Process Mining in Cyber Forensics

- a Methodology for Process Mining in Forensic Investigation of Web Application Activity –

T.T.A (Tim) ter Voert

Snr: ***2108981*** *Anr:* ***552131***

Abstract: *This paper presents an inquiry into combining the research fields of Process Mining and Cyber Forensics. It aims to formulate an answer to the question as to how, and to what extent, Process Mining can be applied in Web Application Forensics. It does so by designing a methodology: PM^2 for Web Application Forensics. The methodology and the approach in general are evaluated and assessed by Expert Panel interviews. The results suggest that the developed methodology can provide a useful guide to apply Process Mining in Web Application Forensics cases. However, there are several limitations to the scope and requirements that need to be in place to make it a successful endeavor.*

January 2024

Internship company:	Joanknecht
Study:	**Master of Information Management (MSc IM)**
Department:	**Tilburg School of Economics and Management (TiSEM)**
University:	**Tilburg University**
Supervisors:	**dr.ir. F. (Francesco) Lelli | Tilburg University**
	dr. Shiva Shekhar | Tilburg University (Co-reader)
	drs. L.R.M. (Lucas) Vousten RA RE | Joanknecht

Table of Contents

Management Summary

Cyber Security has risen to become one of the most influential factors in the modern business environment. However, there is still a long way to go as cyber attacks keep impairing businesses all over. This suggests that the current attack detection and prevention techniques are not able to wear off all users with malicious intent, and thus forensic investigation of attacks becomes ever so important. One of the most favorable targets for attacks are Web Applications, which are widely used by businesses and often contain vulnerabilities. Over the last years, Process Mining has become a more and more widely exploited and used technique in academia and business. Although applied in various fields, there is little coverage on using Process Mining in Cyber Forensics investigations. At Joanknecht, the firm where the internship for this master's thesis is conducted, there is interest in such application. As their IT and forensics teams have started working together recently, they are eager to find out what collaborations are feasible and can add value to offer as services to clients. Therefore, this study aims to exploit this novel area of research by answering to the question as to how, and to what extent, Process Mining can be applied in forensic investigations of malicious activity on Web Applications. It does so by developing a methodology based on an existing Process Mining Project Methodology (PM^2), which is applied and tailored to two Web Application Forensics cases and results in artifact *PM^2 for Web Application Forensics*. To validate the artifact and gain insight on the added value of the approach in general, an Expert Panel is consulted. During interviews the artifact is evaluated, and opportunities and limitations of the approach are assessed by the experts. The results suggest that developed methodology can provide a useful guide to apply Process Mining in Web Application Forensics cases. However, to make it a successful endeavor, the limitations to the scope and requirements that need to be in place, that are mentioned in this study, should be taken into consideration.

Preface

This master's thesis has been written in fulfillment of my MSc Information Management at Tilburg University, between the period of September 2023 until January 2024. First of all, I would like to declare that this full paper is a product of my own writing. Although ChatGPT was occasionally used to clarify concepts for myself or for inspiration, it was not used to generate any argumentation or text that is present in this master's thesis.

After one year of passing the required courses and taking part in the Extended Master's Program (EMP) in the spring 2023 cycle, only my thesis was left to complete in the fall 2023 semester. I conducted my internship for the EMP in the financial- and IT-audit teams at Joanknecht, an audit and advisory firm with offices in Eindhoven and Amsterdam. Fortunately, after the internship I was allowed to stay as a working student while simultaneously write my master's thesis at Joanknecht, for which I am very appreciative. During the challenging, but enlightening process of writing this final piece, there have been several persons that have contributed significantly to the end product. First of all, I would like to thank Francesco Lelli for his continuous critical, but fair, feedback on my work. Secondly, a sincere '*Thank You*' to the Expert Panel participants for their time and valuable insights that helped me answer the research question of this study. Thirdly, I very much appreciate the tips, useful discussions and motivational words from my colleagues at Joanknecht during the last five months. There is one person who deserves an individual mention, which is my supervisor Lucas Vousten. Without his infecting enthusiasm and extensive knowledge, as well as his on-point feedback, writing this master's thesis would certainly not have been possible. Finally, I would like to thank my family for letting me come live home again in '*de Achterhoek*' to write this piece of work, after having lived in Tilburg for a year. Your support, feedback and tolerance of my sometimes-tiresome moods will not be neglected.

Etten, January 12, 2024

T.T.A. (Tim) ter Voert

1. Introduction

In the last two decades, after van der Aalst & Weijters (2004) introduced academia to Process Mining and urged for more research, the body of literature on the data analysis technique has expanded significantly. Process Mining can be seen as *"a means to bridge the gap between data science and process science"*, as the so called Godfather of Process Mining, Van Der Aalst (2016), describes it. This also raised the attention of businesses, who have grown fond of the possibilities Process Mining offers. Popular fields of application vary from Auditing (Accorsi et al., 2013; Accorsi & Stocker, 2012) to Healthcare (Mans et al., 2009, 2009; Munoz-Gama et al., 2022).

In more recent years, Cyber Security has become a notable theme for both academic researchers and businesses. Impactful cyber attacks have become far too common among organizations. One of the most infamous examples is Facebook's data breach in 2018 that lead to 50 million users' personal information being exposed (Isaac & Frenkel, 2018). Efforts to gain more knowledge on the problem resulted in various studies on how Cyber Security can be improved, including studies employing Process Mining. Cyberattacks also are processes over time and, therefore, Process Mining can be used to gain valuable insights. For example, research has been carried out about developing and improving systems that employ Process Mining analyses to prevent Cyberattacks: Intrusion Detection Systems (de Alvarenga et al., 2018; Mishra et al., 2019; Myers et al., 2018). However, reports of businesses being victimized by cybercriminals still flood newspapers all over the world as cybercrime continues to increase. Forbes recently reported that costs related to cybercrime are predicted to hit $8 trillion in 2023 and grow to $10.5 trillion towards 2025 (Brooks, 2023).

Such predictions indicate that measures that aim to detect and prevent cyber attacks are not sufficient to solve the problem. Therefore, past attacks should be investigated to learn from flaws that allowed cyber attacks to be successful, by employing Cyber Forensics techniques. Cyber Forensics is *"a branch of forensic science that focuses on identifying, acquiring, processing, analyzing, and reporting on data stored electronically"* (Interpol, n.d.). There are various techniques and tools used in Cyber Forensics to investigate cyber attacks. However, various challenges have been identified by researchers with respect to Cyber Forensics tools that are currently available, which result in low-quality investigations (Fernando, 2021). Although Process Mining has been applied in preventing and detecting cyber attacks in many studies, literature on Process Mining being applied to investigate cyber attacks is relatively scarce. A study that did employ Process Mining techniques investigated malicious authentication events by finding relations among events (Lagraa & State, 2020). The researchers also pointed out that the body of literature using Process Mining for investigating attacks is thin, as most Cyber Security research focuses on detecting and preventing attacks. Recent literature review studies on Process Mining in Cyber Security confirm this claim, as the most popular area of research is on the detection of various attacks and fraud (Macak et al., 2022; Silalahi et al., 2022).

Web Applications are a favorable target for cyber criminals. They are used widely by businesses as they are known for their accessibility, efficient development, user simplicity and scalability (AWS, n.d.). Due to the intensive use of the web, their servers are targets of attacks, ranging from information leak vulnerabilities to complete infrastructure takeovers (Nazar et al., 2021). This is not without reason, as defense mechanisms for Web Applications often are not secure. Improperly coded filters and misconfigured Web Application firewalls will not block all malicious user input (Huang et al., 2017). *Acunetix*, producer of security scanner software, points out that this is because most Web Applications are custom made, and therefore are less tested regarding their security (Acunetix, n.d.). The software company also highlights that Web Applications are publicly accessible, which makes it vulnerable in case an attacker finds a weakness in the application. Huang et al. (2017) further argue that improving defense mechanisms is a never-ending process, as criminals will continue to find new ways attack Web applications. Finally, a very recent incident, involving *KLM Royal Dutch Airlines,* again proved how vulnerable Web Applications often are. Namely, unauthorized users were able to extract sensitive customer information using very simple techniques (Schellevis, 2023).

This master's thesis aims to contribute to the gap in current literature on Process Mining in Cyber Forensics, in which Web Application Forensics is the scope, in two ways. First, it presents a methodology on how to apply Process Mining in Web Application Forensics cases. Secondly, the extent to which Process Mining in Web Application Forensics can be useful is assessed.

In short, Web Applications are very useful tools for businesses for various purposes. However, due to their vulnerabilities they are a favored target by cyber criminals, which leads to serious harm to businesses. This research aims to contribute to the academic fields of Process Mining and Cyber Forensics by answering the following research question:

"How, and to what extent, can Process Mining be applied in forensic investigation of malicious activity on Web Applications?"

To provide an answer to this research question, the empirical research is composed of the following components. First, the Process Mining Project Methodology (PM^2) by van Eck et al. (2015) is revised to create *PM^2 for Web Application Forensics* based on applying the methodology to two Web Application Forensics cases. Finally, to further validate the methodology and the approach in general, an Expert Panel is interviewed. Therefore, the product of this master's thesis is twofold. It presents both a methodology for applying Process Mining in Web Application Forensics, as well as an assessment of the opportunities and limitations of the approach. This research differs from the research by Lagraa & State (2020), as their focus was on applying Process Mining for malicious authentication attempts while this study focuses on malicious activity on Web Applications. Moreover, this study presents a methodology that can be applied to other cases. Finally, next to a case application, this study validates the approach based on interviews with experts.

The managerial relevance of this research concerns the great deal of companies that make use of Web Applications. In case a business makes use of Web Applications, they naturally have an interest in the security of these applications. Businesses that want to gain better insight in the process behind malicious Web Application activity could take advantage of this research.

This research has academic relevance in terms of expanding knowledge on Process Mining applications and Cyber Forensics. The technique has been applied to detect and prevent cyber attacks, but literature on the application in Cyber Forensics, including subdomains like Web Application Forensics, is scarce. Therefore, a new application for Process Mining is assessed in the literature. Further, this study can inspire researchers to apply a Process Mining approach to other domains or apply this methodology to other cases in Web Application Forensics.

This thesis is written in collaboration with Joanknecht, an audit and advisory firm located in Eindhoven & Amsterdam, the Netherlands. They offer services like assurance, tax advisory, forensics & recovery, IT-assurance, IT-services, and real estate advisory to SME's in a national and international context. Being an innovative company, Joanknecht is eager to find new ways to help their clients improving their businesses. In recent times, the forensics and IT teams have started working together to look at offering IT forensics services. Process Mining is one of the techniques that is being used for such business cases. Consequently, this study could provide proof of instances in which Process Mining can be applied, helping Joanknecht in expanding the possibilities of services that they offer.

The remainder of this paper is structured as follows:

- *Chapter two* introduces relevant concepts, reviews related works and indicates the gap in current literature.
- In C*hapter three*, the Web Applications and data used are described first. Subsequently, the methodology for applying Process Mining in Web Application Forensics is created, refined and presented. Finally, the setup for the interviews is described.
- In *Chapter four*, the results from the interviews with the Expert Panel will be laid out.
- *Chapter five* formulates an answer to the research question, presents limitations and suggestions for future research.
- Finally, this master's thesis is summarized in the conclusion in *Chapter six*.

2. Literature Review

This chapter aims to gain a deeper understanding of the domains that this research interacts with, which are depicted in *Figure 1.* For simplicity, only a few examples of related domains are given for each instance. Moreover, there is no single categorization of the (sub)domains. The depicted categorization is according to the author's view based on the literature. First of all, Process Mining, a combination of Data Analytics and Business Process Management, is the method of analysis in this master's thesis. Within Process Mining, a distinction is often made between process discovery, conformance checking and process enhancement. To this study, process discovery and conformance checking are of relevance. Cyber Security & Forensics are the other domains related to this research. A way to classify Cyber Security is in the three aspects as also shown in *Figure 1*. The technologies related to Cyber Security can be divided in applications that prevent, detect or investigate cyber attacks. The investigation of attacks within Cyber Security is, although not apparent in the figure, closely related to Cyber Forensics. The specific Cyber Forensics domain this thesis focuses on is the investigation of malicious behavior on Web Applications. For this research, the data is from web

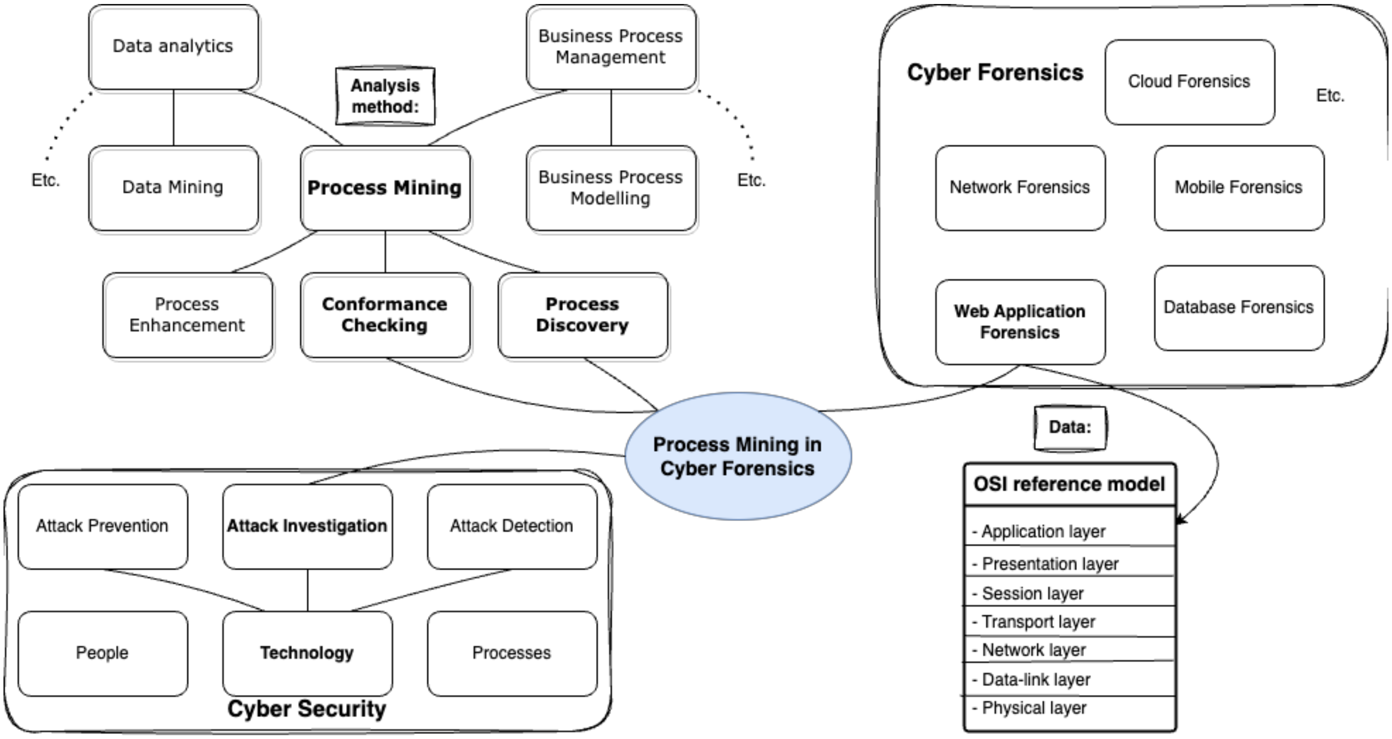


Figure 1: Domain overview of this master's thesis (source: own work)

server access logs, which is application layer data in the *Open Systems Interconnection* (OSI) architecture (Day & Zimmermann, 1983).

This chapter is further structured as follows. *Section 2.1* introduces and discusses Process Mining, followed by the Cyber Security & Forensics domains in *Section 2.2*. Both sections start with book-style domain overviews of the relevant concepts. Subsequently, related works of the concepts are reviewed, mainly including studies that involve multiple of the relevant concepts. Finally, Web Applications and their security issues are introduced. This chapter is wrapped up in *Section 2.4* with a conclusion on the reviewed literature, indicating the research gap that this master's thesis aims to contribute to.

2.1 Process Mining

The following subsections give an introduction to the concept of Process Mining. Further, techniques, algorithms and applicable tools will be discussed and compared in light of this study to identify suitable approaches. Finally, Process Mining applications in other fields are reviewed that give relevant insights and takeaways.

2.1.1 An Introduction to Process Mining

To understand in what way Process Mining bridges the gap between data science and process science, both fields of study and their shortcomings will shortly be explained. A data scientist is someone who aims to turn data into value for organizations by answering data-driven questions. Fields of data science include data mining: *"the discovery of interesting, unexpected or valuable structures in large datasets"* (Hand, 2007 p.621) and machine learning: *"the question how to construct computer programs that automatically improve with experience"* (Mitchell, 1997 p.XV). Process science is by Van Der Aalst (2016 p.16) referred to as *"the broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes".* Both fields miss something that the other complements. To specify, data science tend to be process agnostic as they do not consider end-to-end processes. Process science, on the other hand, often focus on modeling instead of analyzing event data. Process Mining tackles these shortcomings by combining event data with

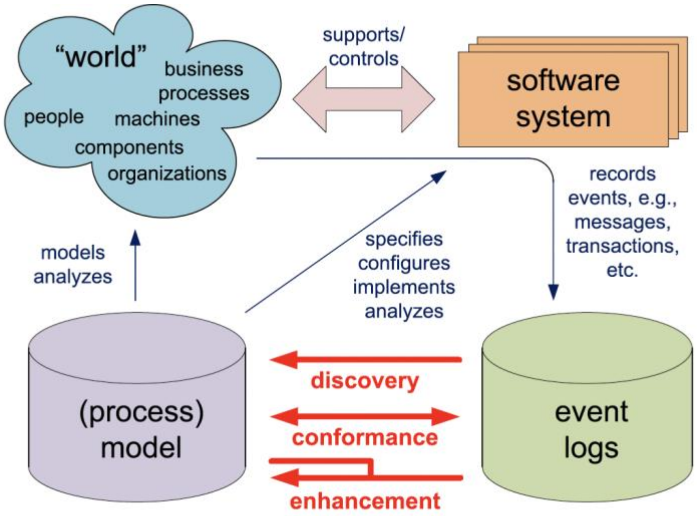


Figure 2: "Process Mining establishes links between the actual processes and their data on one hand, and process models on the other hand" (Van Der Aalst, 2016 p.32)

process models (Van Der Aalst, 2016). Process Mining techniques enable creating process models based on event data and the other way around. *Figure 2* presents a visual of the links between the main components related to Process Mining. Software systems support and control processes from the real world. In doing so, these systems create extensive amounts of data, called event logs. These event logs can be employed to interact with (process) models. Specifically, their relation entails three types of Process Mining, as explained by Van Der Aalst (2016):

- **Process discovery** is a technique that uses the data from an event log to create a model such that it should represent the actual process from the real world. The idea is to discover and visualize how a process One of the first discovery techniques was the α-algorithm, which can create a Petri net out of an event log. These concepts, among others, will be discussed further in *Section 2.1.2*.

- **Conformance checking** is a technique that is used in both ways. In this case, a model of how a process should flow can be compared to data from an event log and vice versa. This technique is used to check whether the actual process, the event logs, follow the path as the process owners designed it. Therefore, it is used to detect deviations from the path that should be followed which can be useful when investigating potential fraud for example. Take the "four-eyes" principle, which requires two people to approve for a certain action. Conformance checking techniques are able to detect when this requirement is violated and, thus, there may be a case of fraud.

- **Process enhancement** uses data from the event logs, which reflect the real process, to extend or improve the existing process model. It is an iterative process aiming for two types of enhancement. One can aim to repair models, i.e., altering the existing process model to better reflect reality. The other type focuses on extension of the model. That is, adding extra information to the process model by showing performance data. For instance, by using time stamps in the processes, it is possible to visualize bottle necks, service levels, throughput times, and frequencies.

In relation to the approach in this study, process discovery and conformance checking techniques will be used. Process enhancement is also called performance analysis, as it is often used to identify activities in the process that are problematic (W. M. P. van der Aalst et al., 2017). This type of Process Mining does not suit this study, as it focuses on improving a process based on a single hand-made model. However, the 'process' in this study is web activity, which is not a single process that needs improvement. Instead, this study aims to investigate paths that malicious users take to gain unauthorized access to Web Applications. Process discovery is fitting as it is supposed to put out a model that exactly represents the real processes, based on the event logs. After the process is discovered, the malicious paths ought to be identified. Using conformance techniques, malicious activity will be filtered out and investigated further. To execute these steps, a Process Mining algorithm is needed. Therefore, the next section will discuss some relevant algorithms relevant for this study.

2.1.2 Process Mining Algorithms

In the paragraphs below, three algorithms will be discussed. It is not the aim to present a comprehensive list of all relevant algorithms, but rather a discussion of a few prevalent algorithms and ones related to this research. First, the α-algorithm is discussed, followed by the Heuristics Miner, which was able to deal with problems traditional algorithms faced. Lastly, the Fuzzy Miner algorithm, closely related to the Heuristics Miner (De Weerdt et al., 2012), will be discussed.

- The **α-algorithm** (W. van der Aalst et al., 2004) was one of the first process discovery techniques that was able to deal with a problem that most classical approaches struggled with: concurrency (Van Der Aalst, 2016). An algorithm that allows for concurrency is able to capture tasks that can be executed in parallel (Cook & Wolf, 1998). The α-algorithm is able to discover a process model from a workflow log and present it in the form of a sound Petri net

(W. van der Aalst et al., 2004). A simple example of a Petri net is shown in *Figure 3*, based on the workflow log from *Table 1*. The process always starts with task A. Then, either B and C are executed, or E is executed. There is parallelism/concurrency between B and C. Both are executed at the same time, but there is no particular order in which finishes first. Lastly, all cases end with task D. With this model, it is possible to replay any of the cases from the workflow log. Although praised for its simplicity, the α-algorithm has problems with noise, defined as infrequent/incomplete behavior, and complex paths, and is therefore not very practical (Van Der Aalst, 2016). For these reasons, the α-algorithm is not suitable for this research, especially since this study investigates infrequent behavior particularly.

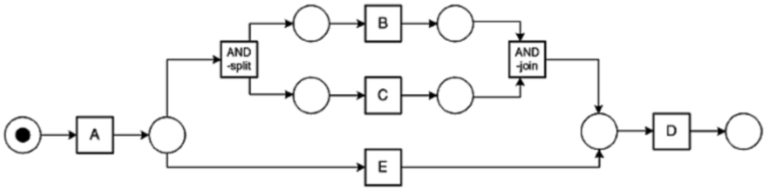
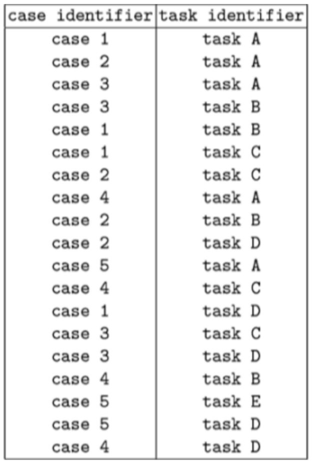


Figure 3: A Petri net based on the workflow log in Table 1 (W. van der Aalst et al., 2004)



case identifier	task identifier
case 1	task A
case 2	task A
case 3	task A
case 3	task B
case 1	task B
case 1	task C
case 2	task C
case 4	task A
case 2	task B
case 2	task D
case 5	task A
case 4	task C
case 1	task D
case 3	task C
case 3	task D
case 4	task B
case 5	task E
case 5	task D
case 4	task D

Table 1: A simple workflow log (W. van der Aalst et al., 2004)

- The **Heuristics Miner** (Weijters et al., 2006) algorithm is a more practical applicable algorithm from a heuristics driven approach. It is particularly good at handling noise in the data and can deal with low frequent behavior. The algorithm is able to focus on the full behavior of an event log, as well as show only the main behavior. It does so by taking frequency of paths into account and by visualizing in the form of Causal Nets. Causal Nets are tailored for Process Mining, using nodes as activities and arcs for causal dependencies (Van Der Aalst, 2016). *Figures 4 & 5* on the next page are Causal Net based on even log L in *Equation 1,* also on the next page. They are the same model, but Figure 5 represents a more clear and intuitive view of the model. Based on a certain threshold, some causal dependencies are not included in the Causal Nets. For example, in 20 of 40 cases, a is followed by b and c concurrently in the models. However, in log L there also is a trace where only c follows a, which is not represented in the model. This is the result of the threshold that determines how extensive the model is and shows how the Heuristics Miner algorithm deals with noise (Van Der Aalst, 2016). This is one of the reasons why a multi-dimensional evaluation study on process discovery algorithms found that in terms of accuracy, comprehensibility and scalability, the Heuristics Miner algorithm is the best applicable for analyses in a real-life context (De Weerdt et al., 2012).

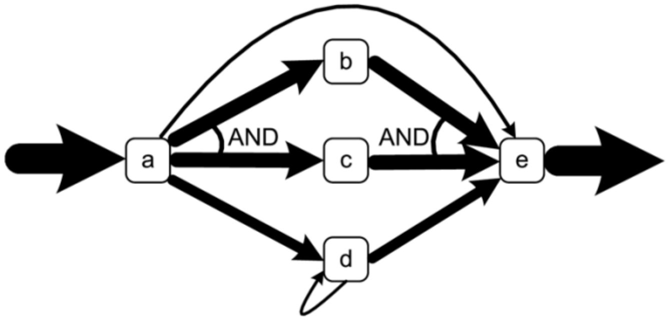


Figure 4: Causal Net based on event log L in Equation 1, with arc thickness representing frequencies (Van Der Aalst, 2016 p.208)

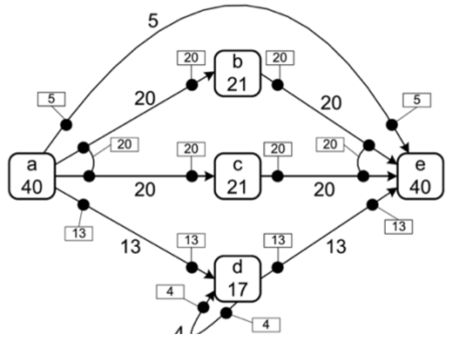


Figure 5: Causal Net based on event log L in Equation 1, with numbers representing frequencies (Van der Aalst, 2016 p.207)

$$\mathrm{L} = [\langle \mathrm{a}, \mathrm{e}\rangle^{5}, \langle \mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{e}\rangle^{10}, \langle \mathrm{a}, \mathrm{c}, \mathrm{b}, \mathrm{e}\rangle^{10}, \langle \mathrm{a}, \mathrm{b}, \mathrm{e}\rangle^{1}, \langle \mathrm{a}, \mathrm{c}, \mathrm{e}\rangle^{1}, \langle \mathrm{a}, \mathrm{d}, \mathrm{e}\rangle^{10}, \langle \mathrm{a}, \mathrm{d}, \mathrm{d}, \mathrm{e}\rangle^{2}, \langle \mathrm{a}, \mathrm{d}, \mathrm{d}, \mathrm{d}, \mathrm{e}\rangle^{1}]$$

Equation 1: Event log including frequencies (Van Der Aalst, 2016 p.202

- The **Fuzzy Miner** (C. W. Günther & van der Aalst, 2007) algorithm is closely related to the Heuristics Miner algorithm, but focuses on dealing with very unstructured data. Often, real-life processes actually are less structured than what is generally thought by people. When applying process discovery techniques, this often results in a 'spaghetti-like model' (C. W. Günther & van der Aalst, 2007). These situations mostly arise in processes in which there is so called self-directed behavior. That is, people who act in less structured environments in which result in a lot of one-of-a-kind cases. Although these logs are harder to analyze, they do offer the most potential in findings (Van Der Aalst, 2016). The Fuzzy Miner algorithm distinguishes itself from others because of its ability to look at all aspects of a process at once, its interactive and exploitive nature, and its integrated simplification algorithm (C. W. Günther & van der Aalst, 2007).

As mentioned already, the α-algorithm not a suitable solution for practical problems for various reason, but rather an easy-to-understand example to gain a first insight in Process Mining algorithms. On the other hand, the Heuristics Miner and Fuzzy Miner algorithms do offer possibilities to be applied in real-life cases. The Heuristics Miner algorithm was found to be best applicable in such cases (De Weerdt et al., 2012). However, this study did not take the Fuzzy Miner algorithm into account as the researchers argue that it is an alternative discovery technique. Still, De Weerdt et al. acknowledge the flexibility of the Fuzzy Miner algorithm as it provides good abstraction capabilities and urged for further research about less restrictive algorithms. Going back to the topic of this study, it can be argued that the Fuzzy Miner algorithm suits best for the approach. Its abstraction capabilities enable analyses from various aspects and deal with unstructured data. The analysis in this study explicitly focuses on extracting specific behavior that occurs less frequent, namely malicious behavior on Web Applications. This behavior is everything but structured, which raises the expectation that the Fuzzy Miner algorithm is the most fitting solution in this case.

2.1.3 Process Mining Tools

To execute the analysis in this research, a data analysis tool needs to be used. Software on data mining and business intelligence, however, rarely include Process Mining techniques (Van Der Aalst, 2016). Fortunately, in the last two decades since Process Mining interest in scientific literature and business rose, various open-source and commercial tools have been developed specifically for Process Mining applications. Van Der Aalst (2016 p.328) characterizes Process Mining tools by asking two questions:

- *"How often is the same analysis repeated?*
- *Can the end-user freely determine the analysis to be conducted?"*

By answering these questions, one is able to derive what kind of Process Mining tool to use. In this study, the data analysis only occurs twice, namely the initial analysis of the generated dataset, and the validation of the approach against the live dataset. Further, the analysis needs to be executed freely in order to find relations among events. Below, two of the most prominent Process Mining tools (Dakic et al., 2019) will be discussed to find a suitable tool to conduct the research.

ProM (van Dongen et al., 2005) currently is the open-source Process Mining tool that is used for the largest part of academic research (Van Der Aalst, 2016). The ProM framework has 1500+ plug-ins available created by various research groups, among which the Fuzzy Miner algorithm that was created as a plug-in for the ProM framework (C. W. Günther & van der Aalst, 2007). The open-source tool was developed so that others could develop and add their own plug-ins into the framework, which is a reason why it has become so popular among researchers. Because of the amount of plug-ins ProM has, it can be applied in most cases.

Disco (C. Günther & Rozinat, 2012) is a commercial Process Mining tool that, according to their creators, makes Process Mining easy and fast. The discovery algorithm used by Disco is based on ProM's Fuzzy Miner plugin, but developed further and improved (Van Der Aalst, 2016). Günther and Rozinat advocate Disco for multiple reasons, like:

- *Automated process discovery is Disco's main functionality. After loading in a data set, the user is directly shown a map of the process, giving an intuitively understandable and 100% truthful map.*
- *Process statistics are generated automatically and can be inspected in another tab. This tab provides information about for example frequency, performance, and additional data attributes from the dataset.*
- *Individual cases and variants of paths can also be inspected in a separate tab, allowing for easy in-depth analyses.*
- *Filtering on various types like variation, endpoints, attributes, or followers can deliver valuable insights in a quick and interactive way.*

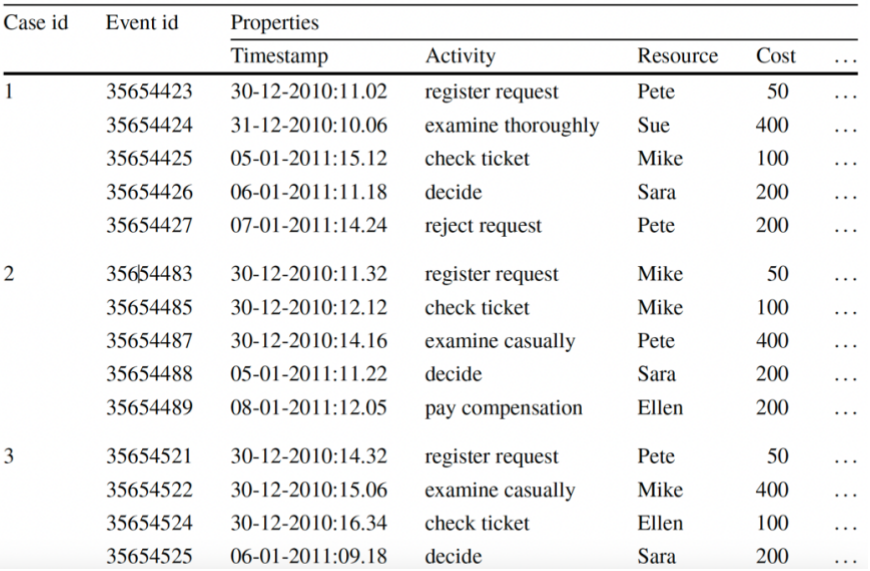
Comparing the two Process Mining tools, one can conclude that the ProM framework is useful when various analysis techniques are employed, and Disco is tailored to quickly discover processes and dive into the (individual) processes using various functions of the tool. Although the Fuzzy Miner algorithm is also compatible with ProM, Disco uses a modified Fuzzy Miner algorithm in combination with practical experiences and user testing (C. Günther & Rozinat, 2012). It is argued that this also makes Disco easy to learn and use, which makes it more accessible to start with Process Mining. Another advantage of Disco is the various filters it has, which basically are conformance checking techniques built in the application. Although ProM also has an conformance checking plugin (Van Der Aalst, 2016), Disco makes the discovery and analysis of the data easy and feasible in the same standard application. Finally, going back to the questions raised by Van Der Aalst (2016) to choose a suitable Process Mining tool, the answers indicate that Disco fits best. In this case, there is a single question that is determined ad-hoc and a flexible analysis should be possible. Van Der Aalst (2016) names Disco as a tool suitable for such style of analysis. Therefore, for the case applications in this study, Disco will be used.

2.1.4 Event Logs

The data that is used for Process Mining are called event logs. Van Der Aalst (2016 p129-130) lists the following assumptions about event logs:

- *“A process consists of cases.*
- *A case consists of events such that each event relates to precisely one case.*
- *Events within a case are ordered.*
- *Events can have attributes. Typical attributes are activity, time, costs, and resource.”*

Table 2 provides an example of what an event log can look like. Columns *‘Case id’* and *‘Activity’* are the minimum requirements to apply the data to some form of Process Mining, but other attributes help in analyzing in different ways. The *‘Timestamp’* variable allows to analyze length of activities or sequence, although in this case *‘Event id’* also indicate sequence within a case. The *‘Resource’* variable indicates what person executed the activity, and the *‘Cost’* indicate the financial costs of the activity. The current standard for the way that event data is stored and loaded into applications is eXtensible Event Stream (XES), an XML-based standard for event logs (Gunther & Verbeek, 2014). The XES format has been standardized by the IEEE (IEEE, 2016). Although the data used in this study is not stored in the XES format, it does not form complications. Most tools, like Disco, also are able to import Comma Separated Values (CSV). After preparation of the data, this is the format used as input for Disco in this study. When imported, one can manually label the case ID and attributes without much effort.



Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	decide	Sara	200	...

Table 2: part of an example event log (Van Der Aalst, 2016 p.129)

2.1.5 Related Process Mining Works

As a result of the many studies and developments in scientific literature around Process Mining, researchers also started to exploit various business applications. Most studied applications are healthcare, ICT, manufacturing, education, finance, and logistics (Garcia et al., 2019). Although the combination of Process Mining, Cyber Forensics and Web Applications has rarely been exploited, other studies using Process Mining can give useful insights. In the paragraphs below, some practical applications and their takeaways will briefly be discussed. Again, it is not feasible to lay out all relevant applications as Process Mining can be applied to almost any process (Garcia et al., 2019). Therefore, a few applications and relevant takeaways for this research will be discussed in the following paragraphs.

Audit was one of the first fields to adapt Process Mining, and currently most large audit companies use Process Mining to increase the reliability of audits (Reinkemeyer, 2020). The investigation of audit trails to validate patterns can be very useful to detect security violations (van der Aalst & de Medeiros, 2005).

Accorsi & Stocker (2012) already pointed out that, previously, security audits failed to detect most violations. Reasons are that audits were based on samples, which do not show the full process, and that there were no tools to adequately analyze workflows. By applying Process Mining techniques like conformance checking, auditors can solve the shortcomings of traditional audits. Namely, Process Mining tools enable analyses to detect fraudulent behavior from full datasets instead of from samples (Accorsi et al., 2013; Jans et al., 2013). Though, some issues were identified that threaten the validity of the approach. Relevant issues were that existing tools required extensive manual work to apply techniques, and that results were hard to interpret for non-experts (Accorsi & Stocker, 2012). In further research, Accorsi et al. (2013) stated that process discovery techniques also provide a solid basis for security audits. Still, this study identified the limitation that, at the time, there were no tools that could analyze the discovered processes. Although, Process Mining tool Disco (C. Günther & Rozinat, 2012) provides proper analysis possibilities through powerful filters, the interpretation of results still requires some domain knowledge. However, this limitation applies to practically every field where Process Mining can be applied. In regard of this study, domain knowledge of the Web

Application and the logged data is required to understand what activities have taken place. Without this knowledge, one is not able to apply the right conformance checking filters to extract and present the malicious behavior properly.

Recent research also points out the possibility to analyze all recorded transactions, which allows auditors to quantify the impact of identified deficiencies (Reinkemeyer, 2020; Werner et al., 2021). However, research also identified further limitations regarding the application of Process Mining in auditing. Werner et al. argue that to create an event log, specific knowledge of the system and its implementation is necessary. Further, only the transactions and procedures registered by the systems can be used as input for analyses (Bahaweres et al., 2021; Jans et al., 2013; Werner et al., 2021). Fortunately, these limitations are not applicable to this study. Unlike in financial audits, relevant activity aside of Web Servers does not occur. Further, most Web Servers, like Apache or Windows IIS, automatically record access logs with relevant fields and administrators can query access logs from web servers relatively easy.

Although Process Mining in analyzing **online user behavior** is not one of the most popular domains in literature (Garcia et al., 2019), it is relevant for this study. Where traditional web analytics tools are not able to display customer behavior on websites, analyzing users' web clicks by applying Process Mining algorithms enables the discovery of the actual paths of users throughout their website visit (Poggi et al., 2013). To do so, logs from web servers are used as datasets, the same source of data that is used in this study. Poggi et al. (2013) found that in order to analyze Web Logs in their case, URLs need to be classified into higher level activities. After doing so, the researchers proposed a few techniques to properly discover customer behavior on an Online Travel Agency's website. Poggi et al. argued that Process Mining Algorithms are designed to show dominant behavior and let the noise out. However, the researchers were interested in buying behavior, which occurred rarely in the dataset. A few mining algorithms were applied to make this behavior more prominent in the process model. Even though this research also fixates on rarely occurring behavior, it does so by investigating the behavior explicitly and not by including it in a high-level overview of processes. Further research on Process Mining in customer journey analysis was carried out by Terragni & Hassani (2018). By analyzing discovered customer journey paths, personalized recommendations were implemented on an advertising web portal, increasing the click-through rate of recommendations significantly. Other studies gain valuable insights by using Process Mining to discover user behavior on newspaper websites (Sarirah Husin & Ismail, 2021) or analyze student behavior in online courses (Van den beemt et al., 2018).

In short, Process Mining has been proven useful in various applications, making it a promising technique for novel approaches. The possibility to discover processes from various types of data, including web logs, and investigate malicious behavior makes it a suitable method of analysis for this study. Additionally, Process Mining tool Disco offers a useful combination of process discovery and conformance checking techniques while also being an intuitive and easy-to-use application.

2.2 Cyber Security & Forensics

As a result of the omnipresence of Information and Communication Technology (ICT) in society, the wellbeing of both organizations and people has become ever so important. The term Cyber Security, subsequently, is being used everywhere. A well-known definition of the term by von Solms & van Niekerk (2013, p101) is phrased as *"the protection of cyberspace itself, the electronic information, the ICTs that support cyberspace, and the users of cyberspace in their personal, societal and national capacity, including any of their interests, either tangible or intangible, that are vulnerable to attacks originating in cyberspace"*. The authors further highlight the difference with Information Security, which only concerns the security of information. Cyber Security, on the other hand, includes protecting the people that operate in the digital world and their information, which is broader and indicates a higher significance of the term Cyber Security.

The following subsections aim to provide an introduction to the fields of Cyber Security & Cyber Forensics in relation to this study. First, the threats leading up to the rising importance of the research fields are discussed. Secondly, related studies and developed tools will be discussed to highlight the need for further research in the field.

2.2.1 A brief History of Cyber Threats

It is argued that the issue of Cyber Security already existed in the *1960s* and that it has developed over the past decades (Warner, 2012). Warner argues that, in the United States, various insights and evidence over time finally lead to more policymaking by the government. In the *1960s*, when most people had not even seen a computer, experts began to acknowledge the need for multiprogramming: a solution that allows multiple users to use computers, without being able to see one another's data. At that point already, the risk of using sensitive data was acknowledged. Ware (1967) pointed out that, accidental or deliberately, human actions can threaten the security of data in systems. The author also mentions the hard- and software vulnerabilities that jeopardize sensitive information. As these risks are not easily mitigated, security had to be improved. Innovations like administrator privileges, file permissions, hashed passwords and data encryption were introduced in the *1970s* (Warner, 2012). As computer networks became globally accessible in the *1980s,* external threats in the form of viruses and hackers came to the attention. The fact that data could be stolen, or manipulated to affect the way systems operate, made data integrity a new part of the Cyber Security matter. Around the 1990s, for example, the Michelangelo virus that caused data loss and a Denial of Service (DoS) attack on internet service provider Panix, gained public attention (Warner, 2012). Ever since, the number of cyber attacks on all sorts of organizations has increased. A report by Sophos (Shier, 2020) looking back on the first two decades of the 21st century, divides main threats in Cyber Security in roughly three periods:

- *the Worm Era* (2000-2004). A computer worm is a *"program that self-propagates across a network exploiting security or policy flaws in widely-used services"* (Weaver et al., 2003). In this period, malware got mainstream media attention and had major impact. For instance, Microsoft introduced '*Patch Tuesday'* to defend its applications against the continuously altering worms.
- t*he Monetization Era* (2005-2012). As the name says, this era monetized cybercrime by making a business out of it and becoming organized. Spam and phishing forced the improvement of e-mail filtering, while malvertising and exploit kits forced the

improvement of web content filtering. Moreover, measures against prepaid cash services were taken as criminals demanded payment in these forms.

- *the Ransomware Era* (2013-present). Although ransomware existed before and is not the only threat in this age, it certainly has the largest destructive impact. Nowadays, many cyber attacks end with releasing ransomware, after which a payment in cryptocurrency is demanded. In 2022, 68% of the globally reported cyber attacks were ransomware (Statista, 2023).

As defined by the US National Institute of Standards and Technology (NIST), ransomware is *"a type of malicious attack where attackers encrypt an organization's data and demand payment to restore access. Attackers may also steal an organization's information and demand an additional payment in return for not disclosing the information to authorities, competitors, or the public"* (Barker et al., 2022 p.II). Recently, NOREA, the Dutch Association of Register EDP-auditors, published an extensive report including a framework on how organizations can increase resilience against ransomware attacks (Gangaram Panday & Zwakenberg, 2023). The authors argue that, due to the quick evolution of ransomware attacks techniques recently, organizations not only need good back-ups, but also to improve security controls. Another example of the developments in ransomware attacks is the creation of Ransomware as a Service (RaaS). RaaS is a business model comparable to Software as a Service (SaaS), where people pay cyber criminals to launch ransomware attacks for them (Baker, 2023). This development led to ransomware becoming even more accessible, even to criminals without technical knowledge.

To gain better insight in the continuously changing ransomware environment, Gangaram Panday & Zwakenberg (2023) present an up-to-date visualization of the most common attack vectors (*Figure 6* on the next page) to map their control framework on. The topic of this master thesis relates to the first the stage of the *NOREA Ransomware Kill chain*, where malicious users gain unauthorized access to systems. More specifically, it focuses on criminals exploiting vulnerabilities as this activity can be logged, in this case in web server access logs. The abuse of weak credentials and phishing employees, unfortunately, are hard to identify from logs, and therefore are out of scope of this study. Further, this thesis does not aim to prevent attacks, like the first step in the model, but rather to understand malicious behavior in hindsight. Particularly, it aims to understand how criminals gained access to systems by exploiting vulnerabilities in Web Applications. The ultimate goal is to understand the paths malicious users took, giving relevant insights in the vulnerabilities of organizations' Web Applications. The NOREA framework, however, is of course not the only measure that has been taken to improve cyber security. The next sections discuss relevant academic research that resulted in the current measures and tools that exist to improve Cyber Security and assist in Cyber Forensics investigations.

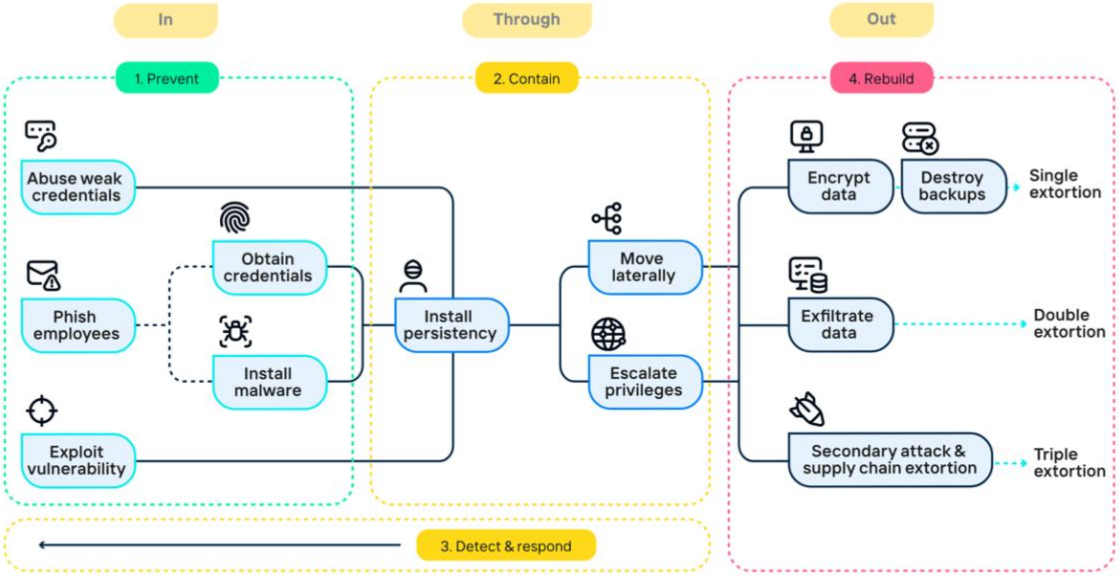


Figure 6: the NOREA Ransomware Kill chain (Gangaram Panday & Zwakenberg, 2023 p23)

2.2.2 Related Cyber Security Works

To counter the threats that cyber criminals pose, various actions have been taken to help organizations improve their cyber security. Although cyber security measures on the people, process and the technical sides exist, the first two will not be discussed. As the empirical research focuses on the technological aspect of Cyber Security, that is, vulnerabilities in code, efforts aiming to protect systems are reviewed. Therefore, this section reviews scientific literature on the improvement of cyber security.

The rise of big data around the end of the 2000s, which led to the transfer of countless streams of information between networks, posed opportunity for both cyber criminals and people who aim to improve cyber security. Connected networks between organizations opened up to new attack methods for attackers. However, improved hard- and software also enabled big data analytics, which resulted in techniques to continuously monitor activity and detect malicious behavior (Mahmood & Afzal, 2013). At the time, research led to various tools like *Intrusion Detection Systems* (IDS). However, the approaches came across issues and challenges (Liao et al., 2013). Although IDS tools have been one of the most discussed techniques in scientific literature over the last decade (Humayun et al., 2020), developing good IDS still remains as hard. As attackers have started using evasion techniques to avoid detection, accurately recognizing intrusions poses a major challenge for academia (Khraisat et al., 2019).

In recent years, Process Mining is also being applied to improve Cyber Security. Where traditional data mining solutions do not focus on business process models, a Process Mining approach takes end-to-end processes in account allowing to check conformance and detect deviations (Mishra et al., 2019). For example, Process Mining has been applied to detect cyber attacks in *Industrial Control Systems* (ICS). Myers et al. (2017) found that process discovery techniques can accurately generate process models of an ICS. In (Myers et al., 2018), the authors used a conformance checking approach to detect deviations from process models generated by process discovery techniques. Their method was able to successfully identify cyber attacks that a widely used open-source IDS/IPS tool was not able to detect. Other research by de Alvarenga et al. (2018) used Process Mining techniques in addition to an IDS tool, aiming to help network administrators to act on IDS alerts. By creating high-level visualizations of attacks strategies based on IDS alerts, the authors presented intuitive models that should help network administrators prioritize vulnerabilities that were detected. However, de Alvarenga et al. (2018) acknowledge the limitations of using only the alerts generated by IDS tools as input data. Attacks that were not detected, logically do not show up in the alerts.

The briefly discussed studies above, among other relevant research on Process Mining for Cyber Security, are included in recent systematic literature reviews by Macak et al. (2022) and Silalahi et al. (2022). Both studies found that the most popular Process Mining application for Cyber Security, similarly to Cyber Security research in general, focused on the detection of various attacks and fraud. Further, both studies identified that various studies employed a combination of process discovery and conformance checking techniques to perform the analyses. Finally, both Macak et al. and Silalahi et al. stated that Disco's fuzzy miner algorithm was the most used approach to discover processes, the same tool that was found most suitable in this case as stated in *section 2.1* of this thesis. Macak et al. (2022) conclude that, as the body of literature on Process Mining for Cyber Security is relatively thin, there probably are many unexploited applications that could be investigated.

This section reviewed the aspects of detecting and preventing malicious behavior mainly, while also pointing out the successful application of Process Mining in Cyber Security. Considering the ever-existing challenges of preventing attacks, the investigation attacks postmortem remains as important. The following section presents an introduction to the domain related to the investigation of attacks and malicious behavior and relevant research in the field.

2.2.3 Cyber Forensics Domains

The branch related to Cyber Security that focuses on the investigation of cyber attacks is Cyber Forensics, also referred to as IT-, Digital- or Computer Forensics. This research field started developing significantly, again, in the 2000s. As people worldwide, including criminals, gained access to computers, the Cyber Forensics field became more specialized. Government agencies and professional organizations started formalizing the branch, resulting in the advanced development of various forensics tools by commercial entities and open-source communities. Moreover, the amount of evidence to work with expanded (Pollitt, 2010). Within Cyber Forensics, there are various subdomains that can be identified. For each domain, different methodologies and tools are used and developed to carry out the specific analyses (Casino et al., 2022; Fernando, 2021). A few of the most prevalent domains in academic literature are cloud-, mobile-, database- and network forensics. To gain a better understanding of the Cyber Forensics domain and the positioning of this study in the domain, the named subdomains are briefly introduced below. Subsequently, the domain in Cyber Forensics that this thesis focuses on is described. It should be noted that not every study classifies the domains the same and some domains have overlap with each other. This classification has been chosen based on various literature studies.

Cloud Forensics is a cross discipline resulting from cloud computing and digital forensics (Ruan et al., 2011). The main characteristics of cloud computing are on demand self-service, ubiquitous network access, multi-tenancy, location independence, rapid elasticity and pay-per-use business models. *Cloud Service Providers* (CSP) typically offer three main service models namely Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) (Almulla et al., 2013). Cloud forensics deals with problems related to the cloud infrastructure and their services. Although there is ample research in the domain about for instance the available data for stakeholders, tools for collecting data and the relation between CSP and clients, there are still many challenges (Manral et al., 2019). A recent review of reviews pointed out that most research focuses on identifying challenges rather than proposing solutions for the problems (Casino et al., 2022). The authors found that the most addressed issues related to the need for tools, fragmentation of data and the lack of mechanisms for forensic readiness.

Mobile Forensics is defined by the NIST as *"the science of recovering digital evidence from a mobile device under forensically sound conditions using accepted methods"* (Ayers et al., 2014 pIII). The increasing heterogeneity and amount of features of mobile technologies makes mobile forensics inherently challenging, while also offering endless streams of valuable data that can be extracted (Chernyshev et al., 2017). Evidence can be extracted from personal produced data, calls and messaging, audio and image, GPS or application data (Alatawi et al., 2020; Chernyshev et al., 2017). Although there are tools available for mobile forensic investigations, various drawbacks still exist. Fernando (2021) identified that most tools have issues with compatibility for properly extracting evidence from various devices, extracting data from cloud services and retrieving deleted or sensitive, protected data. Casino et al. (2022) draw comparable conclusions, while also pointing out that there is no consensus about whether procedures should be tailored for each device, or that standardized guidelines should be created.

Database Forensics focuses on *"detailed analysis of a database including its contents, log files, metadata and data files depending on the type of database used"* (Chopade & Pachghare, 2019 p.180). One of the main advantages of databases is the metadata it includes, providing additional information that cannot be derived from the raw data only (Olivier, 2009). There are various tools available like Microsoft's SQLCMD that are capable of extracting and analyzing from various databases like MYSQL (Fernando, 2021). Chopade & Pachghare (2019) reviewed research on database forensics and found that most research focuses on relational (SQL) databases, leaving a gap for research on the rising interest in unstructured (NoSQL) databases. The authors further identified challenges such as recovering lost data and avoiding data tampering.

Network Forensics primarily focuses on the analysis of network packets, which are groups of bits including data accompanied by control information and are related to the network layer of the OSI model as seen in *Figure 7* (Alani, 2014; Sikos, 2020). Next to the ample tools available for analyzing packets, machine learning-based approaches are proposed to solve issues regarding big network data (Sikos, 2020). The issue of large networks is addressed by Fernando (2021), who also points out the issue of identifying the source of attacks and integrity of log files after attacks in *Software Defined Networks*. Analysis of encrypted traffic is another challenge pointed out in various studies (Casino et al., 2022), which is one of the issues that machine learning studies aim to solve (Sikos, 2020).

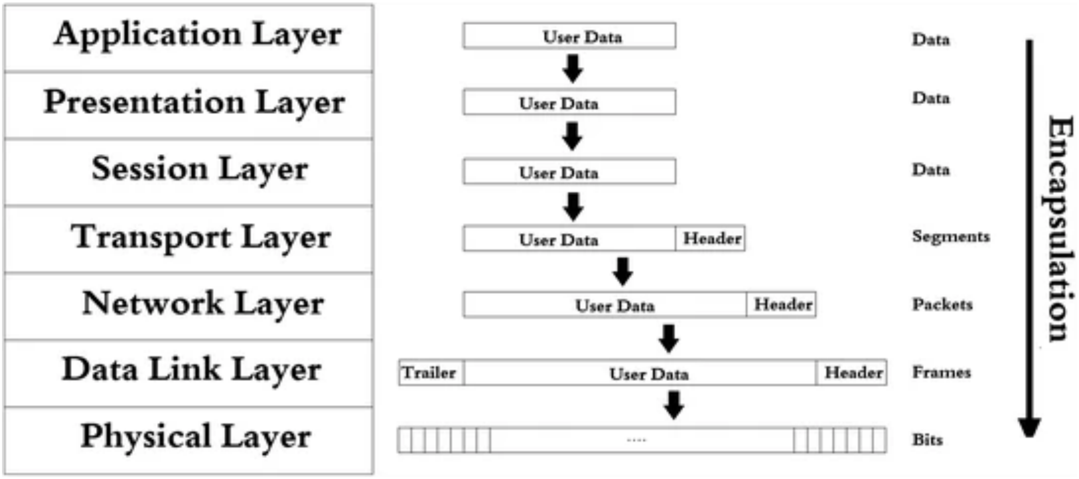


Figure 7: Layers of the OSI model including data flow (source: Alani (2014, p.15))

Web Application Forensics is the subdomain of Cyber Forensics that this thesis is scoped towards. More specifically, it focuses on the analysis of web server access logs. Although sometimes categorized as a subdomain of network forensics next to network packet analysis (Sindhu & Meshram, 2012), the analysis of web server logs is inherently different from analyzing network packets. Where network forensic investigations look at the network layer, Web Application Forensics investigate malicious activity on the application layer of the OSI model

(Deltchev, 2012). *Section 2.3* will discuss the domain, its challenges and current applications in further detail.

Despite the body of literature on Process Mining for detecting and preventing malicious behavior, research on Process Mining in forensic investigation is scarce. To the best of my knowledge, Lagraa & State (2020) are amongst the few to take a Process Mining approach in Cyber Forensics. The authors aimed to understand behavior of users that led to malicious authentication events by looking at the change in event attributes of authentication events. Lagraa & State also claim that the field of investigating and understanding attacks is underdeveloped.

2.3 Web Applications

The web is *"a highly programmable environment that allows mass customization through the immediate deployment of a large and diverse range of applications to millions of global users"*. Two main components of websites are browsers and applications. Web browsers, which are software applications, enable users to interact with Web Applications over the internet. Web Applications are computer programs that allow users to submit and retrieve data to/from databases. A Web application can send queries to a web server, which dynamically generates the web documents in a standard format that is supported by all web browsers. Therefore, Web Applications can be used on any preferred browser and operating system of the user. *Figure 8* on the next page visualizes the way Web Applications work. A user sends a request via their web browser to a web server. A dynamic content generation tool deals with the request and queries the database for content or sensitive customer data. Finally, the requested data is retrieved and presented through the browser to the user (Acunetix, n.d.).

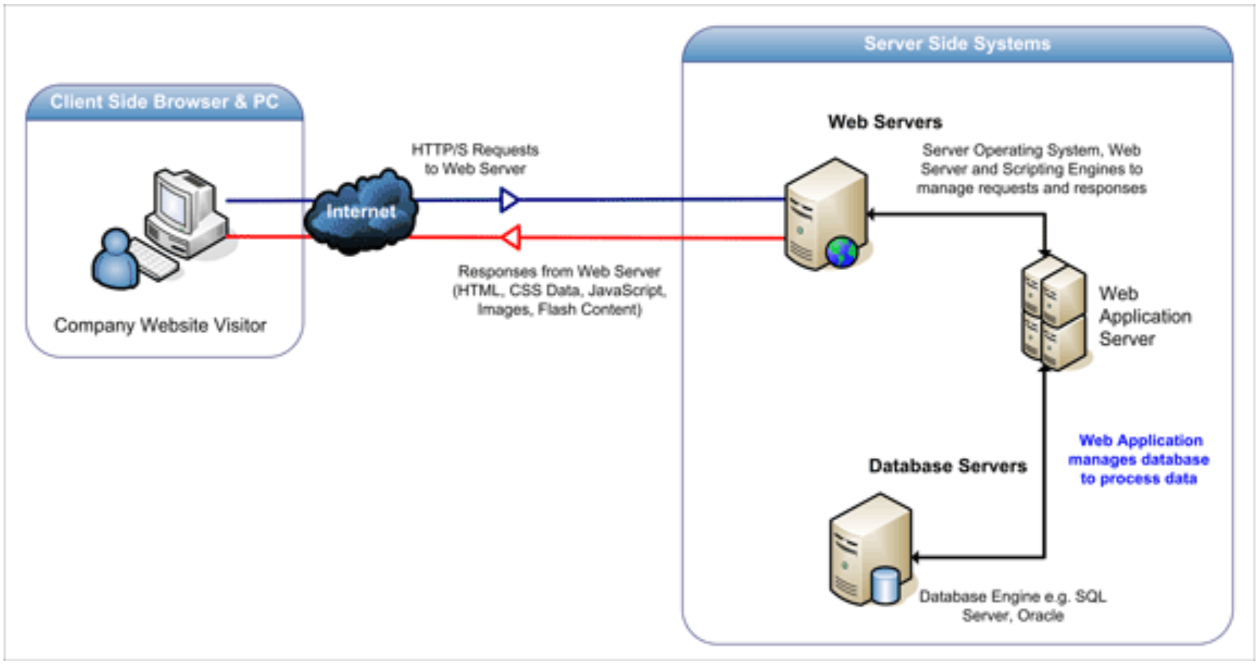

Client Side Browser & PC
Company Website Visitor
Internet
HTTP/S Requests to Web Server
Responses from Web Server (HTML, CSS Data, JavaScript, Images, Flash Content)
Server Side Systems
Web Servers
Server Operating System, Web Server and Scripting Engines to manage requests and responses
Web Application Server
Web Application manages database to process data
Database Servers
Database Engine e.g. SQL Server, Oracle

Figure 8: Web Application architecture (source: Acunetix (n.d.))

2.3.1 Web Application Vulnerabilities

Unfortunately, the public availability of Web Applications also brings risks. Anyone can attempt to connect to the underlying database, which often stores sensitive (customer) data. Moreover, most Web Applications are custom-built, which often involves less testing for vulnerabilities than off-the-shelf software (Acunetix, n.d.). Therefore, insecure Web Applications are a favorable target for people with malicious intent. The *Open Worldwide Application Security Project* (OWASP) is the leading organization that aims to improve Web Application security. Periodically, the *'OWASP Ten'* is published, which is a document to spread awareness about the most threatening Web Application security risks according to a broad consensus of experts in the field (OWASP, 2021). The most recent *Top Ten* is presented and explained in *Table 3.* Although not all categories directly relate to types of attacks, the list does provide a guide to the current challenges that the Web Application security field faces. Moreover, it helps understand the various attack paths that malicious users can take to gain unauthorized access to Web Applications.

Number	Category	Description
A01:	***Broken Access Control***	Insecure access control policy that allows users to act outside of their intended permissions. *Example: Bypassing access control checks by modifying the URL, using force browsing to access authenticated pages as an unauthenticated user.*
A02:	***Cryptographic Failures***	Improper- or non-existing encryption of sensitive data during transit and/or rest. *Example: Transmitting sensitive data like credit card numbers in clear text through protocols such as HTTP.*
A03:	***Injection***	Allowing users to inject data into the application or queries without testing, leading to manipulation of the applications behavior. *Example: Modifying parameters in a URL to request/modify/delete data from the application.*
A04:	***Insecure Design***	Ineffective or missing control design, which is different from insecure implementation. Secure design can be implemented insecurely, but insecure design cannot be fixed by securely implementing the design. Often, a lack of business risk profiling is the cause of insecure design.
A05:	***Security Misconfiguration***	Improperly configured applications leading to access to unintended information or unnecessary features. *Example: Detailed error messages providing overly informative details to users.*
A06:	***Vulnerable and Outdated Components***	Unpatched, outdated or unsupported software on any component of the application. *Example: Failing to update operating software, leaving known vulnerabilities unpatched.*
A07:	***Identification and Authentication Failures***	Insecure identity, authentication and session management of users allowing malicious users to profit from user accounts. *Examples: Allowing for brute-force attacks for guessing credentials or improper session validation settings.*
A08:	***Software and Data Integrity Failures***	Code and infrastructure that do not protect against integrity violations, allowing attackers to intervene. *Example: Use of functionalities from untrusted third parties that may include vulnerabilities.*
A09:	***Security Logging and Monitoring Failures***	Failing to properly log and monitor activity, which has impact on accountability, visibility, incident alerting and forensics. *Examples*: *Not logging application activity or not monitoring suspicious activity on applications*
A10:	***Server-Side Request Forgery***	Allowing to fetch a remote resource without validating the user-supplied URL, making the application send crafted requests to unexpected destinations. *Example*: *An unsegmented network allowing malicious users to map out the internal network to determine open ports.*

Table 3: The OWASP Top Ten explained (source: information from OWASP (2021))

2.3.2 Web Application Forensics Research

The body of literature on forensics investigations of misuse on Web Applications specifically, is slim. Lazzez & Slimani (2015) defined and placed Web Application Forensics in the field of Cyber Forensics, before comparing some of the main considered tools that can aid in forensic investigations on Web Applications. Sindhu & Meshram (2012) proposed an architecture to investigate evidence from file systems and network logs. More specifically, the authors propose datamining techniques to find relations and patterns in files, network packets and access logs. Babiker et al. (2018) also point out data mining techniques as a means to improve forensic investigations, though mentioning that there have been few applications studied in research. However, approaches like data mining do not look at malicious activity from a process point of view. The comparison of forensic tools for Web Application activity by Lazzez & Slimani (2015) also did not mention the capability to analyze the data from a process point of view. The currently available tools were assessed on for example ability to perform in real-time, generate analysis reports and scalability. The analysis of web server logs using AI techniques like Deep Learning is assessed by Nazar et al. (2021). However, AI approaches currently also encounter issues with including time series, i.e. the process, in the models. Although AI offers significant advantages over traditional approaches, technical advancements are required to overcome the current issues (Nazar et al., 2021).

A study that did take the process component of malicious behavior into account was carried out by Gunestas & Bilgin (2016). The authors used a special case of temporal logic in combination with a *Complex Event Processing* (CEP) tool to extract attacks from the logs. Temporal logic is typically used to discover time-based complex patterns. After formalizing various patterns of malicious behavior using temporal logic, the authors ran the queries in the CEP tool to generate the logs that represent the process of the attacks. This research relates to Process Mining approaches in the sense that both take the process of malicious behavior into account.

However, Gunestas & Bilgin's approach requires extensive knowledge of logical constructs to define the behavior, in contrast most Process Mining tools that are relatively easy to use (Van Der Aalst, 2016). Disco, for example, provides easy-to-configure filters that also deliver fast results, making interactive exploratory analyses accessible without requiring significant domain knowledge of Process Mining (C. Günther & Rozinat, 2012). Still, for both approaches it is needed to have domain knowledge of the data that is analyzed. It is important to at least know what kind of behavior you are looking for (Gunestas & Bilgin, 2016), which in the case of Web Application activity also needs to be defined first. However, forensic investigation does not focus on the detection of malicious activity, but rather on the post-mortem investigation. Therefore, at least some information about the attack is known at the time of the investigation.

2.4 Conclusion on Reviewed Literature

In the past sections, an introduction to the domains was presented, and relevant literature was reviewed on the topic of Process Mining in Cyber Forensics. Moreover, it discussed the particular domain of Cyber Forensics that this study focuses on: forensic investigation of malicious activity on Web Applications.

First of all, the concept of Process Mining was introduced. Some relevant algorithms and tools were discussed, leading to the conclusion that Process Mining tool Disco is assumed to be the most fitting for the analysis in this study. Further, some Process Mining applications in other fields and their takeaways have been discussed. The second section started off with a brief introduction to cyber threats over time and the most prevalent research topics in the field of Cyber Security. It also highlighted the studies using Process Mining approaches for the identification or prevention of cyber attacks, which opens the door for applying Process Mining in Cyber Forensics investigation. Despite the considerable body of literature on Cyber Forensics, there is a gap in academic literature on the possibilities of applying Process Mining in forensic investigations. The reviewed literature is categorized in *Table 4* on the next page.

Therefore, this study aims to exploit the possibilities of combining the fields of study. The specific case focused on in this master's thesis, is the forensic investigation of malicious activity on Web Application by performing a Process Mining analysis. The goal hereby is to define how to approach such cases and assess to what extent the approach can be useful. *Chapter 3* goes into the details of the methodology used in this study and how the results are validated and evaluated.

Domain	*(sub)domain*	*Research paper (* = literature review)*
Process Mining	applications	*Garcia et al. (2019)**
	Auditing	*Van der Aalst & de Medeiros (2005), Accorsi & Stocker (2012;2013), Jans et al. (2013), Reinkemeyer (2020), Werner et al. (2021), Bahaweres et al. (2021)*
	Online user behavior	*Poggi et al. (2013), Terragni & Hassani (2018), van den Beemdt et al. (2018), Sarirah Husin & Ismail (2021)*
Cyber Security	applications	*Mahmood & Afzal (2013)*, Liao et al. (2013)*, Khraisat et al. (2019)**
	Process Mining	*Myers et al. (2017;2018), de Alvarenga et al. (2018), Macak et al. (2022)*, Silalahi et al. (2022)**
Cyber Forensics	applications	*Fernando et al. (2021)*, Casino et al. (2022)*,*
	Process Mining	*Lagraa & State (2020)*
	(Cloud)	*Ruan et al. (2011)*, Almulla et al. (2013)*, Manral et al. (2019)**
	(Mobile)	*Ayers et al. (2014), Chernyshev (2017)*, Alatawi et al. (2020)**
	(Database)	*Olivier (2009)* Chopade & Pachghare (2019)**
	(Network)	*Alani (2014), Sikos (2020)**
	(Web Application)	*Deltchev (2012)*, Sindhu & Meshram (2012), Lazzez & Slimani (2015)*, Babiker et al. (2018)*, Nazar et al. (2021), Gunestas & Bilgin (2016)*

Table 4: Reviewed Literature (source: own work)

3. Methodology

The aim of this research is to assess how, and to what extent Process Mining can be applied in forensic investigations of malicious Web Application activity. To do so, the Process Mining Project Methodology (PM^2) by van Eck et al. (2015) is followed, though tailored for the application to Web Application Forensics. First, domain knowledge is applied using a lab setting Web Application to develop the Methodology for Process Mining in Web Application Forensics. Next, a real-life case application is performed for further refining and evaluation of the methodology. To further evaluate the approach, interviews with an expert panel are held. The approach can therefore be regarded as Design Science Research, which is defined as *"designing an artifact that improves something for stakeholders and empirically investigating the performance of an artifact in a context"* (Wieringa, 2014 p.V). Hevner et al. (2004) presented a framework for Design Science in information systems research, which has been applied to this study in *Figure 9*. In the business environment, there currently is no guide to investigate malicious activity on Web Applications, which are vulnerable targets. To determine how PM^2 is to be revised and evaluated for Web Application Forensics, the *Framework for Evaluation in*

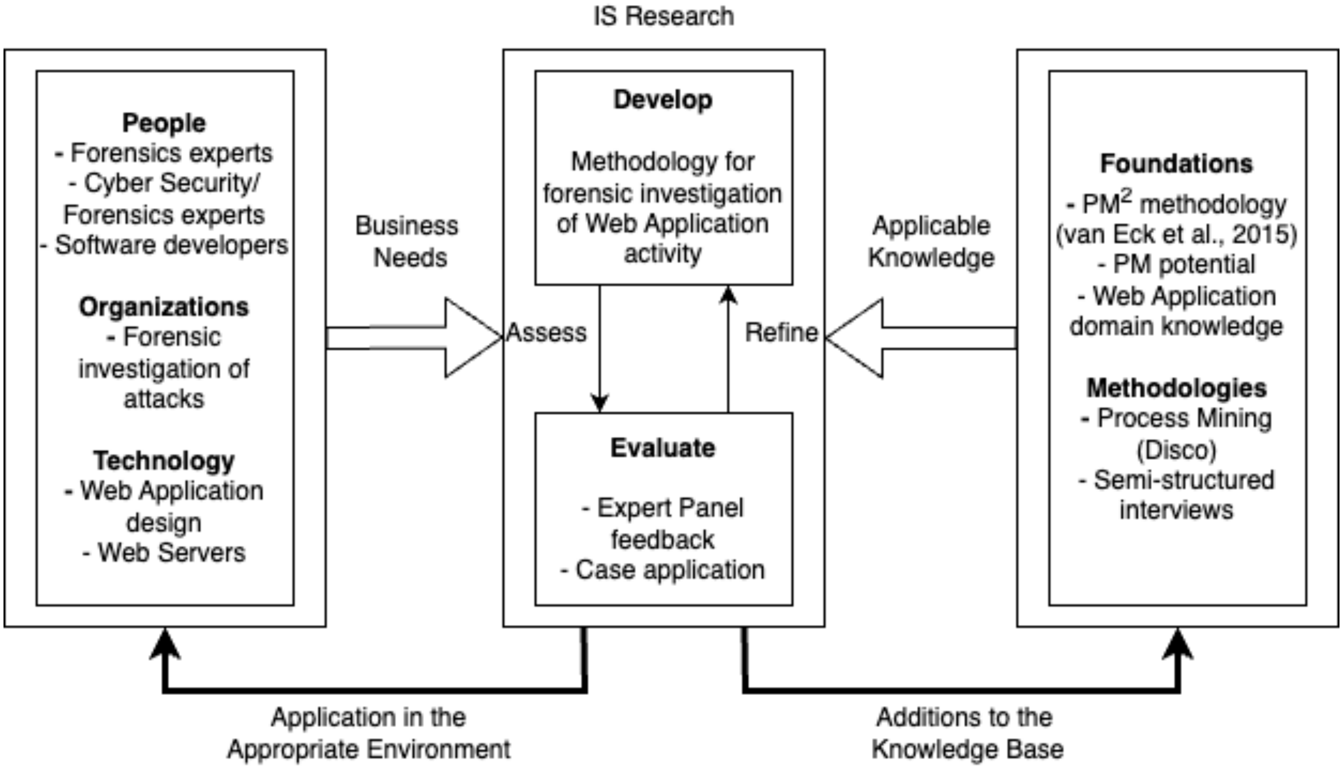

IS Research
People
- Forensics experts
- Cyber Security/ Forensics experts
- Software developers
Organizations
- Forensic investigation of attacks
Technology
- Web Application design
- Web Servers
Business Needs
Develop
Methodology for forensic investigation of Web Application activity
Assess
Refine
Evaluate
- Expert Panel feedback
- Case application
Applicable Knowledge
Foundations
- PM² methodology (van Eck et al., 2015)
- PM potential
- Web Application domain knowledge
Methodologies
- Process Mining (Disco)
- Semi-structured interviews
Application in the Appropriate Environment
Additions to the Knowledge Base

Figure 9: Design Science in Information Systems Research (source: adapted from Hevner et al. (2004))

Design Science Research (FEDS) by Venable et al. (2016) was used. The framework can be consulted to decide when, how and what to evaluate in design science research using two dimensions. First, the framework makes a distinction between artificial and naturalistic evaluation, i.e. in a laboratory setting and real environment. The authors further distinguish formative and summative evaluations: iterative evaluations to improve the artifact during development and (mostly) an evaluation after completing the development respectively. For this thesis, the '*Human Risk & Effectiveness*' strategy is taken as described by *Figure 10*. The approach is chosen as the Process Mining analysis is executed manually, implying that the human factor is paramount. The strategy starts with an artificial formative approach, which refers to the application of PM^2 to the lab setting in this study. Next, Venable et al. propose naturalistic formative evaluations, which refers to the case application. Finally, the authors propose a naturalistic summative evaluation to assess the effectiveness of the artifact. This evaluation refers to the interviews with the expert panel.

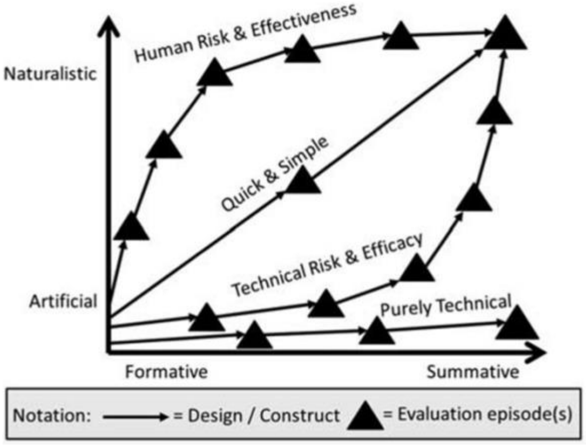

Naturalistic
Artificial
Human Risk & Effectiveness
Quick & Simple
Technical Risk & Efficacy
Purely Technical
Formative
Summative
Notation: → = Design / Construct ▲ = Evaluation episode(s)

Figure 10: *FEDS Evaluation Strategies (source: Venable et al. 2016 p.80)*

This chapter is further structured as follows. First of all, the Web Applications and the data used for designing the artifact and later the case application are discussed in *Section 3.1*. Next, the PM^2 methodology is applied to the lab setting Web Application in *Section 3.2*. After that, the steps carried out for the case application are presented in *Section 3.3*. In *Section 3.4*, *PM^2 for Web Application Forensics* is presented. Finally, the work done related to the interviews is laid out in *Section 3.5*.

3.1 Web Application Activity Data: Access Logs

Activity on Web Applications is automatically logged in web server access logs. Users with privileged access to web servers can retrieve such logs and transform them into a dataset. Access logs are useful input for Process Mining analyses as they always include timestamps. Timestamps, in fact, are required to enable the algorithms to identify processes among the data. The logs further contain information that can give useful insights about the activity of different users that accessed the Web Application and the web server’s responses. The following paragraphs briefly discuss the Web Applications used in this study and the logged data that is available.

3.1.1 Used Web Applications

The first Web Application used was created in a lab setting, after which some activity was generated to demonstrate behavior that is desired on the live Web Application. This lab setting was designed and created in collaboration with the internship supervisor at Joanknecht, using domain knowledge of the live Web Application to make it resemble a simplified version. The created Web Application consisted of the following pages:

- The ***public homepage***, which is the site that most users land on when accessing the Web Application. On the homepage, there is a button that will redirect users to the
- ***Login page***. This page gives users the option to fill in a username and a password. After submitting the credentials, the page redirects users to the
- ***Hidden page***. As the name tells, this is the hidden page containing sensitive information that should only be accessible to authenticated users that logged on to the Web Application.

The second Web Application concerns a real-life Web Application from a client of Joanknecht, the internship company. The business uses the Web Application as a customer portal, where their customers can place orders, look at their order history and follow their placed orders. Before gaining access to the content in the portal, customers need to login. For confidentiality reasons, more information or visuals about the Web Application cannot be provided.

3.1.2 Log Formats and Datasets

The way in which activity on Web Application is logged, depends on the type of web server the application is hosted. Although the exact data that is logged and the logging structure differ per web server, most types have a significant overlap in the logged data. Nearly all web servers logs include for example the IP address of the requester, date and time of the request, name and location of the requested file, and the status of the request (Crowdstrike, n.d.). Although not all logged fields are considered as relevant, for completeness, the log formats used in this research are listed below. This study uses access logs from the following two different web servers.

The created Web Application is hosted on an ***Apache HTTP Server***, which stores requests in the ***Combined Log Format*** with the following fields by default (Apache, n.d.):

- ***%h:*** *IP address of the client*
- ***%1:*** *Identity of the client. However, this information is very unreliable and is often not present. Then, the value will be [-]*
- ***%u:*** *User-id of the person making the request. Often not present*
- ***%t:*** *Time of request, represented in [day/month/year:hour:minute:second]*
- ***\%r\:*** *Request line from the client. It includes the method used (e.g., GET, PUT, POST), the requested resource (e.g., /homepage) and the protocol (e.g., HTTP/1.0)*
- ***%>s:*** *3-digit status code returned by the server. There are various responses: 2XX (successful), 3XX (redirection), 4XX (error by client) and 5XX (error in server).*
- ***%b****: size of object returned. If nothing returned, value will be [-]*
- ***\"%{Referer}i\":*** *Reports site that the client was referred from*
- ***\"%{User-agent}i\":*** *Information about the client's browser*

The dataset was created by generating activity on the Web Application in a controlled environment. By generating requests from multiple PC's resembling activity from different users, the event logs were created. The extracted access log, provided by Joanknecht, consisted of 25 requests on 16-10-2023 & 17-10-2023. It should be noted that session cookies were manually added to the dataset for illustrative purposes. The lab setting dataset with manual additions is included in *Appendix A.1*.

The live Web Application is hosted on an ***Microsoft-IIS Server,*** which stores requests in the ***W3C Log Format*** with the following fields by default (Microsoft, n.d.)***:***

- ***Date:*** date of the request

- ***Time:*** time of the request
- ***s-ip:*** IP address of the web server
- ***cs-method:*** requested verb (e.g., GET)
- ***cs-uri-stem:*** target of the verb
- ***cs-uri-query:*** query the client tried to perform, if applicable
- ***s-port:*** server port number
- ***cs-username:*** name of authenticated user, or [-] for unauthenticated users
- ***c-ip:*** IP address of the client
- ***cs(User-Agent):*** browser that the client used
- ***cs(Referrer):*** site that the client last visited, before being referred to this site
- ***sc-status:*** HTTP status code
- ***sc-substatus:*** substatus error code
- ***sc-win32-status:*** Windows status code
- ***Time taken:*** time that the action took, in milliseconds

The live dataset was extracted from web server of the business owning the Web Application and provided by Joanknecht. The access logs from 06-07-2023 and 26-07-2023 were extracted and consisted in total of 31.937 and 39.122 requests respectively. Again, for confidentiality reasons, the datasets are not allowed to be included in this document.

3.2 Applying PM^2 to Web Application Forensics

In order to answer the research question *"How to apply Process Mining in the context of forensic investigation of malicious Web Application activity?"*, a guide on how to execute such an approach is designed. In this master's thesis, the PM^2 methodology by van Eck et al. (2015) is used and adapted, generating a context-specific methodology for applying Process Mining in Web Application Forensics. The choice for this framework is based on its many citations (380+ according to Google Scholar per November 2023), contribution by well-known Process Mining researcher Wil van der Aalst and its successful application in other domains like logistics (van Cruchten & Weigand, 2018). The application in logistics, for example, also required some context-specific additions to the PM^2 methodology. The need for refining the methodology when applying it to a specific context appears natural, as different fields have different input data and different questions to be answered. Likewise within the Cyber Forensics domain, the tools in various subcategories have different specific goals and use diverse types of data as input

(Fernando, 2021). Therefore, the scope of this study is specifically on the context of Web Application activity. The following subsections describe the steps carried out at each phase of the PM^2 methodology to the lab setting Web Application, as depicted in *Figure 11*. In each section, first the tasks as defined in PM^2 are briefly explained. Next, the proposed steps for the context-specific methodology are discussed. Domain knowledge of Web Applications and the created Web Application is provided by the internship supervisor.

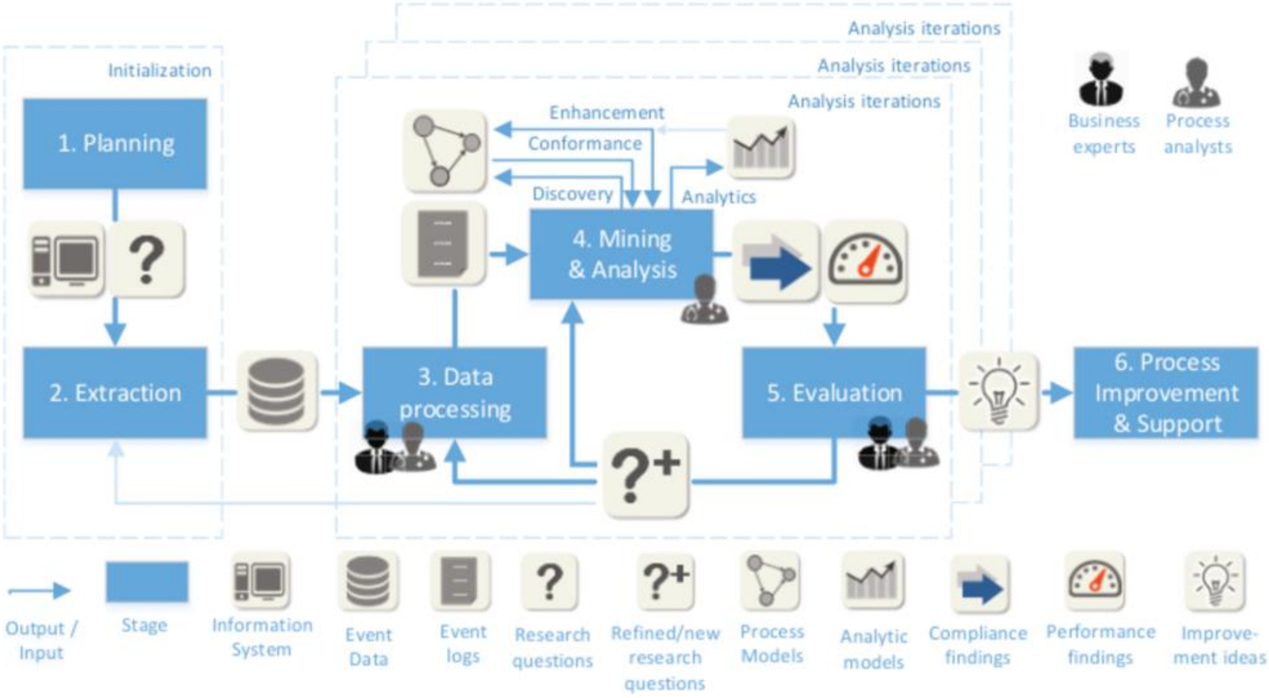


Figure 11: The PM2 Methodology (source: van Eck et al. (2015))

3.1.1 Planning

The first steps in PM^2, as defined by the authors, are as follows. First, the business process and related information system that the project focuses on is selected. Next, a research question is to be identified which can be either concrete or more abstract up front. Lastly, a project team project team consisting of Business Owners, Business Experts, System Experts and Process Analysts should be composed.

The **business process** in this context is not a real 'business process' but concerns reported malicious activity on a Web Application. Therefore, defining the 'business process' should be replaced by defining the reported malicious activity that the investigation focuses on. In the lab setting Web Application, it concerns the report of unauthorized access to a page that should be hidden for users that are not logged in. The information system in this context, logically, always concerns the Web Application and its server.

The **research question** in this context can be phrased relatively concrete, as it concerns the investigation on very specific behavior that is to be extracted. In this case, the following question can be posed: *"In what way was an unauthorized user able to gain access to the hidden page of the Web Application?"***.** This question can be answered by executing a Process Mining analysis that shows the path a user was able to take to gain access.

The **project team**, for this context, should consist of a Business Expert and a Process Analyst. Normally, the Business Owner is in charge of the process, however in this case there is no actual business process involved. Further, the 'process' this context already relates to systems, implicating that expert knowledge of the IT systems equals to being the Business Expert. Therefore, the Business Expert and System Expert are the same person in this case. For the lab setting, the Business Expert is the internship supervisor who has extensive knowledge of the Web Application of Joanknecht their client. The Process Analyst performing the Process Mining analysis is myself, author of this master's thesis.

3.2.2 Extraction

The second phase consists of determining the scope of the data and subsequently extract the log from the information system. Optionally, the business experts and process analysts can create process models after transferring knowledge about the processes and data attributes with each other, which also is useful in the data processing stage.

Regarding **data extraction**, the approach is always the same for Web Application activity. The data is always stored in the web server's access log as described in *Section 3.1.2*. A user with admin access to the web server is able to extract the access log from the specific period in which the malicious activity took place. For the lab setting Web Application, the access logs from 16-10-2023 & 17-10-2023 were extracted.

The step where **process models** are created should be mandatory for Web Application Forensics. As the goal is to extract data that shows very specific behavior, domain knowledge should be applied in advance to define this behavior. Formalization of processes is often represented in Petri Nets, like in security auditing (Accorsi & Stocker, 2012). By defining behavior constraints formally in a Petri Net, the authors picked out violation of the defined process. However, activity on a Web Application is often very complex and allows for countless paths. Therefore, it is very hard to specify allowed behavior in Petri Nets. Although modern computers can create Petri Nets from millions of states (Van Der Aalst, 2016), it may not capture all allowed behavior as some behavior may have not appeared yet. That is why, in the context of Web Application Forensics, a business logic approach is proposed. Using domain knowledge, one can define business rules that represent (dis)allowed behavior on the Web Application. To create the process model, or business rule in this context, domain knowledge of the Business Expert is consulted. This is where the added value of session cookies come into play. A session cookie is assigned to every user that accesses the Web Application and keeps track of the user's session, including for example login status. When the user closes the browser or after a certain amount of time, the session cookie is deleted. For the lab setting Web Application, the following **business rules** are defined:

- a user must be authorized before gaining access to the hidden page, i.e. the user must have accessed the login page in the same session.
- If a user accesses the hidden page in a session during which the user did not access the login page before, it is defined as unauthorized access.

3.2.3 Data Processing

The objective of the third phase is to prepare the data so that the event logs are useful for the following two stages. Van Eck et al. identify four types of activities in the data processing stage:

- *Creating views*: Label the data so that process instances are formed.
- *Aggregating events*: Grouping data into higher levels to reduce complexity.
- *Enriching logs*: Create new data by computing attributes or enrich with external data.
- *Filtering logs*: Slice & dice, group variants or filter on compliance to make the dataset focused on the relevant data.

Creating a view on the data is comparable in most cases as most web server log formats are relatively similar. Applying domain knowledge, a reasonable case identifier is the person who accesses the Web Application, as the goal is to follow a path that users took. The best way to identify the user is by the session cookie that the server assigns to every user. Although log formats like W3C and CLF allow for logging cookies, they are often not logged (Varnagar et al., 2013). The IP-address of the client is the next-best identifier for users. It could be possible that one IP-address actually consists of multiple users accessing the Web Application, but they provide reasonable security that an IP-address represents a user. The activity should represent the path that a user takes, which are the requested web pages in this context. Other useful variables like the method or status code can be labeled as other attributes, which can be used for filtering.

Aggregating events is left out in the Methodology for Process Mining in Web Application Forensics. As the investigation focuses on very specific behavior, the exact events are particularly interesting to understand how the malicious activity was executed.

Enriching logs was not necessary for the lab setting case, as the logs were manually generated to show specific behavior. However, live data can be very complex and presenting the data right can be challenging. Therefore, this step is left in the revised methodology. Especially taking into account that phases 3-5 form an iterative process in PM^2 to keep improving the results, it is expected that the step will be useful for processing live data.

Filtering logs was not applied for the lab setting as the data was specifically generated. Though, slice and dice can be applied to filter out irrelevant events and create a more manageable dataset in a real-life application. Cases can also be grouped based on variance to create simpler processes. This is not relevant as the focus is not on discovering common processes. Compliance filtering is comparable to slice & dice as based on characteristics of cases or events, they can be left out.

In short, even though the data processing stage is less challenging for a dataset that is extracted from a lab setting Web Application, it very likely is important for preparing live data. Therefore, most of the steps are also included in the revised methodology.

3.2.4 Mining & Analysis and Evaluation

In this stage the processed data is loaded in a Process Mining tool and analyzed, using the process models, if created, to check for conformance. Van Eck et al. identify four types of Process Mining: discovery, conformance checking, enhancement and process analytics.

As concluded in *Section 2.1.1*, process discovery and conformance checking suit this context the best. As discussed in *Section 2.1.2*, Process Mining tool Disco is used to perform the analysis. The first step is process discovery. As the refined Fuzzy Miner algorithm used by Disco provides a 100% truthful process map where a user can customize what percentage of activities and paths are shown, this is a relatively simple step. As the lab setting data is a small dataset, all of the activities and paths are shown in *Figure 12*. It should be noted that this, logically, is not possible for a large live dataset with thousands of unique requests. The process map represents the activity on the Web Application on 17-10-2023. The following requests, or activities, can be identified in the figure:

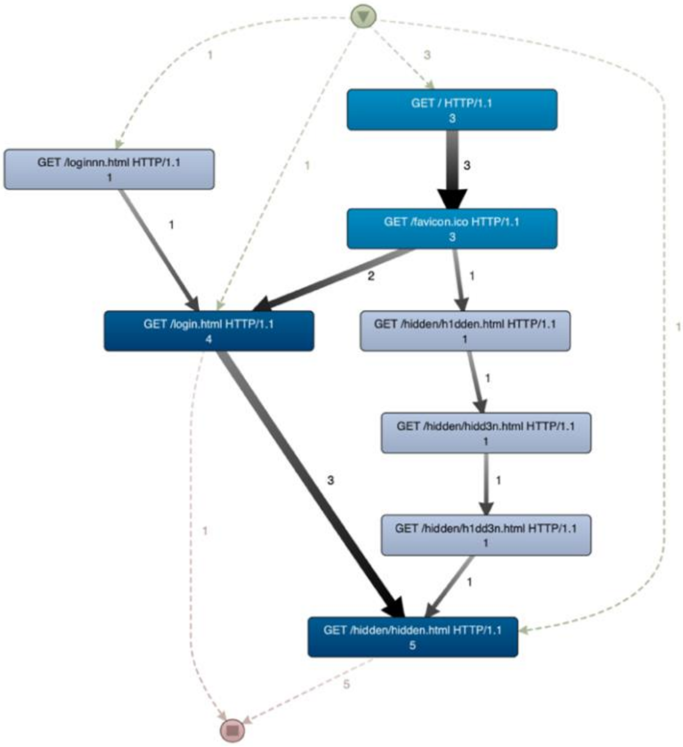

GET / HTTP/1.1
3
GET /loginnn.html HTTP/1.1
1
GET /favicon.ico HTTP/1.1
3
GET /login.html HTTP/1.1
4
GET /hidden/h1dden.html HTTP/1.1
1
GET /hidden/hidd3n.html HTTP/1.1
1
GET /hidden/h1dd3n.html HTTP/1.1
1
GET /hidden/hidden.html HTTP/1.1
5
3
1
1
3
1
2
1
1
1
1
3
1
1
1
5

Figure 12: Activity on lab setting Web Application (source: own work)

- ***GET /HTTP/1.1*** & ***GET /favicon.ico HTTP1.1*** indicate that a user accesses the homepage, as there is no path shown in the request. The homepage also automatically requests the favicon, or the website logo. This provides no extra information but is automatically logged.
- ***GET /login.html HTTP/1.1*** indicates that a user accesses the login page.
- ***GET /hidden/hidden.html HTTP/1.1*** indicates that a user accesses the hidden page.
- ***Any other request*** is a requested page that does not exist on the web server. Checking the status code attribute of these requests, all these requests have a status code that says '*404*'. This means that the requested page cannot be found.

Starting at the top, one can derive six cases, or users, that start with four different requests. Three users first request the homepage of the Web Application, a fourth user goes straight to the login page, the fifth requests a non-existing login page and the last user goes straight to the hidden page.

For conformance checking, the business rules derived from the knowledge of the Web Application is used. First, as the focus is on malicious activity, only cases that access the hidden page are interesting. Therefore, cases that do not access the hidden page are excluded. Secondly, users that gained legitimate access to the hidden page are left out by filtering out users who accessed the login page. The resulting process map, indicating the assumed malicious behavior, is included in *Appendix A.2*.

3.2.5 Evaluation

After the Process Mining analysis, the results ought to be evaluated. PM^2 identifies two steps in this phase: diagnose and verify & validate. Diagnosing means correctly interpreting the results, identify (un)expected results and possibly add/refine research questions for further iterations. Thereafter, the results are to be verified and validated to understand the root causes to extract ideas for process improvements. Van Eck et al. highlight the importance of having experts on the process involved in this phase, who have the knowledge required to draw conclusions.

The **diagnosis** in the lab setting is as follows. Analyzing the process map in *Appendix A.2*, The first case concerns a user that accessed the home page and tried various URLs that were no valid requests, before suddenly getting access to the hidden page. A second user appears to suddenly access the hidden page, which seems strange at first sight.

Subsequently, the described cases are **verified and validated**. At this point, the knowledge of the business expert is consulted to understand the results properly. The first case is clearly a malicious case of a user that tried to access the hidden page directly without authenticating themselves (see upper case in *Appendix A.3* for the full case). Even though the business rules prescribed that only authenticated users should access the hidden page, directly requesting the hidden page by posting the URL is not prevented. The second case, however, requires further investigation to understand it. As the case only contained one event, the access log from 16-10-2023 was added to the analysis. This shows that the same user, within the same session also accessed the Web Application on the day before. When looking at the full case, also included in *Appendix A.3*, one can derive that the user logged on to the Web Application on 16-10-2023, only to access the hidden page on 17-10-2023. In the lab setting dataset, the sessions were manually added to resemble non-expiring cookies. If sessions would expire after, for example, a day, the user should not have been allowed to access the hidden page anymore according to the defined business rules. However, in this case the session cookie was still valid, which makes it an instance where access to the hidden page was authorized.

From the evaluation an answer to the posed research question in phase 2 can be constructed. Although the idea of the Web Application was that it should lead users from the homepage, via the login page to the hidden page, it appeared that it is possible to access the hidden page otherwise. Specifically, unauthorized users are able to access the hidden page by simply guessing the right URL using trial-and-error. This illustrates a vulnerability where users can act outside of their intended permissions, i.e. Broken Access Control.

3.2.6 Process Improvement and Support

Van Eck et al. propose this phase to use the gained insights of the previous phase for improving the actual process. They further name Process Mining as a tool for operational support to assess the results of the implementation.

Although this application concerns an artificial setting, it can be illustrated how this phase is of use. The goal of the project is to gain valuable insights in how to improve the security of the Web Application. Thus, the gained information through the analysis and evaluation should be used to implement a solution for the vulnerability. Further, the implemented solution should be tested to assess whether it needs any further improvements. Process Mining probably does not add value to support the effect implemented solution, as this can better be tested by trial on the Web Application itself.

3.3 Case Application

The application to the lab setting described in the previous section, demonstrates the applicability of the approach. To further tailor the PM^2 methodology towards Web Application Forensics and validate it, a real-life case application is described in this section. The case concerns the live Web Application and corresponding dataset described in *Sections 3.1.1 and 3.1.2*. As described in the introduction of this chapter, the application can be regarded as a formative, naturalistic evaluation of the artifact. The evaluation is implicitly incorporated in the iterative process of applying the PM^2 methodology to the real-life case. Namely, the evaluations consisted of unrecorded discussions with the Business Expert, i.e. the internship supervisor, on the taken approach and required steps. Below, the case application is described by first applying the steps as proposed in *Section 3.2*. Further, the required additional steps resulting from the evaluations are discussed and implemented in the final version of *PM^2 for Web Application Forensics*.

1. Planning

Although the data is significantly more complex in this case compared to the lab setting, the reported malicious activity is comparable. Namely, it concerns an instance where unauthorized access to confidential documents from the company's Web Application took place. Therefore, the research question is also comparable to the lab setting: *"In what way was an unauthorized user able to gain access to confidential documents on the Web Application?"*. The project team also consists of the same business expert and process analyst as in the lab setting. *Figure 13* visualizes the proposed steps in phase 1 of *PM^2 for Web Application Forensics*.

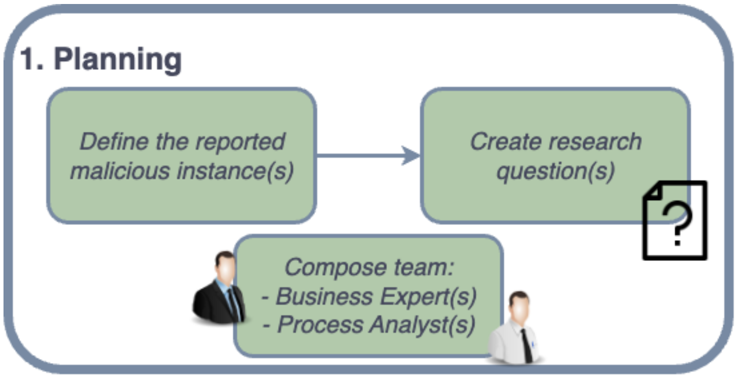

1. Planning
Define the reported malicious instance(s)
Create research question(s)
?
Compose team:
- Business Expert(s)
- Process Analyst(s)

Figure 13: Planning phase in PM^2 for Web Application Forensics (source: own work)

2. Extraction

The data extracted from the access log of the web server are from 06-07-2023 and 26-07-2023, the dates on which the reported malicious activity took place. To define the business rules, the Business Expert's knowledge of the Web Application is consulted. Although the exact business rules related to the Web Application cannot be shared due to confidentiality of the data, a more abstract version is the following:

- Every time a user accesses the Web Application, a *Session Cookie 123* is assigned to the user. After X minutes of inactivity, *Session Cookie 123* is deleted. New activity from the same user results in a new *Session Cookie 456* being assigned.
- If a user successfully requests cs-*uri-stem* ABC with argument *DEF* in the cs-*uri-query*, the login is most likely to be valid and, thus, the user is regarded as authenticated.
- In the same session, a user remains authenticated after logging in. If a session is ended, the user is logged out and therefore unauthenticated.
- *Document A* should only be accessed if a user is logged in.
- File B should not be accessible to any regular user.

Furthermore, the Business Expert transferred useful knowledge on common attack vectors relating to the security vulnerabilities in #1 from the *OWASP Top Ten*: Broken Access Control (*Table 3*). This information will be useful when interpreting, validating and verifying the results from the Process Mining analysis. *Figure 14* visualizes the proposed steps in phase 2 of *PM² for Web Application Forensics*.

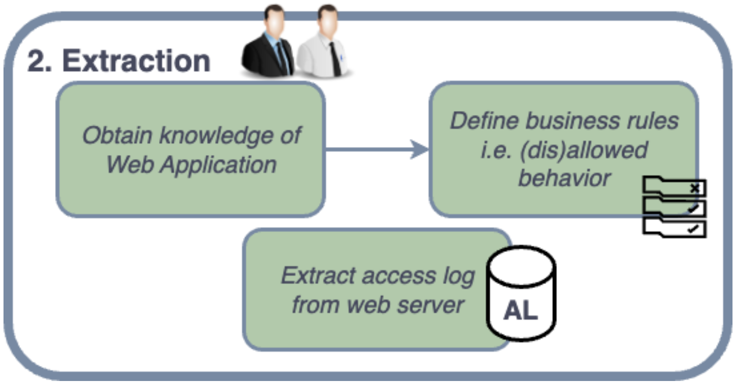

2. Extraction
Obtain knowledge of Web Application
Define business rules i.e. (dis)allowed behavior
Extract access log from web server
AL

Figure 14: Extraction phase of PM² for Web Application Forensics (source: own work)

3. Processing

This phase is significantly more complex with a live dataset compared to a dataset generated in a lab setting. As the Web Application allows for various behavior, the data is very unstructured. Moreover, the iterative process described by van Eck et al. in PM^2 certainly holds for PM^2 in Web Application Forensics as this phase was revisited quite a few times. Below, the created views, enrichments and filters applied in the various cycles are briefly described. The described steps were executed in Knime (Berthold et al., 2009), which is a free open-source platform for with tools for data manipulation and visualization. In one instance, which will be indicated below, Excel was used. The final Knime workflows and used functions in Excel are included in *Appendix B.1 and B.2*.

As described in *Section 3.1.2*, The log format used on the web server unfortunately was not configured to log session cookies. Therefore, in the first iteration the IP-addresses were used as case identifier in the datasets. The request (*cs-uri-stem*) is labeled as the activity and the query (*cs-uri-query*) as an attribute. Enriching the logs was required to create the timestamp for the analysis, by combining the *Date* and *Time* fields. Further, based on the transferred knowledge, the data was filtered. More specifically, only *'GET'* methods were included, some irrelevant request types were excluded, only successful requests were included (*sc-status* = '*200*') and only the columns relevant to the analysis were included.

Unfortunately, the results were insufficient, so the datasets had to be reprocessed. The second time, sessions were manually created by adding the hour and minute to the IP-address to form the case identifiers. However, this still led to insufficient results in both datasets. Finally, after another evaluation with the Business Expert, a new manner to assign sessions to the logs was designed using Excel. For the exact formula and a more detailed explanation, see *Appendix B.2*. *Figure 15* on the next page visualizes the proposed steps for phase 3 of *PM^2 for Web Application Forensics*.

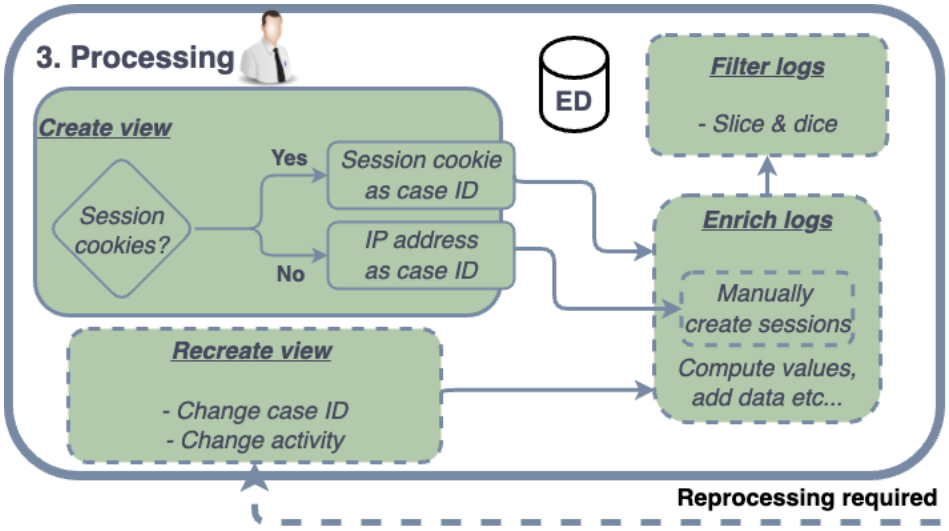


Figure 15: Processing phase of PM² for Web Application Forensics (source: own work)

4. Mining & Analysis

Below, the steps in the first iterations are pointed out briefly, next to discussing the various steps taken in the final iteration. The input for this stage are the prepared datasets and business rules. Like in the lab setting, Process Mining tool Disco is used to execute the analyses.

The access logs from 06-07-2023 and 26-07-2023 concern two separate instances of malicious activity and, therefore, loaded in Disco as individual datasets. As expected, it is not possible to present a comprehensible process map of the datasets. Furthermore, a high-level view of the data does not provide value as it does not show the behavior that is aimed to be extracted. Only the specifically defined behavior is of interest. Therefore, the discovery step of the analysis does not provide any insightful results yet.

The conformance checking step, however, is where the valuable insights can be created. Using the business rules, the IP-addresses that accessed the sensitive documents could be identified in the first iteration already. By filtering on the cases that successfully requested the specified documents, only interesting cases were extracted. However, as there were no sessions present, the results could not provide any other insight. The first attempt at manually creating sessions resulted in inaccurate grouping of events, as described in phase 3. At the next iteration, time between events was taken into account, resulting in more accurate representation of the sessions. Thus, this modification led to more useful results in the dataset of 06-07-2023. However, no results were extracted from the dataset of 26-07-2023. Still, as the business rules were followed, the finding was taken to the next phase. After evaluating with the Business

Expert, this phase was repeated once more. The analysis was executed slightly differently, which did end up leading to a result. The evaluation phase will present the extracted results and discuss them further. *Figure 16* visualizes the proposed steps in phase 4 of *PM^2 for Web Application Forensics*.

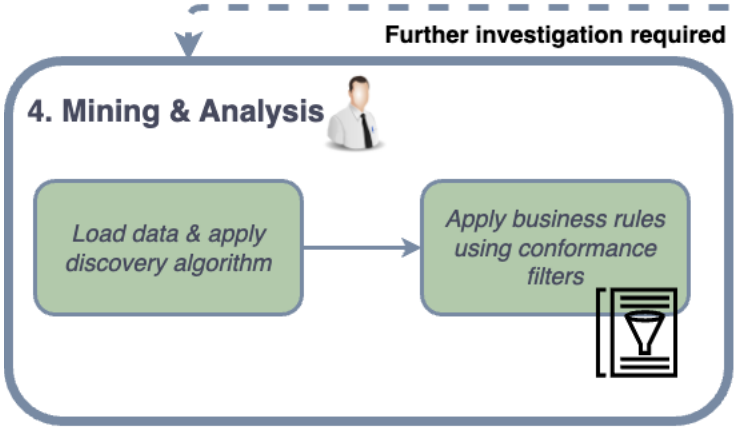

Further investigation required
4. Mining & Analysis
Load data & apply discovery algorithm
Apply business rules using conformance filters

Figure 16: Mining & Analysis phase of PM^2 for Web Application Forensics (source: own work)

5. Evaluation

In this phase, first, the results of the analysis were interpreted to check whether the results are expected. During the first iteration, this was not properly possible as the results were not detailed enough to draw conclusions from. After validating the results in the second iteration, it appeared that the results were not accurate. In both cases, it was necessary to go back to phase 3 to reprocess the datasets. After processing the datasets a third time, the results from the Process Mining analyses were as follows.

In the access log from 06-07-2023 there were two sessions from two different IP-addresses in which *File B* was accessed:

- The first instance concerned a user that started its session on the home followed by the login page. Thereafter, an attempt at retrieving *File B* from the login page was successful. The rest of the session is filled with activity on the login page with numerous suspicious-looking requests. The case lasted 7 minutes and included 32 requests. A reconstruction of the relevant part of the case is included in *Appendix B.3*.
- In the second case, it immediately stood out that it consisted of 6.533 requests in a timespan of 14 minutes, with. In around 650 requests over a few minutes, *File B* was accessed dozens of times. The activity appeared to be a constant stream of various

suspicious requests. A reconstruction of the relevant part of the case is included in *Appendix B.3*.

In the access log from 26-07-2023 there was one session in which *Document A* was accessed.

- It appeared that the access was authorized as the user did adhere to the formulated business rule of a login before accessing *Document A* on two occasions in the same session. Therefore, the case was not included in the results at first, which was an unexpected result. However, after the evaluation with the Business Expert, the repeated analysis resulted in the case being included. The case contains 35 requests over 4 minutes, with the login occurring as the 26th request. Thereafter, *Document A* is accessed twice. Surprisingly, the queries do not fully correspond. A reconstruction of the relevant part of the case is included in *Appendix B.3*.

Subsequently, the findings from both datasets were verified and validated, leading to the following explanations formulated in accordance with the Business Expert:

- *File B* was accessed unauthorized at the first instance by including a malicious query behind a valid stem, which was allowed by the server. So, unauthorized users are not prevented from requesting sensitive files from the server. The second session in which *File B* was accessed, appeared to be some kind of an automated attack with the number of requests in such short time span. However, the Business Expert pointed out that this case concerns a scan. From the log can be extracted that the *File B* is accessed using path traversal techniques. Both manners of access are possible because of a *Broken Access Control* vulnerability in the Web Application.
- *Document A* appeared to be accessed in a valid way, but in fact was not. Although the user logged in to the Web Application in the same session, the second instance where the document was accessed, was malicious. The reason is the difference in the query of the request. The first access is legitimate, however the second access includes "../../" in the query. This, again, implicates path traversal techniques being used to gain access, relating to a *Broken Access Control* vulnerability.

With these findings, an answer to the research question presented in phase 1 can be formulated. Namely, the Web Application has *Broken Access Control* vulnerabilities. Malicious input in the URLs from users is not prevented, which allowed them to gain unauthorized access to sensitive documents. In these cases, users took a brute force approach by guessing some working URLs and using path traversal techniques. Further, the conclusion can be drawn that the right data is required to reach optimal results. The fact that session cookies were not logged, complicated the analysis. Although creating the cookies manually did succeed, it is very hard to reach perfect accuracy. This could be an explanation of the unexpected result from the 26-07-2023 dataset at first.

After completing phase 5, some claims also made by van Eck et al. (2015) are reconfirmed. Corresponding with their finding to iterate phases 3,4 and 5, this case also required going through them several times. Therefore, this loop is also included in *PM^2 for Web Application Forensics* as can be seen in *Figure 17*. Furthermore, Van Eck et al.'s remark to involve right experts in the verification and validation is also found to be paramount in this case. Therefore, there should be extra attention towards the cooperation between the Process Analyst and Business Expert in phase 5.

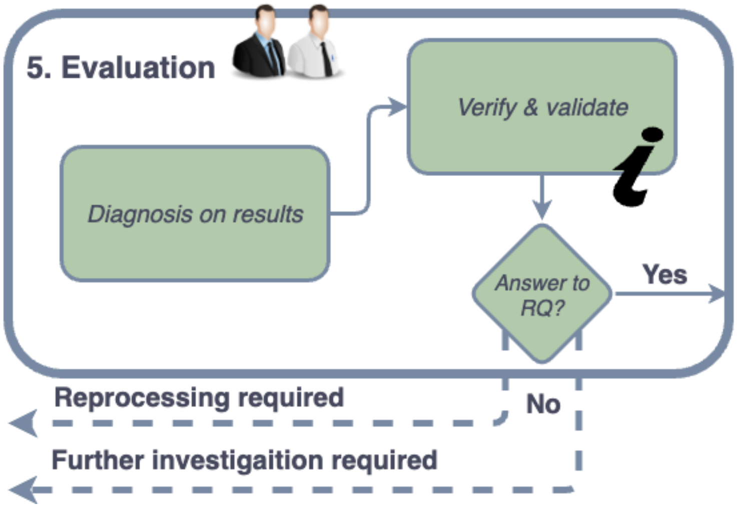


Figure 17: Evaluation phase of PM^2 for Web Application Forensics (source: own work)

6. Process Improvement and Support

Phase 6 is the last, but certainly not least important, in *PM^2 for Web Application Forensics*. Using the findings resulting from phase 5, a solution for the Web Application's vulnerability can be designed. Although this step in the case application is not part of this master's thesis, the Web Application's security could be significantly improved by redesigning the Web Application. *Figure 18* visualizes the proposed steps in phase 6 in *PM^2 for Web Application Forensics*.

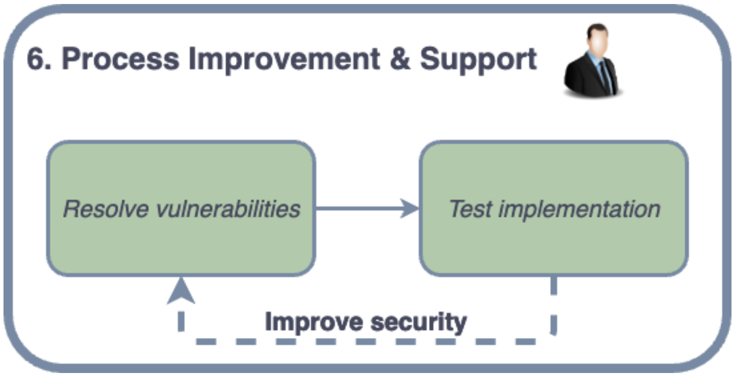


Figure 18: Process Improvement & Support phase of PM^2 for Web Application Forensics (source: own work)

3.4 PM^2 for Web Application Forensics

In the previous two sections, the Project Mining Project Methodology by van Eck et al. (2015) was applied to two Web Application Forensic cases to tailor it for the specific context. Combining the individual phases from *Figures 13-18* results in the final *PM^2 for Web Application Forensics* as visualized in *Figure 19* on the next page. The artifact is complemented with a legend that explains the icons and layouts used in the methodology. More textual explanation on the methodology is included for each phase in *Table 5*, two pages further.

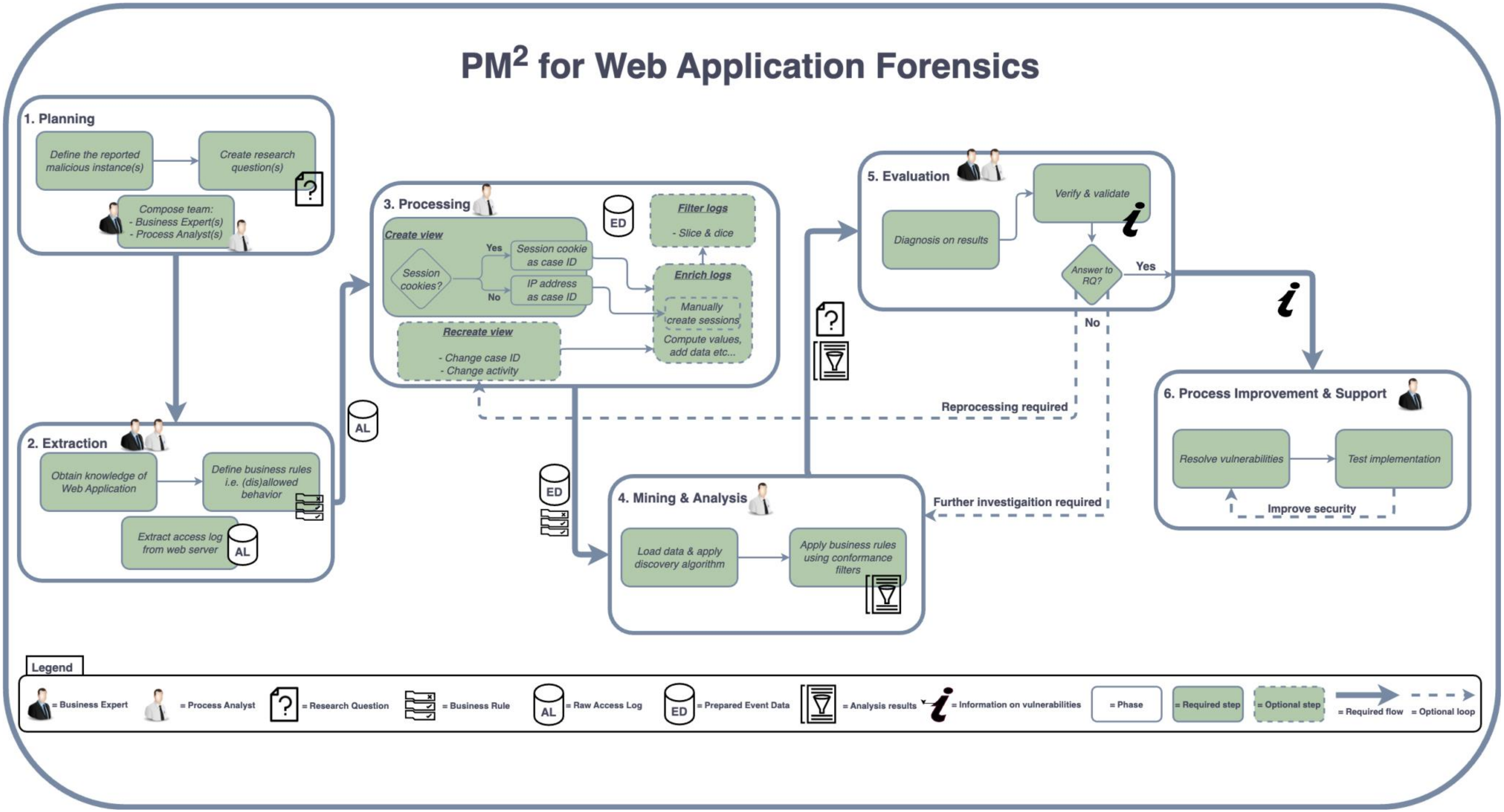


Figure 19: PM² for Web Application Forensics (source: own work)

PM^2 for Web Application Forensics Explained	
1. Processing	- First, define the subject of the forensic investigation. That is, the malicious instance that has been reported → Define what you want to know about the instance by formulating a *Research Question*. - Compose a project team consisting of one or more of both *Business Experts* (knowledge of the Web Application) and *Process Analysts* (Process Mining expert).
2. Extraction	- Let the *Business Expert* transfer knowledge about the Web Application to the team → Define behavior that is allowed or expected, and behavior that is regarded as malicious and formalize them in *Business Rules*. - Extract the Raw Access Log from the web server on which the Web Application is hosted, selecting the relevant timeframe in which the reported activity took place.
3. Processing	**If from Phase 2 → Phase 3:** - Process the data for the analysis, starting with creating a view. → If the log contains session cookies, use them as the case identifiers. → If not, use the client IP-address → manually create *‘sessions’.* - Depending on the defined view, enrich the log and/or filter out irrelevant data to possibly improve the dataset. **If from Phase 5 → Phase 3:** - Based on the Evaluation, define what additional insight is required → recreate a view → Enrich the logs further to gain more valuable input.
4. Mining & Analysis	**If from Phase 3 → Phase 4:** - Load the *Prepared Event Data* in a Process Mining tool that can convert the log into a process model → Based on the formulated *Business Rules*, apply conformance filters so that the malicious behavior is extracted. **If from Phase 5 → Phase 4:** - Investigate the extracted behavior in more detail and include the details in the improved *Analysis results*.
5. Evaluation	- Describe what can be interpreted from the *Analysis Results* → Discuss with the *Business Expert* whether the results are insightful, valid and provide an answer to the *Research Question*: → If yes, advance to *Phase 6*. → If no, → Should the results be further investigated? → Iterate Phase 4. → Is the input data not able to provide sufficient insight? Iterate Phase 3.
6. Process Improvement & Support	- Based on the *Information on vulnerabilities*, the *Business Expert* should aim to have the issue on the Web Application resolved → Test whether the implemented solution is sufficient → If not design new solutions until it solves the vulnerability.

Table 5: Textual clarification to each step of PM2 for Web Application Forensics (source: own work)

3.5 Expert panel Interviews

In addition to applying PM^2 to artificial and naturalistic cases, the added value of a Process Mining approach in Web Application Forensics is also assessed by interviewing an expert panel consisting of experts from the related fields. The purpose of consulting the experts is to evaluate *PM^2 for Web Application Forensics* and incorporate their view on the added value of Process Mining analyses in the context of Web Application Forensics.

The selection of interview candidates was based on their expertise in the fields of Cyber Security, Web Applications, Forensics, Process Mining or a combination multiple. Three interviews were held with colleagues from Joanknecht *[A-C]*, one with a Process Mining software company owner *[D]*, and one with a Cyber Security company owner *[E]*. Finally an interview with four Cyber Security experts from KLM *[F-I]*, who were involved in the investigation of their recent Web Application-related incident (Schellevis, 2023), took place. Beforehand, only one KLM Cyber Security expert was scheduled for an interview. However, per their request, some other experts joined the interview as well. For transparency, all experts are included individually in *Appendix C.1*. With consent of the participants, some more relevant information about the participants and their fields of expertise is also included.

The approach for interviews is semi-structured. Before starting the interview, a brief introduction to the unfamiliar domains of the thesis was given to the participants, if needed. Next, the artifact including textual explanation (*Figure 19* & *Table 5*) were presented to the participants to study for a few minutes. Subsequently, the artifact and the approach were evaluated by asking open questions based on seven metrics: *functionality, completeness, consistency, accuracy, reliability, usability* and *performance*. The metrics are extracted from Hevner et al. (2004), who propose to use these metrics to evaluate an IT artifact in design science research. The open questions related to the artifact itself, as well as the approach in a broader sense. If deemed necessary, follow-up questions were asked based on the response to gain a better understanding of the expert's view. After going through the prepared questions, the participants were asked if they had any questions or remarks left. Almost all experts also asked questions back about the artifact or approach throughout the interview. This resulted in that, sometimes, certain questions or topics were already discussed before explicitly posing the question. It also occurred that an expert brought up a topic that was not directly related to one of the metrics or an expert gave answers that did not relate specifically to the metric that the question related to. These facts play a role in the decision on how the interviews were coded, which is described hereafter.

The interviews were transcribed using Transkriptor (www.transkriptor.com), which is an AI powered tool that can transcribe audio files. After loading the interviews in the tool, all transcripts were reviewed by comparing it to the audio files and finetuning the transcript. Thereafter, the interviews were coded manually in Excel according to the Thematic Analysis method. Kiger & Varpio (2020) state that the data can be analyzed using an inductive or deductive approach, depending on the context. Up front, the idea was to code the interviews deductively by using the mentioned metrics by Hevner et al. (2004). However, after the interviews it appeared that some remarks did not fit in one of the categories. As such, during the coding new groups were created for responses that did not fit the metrics. Therefore, the coding process included both deductive and inductive approaches. After the initial coding, similar answers were merged including notes which experts made the remark. Subsequently, remarks were labeled with keywords. The keywords were then coded into three main groups based on the seven metrics that guided the interviews: the artifact itself, the implementation of the artifact, and the approach in general. In the introduction of *Chapter 4,* the three high-level groups are explained further. It should be noted that, for some individual answers, the initial coding of answers in the metrics (see *Appendix C.2*) differs slightly from where it is presented in *Chapter 4*. There are two reasons for this. The first is that some answers fitted to multiple metrics, but each answer was assigned to only one in the coding process. When merging the answers from all experts, a choice was made where the response fit best. The same goes for labeling the keywords to the three main groups. Some answers labeled to keywords fitted better to a group (and thus one of the metrics) that did not correspond with the initial metric the answer related to. To avoid confusion, the categorization of answers as presented in the results should be followed.

Per request of the participants, the full transcripts are not included in the appendix. However, the coded interviews including questions are included in *Appendix C.2*. To make sure there was no bias in coding the interviews, some measures were taken. Namely, coded interviews and the full transcript were shared with the participants to give them a chance of correcting parts that might have been understood incorrectly. Per e-mail, all experts approved their coded interview and consented to the code being included in the appendix.

4. Results

Before providing an answer to "*how*" and "*to what extent*" Process Mining can be applied in Web Application Forensics, the results of the research are presented in this chapter. Subsequently, the results are discussed, and a conclusion is drawn in *Chapter 5*. The results of this study are extracted from the interviews with the Expert Panel, as described in *Section 3.5*. The results will be presented in the following three categories that are based on the seven metrics that guided the interviews:

- The Artifact: First, the experts' opinions on the methodology itself is presented. It is about the functionality and completeness of the steps and consistency of the methodology when used by different people and applied to different cases.
- The Implementation: Secondly, the factors that determine success of implementing the methodology according to the experts are presented. This is about factors that determine the accuracy and reliability of the results, and ones that determine the usability of the methodology.
- The Approach: Lastly, the thoughts about the approach in a more general sense are presented. The performance of the methodology is assessed by presenting the experts' views on the possible added value and limitations of the approach.

Findings from the interviews are summarized and indicated with ***{x}*** at the end of the summarizing sentence. All findings are included in *Appendix C.3* and form the base for the discussion of the results in the next chapter. If relevant for an expert's claim, their expertise is indicated with an abbreviation of Process Mining (PM), Cyber Security (CS), Web Applications (WA) and/or Forensic Investigations (FI).

4.1 the Artifact

PM^2 for Web Application Forensics has been developed to serve as a guide on how to apply Process Mining in forensic investigations on Web Applications. It is based on the Process Mining Project Methodology by van Eck et al. (2015) and tailored to Web Application Forensics by applying and adjusting it to two relevant cases. To be able to argue that the methodology is adequate to a certain extent, the Expert Panel provided their feedback.

After studying and discussing the artifact, the experts were first asked about the functionality of the steps and completeness of the methodology. That is, to what extent the steps are clearly formulated and useful, and whether important steps might be missing. About the formulation, *Expert [A]* immediately pointed out that understandability depends on the user of the methodology. There is knowledge from Process Mining and Web Applications required before being able to understand the steps in the methodology. Although there were some remarks (and *Expert [D]* did not make an explicit remark about whether the methodology was coherent or not), in five interviews the experts concluded that the methodology overall looks logical and coherent ***{1}***. For example, *CS experts [F]-[I]* work with Web Applications daily and after discussing the methodology one mentioned:

> *"For me, it [the methodology] really feels like an analysis of a Web Application [like we would do it]." – CS/WA Expert [H]*

Regarding the consistency of the methodology, *experts [A], [B] & [D]* mentioned that it would probably be interpreted likewise by different users. Though, they as well as other experts mentioned other factors depending on this, which relates to the implementation and thus will be further discussed in *Section 4.2*. Experts did make some relevant remarks on the applicability of the methodology to other cases, or the generalizability. First of all, *FI Expert [C]* mentioned that extra steps at the final phases are required in case the goal of a forensic investigation is legal prosecution and *Expert [B]* posed where the forensic investigator is put in action in the methodology. In such cases, *CS* Expert *[E]* misses a specific step where for example IP-addresses are normalized. Otherwise, there is a large chance that you do not possess the right data. So, in case of a forensic investigation with a goal of legal prosecution, more specific data preparation and legal steps afterwards are required ***{2}***. Furthermore, the generalizability to other Web Application Forensics cases depends on the type of Web Application you deal with (CS *[E] & [F]-[I]*). Newer Web Applications have more complex architectures, with different applications interacting with each other. This makes it a hard challenge to extract and correlate the data properly from the various servers, according to the Cyber Security Experts. Moreover, CS *Expert [E]* pointed out that modern Web Applications may use other identifiers for users and sessions rather than session cookies. For these reasons, modern Web Application (architectures) may require steps to be added or changed in the methodology ***{3}***. The findings regarding the Artifact and their frequency of mentioning in the individual interviews are visualized in *Figure 20*.

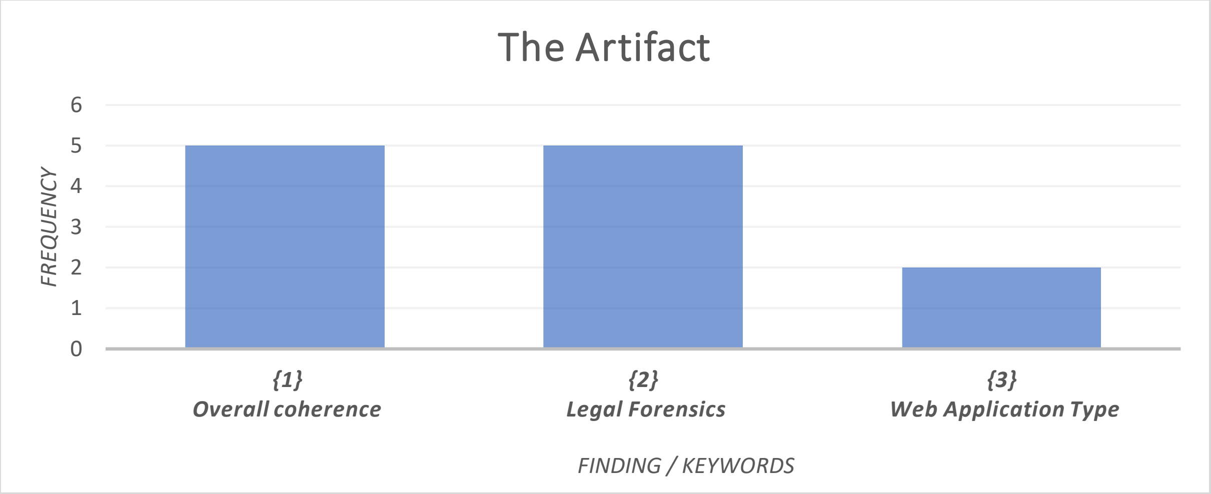


Figure 20: Mentioning frequencies of findings regarding the Artifact

4.2 The Implementation

Although the methodology itself might be regarded as adequate in general, there also can be factors that determine how successful the implementation can possibly be. In the next section, the experts' opinions on the important factors when implementing the methodology are presented.

Regarding the accuracy of the results, experts [A], [C], [D] & [E] explicitly pointed out the importance of starting with a good research question ***{4}***. One said:

> *"It depends on the question: the more specific the question, the more specific an answer can be." – PM/CS Expert [A]*

PM expert [D] mentioned that in practice, the research question often needs to change to get the right results, so you also need to keep all useful data as it might be useful in the analysis phase. Thus, a flexible approach to analyses can improve the results according to the expert ***{5}***. Furthermore, Experts [A] & [D] argue that users should be able to understand what data is available and what the data can tell ***{6}***. *Experts [A]-[E]* also highlighted the importance of the Business Expert's competence to translate their knowledge of the Web Application into usable Business Rules. Further, in two interviews with Cyber Security experts was highlighted that in practice, there often is a difference between someone who knows the processes in the Web

Application and someone who knows about Cyber Security. One expert mentioned the following:

> *"I often encounter product owners or software developers who have no idea about security. They build something that is functional, but do not realize what injecting malicious code in an input field can cause. You need someone with the right mindset [for cyber security]." – CS Expert [E]*

People with knowledge from Cyber Security not only understand the Web Application's functionalities, but also know about attack signatures for example. The same expert also gave the following example:

> *"There are SQL injection patterns that might adhere to a rule, which can lead to the conclusion that there is no deviation [with a Process Mining analysis]. But if you look at signatures of e.g., a timing attack, you have to look at the response times instead of URLs." – CS Expert [E]*

When asked about what knowledge and skills are required to use the methodology, the importance of the Business Expert's knowledge was highlighted again by the experts. *Expert [E]* mentioned that the Business Expert should be someone who has experience with the Web Application and is able to communicate their knowledge well. *PM experts [A] & [B]* also mentioned the skills of the Process Analyst as important. More specifically, *PM Expert [B]* pointed out that a Process Analyst should not only blindly apply the rules, but also be invested in the subject to create meaningful results. In short, the methodology is not useful if there is no expertise available. To make the methodology useful and extract accurate results, the Business Expert should have knowledge of both the Web Application and Cyber Security, and the Process Analyst should be involved in the investigation ***{7}***.

The reliability of the data throughout the investigation also is of importance. First of all, *experts [A] & [D]* made a remark that the reliability of the data depends on whether it was possible to tamper with the data. *PM/CS expert [A]* specifically mentioned that if a user gained rights to log directories, the data can probably not be trusted from the beginning ***{8}***. If this is not the case, data integrity is especially important in forensic investigations with goal of legal prosecution. Various experts raised points related to this: Forensically sound extraction of logs (*[A], [B]*), keeping an original file to go back to (*[C]*) and following the Chain of Custody when handling the data (*[E]*). These results relate to finding ***{2}***, as extra steps are required in such specific cases. Privacy of sensitive data is also a topic that is related to finding ***{2}***, which was brought up by *FI expert [C] & expert [D]*.

Finally, *experts [A] & [B]* pointed out that, when implementing the methodology, users should learn from challenges they face in for example insufficient logging of the data. If there is relevant data missing in the logging for a case, you need to make sure that it is there for a next case where the data is required ***{9}***. One expert said:

> *"Every research question will be different, requiring different data. If you include more data in the logs, there is a bigger chance you will have the right data for the next case." – PM/CS Expert [A]*

The findings regarding the Implementation and their frequency of mentioning in the individual interviews are visualized in *Figure 21.*

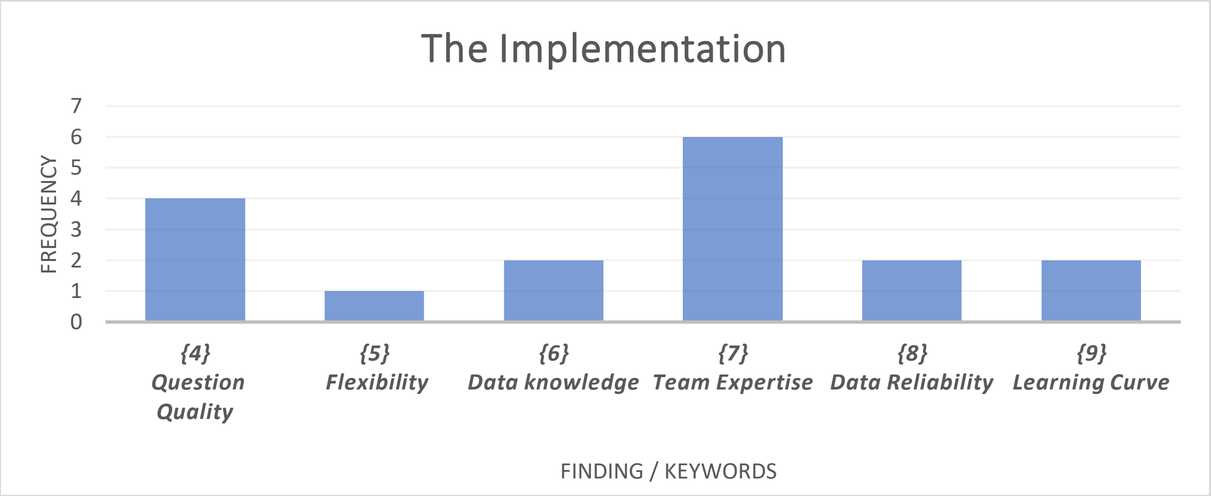


Figure 21: Mentioning frequencies of findings regarding the Implementation

4.3 The Approach

Next to discussing the methodology and its implementation with the Expert Panel, they were also asked about their view on the possible added value and limitations of applying Process Mining in Web Application Forensics in general.

In all six interviews, the experts mentioned that Process Mining can be useful in helping to understand paths of malicious users. One further explains:

> *"Detecting malicious access is not the big win, … , that is a static analysis. … The dynamic part is that you can follow a path and understand how that was possible, that is what the benefit of Process Mining is." – PM/CS Expert [A]*

FI expert [C] & CS expert [E] highlighted that the approach can be useful complementary to other tools. Both mentioned that next to the forensics or cyber security tools, Process Mining can be exploited to dive into data in more detail. One expert concludes:

"I think it is an interesting study what you are doing, which could potentially be useful for our company." – FI Expert [C]

So, Process Mining could be of use in forensic investigations to help understand attacker's paths, for example in support of other tools to dive deeper into the data ***{10}***.

However, after discussing thoroughly with each other during the interview, *CS experts [F]-[I]* concluded that they saw no real-life forensics cases in which they would see an added value for Process Mining. The large company, with complex Web Application Architectures, already employs tools that can deal with this complexity. Although such tools cannot be used by applying business rules to filter cases based on certain behavior, they do run automatically, and queries can be used to search for cases. For their team, it is not worthwhile to manually set everything in place, especially regarding acquiring all the required datasets, for a Process Mining analysis. They argue that their current specialistic market standard tools, though expensive, should be used in their complex environment. On the other hand, the experts acknowledged the following:

"Let's not forget, we are a large corporation with plenty resources and tools. Such [Process Mining] approaches are more useful for smaller companies without expensive tools, who want to look at how something happened." – CS/WA Expert [I]

He adds:

"I think it can add value, in cases for SMEs with simple and small applications. However, when it [Web Application architecture] gets complex, other larger tools are needed" – CS/WA Expert [I]

These findings are also related to finding ***{3}***, as modern, complex architectures require adjustments to the methodology. However, *CS experts [F]-[I]* argue that in large companies, like theirs, Process Mining is probably not a suitable approach. In such cases, other specialistic tools are required. Still, for smaller companies with smaller budgets and simpler architectures they do see added value for the approach ***{11}***.

Though, *CS experts [F]-[I]*, along with *CS experts [A] & [E],* did point out another opportunity where Process Mining could be used. *Expert [A]* described it in the most general way as investigating whether malicious activity happened. The others were more concrete and mentioned threat hunting by testing hypotheses about possible misuse (*[F]-[I]*) or thinking of evil user stories to extract certain behavior (*[E]*):

> *"In the agile methodology [for building software] you have evil user stories (…) you could write down what attackers would want to misuse looking at the confidentiality, integrity and availability of the data (…) and subsequently translate these to business rules." – CS Expert [E]*

Although referring to different concepts, the way the methodology is applied would be alike. All experts refer to use cases where malicious activity is expected or might have occurred, and Process Mining can be used to test it and aim to extract certain malicious behavior according to defined rules ***{12}.*** *PM Expert [D]* pointed out that Process Mining can also be used not for investigation, but rather for communication of something that is already known. By visualizing a path, it can also add value ***[13}***.

Another factor on which the applicability of Process Mining in Web Application Forensics depends is the research question. According to *experts [A], [C] & [D]*, not for every problem a Process Mining approach is the answer. For some questions, a static analysis suits well and thus that approach should be taken. Expert [A] mentioned that when defining the research question, one should also ask oneself what tool fits best to answer the question because there are *"different concepts for different purposes"*, as he got taught some time ago. Another expert said their current analysis tool is always useful and concluded:

> *"I do not think that Process Mining will be the first used tool in every investigation, it really depends [on the research question]." – FI Expert [C]*

In short, Process Mining is not the approach to answer all answers in Web Application Forensics cases. Depending on the research question, a suitable approach should be chosen before blindly taking Process Mining ***{14}***. The findings regarding the Approach and their frequency of mentioning in the individual interviews are visualized in *Figure 22.*

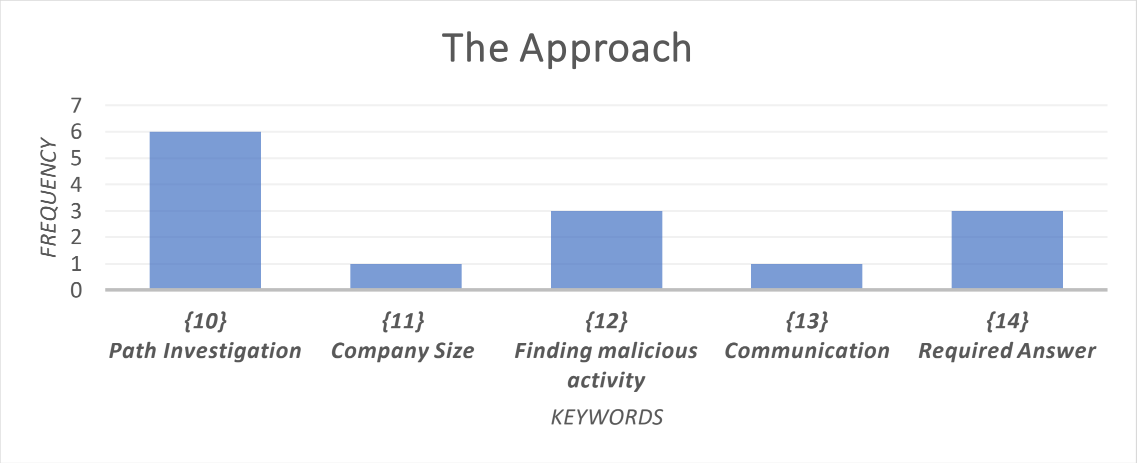


Figure 22: Mentioning frequencies of findings regarding the Implementation

5. Discussion, Limitations & Future Research

In this chapter, an answer to this master's thesis' research question will be formulated based on the findings presented in *Chapter 4*. In the discussion, ***{x)*** referrals will be used to indicate the findings from this study. For further information on each finding, see the corresponding ***{x}*** in the previous chapter. Further, the results are compared to the literature reviewed in *Chapter 2*. Finally, the limitations of the research and suggestions for future research are presented.

This study aims to answer the question *"How, and to what extent, can Process Mining be applied in forensic investigation of malicious activity on Web Applications?"*. To answer this, artifact *PM*^{2} *for Web Application Forensics* was developed and evaluated. Based on finding ***{1}***, one can conclude that the developed artifact presents a proper guide to apply Process Mining in Web Application Forensics. However, other findings indicate that it cannot be applied to any case or without taking notice of some comments. In cases where the forensic investigation aims for legal prosecution, additional steps on top of the ones described in the methodology are required in various phases of the methodology ***{2}***. Moreover, the applicability of the methodology depends on the type of Web Application. Cases concerning modern, often more complex Web Applications, require some steps to be changed and collecting data is harder ***{3}***. Furthermore, when implementing the methodology, there are some requirements according to the Expert Panel. Some indicate that an important starting point is a good research question to get to a good answer ***{4}***. Also a flexible approach for the analysis is advised ***{5}*** and users should have a good understanding of the data they are using ***{6}***. Further, extensive knowledge of the Web Application, Cyber Security and the ability to translate the knowledge to rules is marked as important, next to an invested Process Mining expert to carry out the analysis ***{7}***. Reliability of the data is another requirement, which can depend on what degree of access a malicious user in the case was able to obtain and thus maybe tamper with log data ***{8}***. A suggestion some experts made was to learn from cases where it appeared that the logs were not configured to log important data ***{9}***. This issue also arose during the case application in Section 3.4, where session cookies were not present. Regarding the approach in general, the experts identified opportunities and limitations. Process Mining in Web Application Forensics is regarded useful as a way to get a better understanding of paths behind malicious activity ***{10}***. Moreover, the approach could be used to test whether some defined malicious behavior occurred in a log ***{12}*** or visualizing data for communication purposes ***{14}***. Though, there were also limitations to the added value of the approach identified. In very large companies, there is no significant use for the approach in forensic investigations, as there are other specialistic tools required that can deal with the complexity of their Web Application architectures. Though, it is argued that for smaller companies with smaller budgets and simpler architectures there is an added value for the approach ***{11}***. The fact that Process Mining is probably not the most fitting tool for every Web Application Forensics research question is also highlighted. As a research question is defined, one should consider various approaches and choose the most suitable to answer the research question ***{13}***.

At this point, an answer to the research question can be formulated. *PM^2 for Web Application Forensics* can be regarded as a suitable guide on "*how*" to apply Process Mining in forensic investigations of malicious activity on Web Applications. Though, in cases like legal forensics or concerning modern/complex Web Applications, it should be noted that other steps are required that are not included in the artifact produced in this paper and thus indicate the scope of the methodology. Moreover, one should take into account the prerequisites and suggestion to be able to extract accurate results, which also partly answers the "*to what exten*t" in the research question. The rest of the answer refers to the possible useful applications and limitations of the approach. The results suggest that the approach is useful when aiming to understand malicious behavior, which also appeared from the case applications in *Sections 3.3 & 3.4*. Further, some results point out that the approach can also be useful in checking whether certain behavior occurs using rules, though this application has not been exploited yet. There are also indications of the limitations as to what extent the approach can add value. It looks like in large companies with complex Web Application architectures, there is less need and use for the approach as other expensive specialistic tools are required. It is implied that the use is more for SMEs with simpler architectures and smaller budgets. This is validated by the case application in *Section 3.4*, as this concerned a single-standing Web Application. Finally, it should be noted that a Process Mining approach is not suitable for every problem, as it is very dependent on the research question what kind of approach suits best.

To summarize the answer to the research question: Process Mining can be applied in forensic investigation of malicious activity on Web Applications and add value by employing *PM^2 for Web Application Forensics*, though taking into account the scope, requirements and limitations of the approach.

As there is little research on Process Mining in Cyber Forensics, let alone Web Application Forensics, there is hardly any existing literature to compare the results of this study to. However, the results are in line with the conclusion drawn in *Section 2.4* regarding the expectation that this relatively new application of Process Mining would also yield results. Therefore, this study contributes new insights to both the Cyber Forensics and the Process Mining domains.

Nevertheless, the approach in this paper also has its limitations. First of all, as the topic is relatively new with very little closely related work, the methodology was designed based on two cases. Although the artifact is validated by field experts, application to other real-life cases is required to validate the methodology and approach based on more proof from practice. Another limitation concerns the Expert Panel. Due to the limited time available to execute the research, only six interviews took place. Therefore, some of the results are backed by not too many individual responses. Although all the participants are experts in their fields and thus are expected to express valid arguments, more validation from other experts is needed to gain more consensus on the topic and the claims made. Finally, none of the experts possessed extensive knowledge on all domains, which is understandable regarding the novelty of combining the topics. Still, with little knowledge of the other domains, it is harder to have a strong opinion on an approach that employs multiple.

The results and limitations of this master's thesis also call for future research. Given the originality of the research domain, there is extensive validation of the approach needed to validate these findings further. Starting with *PM^2 for Web Application Forensics*, the artifact could be applied to other cases to identify new issues and refine it further. More specifically, it can be applied to legal forensics cases or cases concerning more complex Web Application architectures to identify the extra required steps more precisely in each case. Another suggestion is to take in more opinions of experts with knowledge of the various fields to get a better grasp of the opportunities and limitations.

6. Conclusion

The objective of this master's thesis was to inquire into combining the academic fields of Process Mining and Cyber Forensics. More specifically, it is an inquiry into using Process Mining in Web Application Forensics. In *Chapter 1*, the relevance of the individual domains and this research as a whole is substantiated. In *Chapter 2*, the relevant domains were further introduced and related research was reviewed. The conclusion was drawn that, although Process Mining is applied successfully in various fields and offers opportunities for more, there exists a gap in current literature on Process Mining in Cyber Forensics. Further, Web Applications were identified as vulnerable targets for cyber attacks, which advocates for Web Application Forensics as the defined scope of the research. In *Chapter 3*, *PM^2 for Web Application Forensics* was developed by applying PM^2 to two Web Application Forensics cases. The chapter also included detailed information on the interviews with the Expert Panel, which were held to evaluate the developed artifact and the approach in general. The results extracted from the interviews were presented in *Chapter 4*. Finally, *Chapter 5* formulated an answer to the question as to how and to what extent Process Mining can be applied in Web Application Forensics. It further discussed the limitations of the research and possible suggestions for future research.

In conclusion, this master's thesis aims to contribute to the gap in scientific literature on Process Mining and Cyber Forensics, focusing on Web Application Forensics. The results suggest that the developed methodology, *PM^2 for Web Application Forensics*, can provide a useful guide to apply Process Mining in Web Application Forensics cases. However, there are several limitations to the scope and requirements that need to be in place to make it a successful endeavor. Future research is suggested to further validate and refine the methodology and assess the applicability of Process Mining in diverse Web Application Forensics cases.

References

1. Accorsi, R., & Stocker, T. (2012). On the exploitation of process mining for security audits: The conformance checking case. *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 1709–1716. https://doi.org/10.1145/2245276.2232051
2. Accorsi, R., Stocker, T., & Müller, G. (2013). On the exploitation of process mining for security audits: The process discovery case. *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 1462–1468. https://doi.org/10.1145/2480362.2480634
3. Acunetix. (n.d.). *Web Application Attack: What Is It and How to Defend Against It?* Acunetix. Retrieved October 9, 2023, from https://www.acunetix.com/websitesecurity/web-application-attack/
4. Alani, M. M. (2014). OSI Model. In M. M. Alani (Ed.), *Guide to OSI and TCP/IP Models* (pp. 5–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-05152-9_2
5. Alatawi, H., Alenazi, K., Alshehri, S., Alshamakhi, S., Mustafa, M., & Aljaedi, A. (2020). Mobile Forensics: A Review. *2020 International Conference on Computing and Information Technology (ICCIT-1441)*, 1–6. https://doi.org/10.1109/ICCIT-144147971.2020.9213739
6. Almulla, S., Iraqi, Y., & Jones, A. (2013). Cloud forensics: A research perspective. *2013 9th International Conference on Innovations in Information Technology (IIT)*, 66–71. https://doi.org/10.1109/Innovations.2013.6544395
7. Apache. (n.d.). *Log Files—Apache HTTP Server Version 2.4*. Retrieved November 2, 2023, from https://httpd.apache.org/docs/2.4/logs.html
8. AWS. (n.d.). *What is a Web App? - Web Application Explained - AWS*. Amazon Web Services, Inc. Retrieved October 9, 2023, from https://aws.amazon.com/what-is/web-application/
9. Ayers, R., Brothers, S., & Jansen, W. (2014). *Guidelines on Mobile Device Forensics* (NIST Special Publication (SP) 800-101 Rev. 1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-101r1
10. Babiker, M., Karaarslan, E., & Hoscan, Y. (2018). Web application attack detection and forensics: A survey. *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, 1–6. https://doi.org/10.1109/ISDFS.2018.8355378
11. Bahaweres, R. B., Trawally, J., Hermadi, I., & Suroso, A. I. (2021). Forensic Audit Using Process Mining to Detect Fraud. *Journal of Physics: Conference Series*, *1779*(1), 012013. https://doi.org/10.1088/1742-6596/1779/1/012013
12. Barker, W., Fisher, W., Scarfone, K., & Souppaya, M. (2022). *Ransomware Risk Management: A Cybersecurity Framework Profile* (NIST Internal or Interagency Report (NISTIR) 8374). National Institute of Standards and Technology.

https://doi.org/10.6028/NIST.IR.8374

13. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2009). KNIME - the Konstanz information miner: Version 2.0 and beyond. *ACM SIGKDD Explorations Newsletter*, *11*(1), 26–31. https://doi.org/10.1145/1656274.1656280
14. Brooks, C. (2023, March 5). *Cybersecurity Trends & Statistics For 2023; What You Need To Know*. Forbes. https://www.forbes.com/sites/chuckbrooks/2023/03/05/cybersecurity-trends--statistics-for-2023-more-treachery-and-risk-ahead-as-attack-surface-and-hacker-capabilities-grow/
15. Casino, F., Dasaklis, T. K., Spathoulas, G. P., Anagnostopoulos, M., Ghosal, A., Bořočz, I., Solanas, A., Conti, M., & Patsakis, C. (2022). Research Trends, Challenges, and Emerging Topics in Digital Forensics: A Review of Reviews. *IEEE Access*, *10*, 25464–25493. https://doi.org/10.1109/ACCESS.2022.3154059
16. Chernyshev, M., Zeadally, S., Baig, Z., & Woodward, A. (2017). Mobile Forensics: Advances, Challenges, and Research Opportunities. *IEEE Security & Privacy*, *15*(6), 42–51. https://doi.org/10.1109/MSP.2017.4251107
17. Chopade, R., & Pachghare, V. K. (2019). Ten years of critical review on database forensics research. *Digital Investigation*, *29*, 180–197. https://doi.org/10.1016/j.diin.2019.04.001
18. Cook, J. E., & Wolf, A. L. (1998). Event-based detection of concurrency. *ACM SIGSOFT Software Engineering Notes*, *23*(6), 35–45. https://doi.org/10.1145/291252.288214
19. Crowdstrike. (n.d.). *What Is a Web Server Log and How to Monitor? - CrowdStrike*. Crowdstrike.Com. Retrieved November 2, 2023, from https://www.crowdstrike.com/cybersecurity-101/observability/web-server-logs/
20. Dakic, D., Sladojevic, S., Lolic, T., & Stefanovic, D. (2019). Process Mining Possibilities and Challenges: A Case Study. *2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY)*, 000161–000166. https://doi.org/10.1109/SISY47553.2019.9111591
21. Day, J. D., & Zimmermann, H. (1983). The OSI reference model. *Proceedings of the IEEE*, *71*(12), 1334–1340. https://doi.org/10.1109/PROC.1983.12775
22. de Alvarenga, S. C., Barbon, S., Miani, R. S., Cukier, M., & Zarpelão, B. B. (2018). Process mining and hierarchical clustering to help intrusion alert visualization. *Computers & Security*, *73*, 474–491. https://doi.org/10.1016/j.cose.2017.11.021
23. De Weerdt, J., De Backer, M., Vanthienen, J., & Baesens, B. (2012). A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs. *Information Systems*, *37*(7), 654–676. https://doi.org/10.1016/j.is.2012.02.004
24. Deltchev, K. (2012, April 20). *Web Application Forensics: Taxonomy and Trends*. https://www.slideshare.net/test2v/web-application-forensics-taxonomy-and-trends

25. Fernando, V. (2021). Cyber Forensics Tools: A Review on Mechanism and Emerging Challenges. *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 1–7. https://doi.org/10.1109/NTMS49979.2021.9432641
26. Gangaram Panday, S., & Zwakenberg, L. (2023). *Ransomware in Control—A study report by NOREA*. NOREA.
27. Garcia, C. dos S., Meincheim, A., Faria Junior, E. R., Dallagassa, M. R., Sato, D. M. V., Carvalho, D. R., Santos, E. A. P., & Scalabrin, E. E. (2019). Process mining techniques and applications – A systematic mapping study. *Expert Systems with Applications*, *133*, 260–295. https://doi.org/10.1016/j.eswa.2019.05.003
28. Gunestas, M., & Bilgin, Z. (2016). Log Analysis Using Temporal Logic and Reconstruction Approach: Web Server Case. *Journal of Digital Forensics, Security and Law*, *11*(2). https://doi.org/10.15394/jdfsl.2016.1377
29. Günther, C., & Rozinat, A. (2012). *Disco: Discover Your Processes*. International Conference on Business Process Management. https://www.semanticscholar.org/paper/Disco%3A-Discover-Your-Processes-G%C3%BCnther-Rozinat/56b596c0762224d612a2cbf9b873423a12ff5b6b
30. Günther, C. W., & van der Aalst, W. M. P. (2007). Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics. In G. Alonso, P. Dadam, & M. Rosemann (Eds.), *Business Process Management* (pp. 328–343). Springer. https://doi.org/10.1007/978-3-540-75183-0_24
31. Gunther, C. W., & Verbeek, H. M. W. (2014). *XES - standard definition*. BPMcenter. org.
32. Hand, D. J. (2007). Principles of Data Mining. *Drug Safety*, *30*(7), 621–622. https://doi.org/10.2165/00002018-200730070-00010
33. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.
34. Huang, H.-C., Zhang, Z.-K., Cheng, H.-W., & Shieh, S. W. (2017). Web Application Security: Threats, Countermeasures, and Pitfalls. *Computer*, *50*(6), 81–85. https://doi.org/10.1109/MC.2017.183
35. Humayun, M., Niazi, M., Jhanjhi, N., Alshayeb, M., & Mahmood, S. (2020). Cyber Security Threats and Vulnerabilities: A Systematic Mapping Study. *Arabian Journal for Science and Engineering*, *45*(4), 3171–3189. https://doi.org/10.1007/s13369-019-04319-2
36. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. (2016). *IEEE Std 1849-2016*, 1–50. https://doi.org/10.1109/IEEESTD.2016.7740858
37. Interpol. (n.d.). *Digital forensics*. Retrieved October 2, 2023, from https://www.interpol.int/How-we-work/Innovation/Digital-forensics
38. Isaac, M., & Frenkel, S. (2018, September 28). Facebook Security Breach Exposes Accounts of 50 Million Users. *The New York Times*.

https://www.nytimes.com/2018/09/28/technology/facebook-hack-data-breach.html

39. Jans, M., Alles, M., & Vasarhelyi, M. (2013). The case for process mining in auditing: Sources of value added and areas of application. *International Journal of Accounting Information Systems*, *14*(1), 1–20. https://doi.org/10.1016/j.accinf.2012.06.015
40. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, *2*(1), 20. https://doi.org/10.1186/s42400-019-0038-7
41. Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030
42. Lagraa, S., & State, R. (2020). Process mining-based approach for investigating malicious login events. *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, 1–5. https://doi.org/10.1109/NOMS47738.2020.9110301
43. Lazzez, A., & Slimani, T. (2015). Forensics Investigation of Web Application Security Attacks. *International Journal of Computer Network and Information Security*, *7*(3), 10–17. https://doi.org/10.5815/ijcnis.2015.03.02
44. Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16–24. https://doi.org/10.1016/j.jnca.2012.09.004
45. Macak, M., Daubner, L., Fani Sani, M., & Buhnova, B. (2022). Process mining usage in cybersecurity and software reliability analysis: A systematic literature review. *Array*, *13*, 100120. https://doi.org/10.1016/j.array.2021.100120
46. Mahmood, T., & Afzal, U. (2013). Security Analytics: Big Data Analytics for cybersecurity: A review of trends, techniques and tools. *2013 2nd National Conference on Information Assurance (NCIA)*, 129–134. https://doi.org/10.1109/NCIA.2013.6725337
47. Manral, B., Somani, G., Choo, K.-K. R., Conti, M., & Gaur, M. S. (2019). A Systematic Survey on Cloud Forensics Challenges, Solutions, and Future Directions. *ACM Computing Surveys*, *52*(6), 124:1-124:38. https://doi.org/10.1145/3361216
48. Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M. P., & Bakker, P. J. M. (2009). Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital. In A. Fred, J. Filipe, & H. Gamboa (Eds.), *Biomedical Engineering Systems and Technologies* (pp. 425–438). Springer. https://doi.org/10.1007/978-3-540-92219-3_32
49. Microsoft. (n.d.). *IIS Log File Formats*. Retrieved November 2, 2023, from https://learn.microsoft.com/en-us/previous-versions/iis/6.0-sdk/ms525807(v=vs.90)
50. Mishra, V. P., Shukla, B., & Bansal, A. (2019). Analysis of alarms to prevent the organizations network in real-time using process mining approach. *Cluster Computing*, *22*(3), 7023–7030. https://doi.org/10.1007/s10586-018-2064-8
51. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
52. Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O. A., Sepúlveda, M., Helm,

E., Galvez-Yanjari, V., Rojas, E., Martinez-Millana, A., Aloini, D., Amantea, I. A., Andrews, R., Arias, M., Beerepoot, I., Benevento, E., Burattin, A., Capurro, D., Carmona, J., Comuzzi, M., … Zerbato, F. (2022). Process mining for healthcare: Characteristics and challenges. *Journal of Biomedical Informatics*, *127*, 103994. https://doi.org/10.1016/j.jbi.2022.103994

53. Myers, D., Radke, K., Suriadi, S., & Foo, E. (2017). Process Discovery for Industrial Control System Cyber Attack Detection. In S. De Capitani di Vimercati & F. Martinelli (Eds.), *ICT Systems Security and Privacy Protection* (pp. 61–75). Springer International Publishing. https://doi.org/10.1007/978-3-319-58469-0_5
54. Myers, D., Suriadi, S., Radke, K., & Foo, E. (2018). Anomaly detection for industrial control systems using process mining. *Computers & Security*, *78*, 103–125. https://doi.org/10.1016/j.cose.2018.06.002
55. Nazar, N., Shukla, V. K., Kaur, G., & Pandey, N. (2021). Integrating Web Server Log Forensics through Deep Learning. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–6. https://doi.org/10.1109/ICRITO51393.2021.9596324
56. Olivier, M. S. (2009). On metadata context in Database Forensics. *Digital Investigation*, *5*(3), 115–123. https://doi.org/10.1016/j.diin.2008.10.001
57. OWASP. (2021). *OWASP Top Ten | OWASP Foundation*. https://owasp.org/www-project-top-ten/
58. Poggi, N., Muthusamy, V., Carrera, D., & Khalaf, R. (2013). Business Process Mining from E-Commerce Web Logs. In F. Daniel, J. Wang, & B. Weber (Eds.), *Business Process Management* (pp. 65–80). Springer. https://doi.org/10.1007/978-3-642-40176-3_7
59. Pollitt, M. (2010). A History of Digital Forensics. In K.-P. Chow & S. Shenoi (Eds.), *Advances in Digital Forensics VI* (pp. 3–15). Springer. https://doi.org/10.1007/978-3-642-15506-2_1
60. Reinkemeyer, L. (Ed.). (2020). *Process Mining in Action: Principles, Use Cases and Outlook*. Springer International Publishing. https://doi.org/10.1007/978-3-030-40172-6
61. Ruan, K., Carthy, J., Kechadi, T., & Crosbie, M. (2011). Cloud Forensics. In G. Peterson & S. Shenoi (Eds.), *Advances in Digital Forensics VII* (pp. 35–46). Springer. https://doi.org/10.1007/978-3-642-24212-0_3
62. Sarirah Husin, H., & Ismail, S. (2021). Process mining approach to analyze user navigation behavior of a news website. *Proceedings of the 4th International Conference on Information Science and Systems*, 7–12. https://doi.org/10.1145/3459955.3460593
63. Schellevis, J. (2023, December 18). *KLM lekte data klanten: Privégegevens eenvoudig te verzamelen*. https://translations.lab.nos.nl/nl/articles/2501984
64. Shier, J. (2020). *CYBERTHREATS: A 20-YEAR RETROSPECTIVE*.
65. Sikos, L. F. (2020). Packet analysis for network forensics: A comprehensive survey.

Forensic Science International: Digital Investigation, *32*, 200892. https://doi.org/10.1016/j.fsidi.2019.200892

66. Silalahi, S., Yuhana, U. L., Ahmad, T., & Studiawan, H. (2022). A Survey on Process Mining for Security. *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 1–6. https://doi.org/10.1109/iSemantic55962.2022.9920473
67. Sindhu, K. K., & Meshram, B. B. (2012). *Digital Forensics and Cyber Crime Datamining*. *2012*. https://doi.org/10.4236/jis.2012.33024
68. Statista. (2023). *Topic: Ransomware*. Statista. https://www.statista.com/topics/4136/ransomware/
69. Terragni, A., & Hassani, M. (2018). Analyzing Customer Journey with Process Mining: From Discovery to Recommendations. *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, 224–229. https://doi.org/10.1109/FiCloud.2018.00040
70. van Cruchten, R. M. E. R., & Weigand, H. H. (2018). Process mining in logistics: The need for rule-based data abstraction. *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, 1–9. https://doi.org/10.1109/RCIS.2018.8406653
71. Van den beemt, A., Buys, J., & Aalst, W. (2018). Analysing Structured Learning Behaviour in Massive Open Online Courses (MOOCs): An Approach Based on Process Mining and Clustering. *The International Review of Research in Open and Distributed Learning*, *19*. https://doi.org/10.19173/irrodl.v19i5.3748
72. van der Aalst, & de Medeiros. (2005). Process Mining and Security: Detecting Anomalous Process Executions and Checking Process Conformance. *Electronic Notes in Theoretical Computer Science*, *121*, 3–21. https://doi.org/10.1016/j.entcs.2004.10.013
73. Van Der Aalst, W. (2016). *Process Mining*. Springer. https://doi.org/10.1007/978-3-662-49851-4
74. van der Aalst, W. M. P., Bolt, A., & van Zelst, S. J. (2017). *RapidProM: Mine Your Processes and Not Just Your Data* (arXiv:1703.03740). arXiv. https://doi.org/10.48550/arXiv.1703.03740
75. van der Aalst, W. M. P., & Weijters, A. J. M. M. (2004). Process mining: A research agenda. *Computers in Industry*, *53*(3), 231–244. https://doi.org/10.1016/j.compind.2003.10.001
76. van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47
77. van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & van der Aalst, W. M. P. (2005). The ProM Framework: A New Era in Process Mining Tool

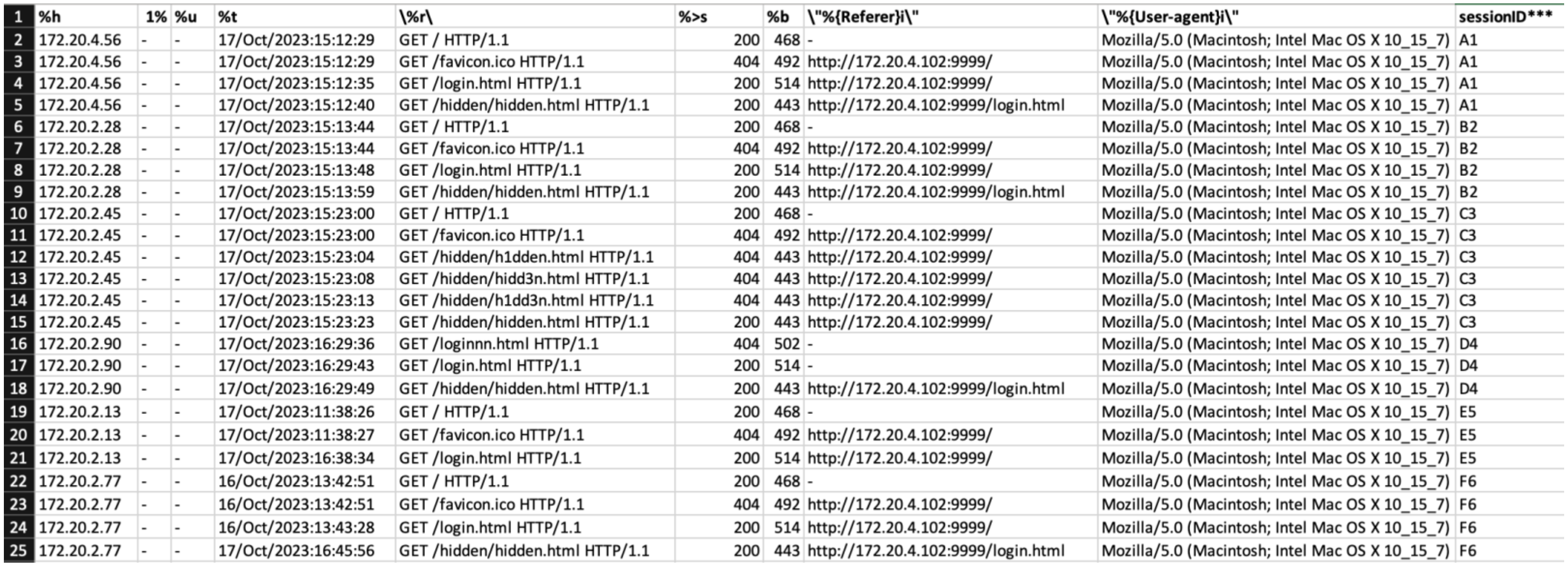
Support. In G. Ciardo & P. Darondeau (Eds.), *Applications and Theory of Petri Nets 2005* (pp. 444–454). Springer. https://doi.org/10.1007/11494744_25

78. van Eck, M. L., Lu, X., Leemans, S. J. J., & van der Aalst, W. M. P. (2015). PM2: A Process Mining Project Methodology. In J. Zdravkovic, M. Kirikova, & P. Johannesson (Eds.), *Advanced Information Systems Engineering* (pp. 297–313). Springer International Publishing. https://doi.org/10.1007/978-3-319-19069-3_19
79. Varnagar, C. R., Madhak, N. N., Kodinariya, T. M., & Rathod, J. N. (2013). Web usage mining: A review on process, methods and techniques. *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, 40–46. https://doi.org/10.1109/ICICES.2013.6508399
80. Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36
81. von Solms, R., & van Niekerk, J. (2013). From information security to cyber security. *Computers & Security*, *38*, 97–102. https://doi.org/10.1016/j.cose.2013.04.004
82. Ware, W. H. (1967). *Willis H. Ware, RAND Corporation, P-3544, Security and Privacy in Computer Systems, April 1967. Unclassified. | National Security Archive*. https://nsarchive.gwu.edu/document/21676-document-01-willis-h-ware-rand-corporation-p
83. Warner, M. (2012). Cybersecurity: A Pre-history. *Intelligence and National Security*, *27*(5), 781–799. https://doi.org/10.1080/02684527.2012.708530
84. Weaver, N., Paxson, V., Staniford, S., & Cunningham, R. (2003). A taxonomy of computer worms. *Proceedings of the 2003 ACM Workshop on Rapid Malcode*, 11–18. https://doi.org/10.1145/948187.948190
85. Weijters, A., Aalst, W., & Medeiros, A. (2006). Process Mining with the Heuristics Miner-algorithm. In *Cirp Annals-manufacturing Technology—CIRP ANN-MANUF TECHNOL* (Vol. 166).
86. Werner, M., Wiese, M., & Maas, A. (2021). Embedding process mining into financial statement audits. *International Journal of Accounting Information Systems*, *41*, 100514. https://doi.org/10.1016/j.accinf.2021.100514
87. Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer.

Appendix A – lab setting Web Application

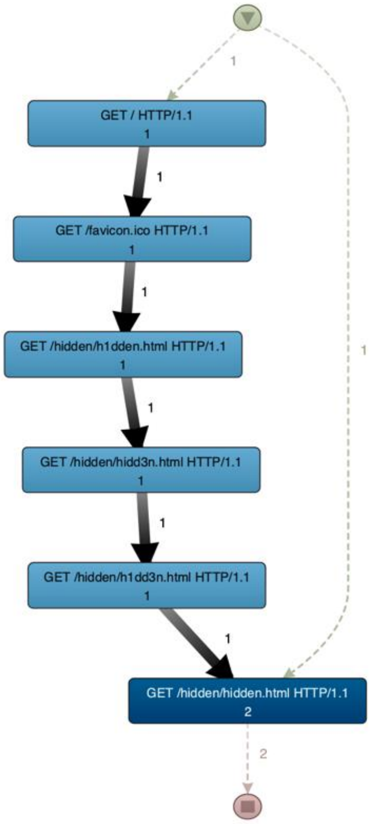
A.1 – Access log

****manually added column*

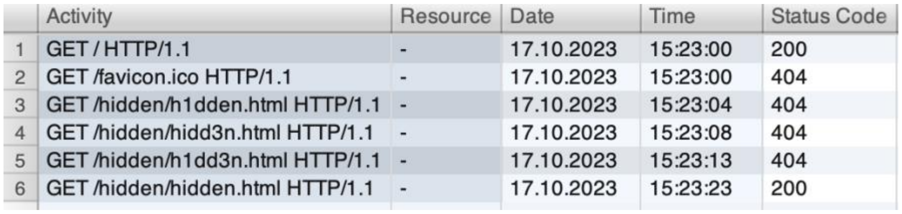


1	%h	1%	%u	%t	\%r\	%>s	%b	\"%{Referer}i\"	\"%{User-agent}i\"	sessionID***
2	172.20.4.56	-	-	17/Oct/2023:15:12:29	GET / HTTP/1.1	200	468	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	A1
3	172.20.4.56	-	-	17/Oct/2023:15:12:29	GET /favicon.ico HTTP/1.1	404	492	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	A1
4	172.20.4.56	-	-	17/Oct/2023:15:12:35	GET /login.html HTTP/1.1	200	514	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	A1
5	172.20.4.56	-	-	17/Oct/2023:15:12:40	GET /hidden/hidden.html HTTP/1.1	200	443	http://172.20.4.102:9999/login.html	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	A1
6	172.20.2.28	-	-	17/Oct/2023:15:13:44	GET / HTTP/1.1	200	468	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	B2
7	172.20.2.28	-	-	17/Oct/2023:15:13:44	GET /favicon.ico HTTP/1.1	404	492	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	B2
8	172.20.2.28	-	-	17/Oct/2023:15:13:48	GET /login.html HTTP/1.1	200	514	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	B2
9	172.20.2.28	-	-	17/Oct/2023:15:13:59	GET /hidden/hidden.html HTTP/1.1	200	443	http://172.20.4.102:9999/login.html	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	B2
10	172.20.2.45	-	-	17/Oct/2023:15:23:00	GET / HTTP/1.1	200	468	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	C3
11	172.20.2.45	-	-	17/Oct/2023:15:23:00	GET /favicon.ico HTTP/1.1	404	492	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	C3
12	172.20.2.45	-	-	17/Oct/2023:15:23:04	GET /hidden/h1dden.html HTTP/1.1	404	443	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	C3
13	172.20.2.45	-	-	17/Oct/2023:15:23:08	GET /hidden/hidd3n.html HTTP/1.1	404	443	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	C3
14	172.20.2.45	-	-	17/Oct/2023:15:23:13	GET /hidden/h1dd3n.html HTTP/1.1	404	443	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	C3
15	172.20.2.45	-	-	17/Oct/2023:15:23:23	GET /hidden/hidden.html HTTP/1.1	200	443	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	C3
16	172.20.2.90	-	-	17/Oct/2023:16:29:36	GET /loginnn.html HTTP/1.1	404	502	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	D4
17	172.20.2.90	-	-	17/Oct/2023:16:29:43	GET /login.html HTTP/1.1	200	514	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	D4
18	172.20.2.90	-	-	17/Oct/2023:16:29:49	GET /hidden/hidden.html HTTP/1.1	200	443	http://172.20.4.102:9999/login.html	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	D4
19	172.20.2.13	-	-	17/Oct/2023:11:38:26	GET / HTTP/1.1	200	468	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	E5
20	172.20.2.13	-	-	17/Oct/2023:11:38:27	GET /favicon.ico HTTP/1.1	404	492	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	E5
21	172.20.2.13	-	-	17/Oct/2023:16:38:34	GET /login.html HTTP/1.1	200	514	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	E5
22	172.20.2.77	-	-	16/Oct/2023:13:42:51	GET / HTTP/1.1	200	468	-	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	F6
23	172.20.2.77	-	-	16/Oct/2023:13:42:51	GET /favicon.ico HTTP/1.1	404	492	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	F6
24	172.20.2.77	-	-	16/Oct/2023:13:43:28	GET /login.html HTTP/1.1	200	514	http://172.20.4.102:9999/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	F6
25	172.20.2.77	-	-	17/Oct/2023:16:45:56	GET /hidden/hidden.html HTTP/1.1	200	443	http://172.20.4.102:9999/login.html	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7)	F6

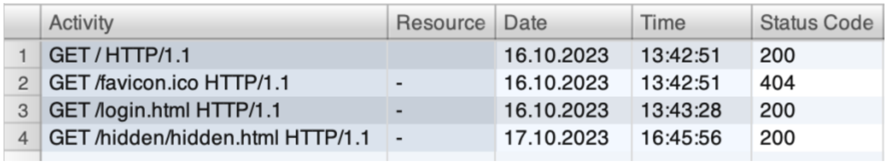
A.2 – Resulting Process Map


1
GET / HTTP/1.1
1
1
GET /favicon.ico HTTP/1.1
1
1
GET /hidden/h1dden.html HTTP/1.1
1
1
1
GET /hidden/hidd3n.html HTTP/1.1
1
1
GET /hidden/h1dd3n.html HTTP/1.1
1
1
GET /hidden/hidden.html HTTP/1.1
2
2

A.3 – Extracted Cases from Analysis



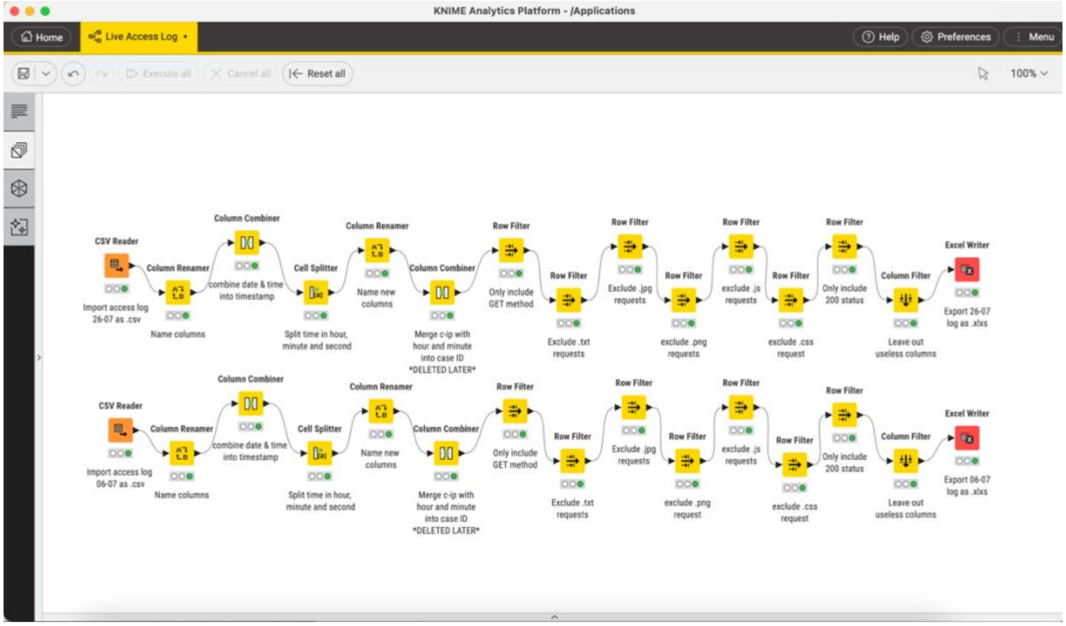
	Activity	Resource	Date	Time	Status Code
1	GET / HTTP/1.1	-	17.10.2023	15:23:00	200
2	GET /favicon.ico HTTP/1.1	-	17.10.2023	15:23:00	404
3	GET /hidden/h1dden.html HTTP/1.1	-	17.10.2023	15:23:04	404
4	GET /hidden/hidd3n.html HTTP/1.1	-	17.10.2023	15:23:08	404
5	GET /hidden/h1dd3n.html HTTP/1.1	-	17.10.2023	15:23:13	404
6	GET /hidden/hidden.html HTTP/1.1	-	17.10.2023	15:23:23	200



	Activity	Resource	Date	Time	Status Code
1	GET / HTTP/1.1		16.10.2023	13:42:51	200
2	GET /favicon.ico HTTP/1.1	-	16.10.2023	13:42:51	404
3	GET /login.html HTTP/1.1	-	16.10.2023	13:43:28	200
4	GET /hidden/hidden.html HTTP/1.1	-	17.10.2023	16:45:56	200

Appendix B – Live Web Application

B.1 Knime Workflows



B.2 Creating Sessions using Excel

To create proper sessions in the datasets, a formula in (the Dutch version of) Excel was inserted as a new column for each row. First, the dataset was ordered based on 1: IP-address, and 2: timestamp. The first event in the log was assigned session '1'. Thereafter, the following formula was inserted for each of the following events:

$$= \mathbf{ALS}(\mathbf{EN}(((\mathbf{A} - \mathbf{B}) < \mathbf{TIJD}(\mathbf{0}; \mathbf{10}; \mathbf{0})); (\mathbf{C} = \mathbf{D})); \mathbf{E}; (\mathbf{E} + \mathbf{1}))$$

A = timestamp of event
B = timestamp of previous event
C = IP-address of event
D = IP-address of previous event
E = session identifier of previous event

The formula, or rule, can be interpreted as follows:

"If the event follows the previous event by less than 10 minutes, and the IP-address of the event corresponds with the IP-address from the previous event, assign the same session identifier to the event as the previous event. If one of the two conditions is not met, create a new session by adding 1 to the previous session identifier."*

**Based on knowledge provided by the Business Expert about the Web Application*

B.3 Extracted Cases from Analysis

As the used data is sensitive to the company (client of Joanknecht), the data is reconstructed by abstracting the sensitive information.

First instance of malicious access to File B

Nr.	*Case ID*	*Activity (stem + query)*	*Date*	*Time*
1	IP-address A + Session B	homepage	06-07-2023	08:15:55
2	IP-address A + Session B	login page	06-07-2023	08:17:03
3	IP-address A + Session B	login page	06-07-2023	08:17:34
4	IP-address A + Session B	login page + ***File B***	06-07-2023	08:17:54
…	IP-address A + Session B	…	…	…

Second instance of malicious access to File B

Nr.	*Case ID*	*Activity (stem + query)*	*Date*	*Time*
…	IP-address P + Session Q	…	…	…
1851	IP-address P + Session Q	homepage	06-07-2023	09:15:47
1852	IP-address P + Session Q	login page + *suspicious query*	06-07-2023	09:15:47
1853	IP-address P + Session Q	login page + ../../../../../../../../../../***File B***	06-07-2023	09:15:47
1854	IP-address P + Session Q	login page + *suspicious query*	06-07-2023	09:15:47
…	IP-address P + Session Q	…	…	…

Malicious access to Document A

Nr.	*Case ID*	*Activity (stem + query)*	*Date*	*Time*
…	IP-address X + Session Y	…	…	…
26	IP-address X + Session Y	login page + *action = login*	26-07-2023	06:40:01
27	IP-address X + Session Y	portal page + *browse page 1*	26-07-2023	06:40:25
28	IP-address X + Session Y	portal page + *browse page 2*	26-07-2023	06:40:41
29	IP-address X + Session Y	portal page + ***Document A***	26-07-2023	06:41:14
30	IP-address X + Session Y	portal page + *suspicious query*	26-07-2023	06:41:27
31	IP-address X + Session Y	portal page + *suspicious query*	26-07-2023	06:41:38
32	IP-address X + Session Y	portal page + *suspicious query*	26-07-2023	06:41:44
33	IP-address X + Session Y	portal page + *suspicious query*	26-07-2023	06:41:50
34	IP-address X + Session Y	portal page + *suspicious query*	26-07-2023	06:41:59
35	IP-address X + Session Y	portal page + *../../***Document A***	26-07-2023	06:42:04

Appendix C – Expert Panel Interviews

C.1 – Expert Panel Participants

	Name	Expertise	Employment	Experience*
[A]	Drs. Lucas Vousten RA RE	Process Mining & Cyber Security	Partner IT audit | Joanknecht	>25
[B]	Lana Ebergen MSc RE	Process Mining	IT-auditor / Manager IT-audit | Joanknecht	4,5
[C]	Drs. Frank Driessen RA	Forensic Investigations	Partner Forensics & Recovery & Forensic Accountant | Joanknecht	>25
[D]	Dr.Ir. Anne Rozinat	Process Mining	Co-founder | Fluxicon Process Mining Software	>20
[E]	Jeffrey Jansen	Cyber Security	Co-owner | Access42 Cybersecurity	12
[F]	Bram van Altena**	Cyber Security	Director CISO | KLM Royal Dutch Airlines	24
[G]	Tako Huisman**	Cyber Security	Deputy CISO | KLM Royal Dutch Airlines	23
[H]	Gertjan Kloek**	Cyber Security & Web Applications	IT Security Specialist | KLM Royal Dutch Airlines	>25
[I]	Jamie Alexander Dekker**	Cyber Security & Web Applications	IT Specialist SOC Analyst | KLM Royal Dutch Airlines	7

* *In years*

** *Interview took place with all four together*

C.2 – Coded Interviews

Abbreviations used in the coded interviews:

PM = Process Mining

WA(F) = Web Application (Forensics)

RQ = Research Question

CS = Cyber Security

Expert [A]

Deductive coding (predefined metrics)

Functionality *(translated & created a sound sentence)*

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
1	00:38-02:04	Are there steps that do not add value or could be described more clear to add more value?	It depends on the reader. Someone without knowledge of PM would not understand it. Further, there should be knowledge up front before being able to understand what's possible with WA logs. But there are no missing steps. **explains importance of defining RQ and relation to other phases**	-

Completeness

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
2	02:55-06:39	Are there steps missing that would add value?	See answer Nr.12	-

Consistency

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
3	06:56-07:16	Would different users interpret the steps in the methodology the same?	In the same cases, different users would interpret the same.	-
4	7:58-9:16	Would the results be comparable in other cases in WAF?	It depends on the question posed, because it can be used for different purposes (file extraction or SQL injection eg). But this model is on meta level, so as a thinking model to base the analysis on. It is useful in finding flows/processes in the data, which is something else than	-

Accuracy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
5	09:40-10:45	What are the most important factors that determine the accuracy of the results?	It all starts with the RQ. [explains example with buses outside]	*10:45: It depends on the question: the more specific the question, the more specific an answer can be.*

6	Q= 9:40
	A= 11:01-13:49
			Knowledge about the web application is also important. An expert should know what is available to log and what the fields mean. Depending on the RQ, you should know the data(and server) you are dealing with. Further, knowledge of how to apply PM is of course important to create meaningful results	-

Reliability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
7	14:17-18.32 (answers are mixed)	What factors are the most important to ensure the reliability of the data?	Depends on the goal. If it is to be included in a legal case, data integrity is paramount. In that case, a forensic sound extraction is	-
8			Also depends on the level of access gained by the attacker. If a user gained rights to access log directories and modify logs, data already is unreliable. Or the user could delete his tracks. So it depends on the degree to which you can say that the data is not manipulated. This depends on IT General Controls to prevent privileged access.	-

Usability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
9	18:52-21:28	What knowledge and skills are equired of the different team members to make the methodology usable?	*same as Nr.6* Knowledge on what kind of behavior is authorized and what not. From there extract what activity to search for.	-

Performance

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
10	21:33-25:00	How can PM improve forensic investigations?	It differs from IDS/IPS, which is live monitoring and reacting. The goal is to find out in hindsight how something happened. By learning this, vulnerabilities can be solved. It can also be used to check whether something malicious happened, or if other malicious activity took place around the case. IP-addresses often are spoofed, proxied etc which is why it doesn't say much. The path is way more interesting.	*23:27-24:08: Detecting malicious access is not the big win, … , that is a static analysis. … The dynamic part is that you can follow a path and understand how that was possible, that is what the benefit of Process Mining is.*
11	25:44-27:36	What are limitations of the approach?	When asking the RQ, there should be a subquestion about which tool is fitting. For some questions, static analyses are better than using PM. So not for every RQ in WAF, PM is the answer.	*26:07: Different concepts for different purposes*

Inductive coding (extracted from responses)

Learning

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
12	03:25-06:39	**not related to a specific question, remark made during discussion**	There should be a learning curve in the model related to the available data for future cases. If you find certain data is unavailable, you need to make sure that for the next time, you make sure it is available. For example including session cookies in the logging.	*06:39: Every research question will be different, requiring different data. If you include more data in the logs, there is a bigger chance you will have the right data for the next case.*

Expert [B]

Deductive coding (predefined metrics)

Functionality *(translated & created a sound sentence)*

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
1	00:38-02:04	Are there steps that do not add value or could be described more clear to add more value?	It depends on the reader. Someone without knowledge of PM would not understand it. Further, there should be knowledge up front before being able to understand what's possible with WA logs. But there are no missing steps. **explains importance of defining RQ and relation to other phases**	-

Completeness

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
2	02:55-06:39	Are there steps missing that would add value?	See answer Nr.12	-

Consistency

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
3	06:56-07:16	Would different users interpret the steps in the methodology the same?	In the same cases, different users would interpret the same.	-
4	7:58-9:16	Would the results be comparable in other cases in WAF?	It depends on the question posed, because it can be used for different purposes (file extraction or SQL injection eg). But this model is on meta level, so as a thinking model to base the analysis on. It is useful in finding flows/processes in the data, which is something else than static analyses.	-

Accuracy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
5	09:40-10:45	What are the most important factors that determine the accuracy of the results?	It all starts with the RQ. [explains example with buses outside]	*10:45: It depends on the question: the more specific the question, the more specific an answer can be.*

6	Q= 9:40
	A= 11:01-13:49
			Knowledge about the web application is also important. An expert should know what is available to log and what the fields mean. Depending on the RQ, you should know the data(and server) you are dealing with. Further, knowledge of how to apply PM is of course important to create meaningful results	-

Reliability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
7	14:17-18.32 (answers are mixed)	What factors are the most important to ensure the reliability of the data?	Depends on the goal. If it is to be included in a legal case, data integrity is paramount. In that case, a forensic sound extraction is required.	-
8			Also depends on the level of access gained by the attacker. If a user gained rights to access log directories and modify logs, data already is unreliable. Or the user could delete his tracks. So it depends on the degree to which you can say that the data is not manipulated. This depends on IT General Controls to prevent privileged access.	-

Usability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
9	18:52-21:28	What knowledge and skills are equired of the different team members to make the methodology usable?	*same as Nr.6* Knowledge on what kind of behavior is authorized and what not. From there extract what activity to search for.	-

Performance

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
10	21:33-25:00	How can PM improve forensic investigations?	It differs from IDS/IPS, which is live monitoring and reacting. The goal is to find out in hindsight how	*23:27-24:08: Detecting malicious access is not the big win, ... , that is a static analysis .... The dynamic part*
11	25:44-27:36	What are limitations of the approach?	When asking the RQ, there should be a subquestion about which tool is fitting. For some questions, static analyses are better than using PM. So not for every RQ in WAF, PM is the answer.	*26:07: Different concepts for differen*

Inductive coding (extracted from responses)

Learning

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
12	03:25-06:39	**not related to a specific question, remark made during discussion**	There should be a learning curve in the model related to the available data for future cases. If you find certain data is unavailable, you need to make sure that for the next	*06:39: Every research question will be different, requiring different data. If you include more data in the logs, there is a bigger chance you*

Expert [C]

Deductive coding (predefined metrics)

Functionality *(translated & created a sound sentence)*

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
1	24:36-25:01	Are there steps that do not add value or could be described more clear to add more value?	**methodology is discussed with questions about the different steps** After the discussion, the conclusion is that the methodology	

Completeness

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable

2	Q= 10:45
	A= 12:27-13:03
		Are there steps missing that would add value?	At the extraction step, it is important to keep an original copy of the extracted log file. A second copy should be used for the investigation. This way, one can always go back in steps to see what has been done.	

Consistency

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
3	34:06-38:07	Would the described steps in the methodology lead to a forensic sound report? **different question than other experts**	As a method it is logical and provides a useful basis. For specific cases, it depends on the RQ and the results: only process improvement or can juridical steps be taken? If juridical procedures are be a goal, the last phase is not only *'process improvement'*. In that case, some legal steps should also be incorporated.	

Accuracy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
4	38:11-39:34	What are the most important factors that determine the accuracy of the results?	By asking the right questions, but also translating them to the technical aspect of the WA is very important. A forensic expert can ask the right question, but the business expert is needed for technical knowledge.	

Reliability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable

5	Q= 10:45
	A= 12:27-13:03
		What factors are the most important to ensure the reliability of the data?	Already answered at Nr.2.	

Usability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
6	42:50-43:15	What knowledge and skills are equired of the different team members to make the methodology usable? **question asked for confirmation, already discussed earlier**	Confirmed that the knowledge about the application of the Business Expert is crucial to translate the RQ in how to execute the investigation.	

Performance

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
7	43:26-47:07 & 49:43	How can PM improve forensic investigations?	Joanknecht recently acquired a forensics tool for organizing and searching through forensic data. PM could add value in terms of helping to understand what happened in an incident or check whether there were more instances.	*49:43: I think it is an interesting study what you are doing, which could potentially be useful for our company.*
8	47:27-49:27	What are limitations of the approach?	PM is indeed not suitable for every investigation. Intella is useful to present an overview, so that is always useful. PM will probably not be used as a first tool in every forensic investigation, it depends on the RQ.	*48:01: I do not think that Process Mining will be the first used tool in every investigation, it really depends [on the research question].*

Inductive coding (new metrics based on responses)

Privacy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
9	31:28-33:04	**not related to a specific question, remark made during discussion**	Question is posed whether privacy plays a role in using the logs for an analysis. Reason is that through data preparation and analysis, people and their activities can be identified. Even if you do not know for sure whether these people committed a crime	

Learning

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
10	39:50-42:25	**not related to a specific question, remark made during discussion**	Remark is made about how to deal with incomplete data in a forensic investigation. If the conclusion is that no useful insight can be extracted from the data, you should make sure that this does not happen next time. That is, enabling more logging to ensure that the next time a user can be identified.	

Expert [D]

Deductive coding (predefined metrics)

Functionality *(translated & created a sound sentence)*

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
1	02:09-05:41	Are there steps that do not add value or could be described more clear to add more value? **question not specifically asked, but answer acquired during discussion**	Part of phase 3 could also be carried out in the PM tool. The advantage is that this makes the analysis really interactive, which enables to look at the data from different views. Advice: include as much relevant data in the tool as possible to enable flexible analyses. Only exclude clearly unnecessary data.	-

Completeness

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
2	08:10-11.04	Are there steps missing that would add value? **question not specifically asked, but answer acquired during discussion**	During PM analysis, it is often that you have to look from different perspectives on the dataset, even if you know what you are looking for. Practice learns that refining the RQ is often required to be able to find the right results.	-

Consistency

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
-	-	Would different users interpret the steps in the methodology the same?	*no answer: step by step execution of example would be needed*	-
4	**later clarified per email**	Would the results be comparable in other cases in WAF	Probably, but comparison of cases would be a step towards generalization	-

Accuracy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
5	**later clarified per email**	What are the most important factors that determine the accuracy of the results?	Understanding and validating the data is important. Also, formulating correct RQ with respect to the data is important.	-

Reliability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
7	**later clarified per email**	What factors are the most important to ensure the reliability of the data?	Whether the data shows what really happened (or whether some data was deleted or modified) lies outside of the process mining tool but is an important part of the data validation.	-

Usability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
8	**later clarified per email**	What knowledge and skills are equired of the different team members to make the methodology usable?	When you run a PM project, you always need different roles and skills present. In this application the 'Business Analyst', which has the use case-specific expertise would need to be someone with knowledge in the cyber / forensics field.	-

Performance

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
9	17:13-19:31	How can PM improve forensic investigations? **question not specifically asked, but answer acquired during discussion**	PM can add value in two ways. It can be the main analysis tool to extract findings from the data to understand if/how/how often something happened. It can also be used to visualize findings in hindsight for communication.	-
10	13:31-17:07	What are limitations of the approach? **question not specifically asked, but answer acquired during discussion**	Before using the methodology and thus PM in a cyber forensics case, think about what cyber security professionals do in these situations. Ask yourself if applying PM even adds value in the case.	-

Inductive coding (new subjects)

Privacy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
11	20:09	**not related to a specific question, remark made during discussion**	Ethical aspect can be important. What data can you use?	*20:09: One of the aspects [that we discuss in our cases] is ethical considerations, so thinking about which data are we allowed to use?*

Expert [E]

Deductive coding (predefined metrics)

Functionality

(translated & created a sound sentence)

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
1	Q: 02:46 A: 11:19-12:08	Are there steps that do not add value or could be described more clear to add more value? **question is posed earlier on, full methodology is discussed inbetween**	Session cookies are related to traditional web applications. Modern ones also use JSON web tokens for example. This step could therefore be described more generally.	-
2	Q: 02:46 A: 12:35 & 13:39		Generally, the methodology looks fine.	-

Completeness

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
3	10:26-10:59	Are there steps missing that would add value?	If you want to use IP addresses (addition Tim: e.g. in legal prosecution. Jeffrey agreed), normalization of the data is required. For example, a reverse proxy or loadbalancer can be put in front of a chain, which might provide another IP-address than the one you are	-

Consistency

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
4	14:55-15:24	Would different users interpret the steps in the methodology the same?	**no answer: argues that he would need to do some research himself before answering**	-
5	15:26-17:04	Would the results be comparable in other cases in WAF	Mainly is dependent on the type of WA. Traditional ones have 1 application that includes everything (including logs). Modern ones (SOA & Kubernetes clusters) can have countless Web Applications and APIs that interact with each other. This means various different logs from different servers. That makes correlating the logs very	-

Accuracy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
6	17:10-18:23	What are the most important factors that determine the accuracy of the results?	Composing the right team with people who can make the translation to form the business rules. Further, asking the right RQ is very important.	-
7	08:47-10:17	**not related to a specific question, remark made during discussion**	To be able to understand attacks properly, only having knowledge of theWA's functionality might not be sufficient. It would be wise to also have knowledge of commong attack patterns (signatures).	*09:46: There are SQL injection patterns that might adhere to a rule, which can lead to the conclusion that there is no deviation [with a Process Mining analysis]. But if you look at signatures of e.g., a timing attack, you have to look at the response times instead of URLs.*
8	12:35-13:30	**not related to a specific question, remark made during discussion**	An often used method for phase 2 is threat modelling. By thinking of evil user stories (what would an attacker want to misuse?), one can also derive various business rules.	*12:54: In the agile methodology [for building software] you have evil user stories … you could write down what attackers would want to misuse looking at the confidentiality, integrity and availability of the data … and subsequently translate these to business rules.*

Reliability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
9	18:30-19:41	What factors are the most important to ensure the reliability of the data?	In forensic investigations, the chain of custody (documented process of data handling) is important.	-

Usability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
10	19:44-23:15	What knowledge and skills are equired of the different team members to make the methodology usable?	The business expert should have good communication skills and have experience and knowledge with the WA.	-

Performance

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
11	25:50-27:28	How can PM improve forensic investigations?	Current tools we use (e.g. Graylog) are useful for identifying anomalies. If we want to look at the anomalies in more detail, by defining the businesss rules with the business expert, this methodology can help. So it could be used as an addition to the tools that can identify anomalies.	-
12	29:52-31:13	What are limitations of the approach?	It depends on the type of attack and where it is executed whether it can be understood from the WA logs. For example, blind SQL attacks are attacks on the WA but are executed on the database. Therefore you cannot see them in the WA logging, but only in the database logging. [refers to complexity of correlating logs]	-

Inductive coding (new subjects)

Project team

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
13	03:00-05:45	**not related to a specific question, remark made during discussion**	Practice learns that the business expert often is not one person. There are product owners (knowledge on business process flow level) and developers (detailed knowledge of WA configuration). With only knowledge of the business process, you might miss out on details important to understand how flows exactly work.	*22:48: I often encounter product owners or software developers who have no idea about security. They build something that is functional, but do not realize what injecting malicious code in an input field can cause. You need someone with the right mindset [for cyber security].*

Experts [F], [G], [H], [I]

Deductive coding (predefined metrics)

Functionality

(translated & created a sound sentence)

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
1	13:55-17:30	Are there steps that do not add value or could be described more clear to add more value?	All agree that the steps make sense and are in a logical order. It looks very much like the approach we take when analyzing a WA. The process is very much the same, only we use other tools where we use queries to test our hypotheses.	*Gertjan on 14:18: For me, it [the methodology] really feels like an analysis of a Web Application [like we would do it].*

Completeness

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
2	13:55-17:30	Are there steps missing that would add value?	See answer Nr.1	-

Consistency

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
3	17:43-18:41	Would different users interpret the steps in the methodology the same?	**answer related more to other category** See answer Nr.6	-
4	19:01-22:18	Would the results be comparable in other cases in WAF? **question not specifically asked, but answer acquired during discussion**	Complexity of WA infrastructure is an important factor in how doable it is to get all the data. In case of a complex architecture, it is way harder to connect the data from the different servers.	-

Accuracy

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
6	17:43-18:41	What are the most important factors that determine the accuracy of the results? **question not specifically asked, but answer acquired during discussion**	How users execute the steps depends a lot on their background knowledge. You can have knowledge of CS, but knowing how the WA's exactly work is something completely else. Not everyone possesses sufficient knowledge from both fields.	-

Reliability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
-	-	What factors are the most important to ensure the reliability of the data?	**question not specifically discussed because of time**	-

Usability

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
-	-	What knowledge and skills are equired of the different team members to make the methodology usable?	**question not specifically discussed because of time**	-

Performance

Nr	timeframe	Question (translated)	Summary of response	Quote *if applicable
7	26:46-41:31	What are limitations of the approach? **question not specifically asked, but answer acquired during discussion**	*[Discussion with everyone about hypothetical and real-life cases that the experts deal with and how PM fits]* Generally, there was no consensus among the experts on a specific forensics case in which PM could be used for KLMs CS team.	-
8	41:38-54:02		*[Comparing different tools and their capabilities to PM]* A tool that is currently used is capable of taking in all data related to the complex WA architectures. Where it would be very complex to do this manually for PM, this tool does this automatically. Data can also be queried to search for cases. The only thing is that such a tool can not be used to apply business rules to filter certain cases.	-
9	54:09-56:19	How can PM improve forensic investigations? **question not specifically asked, but answer acquired during discussion**	The value of the approach [for KLM] is possibly more be in threat hunting [Tim: evil user stories. Expert: Yes] rather than forensics. Not by acting on alerts, but proactively. You think of hypotheses what could be misused and subsequently test whether such flows exist.	-
10	56:46-59:10		KLM is a large company with multiple resources. PM could be more useful for smaller companies who don't have access to expensive tools. For more complex environments, other other specialistic market standard (and expensive) tools should be used [for forensic investigations].	Alex on 56:46: Lets not forget, we are a large corporation with plenty resources and tools. Such [Process Mining] approaches are more useful for smaller companies without expensive tools, who want to look at how something happened.
11				*Alex on 58:51: I think it can add value, in cases for SMEs with simple and small applications. However, when it [Web Application architecture] gets complex, other larger tools are needed.*

C.3 - Findings extracted from Interviews

Table with all findings listed, including their identifier that is used in the text, the keywords from the bar charts and frequency of mentioning.

{Identifier}	Category	Finding	Keywords	Frequency*
{1}	*Artifact*	*The methodology overall looks logical and coherent.*	*Overall coherence*	5
{2}	*Artifact*	*In case of a forensic investigation with a goal of legal prosecution, more specific steps regarding data extraction, handling, preparation and steps after the final phase are required.*	*Legal Forensics*	5
{3}	*Artifact*	*Cases regarding modern Web Application (architectures) may require steps to be added or changed in the methodology.*	*Web Application Type*	2
{4}	*Implementation*	*It is important to start off with a good research question to get a good answer.*	*Question Quality*	4
{5}	*Implementation*	*Keeping a flexible approach to the analysis might lead to improved results.*	*Flexibility*	1
[6}	*Implementation*	*users should be able to understand what data is available and what the data can tell*	*Data knowledge*	2
{7}	*Implementation*	*To make the methodology useful and extract accurate results, the expertise of the business expert and process analyst is of great importance.*	*Team Expertise*	6
{8}	*Implementation*	*If a user gained rights to log directories which allowed them to tamper the access logs, the data already cannot be trusted before the analysis starts.*	*Data Reliability*	2
{9}	*Implementation*	*If there is relevant data missing in the logging for a case, you need to make sure that it is there for a next case where the data is required.*	*Learning Curve*	2

{10}	*Approach*	*Process Mining could be of use in forensic investigations to help understand attacker's paths, for example in support of other tools to dive deeper into the data.*	*Path Investigation*	6
{11}	*Approach*	*In very large companies, Process Mining is probably not a suitable approach. In such cases, other specialistic tools are required that can deal with complex architecture of the Web Applications. There is more added value for SMEs with smaller budgets and less complex architectures.*	*Company Size*	1
{12}	*Approach*	*Process Mining can be useful in use cases where malicious activity is expected or might have occurred, by testing it to aim to extract malicious behavior defined by certain rules.*	*Finding malicious activity*	3
{13}	*Approach*	*Process Mining tools can also be used to visualize something that is already known but making it better understandable.*	*Communication*	1
{14]	*Approach*	*Process Mining is not the approach to answer all answers in Web Application Forensics cases. Depending on the research question, a suitable approach should be chosen before blindly taking Process Mining.*	*Required Answer*	3

* Number of interviews in which a remark related to the finding was made