Werkplan Masterproef

Titel	Capturing realistic cyberattacks in controlled, containerized environments for the purpose of dataset creation.
Naam Student	
Email	
Bedrijf/ Onderzoeksgroep	IDLab, University of Ghent
Promotoren	
Begeleiders	

Opsplitsing per semester:

*Duid aan hoeveel studiepunten (vakken) je in elk semester volgt (behalve de masterproef). Verdeel ook de **18** studiepunten van de masterproef over de twee semesters. (Eén studiepunt komt overeen met 25-30 uren)*

	Semester 1	**Semester 2**
# studiepunten vakken	27	15
# studiepunten masterproef	4	14

Bestaande situatie en probleemstelling

Geef duidelijk aan wat er nu gebruikt wordt, en waar het probleem zich situeert (minstens een halve pagina).

An intrusion detection system is a piece of software or a device that captures traffic flowing through the network, often used as a means to look for malicious activities. Two types of IDSs exist: signature-based and anomaly-based.

As the name suggests, signature-based IDSs use a signature, a certain attack pattern which is often stored in a database, in order to identify an attack. This approach is effective against well-known attacks, but rather ineffective against new attacks, or even well-known attacks that were modified.

The goal of an anomaly-based IDS is to distinguish between normal and malicious network traffic through extensive monitoring. Typically, the first step is to put the IDS in the real work environment. The IDS will monitor the normal traffic for a certain period, this could be days or even weeks. The monitored traffic will be used as a basis for the benign activities in the network. Once the IDS is deployed, any irregularities in the network traffic will be flagged for review. This sort of IDS is safer against new and obfuscated attacks.

The problem with anomaly-based IDSs is the amount of good datasets readily available, due to the lack of a well-defined and structured approach in generating the datasets. There are certain requirements that must be met in order for a dataset to be considered useful.

First of all, the dataset must be recent or at least up to date with the latest technologies. Network traffic evolves as new technologies are created, or new attacks are discovered.

The creator(s) of the dataset should consider generating the data in a real work environment, instead of a virtual environment, to ensure the traffic is realistic. There is an abundance of factors that go into what makes an environment what it is, which would make it nearly impossible to be achieved using a virtual environment, e.g. response times, downtime, erratic behavior (of a user or system), working hours and breaks (to get coffee or work on an offline document).

The network topology might just be the most important factor. This entirely depends on the internal structure of the network that the IDS will be deployed in and differs from network to network. If the goal of an IDS is to be deployed in multiple environments different environments, one must consider the differences in operating systems, router software, network protocols used, etc.

A realistic combination of normal and malicious traffic is essential. The normal traffic is generated by the normal users, while the malicious traffic can be generated by an active attacker inside the network or even outside the network if there is a connection with an external server. Both forms of traffic can be generated using scripts, but with careful consideration of the factors that make it realistic which were mentioned earlier.

The type of attacks performed in the network could vary. If a certain type of attack is prominent in the network, the possible attack scenarios should be exhausted to provide a big enough safety net. If nation state resources are required to perform a certain attack, don't invest too much time into this.

Datasets must contain data that is properly labelled if it is to be used for training and testing of a machine learning algorithm, which is often the case with anomaly-based IDSs.

The dataset should be publicly available. This is often not possible due to privacy concerns, which means a certain degree of anonymisation is required. Anonymisation could reduce the quality of the dataset if done improperly.

Doelstelling van de masterproef

Beschrijf de concrete opdracht voor de masterproef – dit moet meer gedetailleerd zijn dan in het voorstel op Plato (geen copy/paste – minstens een halve pagina).

The goal of this dissertation is to capture a variety of realistic cyberattacks in a well-defined, controlled and containerized environment for the purpose of creating a useful, labelled dataset.

The focus will be on internal clients performing insider attacks, such as port scans, brute force and denial of service. Possible expansion could be made to attacks against specific network protocols or web applications.

The attack methods consist of two categories. Some attacks will be manual attacks performed by an attacker from within the network. This could be a malicious employee or someone unknowingly performing actions outside of the predefined set of actions assigned to that user based on misuse of company policy and/or broken access control. On the other hand, these attacks could be made by an active attacker suitably placed in the network, or an attacker that has gained access to an internet-facing system, using a wide variety of tools most commonly

used in contemporary attacks. Most of the attacks will be scripted attacks, which can be automated and set to be executed at certain points in time on predefined, but varying, amount of systems. These systems will have different configurations and security measures in place and different operating systems in order to ensure heterogeneity in network topologies.

The attacks will be contained within the internal network. There will not be any external servers that have access to the internal network, therefor disallowing attacks from the outside to seep into the network or inside attacks to propagate outside the internal network.

As most attacks will be scripted, the generated malicious traffic will be easier to label as the traffic flows will be extensively logged. The labelling process consists of classifying the type of attack, identifying the attack type and providing a description of the attack. The temporal sequence of the captured traffic will prove useful to distinguish certain events.

The benefit of using a simulated environment is that the traffic does not have to be anonymized, so once the traffic is labelled, the dataset can be documented and published for future use.

Planning en mijlpalen

Definieer enkele duidelijk afgebakende taken. Vermeld voor elke taak het aantal weken dat je hiervoor nodig zal hebben en een concrete deadline. Geef een concrete omschrijving van het doel en wat er op het einde van de taak afgeleverd zal worden. Een typische taak neemt 3 tot 5 weken in beslag. Taken kunnen overlappen (bv. Schrijven van de scriptie en uitvoeren van aanvullende experimenten). Hieronder een aantal voorbeelden.

Task 1	*1 week*	*Deadline: 20/10/2023*	**Work plan**
Content			
Write the history and the current situation regarding this topic and the problem that will be tackled in this thesis. Scheduling the different sections that go into writing the thesis and performing research.			
Results, deliverables and insights:			
The main result is the work plan, but also gaining insight in the different tasks that have to be performed and approximating how long these will take.			

Task 2	*4 weeks*	*Deadline: 01/11/2023*	**Exploratory literature study**
Content			
Reading papers about dataset creation for anomaly-based intrusion detection systems and writing down my thoughts, while comparing different papers.			
Results, deliverables and insights:			
Sufficient knowledge about how to (not) create a good dataset, based on experiences from researchers in this particular field. The knowledge obtained will be summarized and subsequently discussed with the mentors in order to construct a baseline for further research. Exploring the benefits of certain methodologies will prove useful for the next task.			

Task 3	*2 weeks*	*Deadline: 15/11/2023*	**Technology exploration**
Content			

Researching certain tools and methodologies and experimenting with technologies to get comfortable with their use. The platform used to provide the containerized environment will also be researched and experimented on using a publicly available dataset.
Reading more papers for the literature study and writing down my thoughts.

Results, deliverables and insights:

A thorough understanding of the technologies that will be used throughout this thesis. It's evident that some methodologies will prove to be more useful than others. Findings and perhaps uncertainties will be summarized and briefly communicated with the mentors during these two weeks and a conclusion should be reached at the end of this task as to which technologies will be used. If this approach proves to be ineffective along the way, this can still be adapted to certain needs.

Task 4	*2 weeks*	*Deadline: 29/11/2023*	**Intermediate presentation about the environment's creation**

Content

This period marks the start of the implementation of the structure that will be used in the containerized environment, this includes configuration of the systems (such as intermediary devices, clients and servers) and setting up the network.
Reading more papers for the literature study and writing down my thoughts.

Results, deliverables and insights:

A brief presentation to the mentors will be given regarding the setup that has been created so far. Insights will be gained about the design choices and why they are ideal or could be improved:

- Misconfiguration of systems or security measures
- Illogical structure of the network
- Myriad (or the lack of) systems or services to reduce (or increase) the scope
- …

Task 5	*2 weeks*	*Deadline: 13/12/2023*	**Finalization of the environment's creation**

Content

During this period, the insights gained from the presentation at the end of the previous task will be taken into consideration to adapt or further expand the current environment. Any uncertainties will be communicated with the mentors.
Reading more papers for the literature study and writing down my thoughts.

Results, deliverables and insights:

The finished environment will be in place and a brief presentation will be given to the mentors how the network is essentially structured and its inner workings.

Task 6	*1 week*	*Deadline: 20/12/2023*	**Attack plan research**

Content

Researching which attacks will be used to create the malicious traffic and how these attacks work in a general sense.
Start putting the notes from the papers together and phrasing them in a proper manner.

Results, deliverables and insights:

A list of possible attacks and knowledge about these attacks.

Task 7	*1 week*	*Deadline: 30/12/2023*	**Intermediate presentation**

Content

Giving a presentation to the mentors and the promotors about the current state of the literature study and the overall research I've done up until that point. Touch on the subject of the technologies explored and used. Frame the work that will be performed in the second semester based on the experiences in the first semester and the insights gained.
Continue writing the literature study part of the thesis.

Results, deliverables and insights:

Feedback on what I've already accomplished so far will give me the motivation to keep doing what I'm doing or switch it up entirely, probably somewhere in between.

Task 8	*2 weeks*	*Deadline:* 28/02/2024	**Writing/modifying attack scripts**

Content

Researching and writing attack scripts or modifying existing attack scripts and trying these out on the containerized environment in order to test their effectiveness.
Continue writing the literature study part of the thesis.

Results, deliverables and insights:

Hopefully not a broken environment. Some insight should be gained in the workings of these attacks and how it's possible to tell from certain flows if they were (un)successful. If some scripts don't work to a certain extent, they can be modified or used as an example of a failed attack, only if they were to reach an appropriate stage, it would be less interesting to see an attack fail in the first stage, e.g. due to an unavailable host or simply a wrong IP. These results will be discussed with the mentors.

Task 9	*2 weeks*	*Deadline:* 13/03/2024	**Researching feature extraction**

Content

A lot of traffic will be generated by certain attacks, features need to be extracted to be able to label these traffic flows. Knowledge about feature extraction from raw data will be gained during this period. The insights gained will be summarized and discussed with the mentors.
Continue writing the literature study part of the thesis.

Results, deliverables and insights:

The methodology used for feature extraction will be clear. The details for the feature extraction itself will be decided upon during the discussion with the mentors.

Task 10	*2 weeks*	*Deadline:* 31/03/2024	**Thesis – first 25 pages**

Content

All of the knowledge and documented implementation specifics gathered during the first semester and the second semester up until now will be written down.
Finish writing the literature study part of the thesis.

Results, deliverables and insights:

The first 25 pages should give an insight regarding the history in this particular field, the problem this thesis wants to tackle, the design choices regarding the containerized

environment and the selection of attacks. The literature study will be finished at this point, so that this won't be in the back of my mind moving on.

Task 11	*1 week*	*Deadline:* 01/04/2024	**Administrative data**

Content

Fill in promotor + co-promotor fields, the final title of the thesis and the language of the thesis.

Results, deliverables and insights:

Satisfaction from filling in some fields.

Task 12	*2 weeks*	*Deadline:* 10/04/2024	**Intermediate presentation about feature extraction and data cleaning**

Content

Features will be extracted from the generated traffic as preparation for the labelling.
Start writing research part of the thesis in a proper manner.
The company visit is likely going to take place during these two weeks, however this is still uncertain.

Results, deliverables and insights:

The extracted features will be cleaned up and ready to be labelled. It's important to think about the effects that this will have. If the data cleanup is not properly done, this will almost certainly have an adverse effect on the quality of the dataset. The presentation should be a way to convey the reasoning behind the feature extraction and cleaning up until that point, in order for the mentors to provide constructive criticism.

Task 13	*1 week*	*Deadline:* 17/04/2024	**Finishing feature extraction and data cleaning**

Content

The improvements mentioned during the last presentation will be used to extract better features or perform a certain subtasks of data cleaning better.
Keep writing the research part of the thesis.

Results, deliverables and insights:

The extracted features and cleaned up data, ready to be labelled.

Task 14	*1 week*	*Deadline:* 24/04/2024	**Labelling**

Content

Most of this week will be spent labelling the data in order for the machine learning algorithm to know which type of attack it's dealing with, this way it will be able to learn the underlying patterns and distinguish the type of attack on new data.
Keep writing the research part of the thesis.

Results, deliverables and insights:

Primarily labelled data, but also insight in the effect certain features have on the result that ultimately characterizes the type of attack.

Task 15	*2 weeks*	*Deadline:* 08/05/2024	**Visualisations and statistical analysis**

Content Work on the visualisations that are to be used in the thesis. This could be graphs or other statistical visualisations, such as boxplots or scatterplots. Which one is to be used depends on the type of data and the distribution.
Results, deliverables and insights: Correct and comprehensible visualisations, e.g. to get an idea why X is better than Y in every case where Z presents itself.

Task 16	*4 weeks*	*Deadline:* 25/05/2024	**Deliver 95% version of thesis**
Content Keep working on the thesis until it's almost finished, like adding a conclusion, an extended abstract, a normal abstract, etc. If something small were to be forgotten this can still be added during the next two weeks.			
Results, deliverables and insights: The 95% version of the thesis and getting an idea of what still has to be added or changed before finalizing the thesis and submitting the result.			

Task 17	*2 weeks*	*Deadline:* 06/06/2024	**Finalizing the thesis**
Content Finalizing the thesis.			
Results, deliverables and insights: The final thesis.			

Task 18	*3 weeks*	*Deadline:* 26/06/2024 – 28/06/2024	**Final presentation**
Content Preparing the final presentation. The presentation should cover the final state of the literature study and the overall research I've done. The visualisation and statistical analysis will help to bring my point across.			
Results, deliverables and insights: The final presentation.			

Contactmomenten

Vermeld duidelijk hoe de tussentijdse communicatie met de begeleiders en promotoren zal verlopen. Geef bijvoorbeeld aan of er een wekelijkse/ tweewekelijkse/ maandelijkse rapportering zal zijn en of die fysiek, digitaal of via e-mail zal doorgaan.

In the first semester, an email will be sent biweekly with the progress made during the previous two weeks, bullet point style. In the second semester this email will be sent weekly. If there are any questions, these can be sent using email as well. If there are any complications with implementation or any misconceptions, an online meeting can be arranged.

Gantt chart:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Week 1: 25/09																		
Week 2: 02/10																		
Week 3: 09/10	X	X																
Week 4: 16/10	X	X																
Week 5: 23/10		X																
Week 6: 30/10		X																
Week 7: 06/11			X															
Week 8: 13/11			X															
Week 9: 20/11				X														
Week 10: 27/11				X														
Week 11: 04/12					X													
Week 12: 11/12					X													
Week 13: 18/12						X												
Kerstvakantie:							X											
Examenperiode januari																		
Week 1: 12/02								X										
Week 2: 19/02								X										
Week 3: 26/02								X										
Week 4: 04/03									X									
Week 5: 11/03									X									
Week 6: 18/03										X								
Week 7: 25/03										X								
Paasvakantie											X	X						
Week 8: 15/04													X					
Week 9: 22/04														X				
Week 10: 29/04															X			
Week 11: 06/05															X			
Week 12: 13/05																X		
Week 13: 20/05																X		
Week 14: 27/05																	X	
Week 15: 03/06																	X	
Week 16: 10/06																		X
Week 17: 17/06																		X
Week 18: 24/06																		X