margin in the time. *Acceptable error* refers to the tolerated error of an analyst, measured as the reported time of detection minus the actual time of the compromise. We take the reported time of the compromised machines as the measure of detecting a compromise, then the accuracy of the defence of a group can be defined as the fraction of total number of *accurately* reported compromised machines to the total number of compromised machines. We measured the accuracy of the reported time of the compromise for range of acceptable errors. As we can see from Figure 2 that even after increasing the acceptable error interval from 0 to 40, the participants have not reported single compromised machine with 100% accuracy. However, we cannot draw any conclusion based on 30 reports submitted in total for all the five days.

Figure 2. The accuracy in identifying the compromised machines



4 Conclusion

The Cybon CDX presented a great opportunity to demonstrate the capability of the red team automation tool Lore. Furthermore, Cybon was run around the clock for five days, which highlighted the advantage of having an autonomous red team tool such as Lore that did not require manual work for its operation. Most importantly, from the results we see that the participants' experiences of Lore were positive in terms of realism, pedagogical value, active learning and understanding of cyber adversarial actions. Furthermore, although the type of threats executed by Lore varied in terms of threat severity across the five days, the participants' perception of Lore remains the same.

However, there are few limitations in the study and they are:

i) All the questionnaires that were to be answered after the end of each session were in fact answered on the same day,
ii) The participants only answered the "participants' experience of Lore" questionnaire for the first three days, not for all the five days.
iii) The participants did not maintain their pseudo-ids throughout the three days data collection period, which means that the conclusion drawn about the analysts' perception of Lore when the severity of attacks varied needs further testing.
iv) The incident report template used by the participants was slightly different from the template prepared for this study (the prepared template is shown in Section 2.1).

5 References

[1] Försvarsmakten, “Cybersoldaternas första repövning,” Försvarsmakten. https://www.forsvarsmakten.se/sv/aktuellt/2022/09/cybersoldaternas-forsta-repovning/ (accessed Dec. 15, 2022).
[2] D. Granåsen, T. Sommestad, and P. Lif, “Informationselement i incidentbeskrivningar. Framtagning och utvärdering under övningen iPILOT,” FOI Totalförsvarets forskningsinstitut, FOI-R--4501--SE, Feb. 2018.
[3] T. Sommestad, “SAFE Cyber 2020, genomförande av distribuerad övning,” FOI Totalförsvarets forskningsinstitut, FOI Memo 7522, Apr. 2021.
[4] H. Holm, M. Ekstedt, and D. Andersson, “Empirical Analysis of System-Level Vulnerability Metrics through Actual Attacks,” IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 6, pp. 825–837, Nov. 2012, doi: 10.1109/TDSC.2012.66.
[5] H. Holm, “Lore A Red Team Emulation Tool,” IEEE Transactions on Dependable and Secure Computing, pp. 1–1, 2022, doi: 10.1109/TDSC.2022.3160792.