BUSINESS PROCESS REENGINEERING

USING PROCESS MINING

YEŞİM YILMAZ

BOĞAZİÇİ UNIVERSITY

2019

BUSINESS PROCESS REENGINEERING USING PROCESS MINING

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Management Information Systems

by

Yeşim Yılmaz

Boğaziçi University

2019

DECLARATION OF ORIGINALITY

I, Yeşim Yılmaz, certify that

I am the sole author of this thesis and I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;

This thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;

This is a true copy of thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date.....22.07.2019.....

ABSTRACT

Business Process Reengineering Using Process Mining

In today's competitive business world, organizations aim to ensure their position by continuously improving their processes to adapt the rapidly changing environment. However, continuous improvement of processes with the traditional approaches require long and costly processes. With the advances in digitalization, a new approach, process mining has been developed to overcome these challenges by using transactional data in information systems. Process mining aims to discover, monitor and improve processes by acquiring knowledge from the event logs. Within the scope of this study, process mining techniques are analyzed to confirm the role of process mining in process improvement by using real process data of an international company in construction sector. These analyses are performed using two different process mining tools: ProM and Celonis. Moreover, these tools are compared using the framework that is generated for the comparison of process mining tools.

ÖZET

Süreç Madenciliğini Kullanarak İş Süreçlerini Yenileme

Bugünün rekabetçi iş dünyasında, kuruluşlar, hızla değişen ortama adapte olmak için süreçlerini sürekli geliştirerek mevcut konumlarını korumayı amaçlar. Ancak, geleneksel yaklaşımları ile süreçlerin sürekli iyileştirilmesi uzun ve maliyetli süreçler gerektirmektedir. Dijitalleşmedeki gelişmelerle birlikte, bilgi sistemlerindeki işlem verileri kullanarak bu zorlukların üstesinden gelmek için yeni bir yaklaşım olan süreç madenciliği geliştirildi. Süreç madenciliği, olay kayıtlarından bilgi elde ederek süreçleri keşfetmeyi, izlemeyi ve iyileştirmeyi amaçlar. Bu çalışma kapsamında, inşaat sektöründeki uluslararası bir şirketin gerçek süreç verilerini kullanarak, süreç madenciliğinin süreç iyileştirmedeki rolünü doğrulama amacıyla süreç madenciliği teknikleri incelenmiştir. Bu analizler iki farklı süreç madenciliği aracı; ProM ve Celonis kullanılarak gerçekleştirilmiştir. Ayrıca, bu araçlar, süreç madenciliği araçlarının karşılaştırılması için oluşturulan çerçeve kullanılarak karşılaştırılmıştır.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Prof. Aslı Sencer, my thesis advisor, for her patient guidance, enthusiastic encouragement and useful critiques of this research work.

I would also like to thank to Information Technology and R&D Director of the company who supported this research by sponsoring the case study.

In addition, I would also like to the project team members who shared their precious time and experience during the case study.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BPM	Business Process Management
CRM	Customer Relationship Management
ECC	Enterprise Central Component
ERP	Enterprise Resource Planning
IS	Information System
IT	Information Technology
IvM	Inductive Visual Miner
SSP	Service Sales Process

CHAPTER 1

INTRODUCTION

With the developments in technology and digitalization, the scale and scope of information systems (IS) have grown rapidly to include Enterprise Resource Planning, Workflow Management, Web Services, and Mobile Applications tools that support users in various processes within the organization. This widespread use of IS results in huge amount of process data recorded in the created databases. The data stored in these databases can be used to discover valuable information in various areas including discovering user patterns, detecting causal relations, and generating process models (Frawley, Piatetsky-Shapiro & Matheus(1992), Feyyad, 1996, Weijters & Van der Aalst, 2001). Process model discovery is an outstanding and promising research area in business process management (BPM) since it gives way to monitor and improve processes by the use of process data analytics techniques summarized under the new term “process mining”. These approaches fulfill the gap between data mining and business process modeling.

Process mining has found a growing interest among researchers in the last decade. Under the leadership of The Institute of Electrical and Electronics Engineers (IEEE) Task Force on Process Mining, researchers from 20 countries including Germany, Italy, The Netherlands, and USA have signed the process mining manifesto to promote the subject of process mining (Adriansyah et al., 2012). This group represents software vendors, consultancy firms, customers, and research institutes. Adriansyah et al., (2012) state that the growing interest in process mining is driven by two facts: the availability of detailed historical data of processes, and the necessity of process improvement in the competitive business world.

The interest on process mining continues increasingly as process mining applications are capable to support traditional process improvement approaches such as Continuous Process Improvement, Business Process Improvement, Total Quality Management, and Six Sigma. The first international conference on process mining (ICPM) held in Aachen, Germany in 2019 attracted hundreds of international researchers showing the growing interest in this area. While academic research on process mining continues, the tools that provide process mining applications have been extended to include new capabilities.

Process mining is realized under three headings: process discovery, conformance checking, and enhancement (Van der Aalst, 2012). The first type is process discovery aiming to reveal the real executions of the business processes. For instance, a sales process includes several sub processes and the process owner would like to know the frequency of the instances to identify the most recurring sub processes, rework loops or failures. There are several algorithms generated for discovering processes in various aspects; alpha miner, fuzzy miner, and inductive visual miner (IvM) are among these to be explored in the current study. The second type, conformance checking compares the event transactions data with the as-is process model of the organizations. For example, the owner of the sales process would like to know how much the as-is model is followed by the actual transactions that took place in the previous period. Third type of process mining is process enhancement that aims to improve processes using the event transactions data and the results of the process discovery and process conformance. In general, process mining techniques aim to support business owners for business process reengineering by showing bottlenecks, throughput times, reworks and delays.

As mentioned above, there are specialized tools to perform the process mining algorithms. Right after the introduction of process mining, these tools have been developed to be used for process mining and related analyses. The earliest tool for process mining, called ProM is presented by the study of Dongen et al. (2005). This is an open source framework which is still widely used for academic research. Currently, there are two open source tools, namely ProM and Apromore. In addition to open source tools, there are also many commercial process mining tools. Commercial tools include ARIS Process Performance Manager, Celonis, Disco, Minit, myInvenio, ProcessGold, QPR Process Analyzer. All these tools provide solutions using various process mining algorithms and highlight their distinguishing features.

Inspired by these developments and the growing interest in the area, the aims of this thesis study are (i) to observe the role of process mining in process improvement, (ii) to examine process mining tools and compare two of them by generating an evaluation framework. The experiments of the research are made on real process data obtained from an international company in the construction sector.

The remainder of this thesis is organized as follows. In the next chapter, we provide a literature survey on the use of process mining in process improvement. In Chapter 3, we provide a background on the process mining project methodology and the theoretical algorithms used in process discovery in process mining. In Chapter 4, problem definition is provided. Chapter 5 includes the application of process mining steps to discover the process, check conformity and improve the current process. In Chapter 6, the comparison of the tools is provided with the evaluation framework generated. Finally, in Chapter 7 conclusions are provided and recommendations for further studies on the subject are stated.

CHAPTER 2

LITERATURE SURVEY

In this section, we first explore the new trends in business process reengineering that come with the advances in digitalization, and then analyze in depth the studies in process mining.

New Trends in Business Process Reengineering

After the first use of "Industry 4.0" term in 2011 by Merkel, several studies have been performed in BPM area. Krumeich et al. (2016) emphasize in their study that companies will be ahead in the competition from their competitors if they can understand their processes, optimize the outcome and improve them with knowledge. They suggest generating event-based process predictions to dynamically adapt process instance by viewing each business situation individually. They also analyze the benefits of using big data on the success of prescriptive control for business processes.

Lederer et al. (2017) state that companies prefer combination of process-oriented organizations and modern technologies rather than product- or function-oriented organizations. The aim of their study is to investigate new trends in BPM in terms of the catchwords and their research popularities. It is stated that with the advances in digitalization, the use of IT (Information Technology) for assisting workflows has recently become a core principle of BPM. It is also emphasized that some topics of BPM is renamed and restated within Industry 4.0. Examples include using "smart home" instead of customer interaction in service processes, or using "digital factory" instead of IT-enabled production processes. Lederer et al., (2017) suggest that the new trends that include "BPM in the cloud", "process mining",

"social BPM" and the "reuse of processes" can be collected under three titles for data-driven, case-driven, and social-driven BPM. Data-driven BPM makes use of enormous amount of data to create information for decision making. Case-driven BPM is the result of the need for an agile, flexible and adaptive process management in a rapid changing environment. Social-driven BPM benefits from the use of social software tools to get valuable inputs from users in an efficient way and optimize processes with human involvement. Table 1 summarizes the catchwords that appear under these new trends in comparison with the catchwords of the traditional BPM. It follows that "Process Mining" under data-driven BPM appears to be the most popular catchword among the most prestigious conferences in the IS area. The second next popular catchword is "Automated discovery" is again under data-driven BPM. However, it attracts significantly less interest, leaving the most attractive in BPM area to process mining.

Table 1. Systematization of BPM Trends and Digitalization Catchwords Extracted from the Analyzed Proceedings (Table 1 in Lederer et al., 2017)

		Conferences									Number
Trend	Catchwords	BPM	MKWI	WI	INF	ECIS	SEAA	CBI	S-BPM	BIS	
Data-driven BPM	Process Mining	1	3	4					1	4	13
	Automated discovery	8									8
	Conformance checking	2									2
										Subtotal	23
Social-driven BPM	BPM 2.0				2		1				3
	S-BPM								7		7
	Process Culture					1			1		2
										Subtotal	12
Case-driven BPM	Collaborative Case Management							1	1		2
	Emergent Case Management			1					1		2
										Subtotal	4
Traditional BPM	Advanced Modeling		4	4	3		1				12
	Model Foundations	3									3
	Related Models				1		1				2
	Execution Automation		2	1							3
	Standardization		2		1						3
	Goal modeling					1	1				2
	Process Translation	1					1				2
	Process performance	2		1		2					5
	Reference Models			1	1						2
	Digitalization Frameworks		2								2
										Subtotal	36
										Total	75

With a different perspective, some articles indicate that although several studies have been made, theoretical research in BPM is far from being aligned with the real business needs (Hull and Nezhad, 2016, Lederer et al., 2017). At this point we dwell on the studies in process mining that can certainly be used to bridge the gap between the theoretical research and real business needs in BPM.

Process Mining

In one of the most comprehensive books in this area, Van der Aalst (2012) states that process mining acts as a bridge between data mining and business process modeling. As we all know, data mining is about value creation from data to be used in decision making in marketing, sales, production and many others. Similarly, process mining is about value creation from business process transaction data to be used in business process modeling.

When a business process is executed, the trails of its subprocesses are left in the IS in the form of logs. These log files include valuable data related to transactions of processes. There are three process mining types that use an event log. These are listed as discovery, conformance checking, and enhancement (Van der Aalst, 2012). Discovery is the technique that extracts the business process model without any other information. Conformance determines the fitness of event log to the process model of the organization by delta analysis. The third type, enhancement aims to improve the current process model using both process model and event log. Apparently each one of these concepts include the application of process mining theory on the real process data, and hence strengthen the bridge between theory and practice of BPM.

Process mining case studies are conducted in various industries including IT, healthcare, government, finance, manufacturing, energy, and education (Dakic et al., 2018). Dakic et al. (2018) analyze several BPM case studies and find that although

all of them include process discovery applications, only a small proportion of these studies carry out conformance checking. There are a moderate number of studies in enhancement when compared to discovery and conformance.

Now, we are ready to explore the studies in each type of process mining, respectively.

Algorithms for Process Discovery

Studies in process discovery algorithms start with discovering workflows. In an early study, Van der Aalst et al. (2002) study to discover workflow process models by using sequentially ordered log that do not include any information on the timing of the activities. Indeed, modeling workflows require knowledge on workflow modeling language and workflow practice itself. So, knowledge transfer is required from business owners and experts to business analysts (Van der Aalst et al, 2004). As it can be seen, this is quite a lengthy process that may also include several inefficiencies in information transfers between business owners, experts and business analysts. Hence, collecting digital information from processes as it is happening might serve as a preferred alternative to modeling the workflow manually. Early studies in process discovery are only capable of creating models for structured processes e.g. approval of loans. Unlike the early works in process discovery, Hwang et al. (2004) study to create methods to mine more unstructured processes that contain many temporal variants. As the study aims to extract temporal behavior in the execution of processes, the method they develop can be used to detect fraudulent cases.

Several algorithms have been generated in the literature for process discovery. In the aforementioned studies, Van der Aalst et al. (2002, 2004) introduce α algorithm to discover a workflow. This algorithm is the starting point for process

discovery. However, it has some critical requirements on the event log; it has to be complete without short loops or duplicate tasks. α algorithm has been extended by Alves de Medeiros et al. (2004) and Wen et al., (2007) as α+ and α++ algorithms to overcome these drawbacks. α+ miner can solve event logs with short loops and α++ further improves it by introducing the ability to identify implicit dependencies that appear in non-free-choice tasks.

Weijters et al. (2006) introduce Heuristic Miner which is an easily executable algorithm that can cope with noise and reveal common patterns by excluding exceptions and details. Weijters and Ribeiro (2011) represent an enhanced version of Heuristic Miner as Flexible Heuristic Miner where a new set of parameters are introduced. In application, first these parameters are selected and then the model is created.

Van der Aalst et al. (2005) present a new algorithm called Genetic Process Mining that can deal with duplicate task and silent transitions as well as incompleteness and noise. This study starts with developing a causal matrix and then maps the relation between causal matrix and petri net.

Güntner and Van der Aalst (2007) introduce Fuzzy Miner to become one of the most commonly used algorithms in process discovery. Its significance is due to the fact that fuzzy miner can deal with unstructured processes. Furthermore, unlike its predecessors, the interference of user to select the threshold values is allowed in this methodology, so that oversimplicity or over fitting can be prevented.

Chinces and Salomie (2013) present three new algorithms as Ant Colony Optimization (ACO Miner), Guided Local Search Miner (GLS Miner), and Iterated Local Search Miner (ILS Miner). GLS Miner and ILS Miner employ local search methods whereas ACO Miner benefit from artificial intelligence. It is found that

ACO Miner performs well with simple models meanwhile ILS Miner produces better models with loops.

Leemans et al. (2014) introduce IvM which serves as one of the most comprehensive algorithms in process discovery. It enables interactive and iterative process discovery that further leads to a more valid process model. IvM is tested and compared with other miners using real event logs of example processes. It is shown that IvM outperforms its competitors in terms of semantics, speed, and evaluation of model, activity, and event level.

Although many algorithms have become available for process discovery, there has not been a common methodology to compare the outputs of these algorithms. Rozinat et al. (2007) study on how to validate models created by process mining algorithms. To create an evaluation framework, two approaches, metrics and k-fold cross validation are initiated. Metrics approach utilizes four evaluation criteria; fitness, precision, generalization and structure to compare discovered process models. The details of this evaluation are explained in Chapter 3 where the background on Process Mining is provided. K-fold cross validation is a machine learning approach that divides the sample into k groups and tests the algorithm by using k different test data while using the remaining data to train the algorithm. By using the metrics approach, a tool called Control Flow Benchmark is developed. This tool includes a framework to evaluate and compare process models that are discovered using various process mining algorithms.

Applications of Process Discovery

In various industries, many case studies on process discovery are made using real life data. Van der Aalst et al. (2007) conduct one of the earliest case studies in process mining where real data from Dutch National Public Works Department is

used to discover the invoice handling process. Event log is analyzed from process, organizational and case perspectives with various process discovery plugins in the open source process mining framework ProM. As a result, some significant insights on how process is handled and where the delays occur in the process are gained. As Van der Aalst et al. (2007) state, these insights may assist the management while setting targets and improving the process.

Mans et al. (2008) conduct a study in healthcare sector which is one of the most common process mining research areas. They aim to prove that nontrivial healthcare processes can be mined with process mining algorithms to acquire knowledge. Process discovery study on various perspectives reveals important information on the process. In the research, both heuristic miner and fuzzy miner are used. In addition to precious knowledge gained for improvements, it is also proved that process mining is applicable and essential for unstructured processes like healthcare processes. However, it is noticed that mining algorithms should be improved to handle unstructured processes better. Almost a decade later, another process discovery study in healthcare is conducted by Ganesha et al. (2017). The aim of this research is to acquire knowledge by evaluating available patient treatment data using IvM plugin of ProM which Leemans et al. (2014) introduce to discover unstructured processes. Discovered model is able to display the bottlenecks and delays. With these findings, resources can be utilized to decrease the waiting times.

Rubin et al. (2014) show that process mining is not only usable for organizational functions but also for software systems. Accordingly, they conduct two case studies on ticket reservation processes. First one is a computer reservation system used by travel agencies and the second one is a web portal used by customers. The process of former is discovered using a commercial tool Disco and the latter one

is discovered using heuristic and fuzzy miner plugins of open source framework ProM. The findings are remarkable for management as well as for developer teams as they observe user behavior that are neither designed nor expected. It is emphasized that process mining can be used to analyze user behavior in software systems and researches in this area should be motivated.

Conformance Checking

The second type of process mining is conformance checking that aims to compare the process reality with the reference model. The main practice in conformance techniques is to highlight the difference between event log and the reference process model.

Rozinat and Van der Aalst (2008) study conformance checking to evaluate the fitness of event log to the business model of the organization. Fitness analysis displays the deviations of event log from the model which is referred to as delta analysis.

Conformance checking can also be used to audit the business processes. Van der Aalst and Alves de Mederios (2005) state that proper audit trails can be used to create a desired process model with α-algorithm. New trails, i.e., event log, can be assessed to check if there is any instance that does not comply with that process model.

The performance of conformance checking is yet another interest of researchers. Muñoz-Gama and Carmona (2010) and Van der Aalst et al. (2012) suggest measuring the performance of conformance checking by the criteria initiated by Rozinat et al. (2007), namely fitness, precision, generalization and structure. Munoz-Gama and Carmona (2010) develop a technique to measure the precision of a model by determining the escaping edges that correspond to the difference of the

model from the event log. This technique assumes that the fitness of the model with the event log is perfect, i.e., the process model covers all traces. If this assumption isn't valid, it's recommended first to calculate the fitness level of the model then measure the precision of it.

With a different perspective, Van der Aalst et al. (2012) indicate that it is possible to measure the level of conformance numerically between zero and one. The terms "move in log" and "move in model" are introduced in (Van der Aalst et al., 2012) where the former means the activity is only observed in the event log but not in the model, and the latter is vice versa. The algorithm is based on aligning the event log and the model. Here, computation is performed by replaying the event log on the process model. In addition to conformance checking, performance evaluation and bottleneck analysis are also possible with this algorithm as event logs contain timestamps.

Process Enhancement

In general, outputs of conformance checking applications constitute the inputs for process enhancement, the third type of process mining. Deviations and inconsistencies found with conformance guide the process owner for improvement by changing either process model or the execution itself. On the other hand, delays and bottlenecks may indicate that process model should be revised so that process can be enhanced for a better performing version.

In their case studies for process enhancement, Maruster and van Beest (2009) and Rozinat et al. (2009) emphasize the significance of discovery phase. They state that the process miner for discovery should be chosen according to the structure of the process, i.e., if the process is unstructured, methods that deal with noise and

exceptional cases should be employed, otherwise a traditional miner such as alpha, alpha+ miner can be chosen.

Maruster and van Beest (2009) introduce a methodology to be applied in process reengineering exercises. This methodology is based on simulation and uses bottom-up approach. In other words, it uses process data to diagnose process improvement areas. First process model is created with Heuristic Miner, then conformance check is applied to determine throughput times and bottlenecks for performance analysis. As-Is and To-Be process models are prepared via a Colored Petri Net Tool and the throughput time distribution model and bottlenecks are simulated. The comparisons are made between initially mined process model and as-is model, and between as-is and to-be models. Implementations in a gas company and a governmental institution show that processes can be improved with bottom-up approach if improvement expectations include reducing throughput time and removing bottlenecks.

Another case is studied by Rozinat et al. (2009) to enhance the test process of a wafer scanner producer. After comparing the discovered model with the existing model, findings are considered for business improvement. An important inference of the study is that if a process changes constantly by its dynamic nature, process enhancement application by process mining should be performed in an iterative manner, i.e., findings of the previous process version should be assessed for improvement of the next process version. To be able to implement the findings, there should be the mutual subprocess. For example, in the wafer scanner case, the products are produced in small amounts and different types, and the test procedure needs to be changed frequently for each product type. As a result, although the

findings of the study are valuable, some of them may not be applicable for the new test process.

Tools for Business Process Mining

Most of the process mining algorithms and techniques have been introduced above are implemented using process mining tools. One of the most popular tools is ProM which is an opensource framework developed by Dongen et al. (2005). This framework creates an environment for process mining applications where researches can continuously improve. Moreover, newly developed algorithms can be plugged in without changing the main framework. The first version of ProM had 23 plugins including alpha miner. Today, ProM includes 2,000 plugins developed by many researchers (ProM 6.8 Plug-ins, 2019).

A few years after ProM has developed, Apromore which is another open source application is introduced by Fauvet et al. (2010). This tool is available as cloud or on-premise solutions. There are more than 50 plugins available in Apromore.

In addition to open source tools, there are also many other commercial tools available today; Celonis, Disco, ProcessGold, Minit, myInvenio, QPR Process Analyzer, ARIS Process Performance Manager, to name but a few. Among these, the rapidly growing and award-winning German company Celonis is an industry leader, realized by its integrated solutions with SAP. It holds the cloud award for the best SaaS in U.S. in 2018-19 (The Cloud Awards, 2019). Celonis provides process discovery as well as conformance checking solutions. It may also be integrated to ERPs of organization to make predictions and enhance processes. It also offers academic license for research.

The process mining analysis of this thesis study is carried out on two aforementioned platforms ProM and Celonis. ProM is widely used in academic studies since it is an open source tool that provides diverse plugins. We also experiment with Celonis as it appears as the most commonly used commercial tool. Using two widespread tools provide us the ability to compare their performances as well as to reach a broader variety of plugins.

CHAPTER 3

BACKGROUND ON PROCESS MINING

In this chapter, the background on process mining terminology, its steps, and related algorithms are provided mostly based on the most comprehensive book in this area by Van der Aalst (2012).

Figure 1 provides a general understanding of the types of process mining. As seen in Figure 1, there are four components that consist the realize the businesses functions. In real world, business functions are executed by many parties including people, machines, and organization. These executions are modelled by process models and supported or controlled by software systems. Moreover, process models sometimes analyze the real world to configure software systems. The operations that are supported by the software system are recorded to the databases as event logs. The relations between event logs and process models express three process mining types: process discovery, conformance checking and enhancement. These process mining types are explained in the previous chapter.

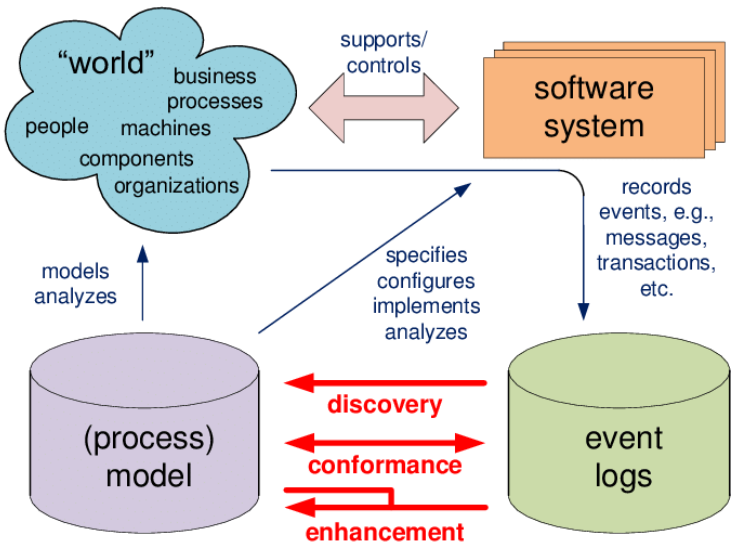


Figure 1. The types of process mining (Van der Aalst, 2011, p.262)

Process discovery aims to discover the business process models by using various process mining algorithms and the data that is recorded in IS during the execution of process tasks. Thus, large data sets that contain all necessary information about how processes are handled, are stored by different entities.

The starting point of process mining is sequencing the tuples in the event log to create a process model. An example event log is provided in Table 2. Event log should at least contain case, activity, timestamp information for process mining. Van der Aalst (2012) defines a case as an instance which consists of events to execute a process. An activity represents a well-defined task whereas an event can be defined as the record of the executions of an activity for a specific case. Timestamp is defined as the time of the event. Depending on how the IS records the execution time, timestamp can be the start of an event as well as the completion time of an event or both. Resource or originator or any other available information can be used for further analysis using various process mining techniques.

The output of process mining is a model that aims to explain how the real process is executed. This is called process discovery. To mention a few, there are many algorithms for process discovery like alpha miner, fuzzy miner, genetic miner, heuristic miner, inductive miner.

The resulting process models of different algorithms may be evaluated in terms of different dimensions. Rozinat et al. (2007) define these dimensions as fitness, precision, generalization and structure. Fitness shows how much of the event log can be replayed by the model. Precision evaluates the ratio of the replayed behavior to all behavior that can be allowed by the model. The higher precision ratio means larger proportion of the allowed behavior by the model is observed in the log. Generalization checks that the discovered model describes the observed behaviors,

but it is not specific to them, i.e., it is not overfitting. In other words, a generalizable model can easily be updated to include a new behavior that do not exist in the log. Finally, structure checks the modelling language. High structure level means a compact model, e.g., a model that doesn't display the same task more than once or a task that is not observed in the log. Some researchers as Buijs (2014) prefer to use simplicity criterion in place of structure to evaluate how simple and easy to understand the model is by humans.

Table 2. An Example of Event Log

Case ID	Activity	Timestamp
1001	register application	16.04.2013 10:10
1001	check credit	16.04.2013 10:16
1001	calculate capacity	16.04.2013 10:16
1001	check system	16.04.2013 10:20
1001	accept	16.04.2013 10:24
1001	send decision e-mail	16.04.2013 10:29
1002	register application	16.04.2013 10:15
1002	check credit	16.04.2013 10:22
1002	calculate capacity	16.04.2013 10:23
1002	check system	16.04.2013 10:29
1002	accept	16.04.2013 10:37
1002	send decision e-mail	16.04.2013 10:44
1003	register application	16.04.2013 10:19
1003	check credit	16.04.2013 10:23
1003	calculate capacity	16.04.2013 10:28
1003	check system	16.04.2013 10:33
1003	accept	16.04.2013 10:42
1003	send decision e-mail	16.04.2013 10:50
1004	register application	16.04.2013 10:28
1004	check credit	16.04.2013 10:32
1004	calculate capacity	16.04.2013 10:37
1004	check system	16.04.2013 10:40
1004	accept	16.04.2013 10:45
1004	send decision e-mail	16.04.2013 10:48

Figure 2 compares four different models created by referring to the log table in Figure 2. Here (a) shows the numbers instances for each log trace or behavior. (b,c,d,e) are the models generated by several algorithms. These models are evaluated with respect to the above four criteria. The evaluations are as "-" for poor performance and "+" for good performance.

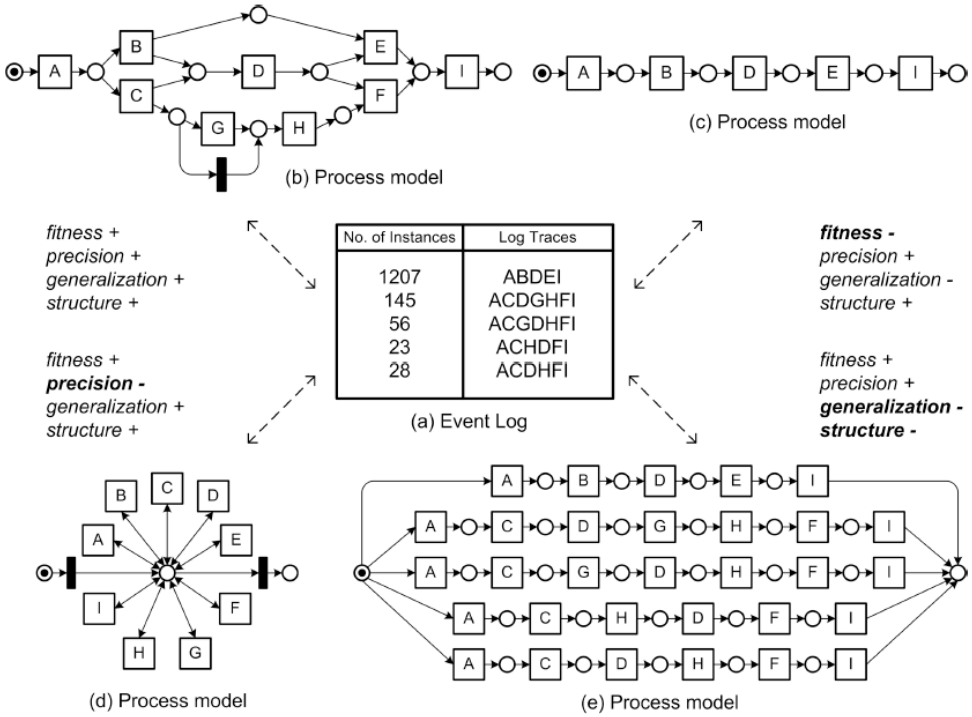


Figure 2. Evaluation of four difference models in terms of four dimensions; fitness, precision, generalization, and structure (Rozinat et al., 2007)

In addition to process discovery, process mining also aims to compare the discovered model with a given reference model and generate an improved process model. These are referred to as conformance checking and enhancement, respectively as mentioned in the Literature Review in Chapter 2.

In the following parts of this chapter, first, project steps for process mining where the current thesis study is based on are explained. Then, commonly used

process discovery algorithms that are considered in the current thesis study are described in detail.

3.1 Process mining project steps

The most well-known researchers in process mining area introduce a process mining project methodology called PM^2 in (van Eck et al., 2015). PM^2 methodology fulfills the need for an iterative, tailor-made methodology for projects that aim business process improvement and reengineering with process mining. As the overview of the methodology shows in Figure 3, there are six stages as Planning, Extraction, Data Processing, Mining and Analysis, Evaluation, and Process Improvement and Support in PM^2.

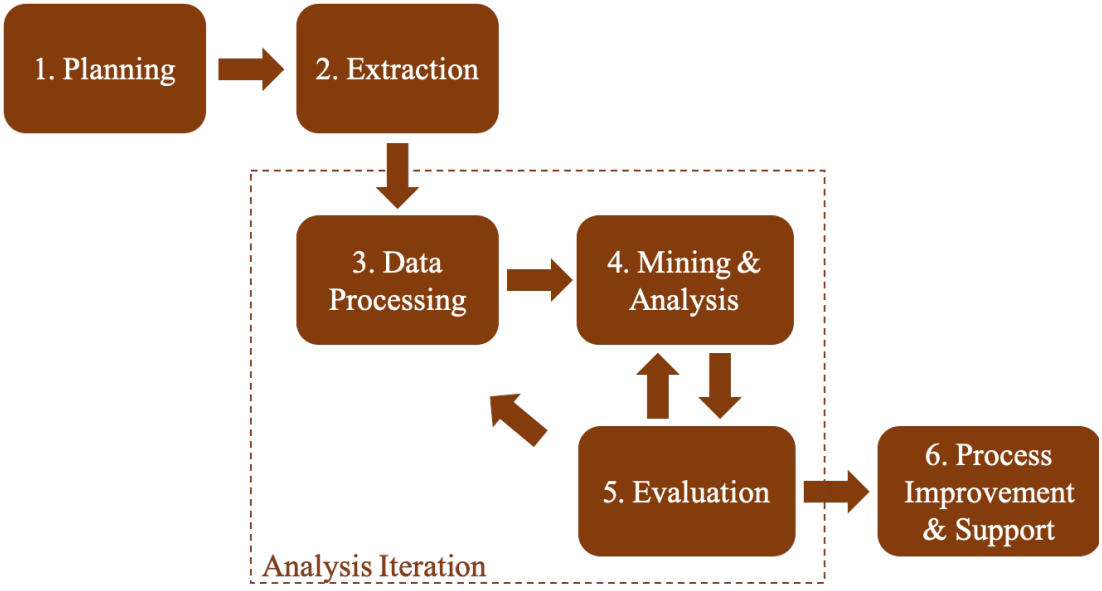


Figure 3. PM^2 Methodology

Planning includes deciding on the business process, determining the questions that need to be answered in later stages, and building project team. In extraction phase, the data is extracted according to decided project scope. In this

phase, information about business process and process mining is also shared within the project team so that all the necessary data can be prepared appropriately. The aim of the third stage, data processing is to prepare event log from the extracted data. Logs can be enriched by adding various attributes or can be filtered to lessen complexity. Cases and activities that will be used in the further analysis are determined in this stage.

The next phase, Mining and Analysis includes process discovery, conformance checking, enhancement, and process analytics. The first three items of this stage have been described in Literature Survey. The fourth item, process analytics includes data mining and visual analytics that are used to generate a better understanding about the business process. Evaluation phase includes the diagnosis, verification and validation of the outputs and findings of the previous stage, Mining and Analysis (Eck et al., 2015).

Data processing, mining and analysis, and evaluation stages are further grouped as analysis iterations as seen in the Figure 3. Outputs of these stages should be inspected several times to explain all questions including the ones that arise in any of these stages.

Finally, Process Improvement and Support phase proposes process changes from improvement ideas based on the finding in the previous stages. Subtasks of this stage are supporting operations and implementing improvement ideas. The former covers supporting daily operations by detecting problematic instances using live data whereas the latter usually leads to another project which requires a different expertise.

3.2 Process mining algorithms for process discovery

In this section, commonly used process discovery algorithms, Alpha Miner, Fuzzy Miner, and IvM that are also available as Plug-Ins in ProM Framework are explained in detail.

3.2.1 Alpha miner

One of the earliest process discovery algorithms is alpha miner which has been proposed by van der Aalst et al. (2004). This algorithm visualizes the process by creating a Petri Net from the event log. An example Petri net model is given in Figure 4. Petri net has a start place as well as an end place. Tasks are shown as transitions in squares and causal dependencies are shown as places in circles with its arrows. AND-splits and AND-joins are also modeled by transitions in squares. Tasks divided by AND-splits are performed in parallel. Likewise, OR-splits and OR-joins are modeled by places in circles. They indicate the choices in a process flow. In Figure 4, the process starts with task A and ends with task D. The place in circle after task A has one in and two out arrows. This means is this is a choice place and one of the two arrows should be followed, i.e., either task E, or task B and C in parallel is followed task A. AND-split is used to indicate the start point of two parallel tasks B and C. D can be performed when E or both B and C are completed.

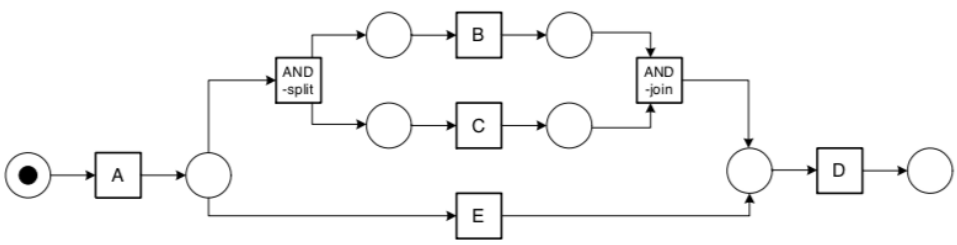


Figure 4. A petri net for an example workflow process (Van der Aalst et al., 2002)

Alpha miner algorithm assumes that the log is complete, i.e., it uses all instances without excluding non-frequent or incomplete instances. It uses either the order or the timestamps of events to create a model. It checks the causal dependencies of an activity, e.g. if a task is always observed after another task, it is expected that there is a causal relation between the first and second tasks. This dependency is shown with a connection between two places.

Parallelism and choice are also evaluated by the algorithm. If two tasks are followed directly by each other in any order, then they can be modeled as parallel activities. For instance, let A, B, C, D denote the tasks in a process. If we have two instances ABCD and ACBD in the process log, then B and C tasks are modeled as parallel tasks.

On the other hand, two tasks can be modeled as transitions after a choice place. For instance, if we have only ABD or ACD as instances in the process log, tasks B and C are modelled as choice activities since two tasks B and C never occur in the same trace between A and D. Obviously parallelism refers to an AND relation, whereas choice refers to an OR relation between tasks.

Van der Aalst et al. (2004) study various cases to discover petri net from event logs. The derived models show that α-algorithm can handle large loops, but it is not possible to rediscover short loops with alpha miner.

As this algorithm is based on the causality of tasks, complexity of output is exponentially proportional to the number of tasks. This complexity is limited with the number of tasks, but the time needed to calculate a result is proportional to the size of the log for large scale implementations.

Alves de Medeiros et al. (2004) study further to address issues faced with α-algorithm and to improve it by dividing the discovery into three parts. First, all

length-one loop traces are removed from the event log. Then, process model is discovered using simplified log. Finally, length-one loop tasks are placed to their specific position on the discovered model.

It is important to restate that α and α+-algorithms require a complete log without noise. They are also only capable to discover structured processes. These can be considered as drawbacks since event logs of most of the real businesses are unstructured, e.g. customer relationship management (CRM) or IS change request processes. Furthermore, these unstructured processes contain noisy and exceptional traces that the algorithm fails to ignore. On the other hand, if the process is well-defined and structured like a credit evaluation or quality control process, these algorithms can be employed to discover process models unhesitatingly.

Although the discovered process models of alpha algorithms are convenient and comprehensible for structured problems, the outputs of these miners do not reveal the frequency or density levels of events. Thus, there is no chance to distinguish infrequent behavior from the petri net. Moreover, they are not flexible enough to allow the user to change some parameters that may affect the quality of the output.

3.2.2 Fuzzy miner

Güntner and Van der Aalst (2007) propose a new approach for process discovery which aims to address the complexity issues of other process algorithms. They state that the main problem with the existing algorithms is the assumptions about logs. Many algorithms assume that there is one exact perfect process model and the log is accurate. These assumptions cause nonrealistic and complex models that are referred to as spaghetti models as seen in Figure 5.

Fuzzy miner aims to discover processes even if they are unstructured and complex, simply by filtering the activities according to their significance defined by the user. Here significance is defined as the relative importance of action. In other words, the degree of interest is in the frequency of the activities or the order of their occurrence. If the significance level is increased by the user, then higher filtering is applied, and simpler models are generated.

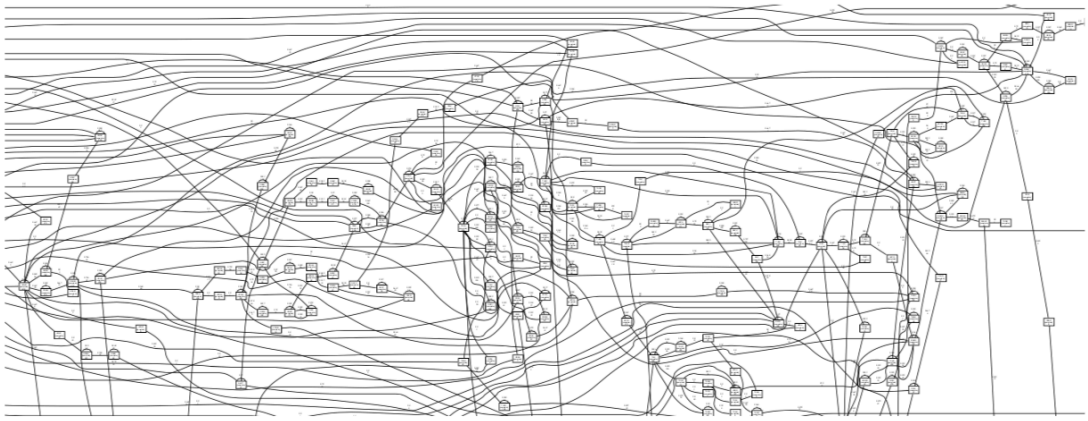


Figure 5. Small portion of a typical “spaghetti” process model (ca. 20% of complete model) (Güntner and Van der Aalst, 2007)

Figure 6 provides example models created by Fuzzy Miner. Activities are specified by rounded rectangles and named as “nodes”. Significance of the nodes are written below the activity name. Connections are specified by arrows and named as “edges”. The thickness of arrows specifies the significance of the order of occurrences of connected nodes, i.e., thick arrows imply high significance whereas thin arrows imply low significance. Figure 6 (left) displays a model that has many edges as the selected significance thresholds for edge and node are low. Many process variants can be seen in this model. When the edge threshold and node threshold is increased as in the right, some nodes are clustered, and many edges are removed from the model.

In the output of fuzzy miner, the activities with higher significance are kept whereas less significant but highly correlated activities are aggregated, i.e., hidden in clusters. An example cluster appears in Figure 6 (right) with the blue background. The remaining activities that are less significant and lowly correlated are abstracted from or filtered in the simplified model. The resulting models give emphasis on the most significant parts of the process to prevent investigating complex and vague process models.

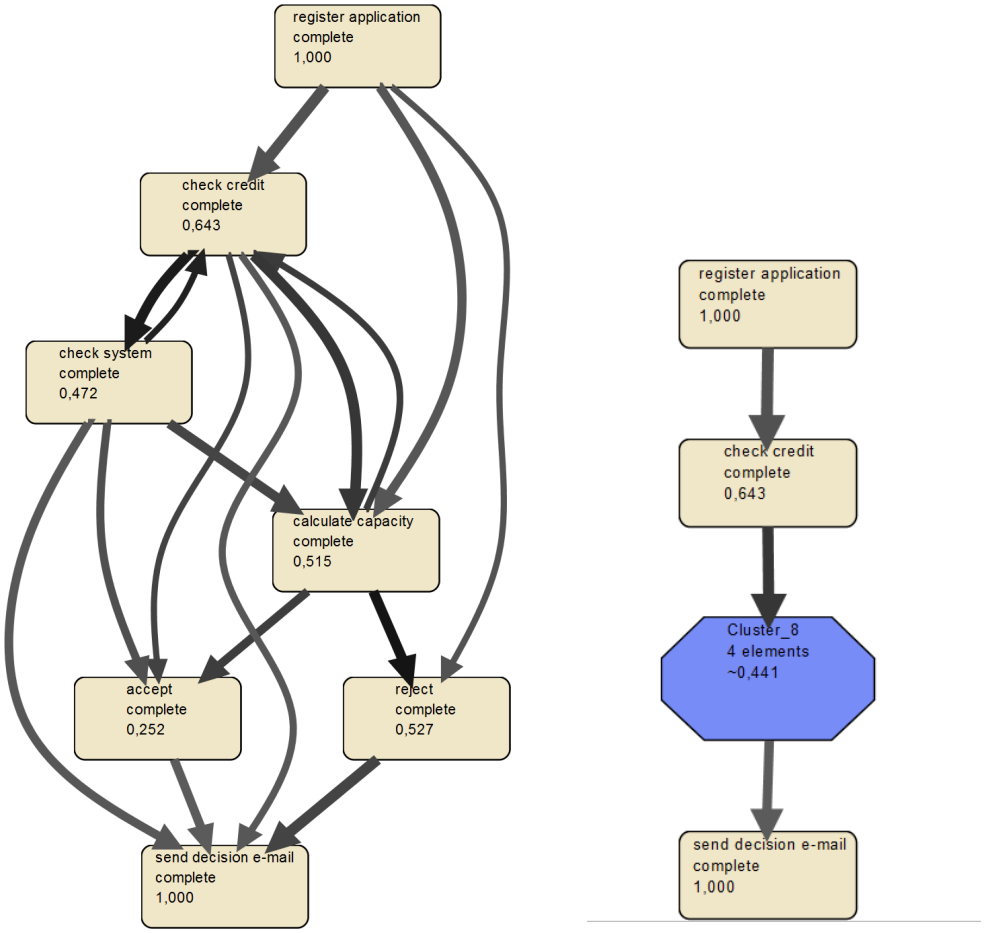


Figure 6. Example of a process models created with Fuzzy Miner by low edge threshold (left) and high edge and node threshold (right)

This approach allows to create a less complicated process model not only by evaluating the activities but also filtering the edges. This calculation is performed for each node. The edges above the significance threshold are drawn in the model.

Different threshold values can be chosen for edge and node filters. The user has the opportunity to analyze the processes for various levels of simplicity e.g. the higher edge threshold means the simpler the model or the lower node threshold means more inclusive model. Obviously, in Figure 6, the model on the right is simpler where the model on the left is more comprehensive.

Leemans et al. (2014) state that this approach is the starting algorithm for the commercial process discovery tools.

One of the most important advantages of fuzzy miner is that user can select the significance thresholds depending on the process characteristics or business needs. When the process is structured, and event log is complete, low values are preferred for significance thresholds to reflect the entire event log. In the meantime, when the event log is incomplete and noisy, and process is unstructured, higher significance thresholds are considered to filter out noisy data and highlight the common variants.

A critical disadvantage of the miner is its failure to represent parallel activities. With fuzzy models, it is not possible to discover whether two arrows coming out of an activity represent parallelism or choice. Moreover, since there is no chance to have a petri net as an output, the generated process model cannot be directly configured in the IS system or used in process conformance algorithms.

3.2.3 Inductive visual miner (IvM)

The third process discovery method is called inductive visual miner. Leemans et al. (2014) introduce IvM to fulfill the need for fast, functional, and user-friendly process discovery tool.

IvM uses the same type of event logs as other process miners. Figure 7 displays the steps of execution. Leemans et al. (2013) state that the first step of mining process is

to prepare log activity in which artificial start and end events are added to each instance.

The second step is filtering activities by setting a threshold to exclude nonfrequent activities from the output of IvM. The next step is setting a threshold for paths to filter out noisy instances. The default values for activity filter and path filter are 1.0 and 0.8 respectively. The output of this step is the business process model which not only displays paths but also parallel activities and choices. The created model is aligned with the log and paths are shown with the frequencies as default option. The fifth step is optional which allows user to select any activity to highlight the paths that goes through that node. The final step is the animation of instance which gives information about bottlenecks and intensity. If a user changes a parameter, the necessary tasks restart immediately, eliminating the need to rerun the chain of tasks in Figure 7 from the beginning.

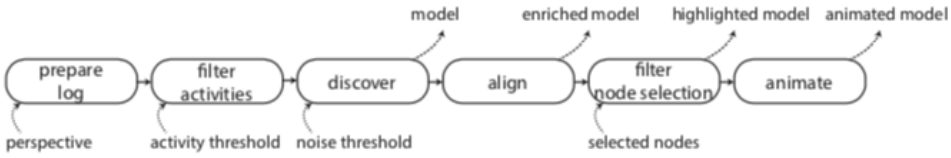


Figure 7. Chain of tasks, their parameters (bottom) and their visual results (top)

Figure 8 provides an example of process map generated by IvM. It is seen that, there are adjustment options on the right of the screen to play with the model created. The output of IvM includes a model of the business as a petri net, a process design and an animation that shows the paths each instance follows. Process miner options such as default miner, all operators miner, directly-follows miner, lifecycle miner are available so that user can choose one of them depending on the needs of the discovered model. All operators miner studies the model in terms of users who

execute the task. Directly-follows miner, the next option, results in a model that connects task if one task follows straight the next task (Augusto et al., 2019). Finally, Lifecycle miner discovers the business process by evaluating the life cycle transitions of tasks such as start and complete (Leemans, Fahland & Van der Aalst, 2016). This allows to compute both the sojourn time and service time.

The resulting model can be evaluated by the user and more filters may be used to exclude unnecessary traces from the model. These filters can also be used to display only the selected instances for more comprehensive investigation of these instances. The selections for display option enable to evaluate the process in various ways such as frequency, sojourn time, service times, and deviations from the model. Here, sojourn time refers to the waiting duration between two tasks.

In Figure 8, blue rounded rectangles are activities. The numbers in rectangles and on arrows imply how many traces are visited these activities or connections. Darkness of these rectangles and arrows express the relative frequency levels, i.e., the darker the blue is the higher the relative frequency of the item. This process starts with “register application”. And then, there are three parallel activities; “check credit”, “calculate capacity” and “check system”. These are indicated with a parallelism symbol with a plus sign. The next branch with a point symbol means a choice between following activities “accept” and “reject”. After this choice, process ends with the activity “send decision e-mail”.

Numbers indicating the frequency of activities are further explained in Figure 9 where three parallel activities are shown. In theory, parallelism implies that all activities that enter the parallelism symbol should pass from the all the parallel branches. Accordingly, one would expect 100 cases on the parallel activities “check credit”, “calculate capacity” and “check system”. However, it is seen that although

only ninety cases out of hundred passes from the activity “check system”, IvM still shows these three activities as parallel activities. Doing this enables the user to discover the process and gain insight effortlessly.

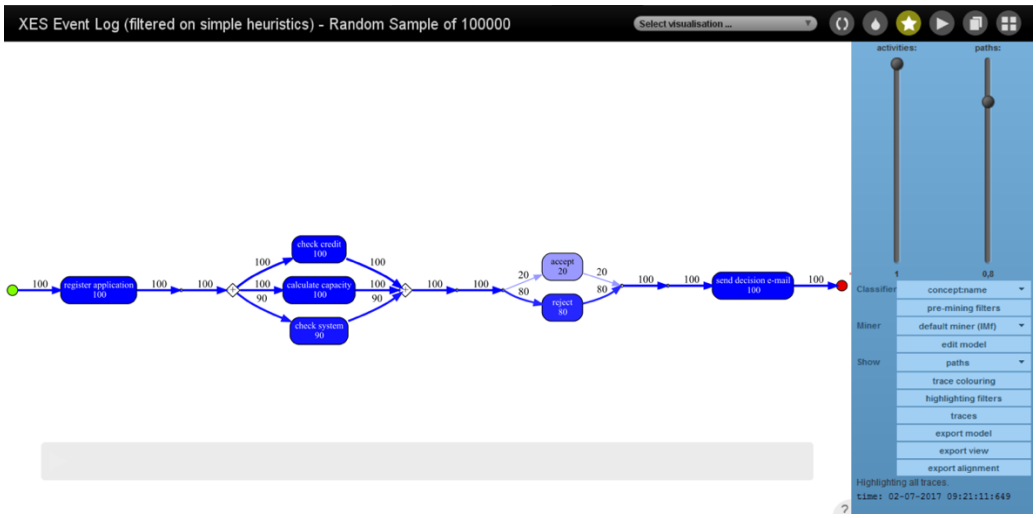


Figure 8. An example process discovered with IvM

Secondly, in the process discovery step, the output of IvM shows the number of cases that goes through an activity or path as well as the throughput time of an activity. These are important features since users can analyze the frequencies and durations to evaluate the reality of the processes. Moreover, various filtering options enable users to examine the processes in detail and to gain insights.

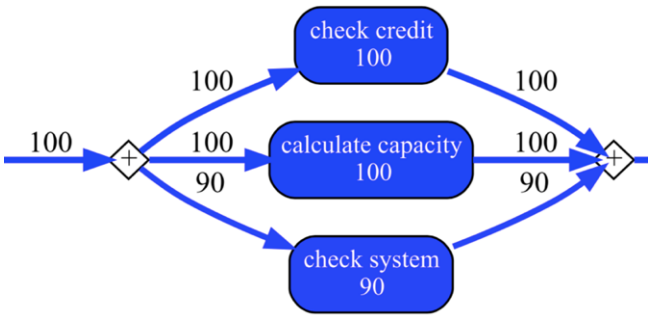


Figure 9. An excerpt of process model created that displays parallelism.

Finally, IvM may also create a petri net model so that the output of this miner can be used by other conformance checking and enhancement applications. On the other hand, although IvM has many advantages, execution time is much larger than other mining algorithms especially for large event logs. Nevertheless, as IvM is the most beneficial plug-in among other miners explained in this section, it is used in this thesis study.

CHAPTER 4

INTRODUCTION OF THE CASE STUDY

The inspiration of this thesis originates from a real process problem in an international distribution company employed in construction machinery sector in Turkey. In this chapter, we introduce this company and its related business process problem that constitutes the basis for our case study in this thesis.

The company provides many solutions including sales, leasing and after sales support services for the construction machinery they distribute. Providing maintenance and repair services for their machinery are the critical services provided by the distribution company. First of all, failures of construction machines are very costly for the construction companies since the delays in meeting the project completion deadlines are subject to severe fines. Furthermore, noting that the number of occurrences of these repair and maintenance processes is very high, these services are critical for construction machinery producers and distributors.

The maintenance and repair service in the distribution company is referred to as service sales process (SSP) as it includes sales activities. SSP is initiated by the customer by making a call to the call center. This process is referred to as ticket creation where the call center employee receives and evaluates this call. If the problem can be solved remotely during the call, then the ticket is closed and information is not transferred to the ERP system, SAP. Otherwise, a work order (WO) is created for further analysis where a service specialist checks the issue on-site. According to the findings of the service specialist, new items may be added to the work order. After the approval of the customer, a sales order is created and transferred to SAP Enterprise Central Component (ECC) for the following sub

processes such as the supply of the spare parts and planning. With the execution of maintenance and repair services, service confirmation is created. After the completion of service sales actions, the necessary sub processes on payment and invoices are executed in SAP.

As mentioned before, the maintenance and repair services should be provided seamlessly since direct communication with the customer is critical for customer satisfaction. Therefore, SSP is executed with the support of IS which requires to follow up the process model. Even though the business process model for SSP is defined and implemented to SAP, the process owners and IT consultants want to discover the real-world practices of their SSP and check whether they match with the proposed process model or not.

The urge of process owners to decrease the throughput time and increase the performance lead them to discover the process using the data that is available in the ERP system (SAP CRM and ECC). The yearly transactions data obtained from the ERP system is used throughout the course of the study. These transactions belong to the SSP operations that took place in Turkey in a selected previous year.

In the next chapter, we analyze the selected SSP case under study and then we apply process mining procedures to discover the process model, check its conformance and generate ideas for enhancement.

CHAPTER 5

PROCESS MINING APPLICATION

In this chapter, we reengineer the business case introduced in the previous chapter by following the steps of PM^2 methodology explained in Chapter 3.

The Mining and Analysis step of PM^2 methodology is performed by using two different tools; ProM and Celonis. In the next Chapter 6, we provide a comparison of the performance of these tools. Details of the methodology and its execution are explained in the following subsections.

5.1 Planning

The planning stage includes the selection of the process, determination of the research questions, and composing the project team. The process that is selected for this thesis study is already introduced in Chapter 4. In this section, the details of the selection process are explained. Moreover, the research questions are summarized, and the project team is constituted.

For the process selection, prioritization analysis is performed according to the study of Page (2015). Page (2015) states the categories for evaluating the priority of the processes as impact, implementation, current state, and value. These main categories can be scored using criterion as listed in Table 3. The scales and the weight of the categories are selected by considering the objective of the case study and process mining applications. As seen in Table 3, only one process is evaluated, as there is not any other process to compare for prioritization. As SSP is one of the main solutions that are provided by the organization, the impact is in terms of the number of people affected. As the internal customer of this process is the

management team and below, the client level is scored as 2. The evaluation on implementation criterion namely time to market, funding and timing of next cycle shows that the process has almost average score. The reason is that the process is supported by SAP CRM and SAP ECC and implementations in SAP need time and funding. Assessing the current state reveals that customer satisfaction is very important for this process whereas the pain level is also high. The score for existence of the process is zero. In the value category, the benefit of the process is high as the process is a part of sales processes and has a high effect on the organization's success. With considering the weights of the categories, the total score is calculated as 5. The possible scores of this process prioritization framework are between 2.05 and 6.45. Thus, the total score of SSP confirms that it has high priority.

Table 3. Process Prioritization Table

Category	Criteria	Scale	SSP
Impact (35%)	Number Affected	3 =large 2 = medium 1 = small	3
	Client Level	3 = senior 2 = management 1 = other	2
		Subtotal	5
Implementation (20%)	Time to Market	3 = short 2 = average 1 = long	2
	Funding	3 = low 2 = medium 1 = high	1
	Timing of Next Cycle	3 = close 2 = intermediate 1 = far	2
		Subtotal	5
Current State (30%)	Client Satisfaction	3 = low 2 = medium 1 = high	3
	Pain Level	3 = high 2 = medium 1 = low	3
	Process Exist	1 = no 0 = yes	0
		Subtotal	6
Value (15%)	Benefit / Return	3 = high 2 = medium 1 = low	3
		Subtotal	3
		Total Score	5

Secondly, there are two main research questions that direct the case study: what the reality of the SSP is and how the throughput time can be shortened.

Finally, the project team including a business analyst, process experts, and an IT consultant is constituted. The roles of the members are defined. The business analyst conducts the data preparation and process mining applications. The process experts act as both process owner and process expert, and offer consulting service for their process know-how upon request. The IT consultant supports the extraction stage and responds other IT related requests.

5.2 Extraction

Extraction step aims to acquire data that are needed for process mining applications. For this purpose, scope determination and process knowledge transfer are required should be transferred. Thus, scope is determined with the process owners in the company and knowledge related to selected business process is shared by process experts with us.

To perform the data extraction in the project, first the IS of the organization is explored. The organization uses SAP ECC as ERP system to execute most of the main processes. SAP CRM is also used by the sales and service teams. So, the information for service order process is stored both in SAP ECC and CRM. First part of the process, before the creation of a sales order is handled and recorded in SAP CRM. The rest of the process is executed and recorded in SAP ECC system. Therefore, data from both systems are extracted.

SAP stores transactional data in various tables. Every action is stored in a different table to increase the performance of database management. For the selected process, activities start in SAP CRM. Documents are created and saved as service

orders in CRM and then transferred to ECC as sales order documents. The transactional data are stored in Sales Document Header Data (VBAK). This table records only header information that is common for every item in the order. Item data are stored Sales Document Item Data (VBAP). These tables have the latest version of documents. The field changes and status changes are saved in other tables.

In the extraction phase of the thesis study, the document tables as well as the change and status tables from both ECC and CRM are extracted. These extracted versions of the tables include the transaction data for a span of one-year. One-year time interval for data extraction is selected in consideration with the expected duration of the sales process. In practice, the expected duration of the selected sales service process is more than a month. There may also be seasonal changes in the process execution that cause fluctuations in the service durations. Thus, data collected in one year give us the chance to obtain a valid estimate of the expected duration for the selected sales process.

Next, we provide the following numbers to give an impression of the size and the complexity of the big data used in this study. Data from more than hundred tables are extracted. A new database is created using Microsoft SQL Server to handle the huge amount of data. Total size of database is approximately 200 gigabytes. When we searched through descriptions of the tables and their column names within the extracted tables, it is decided that 34 tables are worth to analyze further to generate the event log as it can be seen in Table 4.

Table 4. Extracted SAP Table List

SAP Table Name	Description	Number of Columns	Number of Rows
CDHDR	Change document header	15	576.368
CDPOS	Change document items	16	15.317.543
COMM_PRODUCT	Master Table for Product	14	2.266.393
COMM_PRSHTEXT	Product Description	10	4.527.151
CRM_JCDO	Change Documents for Status Object	9	13.558
CRM_JCDS	Change Documents for System/User Statuses	10	56.022.946
CRM_JEST	Individual Object Status information	5	150.009.895
CRM_JSTO	Status Object Information	6	72.323.535
CRMD_BINREL	Interlinkages Between CRM Application Objects	6	11.194.029
CRMD_BRELVONAE	Additional attributes: Object interlinkage VONA	19	11.266.724
CRMD_LINK	Transaction - Set - Link	5	88.681.569
CRMD_ORDERADM_H	Business Transaction Header	152	483.157
CRMD_ORDERADM_I	Additional Site Details at the Item Level of a Service Contract	135	4.520.012
DD02T	SAP DD: SAP Table Texts	5	113
DD03M	Generated table for view	51	4.873
DD07T	DD: Texts for Domain Fixed Values	9	220.925
JCDS	Change Documents for System/User Statuses	12	3.165.050
LIKP	SD Document: Delivery Header Data	220	164.108
LIPS	SD document: Delivery: Item data	304	1.322.953
SRRELROLES	Object Relationship Service: Roles	7	16.023.605
T001	Company Codes	79	32
T003	Document Types	38	70
T003T	Document Type Texts	4	71
TJ02T	System status texts	4	3.544
TJ30	Status	11	1.048
TJ30T	Texts for User Status	7	3.436
TSTCT	Transaction Code Texts	3	360.234
TVAK	Sales Document Types	173	284
TVAKT	Sales Document Types: Texts	5	1.005
VBAK	Sales Document: Header Data	232	205.767
VBAP	Sales Document: Item Data	361	1.923.173
VBFA	Sales Document Flow	43	6.566.172
VBRK	Billing Document: Header Data	151	164.601
VBRP	Billing Document: Item Data	257	1.835.405

5.3 Data processing

The aim of data processing is to generate the event log which will serve as the basis for further analysis. As introduced in Table 2, an event log includes headers for cases, activities and timestamps. To create the event log from the data extracted in the previous section, we need to select the key for the cases and identify the activities to be included for further analysis.

The selection of the key for the cases requires that a case identifier should be determined at this phase to serve as the primary key in merging various tables obtained in the extraction phase. Eck et al. (2015) state that case notion, in other words case identifier should be decided before processing the data. Indeed, this is a very critical decision since the all remaining analysis will be based on the selection of this case identifier.

Figure 10 summarizes subprocesses for the sales process together with the SAP ECC and CRM that holds the related data. The data stored in CRM side is related to three subprocess for work order, sales order and confirmation. CRM transactional data include some cases that haven't reached to sales order subprocess. Hence, these incomplete transactions should be ignored or given special care. This can be achieved by using the case id of sales order as the base of the further analysis. Sales order id is also used in SAP ECC as the reference for the documents that will be generated after sales order subprocess. This shows that sales order id appears in all transaction data stored in ECC. So, if sales order id is the key to track a trace both in CRM and ECC. Hence, sales order id is a good candidate for case identifier.

Yet, a comprehensive evaluation on sales order id shows that a sales order may contain more than one item and these items can be processed independently from each other. For example, a sales order may include more than one item type.

Although all item types all have the same sales order id, they may follow different business process models. In this case, sales order id is not enough to trace the transactions of different item types that appear under the same sales order. Thus, to analyze business process properly in the item level, sales order item id is chosen as case identifier. Although selecting the identifier in the item level will significantly increase the number of cases to be included in the analysis, the findings will be more valid and more applicable.

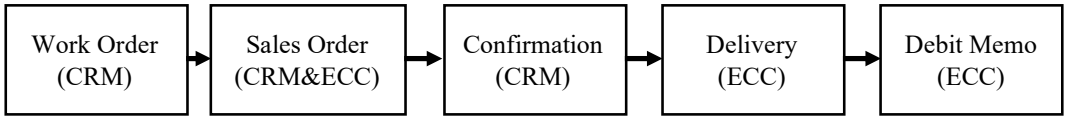


Figure 10. Subprocesses of SSP

The second decision in the data processing phase is to identify the activities in SSP to be included in further analysis. The first step for activity identification is discovering the documents that are linked to the SSP. The term document is used in SAP for a record of a posting. For example, work orders are documents that are created by entering necessary information about service order requests of customers. SAP has "following document tables" that record the linked documents and their document types. A document type is a unique code given to a document that corresponds to a subprocess. This enables us to list the documents related to the SSP cases that are considered in this case study. As seen in Table 5, fifteen document types which indicate various subprocesses are found. The corresponding SAP table names of these document types are listed as the third column. Using this list, tables that are extracted in previous stage are examined and aggregated or filtered when necessary. The tables that are listed in Table 5 are used to collect information on

subprocesses to form the event log. Only RESB table that stores information related to goods movement subprocess is not used since we were not able to extract this data from the company. It is important to state that this exclusion do not affect the duration of cases but only hides the details about goods movement subprocess.

Table 5. Document Types Included in SSP

SAP Document Type	Document	Related SAP Tables
C	Sales Order	VBAP
H	Returns	VBAP
J	Delivery	LIPS
K	Credit Memo Request	VBAP
L	Debit Memo Request	VBAP
N	Invoice Cancellation	VBRP
O	Credit Memo	VBRP
P	Debit Memo	VBRP
R	Goods Movement	RESB
S	Credit Memo Cancellation	VBRP
T	Returns Delivery for Order	LIPS
BUS2000115	CRM Sales Order	CRMD_ORDERADM_I
BUS2000116	Service Order	CRMD_ORDERADM_I
BUS2000117	Confirmation	CRMD_ORDERADM_I
BUS2000223	Service Incident	CRMD_ORDERADM_I

After deciding on subprocesses, tasks that are relevant to our study are identified. A task is an operation performed on any entry in a document. This can be either an Insert (Create) or a Change (Update) or a Delete operation. The identification of the tasks corresponding to each document type in Table 5 is made by using the information in the extracted tables CDHDR (Change document header), CDPOS (Change document items) that appear in Table 4. These tables include the information for the time and the type of the task applied to a transactional data in a document. Additionally, several status changes table that appear in Table 4 are used

to identify the status changes in each document type. Examples include changing a confirmation status from open to complete or cancelled internally.

As a result of the consolidation of all the tasks and fourteen subprocesses determined above, 28 activities that are considered to form the event log are listed in Table 6.

Table 6. Activity List to Be Included in the Event Log

	Activity
1	Confirmation Status Change to Billing Cancelled (I1095)
2	Confirmation Status Change to Cancelled internally (I1096)
3	Confirmation Status Change to Completed (I1005)
4	Confirmation Status Change to for billing (I1072)
5	Confirmation Status Change to Fully Billed (I1073)
6	Confirmation Status Change to Open (E0002)
7	Confirmation Status Change to Open (I1002)
8	Create Credit Memo
9	Create Debit Memo
10	Create Return for Item
11	Create Sales Order Item
12	Create Service Incident
13	Create WO Item
14	Credit Memo Cancellation
15	Credit Memo Request
16	Debit Memo Request
17	Delivery
18	Invoice Cancellation
19	Return Delivery for Order
20	Service confirmation
21	Update Item
22	Update Item Net price
23	WO Header Status Change to Cancel (E0004)
24	WO Header Status Change to Closed (E0006)
25	WO Header Status Change to Estimate (E0001)
26	WO Header Status Change to Invoice (E0008)
27	WO Header Status Change to Open (E0002)
28	WO Header Status Change to Part Invoice (E0007)

Next, the activities in Table 6 are discussed with the process experts in the company in terms of their validity and significance. The experts suggested that some

of these activities might be excluded since they are system updates that do not affect process execution. Accordingly, the activities that imply status changes in confirmation subprocess such as for billing (I1072), fully billed (I1073), and open (E0002) in Table 6 are excluded. Hence, we have 25 activities to be included in the event log.

The second suggestion made by process experts is introducing the Update Item activity in Table 6 which is indeed composed of fifty different update activities about changes in sales order items such as Order Quantity, Business Area, Loading Time, and many others. These numerous update activities that are executed similarly are grouped as "Update Item" to prevent overfitting and reduce complexity in the process mining and analysis. However, the activity "Update item net price" is not included in this grouping as price change in service sales orders may indicate a different behavior than updates in any other fields.

After the selection of the key for the cases and finalization of the activities to be included in further analysis, the database tables in Table 4 are mapped to form the event log.

At this point, we needed to filter some irrelevant data. The company has branches in some countries among which process execution may differ. To exclude the variances due to the country, data from Turkey are filtered and added to the event log. Moreover, the tables that appear in Table 4 include all types of transactional data related to SSP cases that are not essential for process mining purposes of the current study. Hence, to reduce the complexity of the big data on hand, only the case identifier, activity, and timestamp variables are used to create the event log.

Table 7 lists an excerpt of the event log which consists of 471,325 cases with 3,991,502 events.

Table 7. An Excerpt from the Event Log (YYYY Is the Hidden Year Entry in the Timestamp Data)

CASE	ACTIVITY	TIMESTAMP
Case 1	Create WO Item	YYYY0103060530
Case 1	WO Header Status Change to Estimate(E0001)	YYYY0117075317
Case 1	Create Sales Order Item	YYYY0117075325
Case 1	Service confirmation	YYYY0524075929
Case 1	Confirmation Status Change to Completed(I1005)	YYYY0524105930
Case 1	Debit Memo Request	YYYY0524105943
Case 1	WO Header Status Change to Open(E0002)	YYYY0524110006
Case 1	WO Header Status Change to Closed(E0006)	YYYY0524110006
Case 1	Create Debit Memo	YYYY0524115034
Case 2	Create WO Item	YYYY0103114632
Case 2	WO Header Status Change to Estimate(E0001)	YYYY0111164642
Case 2	WO Header Status Change to Open(E0002)	YYYY0111164642
Case 2	Create Sales Order Item	YYYY0111164703
Case 2	Service confirmation	YYYY0124083934
Case 2	Confirmation Status Change to Completed(I1005)	YYYY0124113935
Case 2	Debit Memo Request	YYYY0124113952
Case 2	Delivery	YYYY0125122427
Case 2	Create Debit Memo	YYYY0125122442
Case 3	Create WO Item	YYYY0824051813
Case 3	WO Header Status Change to Estimate(E0001)	YYYY0824081815
Case 3	Create Sales Order Item	YYYY0824081823
Case 3	Service confirmation	YYYY0926083208
Case 3	Confirmation Status Change to Completed(I1005)	YYYY0926113413
Case 3	Confirmation Status Change to Open(I1002)	YYYY0926113413
Case 3	Debit Memo Request	YYYY0926113440
Case 3	WO Header Status Change to Closed(E0006)	YYYY0926114057
Case 3	WO Header Status Change to Open(E0002)	YYYY0926114057
Case 3	Create Debit Memo	YYYY0926185324
Case 4	Create WO Item	YYYY0130173108
Case 4	WO Header Status Change to Estimate(E0001)	YYYY0130203119
Case 4	Create Sales Order Item	YYYY0130203126
Case 4	Service confirmation	YYYY0208084611
Case 4	Confirmation Status Change to Completed(I1005)	YYYY0208142239
Case 4	Confirmation Status Change to Open(I1002)	YYYY0208142239
Case 4	Debit Memo Request	YYYY0208142256
Case 4	WO Header Status Change to Closed(E0006)	YYYY0208142933
Case 4	WO Header Status Change to Open(E0002)	YYYY0208142933
Case 4	Create Debit Memo	YYYY0208160528
Case 4	Invoice Cancellation	YYYY0208160911
Case 4	Create Debit Memo	YYYY0209084335
Case 5	Create WO Item	YYYY0914064934
Case 5	Service confirmation	YYYY0914074305
Case 5	Create Sales Order Item	YYYY0914094955
Case 5	Confirmation Status Change to Completed(I1005)	YYYY1012160839
Case 5	Confirmation Status Change to Open(I1002)	YYYY1012160839
Case 5	Debit Memo Request	YYYY1012160902
Case 5	Create Debit Memo	YYYY1012163956
Case 5	WO Header Status Change to Open(E0002)	YYYY1013173733
Case 5	Update Item	YYYY1013173749
Case 5	WO Header Status Change to Closed(E0006)	YYYY1027112131

In essence, in data processing stage, almost 450 million lines of extracted data from SAP are processed to form an event log of four million lines.

5.4 Mining and analysis

This stage starts with the process discovery and then follows with the comparison of the discovered process model with the organization's as-is model. After that, enhancement for process improvement is applied. Finally, some visual analytics are prepared to display various characteristics of the SSP.

Execution of process mining types are performed by using a commercial tool Celonis Academic Cloud and an open source tool ProM version 6.8 as explained in Chapter 3. Mining and analysis by using personal computer with 2.20 GHz Intel Core i5 processor and 8 GB memory.

To begin with the application, the event log as in Table 8 is extracted in CSV file format from Microsoft SQL Server. Then, it is uploaded to both ProM and Celonis. In the following, we will explain the execution of Mining and Analysis as introduced in PM^2 Methodology (Eck et al., 2015) in Chapter 3.

5.4.4 Process discovery

As explained in Literature Survey in Chapter 2, process discovery aims to discover the business process model using the event log. Here, we import the same event log to ProM and Celonis to model SSP using plugins in these tools. Process discovery is performed by using the process explorer application in Celonis. Moreover, several plugins including IvM are used for process discovery in ProM.

Process Discovery by ProM

The first tool that will be utilized here is ProM Framework. The input file format for ProM process mining plugins is XES (eXtensible Event Stream). It is an XML-based format that is used for event logs and follows the universal structure of event log information. As we have extracted the event log in csv file format, we first convert the file using “Convert CSV to XES” plugin as seen in Figure 11.

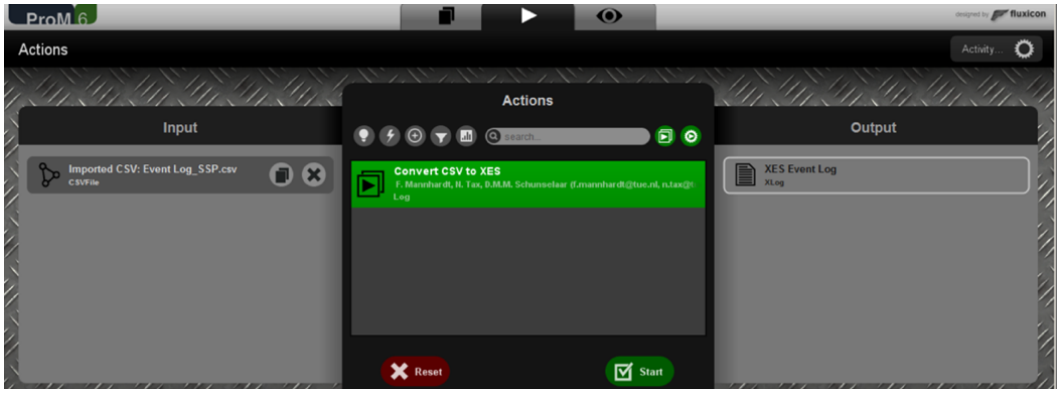


Figure 11. Screenshot of plugin selection in ProM

The output of this tool shown in Figure 12 displays general information on the event log. Accordingly, there are 471,320 cases and 3,991,503 events in our event log. Noting that many activities occur several times in many cases, the distinct count of activities is 25. The graphics on the top displays the minimum (2), maximum (41) and average number of events per case as the height of the bars. The width of the bars corresponds the frequency of occurrence of that number of events in a case. Additionally, the graphics on the bottom shows the minimum (2), maximum (17) and average (8) number of event classes per case, i.e. activities per case, as well as their frequency of occurrence in a case. In a case there can be at most 41 events. Moreover, in a case there can be at most 17 different activities. The frequency of occurrences of 41 events and 17 activities can be followed from the

width of respective bars in the above and below graphics which indicate that these extremes are very infrequent. Indeed, we deduce that most of the cases include 8 events and 8 activities.

Figure 12. Screenshot of dashboard that is the output of XES converter plugin in ProM

As it is seen on the left side of the Figure 12, there are two other tabs than dashboard; inspector and summary. Inspector screen shows detailed information for each case in the event log. Due to the privacy concerns of the organization regarding transactional data of orders, we will not display the picture of it. The last tab summary lists all 25 activities, start activities, and end activities with their frequency of occurrences and frequency of relative occurrences. Part of this screen is given in Figure 13 and the complete report is given in Appendix A, Table A1.

In Appendix A, Table A1, all activities that are observed in the event log, start activities and end activities of the cases are listed. Start events and end events indicate the first and last tasks of the cases, respectively. At this point, we shared the list for all activities, start activities and end activities with the process experts.

- In the report, it is seen that there are start activities other than the expected start activity “Create WO Item”. As the event log is prepared for the cases that have already passes work order creation step, it can be understood that there are some cases in which other activities are executed before the activity “Create WO Item”. This is noted as FACT1 for further analysis in enhancement and evaluation stages.

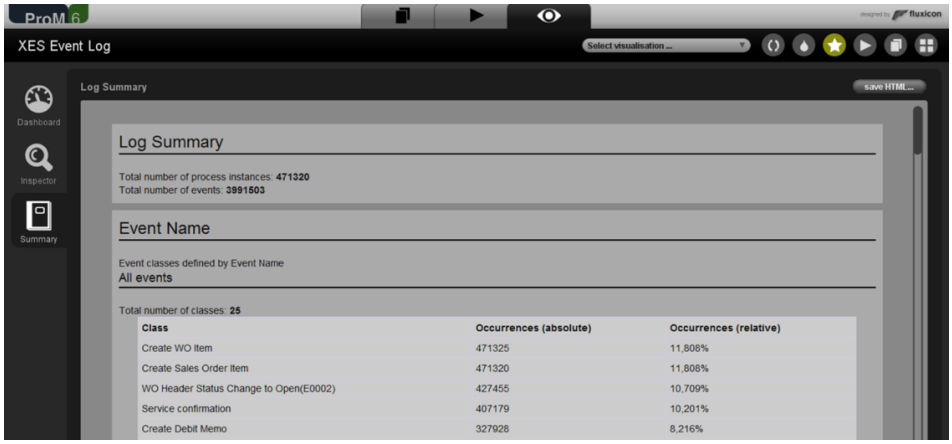


Figure 13. Screenshot of log summary in ProM

With the investigation of the end activity list in the summary report, it is found and stated that some cases in the event log are incomplete and they may lead to incorrect results in the following analyses. So, end activities are analyzed with the process experts to distinguish unfinished cases that are needed to be excluded from further analysis. As seen Table 8, cases that ends with one of the seven end activities marked as “excluded” are not included in further analysis.

- So, the cases with end activities namely “Create Sales Order Item”, “Update Item”, “Confirmation Status Change to Open(I1002)”, “Service confirmation”, “WO Header Status Change to Open(E0002)”, “WO Header Status Change to Estimate(E0001)”, “Update Item Net price” are

excluded from further analysis. The high number of incomplete cases is also noted for further analysis as FACT2.

The exclusion task is performed by using Filter Log Using Simple Heuristics plugin in ProM as this plugin gives the options to filter out with respect to start activity and end activity. A filtered event log that consists of 348,712 cases with 3,199,724 events is created as seen in Figure 14. It can be seen that after filtering some cases, maximum number of events per case decreases to 39 from 41 whereas maximum number of activities per case is unaffected. The average number of events per case increases from 8 to 9 likewise the average number of activities per case becomes 9 as well. The minimum values for the number of events and activities also increase to 3 from 2.

Table 8. End Activities Analysis List

Activity	Occurrences (absolute)	Exclusion Evaluation
Create Debit Memo	190427	included
WO Header Status Change to Closed(E0006)	109206	included
Create Sales Order Item	45260	excluded
Update Item	43288	excluded
WO Header Status Change to Invoice(E0008)	39424	included
Confirmation Status Change to Open(I1002)	23053	excluded
Service confirmation	5835	excluded
WO Header Status Change to Open(E0002)	4469	excluded
WO Header Status Change to Part Invoice(E0007)	2347	included
Delivery	2035	included
WO Header Status Change to Cancel(E0004)	2013	included
Confirmation Status Change to Cancelled internally(I1096)	1150	included
Create Credit Memo	627	included
Confirmation Status Change to Completed(I1005)	589	included
WO Header Status Change to Estimate(E0001)	517	excluded
Invoice Cancellation	509	included
Debit Memo Request	313	included
Update Item Net price	186	excluded
Credit Memo Cancellation	40	included
Credit Memo Request	32	included

Using the filtered event log, we tried to implement IvM for process discovery. However, the process model could not be created with IvM since the data size was too large. Due to this performance issue, the size of the event log is needed to be reduced by using “Extract sample of traces (Random)” plugin in ProM. This plugin enables us to randomly select a given number of cases. After several iterations, the maximum number of cases that can be processed in IvM is identified as 100,000 and rest of the study is made by using this randomly filtered event log.

Figure 14. Screenshot of dashboard for the filtered event log in ProM

Figure 15 displays the dashboard of the filtered and reduced version of the event log. The details of the event log are given in Appendix, Table A2. In the final version or our event log there are 100,000 cases with 918,026 events in 24 activities. The activity “Confirmation Status Change to Open (E0002)” does not appear in the filtered event log however this is only observed once in the original event log as seen in Appendix, Table A1. Moreover, as discussed by the process experts, excluding this activity doesn’t introduce a major issue in the further analysis. In the final event log, the minimum, average and maximum events per case are 3, 9, and 36,

respectively. The minimum, average and maximum activities per case are 3, 9, and 16, respectively. This shows that on the average there are nine distinct activities in a case whereas a case may contain maximum 16 distinct activities. However, in some cases, the maximum number of events might be as large as 36 showing that several activities are unnecessarily repeated.

- The recurrence of the activities within a case is noted as FACT3. These will be discovered in enhancement phase.

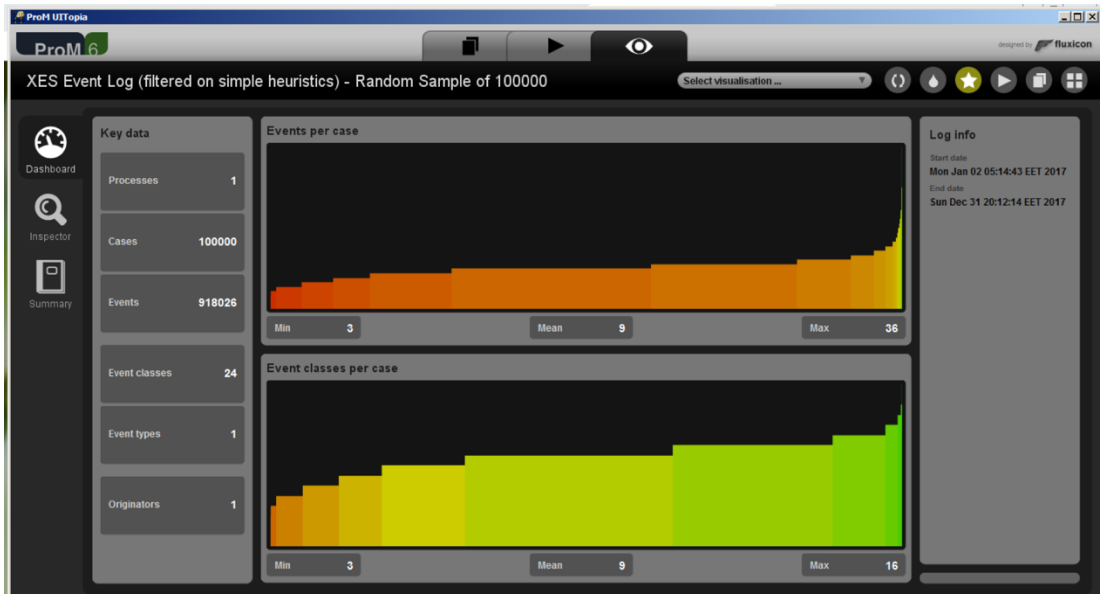


Figure 15. Screenshot of ProM dashboard for 100,000 cases

At this point, our event log is finalized, and we are ready to apply process discovery algorithms. There process discovery algorithms; alpha miner, fuzzy miner, and IvM that are available in ProM are analyzed and compared as explained in Background on Process Mining, in Chapter 3. As IvM is found to be the most comprehensive and beneficial miner among three miners, IvM plugin in ProM is used for process discovery.

IvM has several miner and display options as explained in Chapter 3. For process discovery, we choose to use its default miner with paths and sojourn times display option. As explained in Background on Process Mining in Chapter 3, the default activity and path filters are 1.0 and 0.8 respectively. We have used this default values in our case study. The resulting model is shown in Figure 16 to give a general impression on how the model is. The proposed model is zoomed out in three parts (See Appendix B).

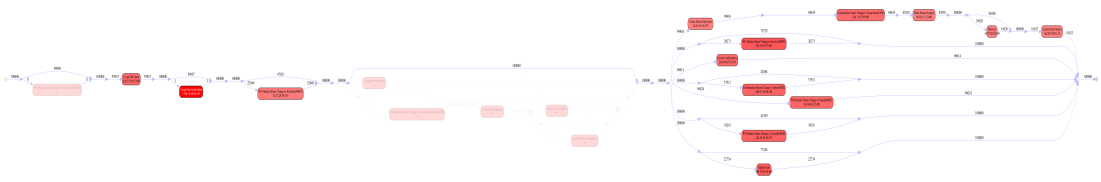


Figure 16. The process model with paths and sojourn time created by IvM with 1.0 and 0.8 activity and path filter levels respectively

In Figure 16, red rounded rectangles are activities. The numbers in rounded rectangles imply the average sojourn times and the numbers on arrows entering an activity denote the number of occurrences of that activity in event log. As we only have the completion time of the task in the event log, sojourn time refers to the sum of waiting duration between two tasks and the duration of the second task. The red color intensity of these rectangles expresses the length of sojourn time, i.e., the darker the red color is, the longer the average sojourn time of the activity will be realized.

- The darkest red colored activity is “Create Service Incident” (See Appendix B). The average sojourn time of this activity is almost 176 days. Although this duration is very high, the number of cases that are exposed to this duration is only three. These extreme cases are noted as a

fact in the process flow to be discussed with the process owners in the enhancement phase as FACT4.

- Another activity “Create WO Item” has also almost 23 days of an average sojourn time. As the expected first activity is “Create WO Item”, this loss of time will be discussed as FACT5.
- The average sojourn times of remaining activities that increase the throughput time of the cases are noted and will also be discussed for process improvement as FACT6.

Moreover, with the selection of path filter as 0.8, four activities are not displayed in the model as their entering and leaving arrows are below the significance level. The other 20 of 24 activities are displayed in the model where six of them are shown transparent as listed in Table 9. See Appendix B, for the activities that are transparent and do not have average sojourn times. For example, the activity “WO Header Status Change to Part Invoice (E0007)” is transparent since it is the first in 11 cases. Hence, their sojourn times cannot be computed.

Table 9. Transparent Activity List in Figure 16

WO Header Status Change to Part Invoice (E0007)
Update Item Net Price
Confirmation Status Change to Cancelled internally (I1096)
Credit Memo Request
Create Credit Memo
Credit Memo Cancellation

Additionally, there are activities that are transparent, and their number of occurrences are zero (See Appendix B). These show that the activity is a part of the rework loop. Proposed model by IvM prevents rework and recommend a position in the map for these kinds of activities. In our study, the number of cases that follow

these paths are low. Thus, they are shown transparent. We know that the transparent activity “Update Item Net Price” is observed in the event log as seen in Table 8. So, we analyze further the reason why this activity might be displayed as transparent. It is found that in the event log, this activity follows seven of the activities that are modeled after this activity and one activity that is modeled before this activity in Figure 16. After the activity “Update Item Net Price”, three activities namely “Update Item”, “Service Confirmation”, and “Confirmation Status Change to Open (I0002)” that are modeled after it in Figure 16 are returned. As the proportion of the cases that include this activity and enter to the loop is below the limits (1.0, 0,8) of process discovery, they are shown as transparent in the proposed model. So, in the model, “Update Item Net Price” is illustrated before the next activities in Figure 16 to propose the removal of rework loop.

- The transparent activities in Table 9 are noted as FACT7 to evaluate the reasons for executing them at different positions than the expected positions in the model which causes longer throughput times.

In the right of Figure 16, there are samples of parallel activities and choices that are explained in Chapter 3. For instance, there are seven arrows that comes out of parallelism sign. All activities that are on this followed by these arrows are performed in parallel, i.e., these activities are performed in any order according to the cases in the event log that.

- One of parallel activities WO Header Status Change to Open (E0002) has a longer average sojourn time than many other parallel activities. As it is expected that this status should be observed before activities related to Service Confirmation, this high sojourn time is noted as FACT8 to be discussed with the process owners in the enhancement phase.

Moreover, there are also many examples of choice branch with the point symbol. In Figure 16, these choices are between an activity and an activity free path. For instance, the activity “Update Item” occurs in 22,730 cases and in the remaining 77173 cases this activity is not observed.

- The choices may allow to skip some activities that are identified as mandatory by the process experts such as WO Header Status Change to Closed (E0006), WO Header Status Change to Estimate (E0001). These activities are noted as FACT9 to further analyze the reasons that cause to avoid them.

Celonis

As the second tool, we will utilize Celonis to discover the SSP. Celonis has an academic license version that can be used on cloud. It supports CSV, XLSX, XES and DBF file formats as event log. Therefore, we can either use the original event log extracted from database or the filtered version that has analyzed with process experts. As this study aims to compare two tools, we decide to use the same filtered and reduced event log for process discovery. However, since event log size is reduced because of the performance issue with ProM, we first try to discover process by using filtered event log to observe the capacity of Celonis. As seen in Figure 17, Celonis can discover a process model using 348,712 cases within seconds. This capability of Celonis is noted for comparison of the tools but further mining and analysis steps will be performed with the reduced event log that consists of 100,000 cases as introduced above.

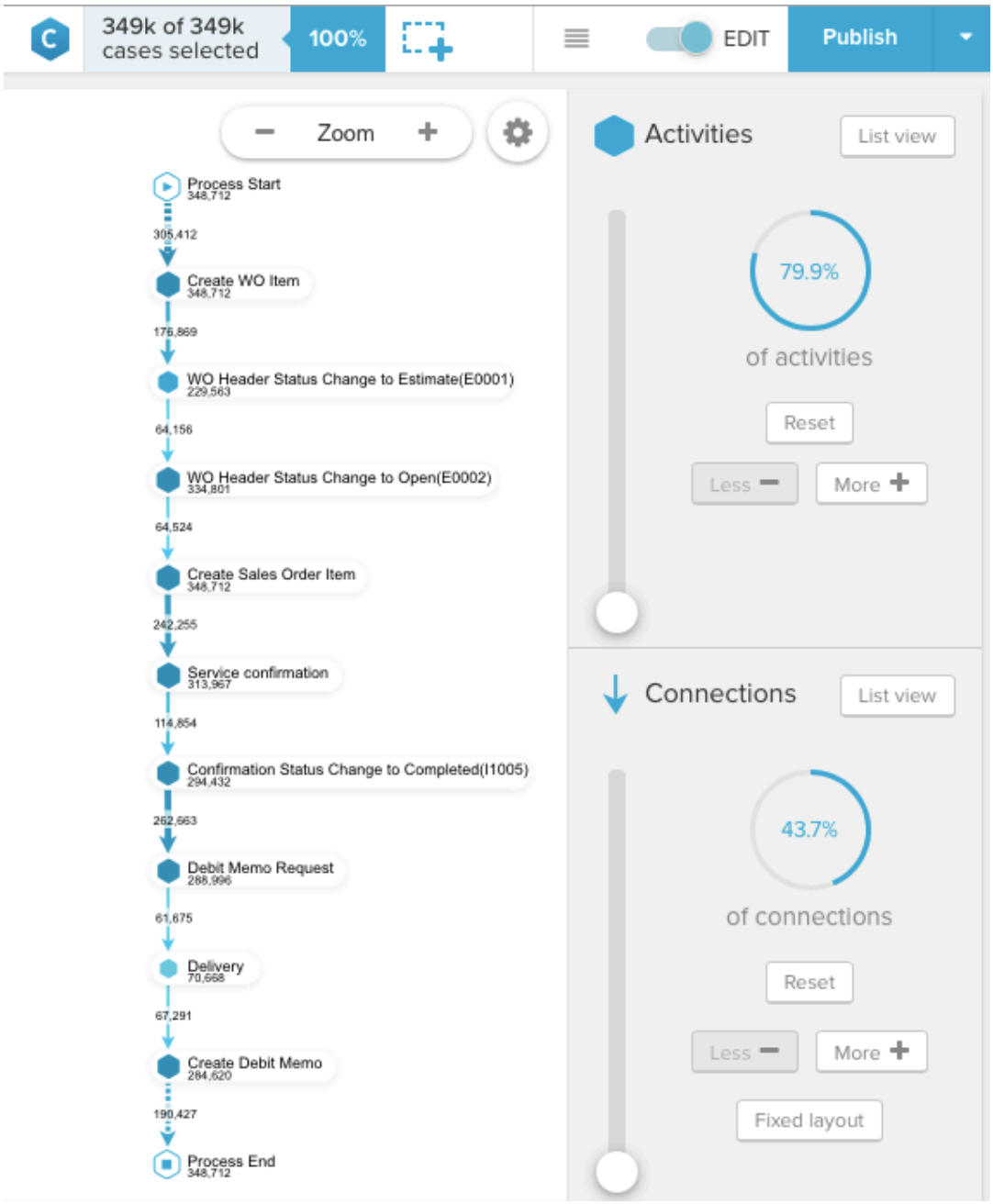


Figure 17. Screenshot of Celonis process explorer for SSP by using 348,712 cases

In Celonis, paths are referred to connections. Similar to ProM, filters can be applied to the activity and connection levels to obtain a lean process map. Figure 18 displays the discovered process model for the SSP by using the reduced event log, i.e., 100,000 cases, with the filter levels suggested by Celonis to discover a lean process flow. As seen on the right of the Figure 18, the filter levels for activities and

connections are 79.9% and 43.7%, respectively. They can be changed by the user to reach the preferred level of inclusion.

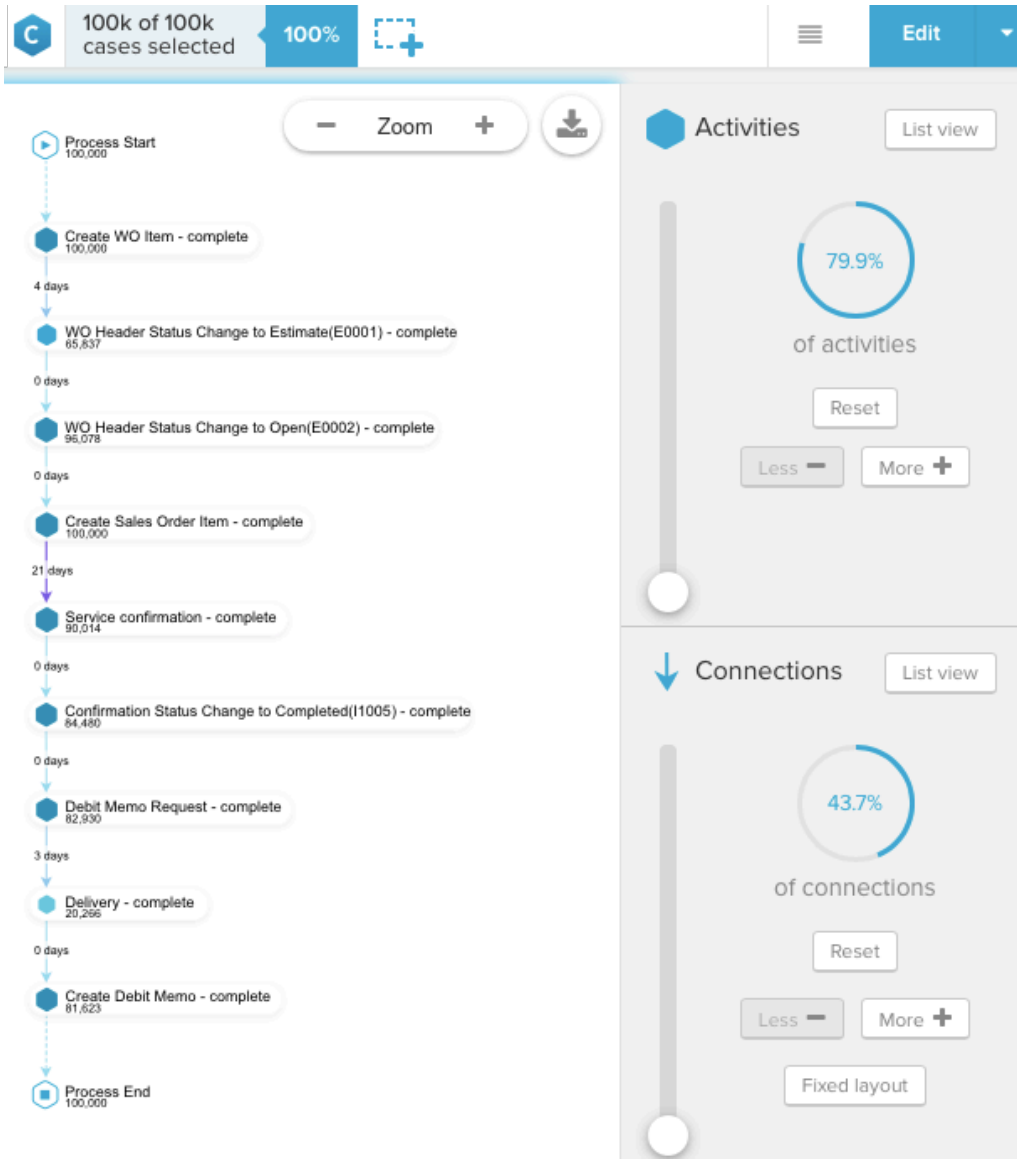


Figure 18. Screenshot of Celonis process explorer with 79.8% of activities and 43.7% of connections

On the left of the Figure 18, the process flow is displayed with nine of the 25 activities in the event log. Activities “Process Start” and “Process End” are added to the flow as first and last activities, respectively. The numbers on the arrows express the average throughput time that is the same with the sojourn time of IvM in our case

study. For instance, it takes 26 days from the completion of the activity “Create Sales Order Item” to the completion of the activity “Service confirmation”. Moreover, the numbers under the activity names denote the number of cases that includes that respective activity. For instance, the process starts with 100,000 cases but the activity “Service confirmation” occurs in 90,014 cases after the activity “Create Sales Order Item”.

As the filter levels of Figure 18 are much less than the IvM filter levels in ProM that are 100% for activities and 80% for paths, we also provide the process model that has the same filter levels with IvM. Figure 19 shows the process model that includes all the activities and 80% of the connections. Apparently, the process model in Figure 19 is more complex than Figure 18. It is not very reasonable to interpret this complex process model in Figure 19. We present this model for the comparison of the process discovery performances of ProM and Celonis with the same filter levels.

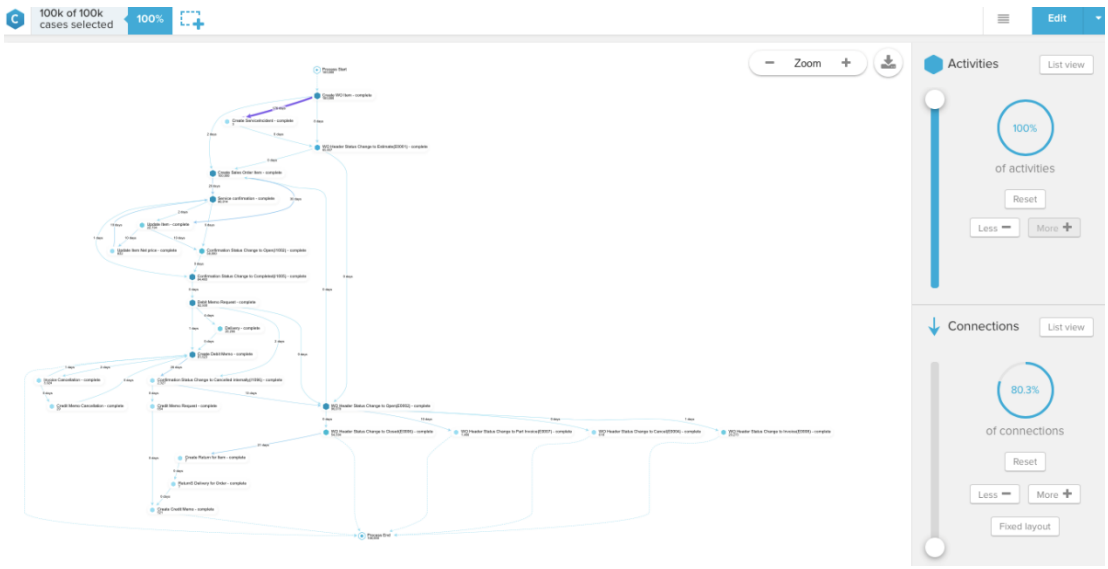


Figure 19. Screenshot of Celonis process explorer with 100% of activities and 80.3% of connections

In addition to the process discovery application shown in Figure 18 and Figure 19, Celonis provides an option to discover the process model with respect to variant selection as seen in Figure 20. On the right of the Figure 20, variants are sorted from most common to least common. The ratio of the covered cases by the variant is displayed with the length of bar and the median throughput time of the variant is written on the right of the bar. The user can select the variants that they want to include to the process discovery. On the bottom of the right side, the number of selected variants and the proportion of covered cases are shown. In Figure 20, the most common variant of the 5,581 variants is selected to discover the process. The median throughput time of this most common variant is 14.1 days. It covers 7% of the cases in the event log. On the left of the Figure 20, the process flow for this variant is displayed where the numbers under the activity is the number of the covered cases and the numbers on the connections are the median throughput time. The connections with longer median throughput time are highlighted by purple arrows.

It is important to mention that although the process flows in Figure 18 and Figure 20 is the same, they do not represent the same number of cases on graphs. In Figure 18, most common path is given as process flow and the cases that fit to this model in any part of the flow is added. However, in Figure 20, only the cases that perfectly fit to the selected variants are displayed. Thus, the number of cases is different in the figures.

Moreover, Celonis provides an algorithmic happy path. A happy path is defined as the flow from the most frequent process start activity to the most frequent process end activity. As seen in Figure 20, only the 7.38% of the cases, in other words 7,375 of 100,000 cases, follow the happy path. The average throughput time

of the cases in happy path is 15 days which 25 days less that the average of the whole 100,000 cases.

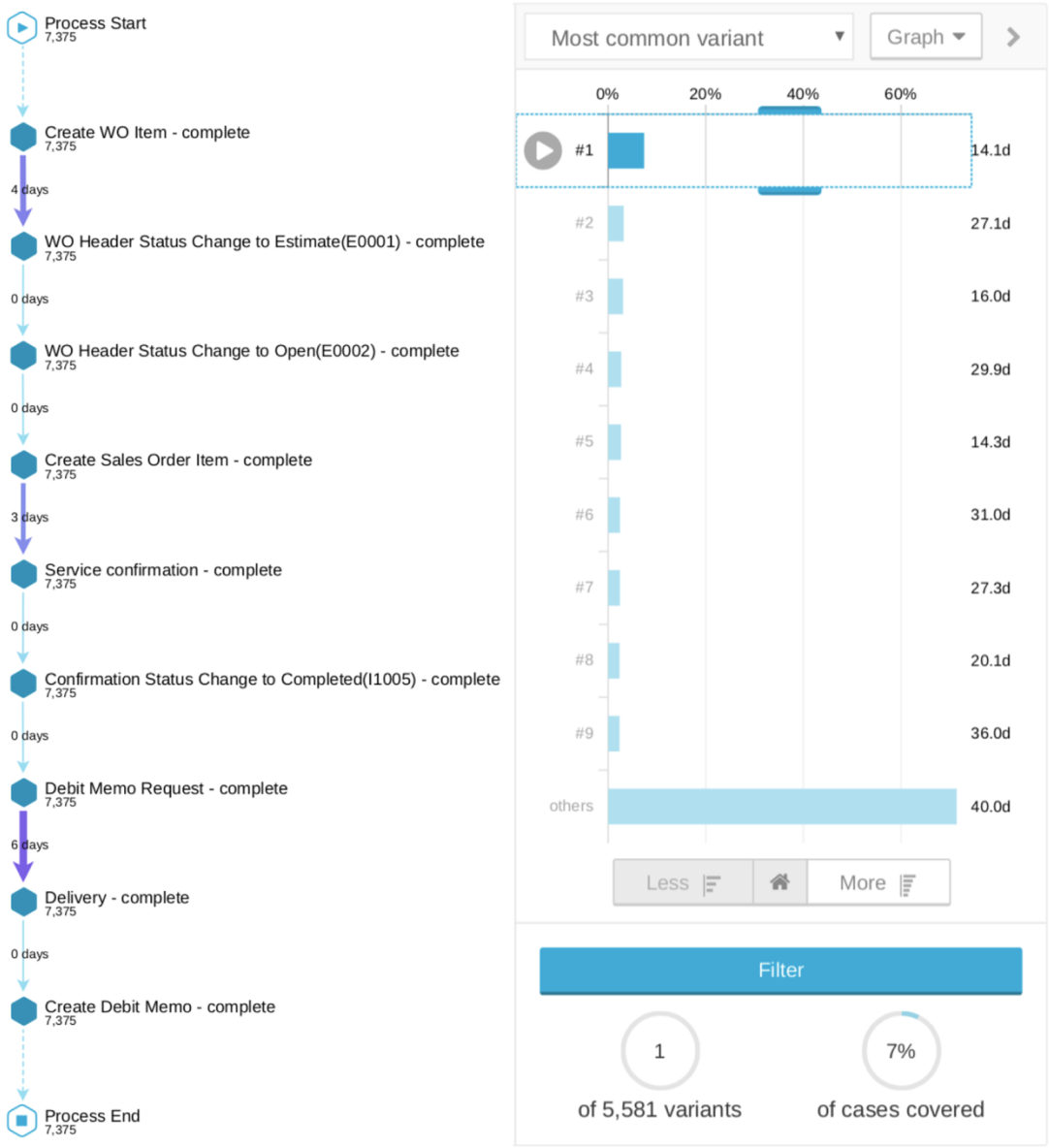


Figure 20. Screenshot of Celonis variant explorer with most common variant filter

Moreover, Celonis provides an algorithmic happy path. A happy path is defined as the flow from the most frequent process start activity to the most frequent process end activity. As seen in Figure 21, only the 7.38% of the cases, in other words 7,375 of 100,000 cases, follow the happy path. The average throughput time

of the cases in happy path is 15 days which 25 days less that the average of the whole 100,000 cases.

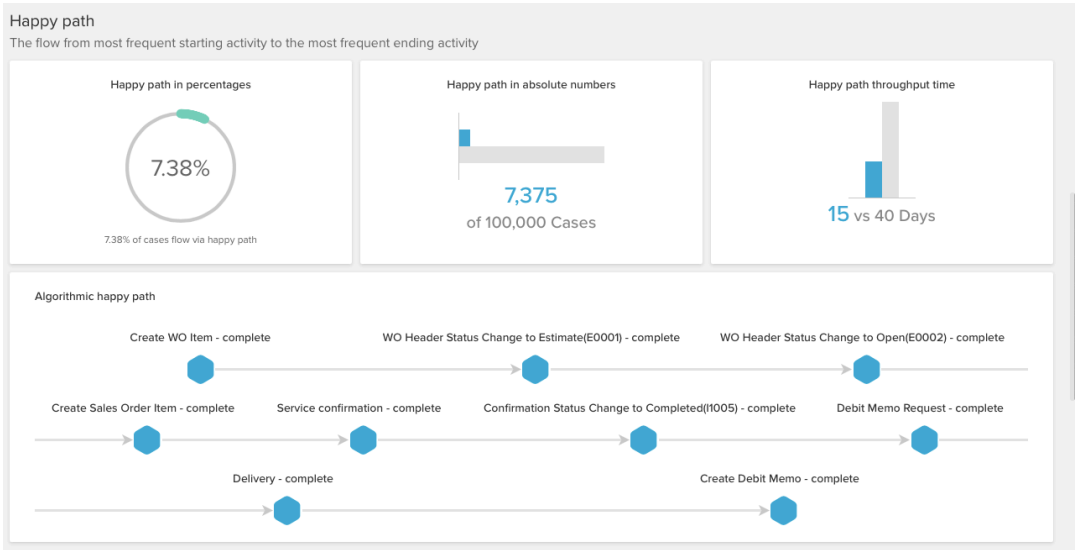


Figure 21. Screenshot of happy path discovered by Celonis

According to the happy path, Celonis computes and displays bottlenecks that increase the throughput time of the process. There are 20 bottlenecks that affect various percentages of the cases listed defined by Celonis. Five of the bottlenecks can be seen in Figure 22. As seen in Figure 22, the delay caused by the activity “Item Update” appears as the longest. However, this delay affects only 11% of the cases with a relatively small impact on the total event log.

- Moreover, the bottleneck between the activities “Create Sales Order Item” and “Service Confirmation” comes second with the highest percentage (69%) of cases affected. This bottleneck is noted as FACT10 for further analysis in enhancement.
- There is another bottleneck between Create WO Item and WO Header Status Change to Estimate (E0001) that is highlighted by the tool. The

proportion of the cases that are affected is 51%. This means that almost half of the cases include this activity. This is also noted as FACT11 for further analysis in enhancement.

Remaining bottlenecks are provided to the process experts for further evaluation.

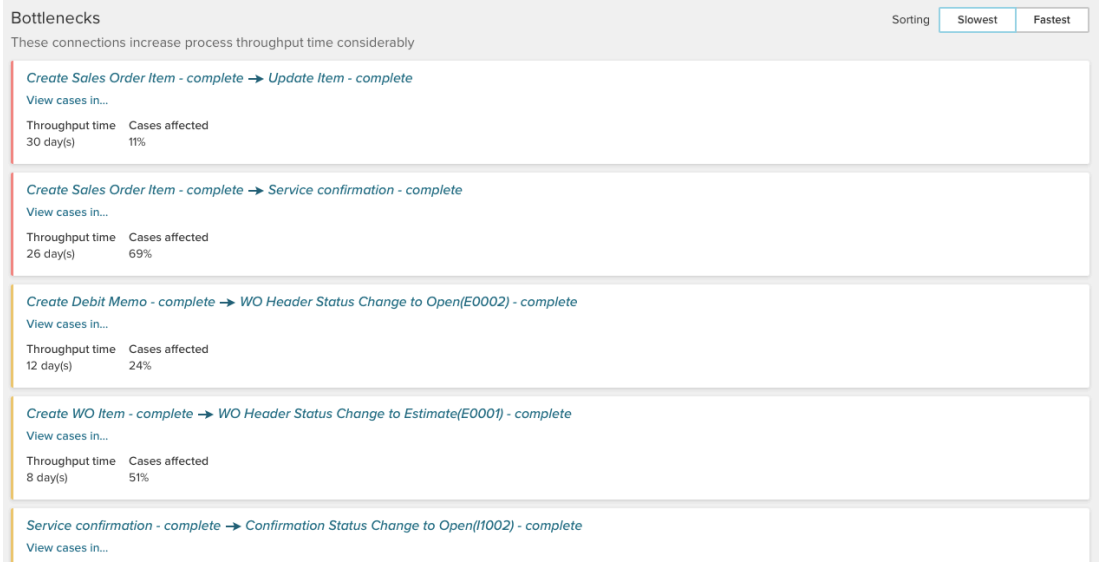


Figure 22. Examples of bottlenecks provided by Celonis

5.4.5 Conformance checking

Conformance checking is the step where the event log and the reference process model is compared to identify the inconsistencies between them. Usually, the as-is model of the process is used as the reference model. However, when there is not an as-is model available, a reference process model can be mined using proper cases in the event log by process discovery tools. In the current study, we will use the company as-is model generated by using proper cases in the event log as our base model to be used in conformance checking. This model is provided below in Figure 23. Next, conformance checking will be between company as-is model and the event log by using ProM. Similarly, the same analysis should be performed in the Celonis environment. However, due to a software bug in Celonis, the company as-is cannot

be used in conformance checking. Hence, we will use the proposed model generated by Celonis for the conformance checking.

Figure 23 displays the petri net as-is model of the company. This model includes 13 of the 24 activities that are observed in the log. The flow starts with the activity “Create WO Item” and ends with the activity “WO Header Status Change to Closed (E0006)”.

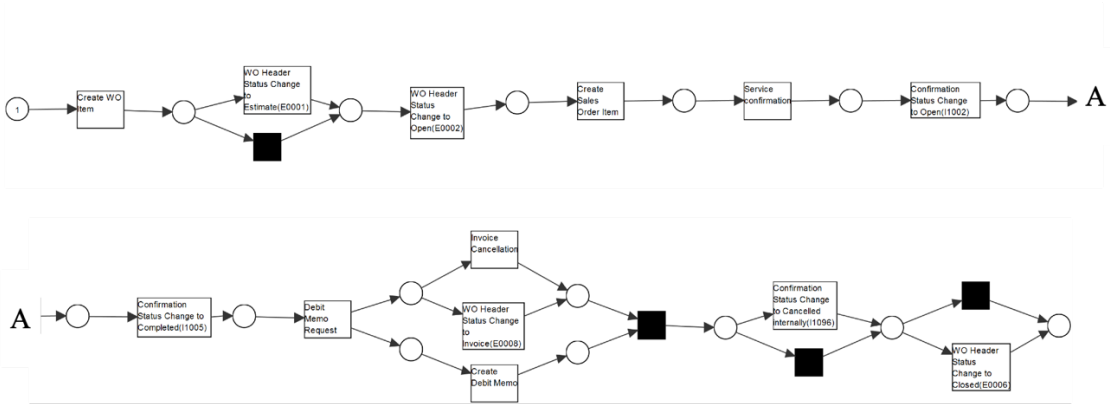


Figure 23. Petri net for SSP created for conformance analysis

ProM

The conformance checking plugin “Replay the event log for performance/conformance” requires an event log and a petri net model to create the conformance model and evaluate the statistics as seen in Figure 24. In this section, event log that has 100,000 cases and the process model in Figure 23 are used to perform the conformance checking. The output of this conformance checking plugin is a model that is in the form of petri net. The darkness of blue rectangles indicates level of the frequency of the activities, i.e., the higher the frequency is, the darker blue the color is. Three types of behavior are displayed in this model; synchronous moves, move on model, and move on log. Synchronous move indicates the alignment of the model and the case in the event log, i.e., the activity is executed

according to the model. A move on model implies an activity that exists in the model but not observed in the case whereas move on log implies an activity that is not modelled but observed in the case. In Figure 24, the green parts of the bars at the bottom of the rectangles imply the ratio of synchronous moves between model and log whereas pink part implies the ratio of move on model, i.e., the ratio of the cases that do not follow model at this point. Similarly, the numbers above the bar express the frequency of synchronous moves and move on model, respectively. For example, the activity “Service confirmation” includes 79523 synchronous moves and 20477 moves on model.

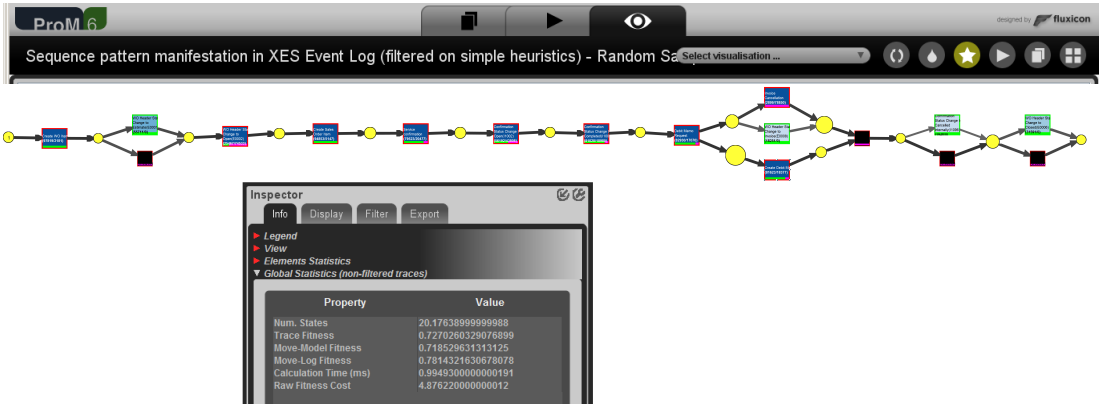


Figure 24. Screenshot of “Replay the event log for performance/conformance” plugin in ProM

Moreover, yellow circles display places where there are moves on log and the size of the circle increases with the number of occurrences of moves on log. As seen in the Figure 24, trace fitness is calculated as 0.727 with the conformance checking. It should be noted that synchronous moves indicate that the to-be process model is followed. Thus, the higher ratio of synchronous moves is the desired result. In this study, it is found that six activities namely “Create WO Item”, “WO Header Status Change to Estimate (E0001)”, “Create Sales Order Item”, and “WO Header Status

Change to Invoice (E0008)", "Confirmation Status Change to Cancelled internally (I1096)", and "WO Header Status Change to Closed (E0006)" have a pattern of synchronous moves in the rate of almost 100%. On the other hand, other activities with high ratios of moves on model and places with high ratios of moves on log express undesired behaviors that lower the trace fitness.

Figure 25 shows a small proportion of the conformance checking results. The first three activities namely "Confirmation Status Change to Open (I1002)", "Confirmation Status Change to Completed (I1005)", and "Debit Memo Request" show a linear flow. As the bars in the symbols of these activities have pink parts, it can be said there are a significant amount of cases that do not follow this path. Moreover, the yellow circles between these events reveals that some activities are included in the process besides the activities in the model. After the activity "Debit Memo Request", there are two parallel paths. One of them leads to the activity "Create Debit Memo" and the other one leads to a decision place that is followed by either the activity "Invoice Cancellation" or the activity "WO Header Status Change to Invoice (E0008)". However, the numbers and the pink parts of the bars imply that not all cases follow these paths.

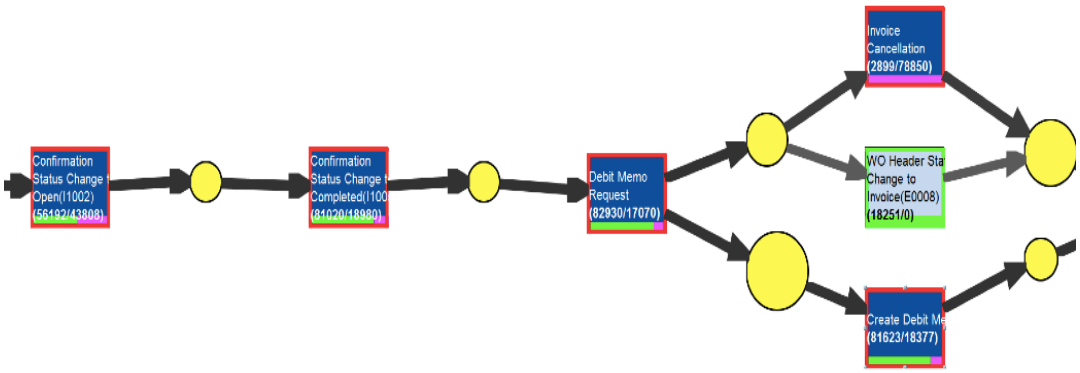


Figure 25. A small proportion of screenshot of "Replay the event log for performance / conformance" plugin in ProM

- For example, 18377 cases do not include the activity “Create Debit Memo” as shown in the model. It is noted that although 81623 cases include the activity “Create Debit Memo”, only 18251 cases include the activity “WO Header Status Change to Invoice (E0008)”. This gap is noted as FACT12 for further analysis with the process experts.
- The three activities “WO Header Status Change to Open (E0002)”, “Service Confirmation”, and “Confirmation Status Change to Open (I1002)” has ratio of synchronous moves lower than 80%. As these activities are expected to be performed on the exact position in the model, the existence of nonconforming cases is noted as FACT13 for further analysis in enhancement stage.

Celonis

Celonis has a predefined conformance checking application that is called Conformance PI. This application compares the event log with a reference model. Either a model can be added by the user or the model that is suggested by the application can be used for the analysis. Figure 26 shows the reference model that is suggested by the algorithm of Celonis to be used in the conformance checking. This model includes 11 of the 24 activities that are observed in the event log. The model starts with the activity “Create WO Item” which is followed by the activity “WO Header Status Change to Estimate (E0001)”. The remaining parts of the model include many exclusive gateways that are shown with the symbol cross and returns. For example, there is a return path at the end of the model that goes back to the exit of the second activity in the model ““WO Header Status Change to Estimate (E0001)”. Thus, this model allows to execute some variants that are seen in the event log.

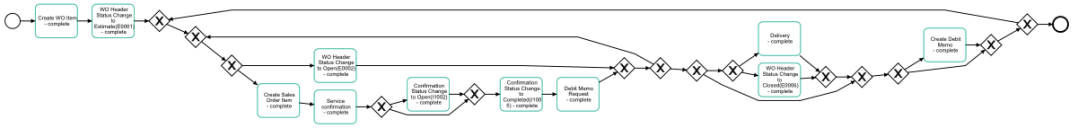


Figure 26. The process model discovered by Celonis for conformance checking

Figure 27 shows the final conformance checking report generated in Celonis after several iterations. As can be followed from the Figure 27, the report includes several measures on the conformance of reference model to the event log. It is found that only 45% of the cases fully complies with the suggested model. This corresponds to the 45.0K cases of 100K cases in the event log.

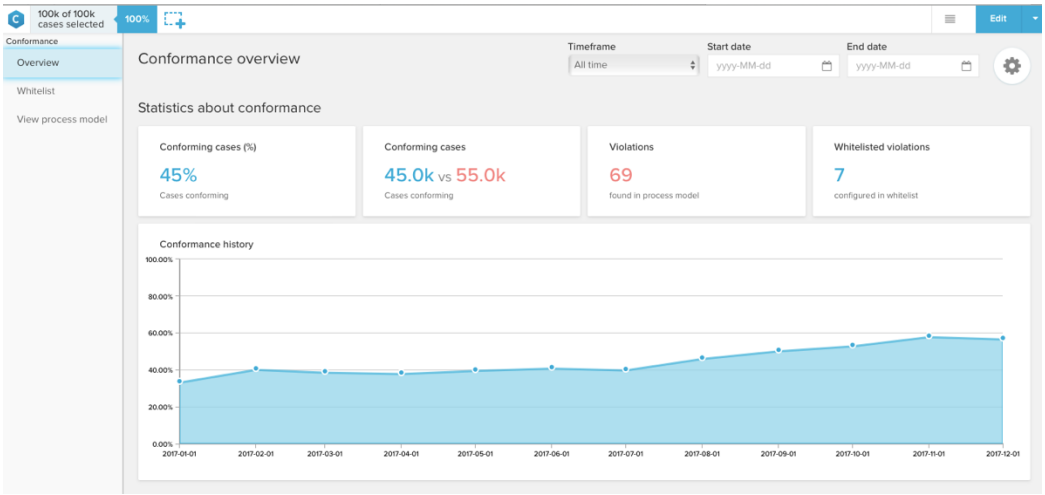


Figure 27. Screenshot of conformance checking using Celonis PI

There are 69 model violations identified in the proposed model. Examples include identification of the undesired activities that cause significant delays in process throughput time or identification of undesired activity orders that do not conform the proposed process model. Each one of these violations can be further analyzed to place them in the white list so that they are reconsidered in the process model.

When a violation is added to the white list, percentage of conforming cases will increase accordingly. For instance, Figure 28 shows that the whitelisted violation that is the following the activity “Create WO Item” by the activity “Create Sales Order Item” increases the percentage of conforming cases from 27 to 36.

- The violations that are not eligible for the white list are noted as FACT14 will be evaluated for the enhancement stage.

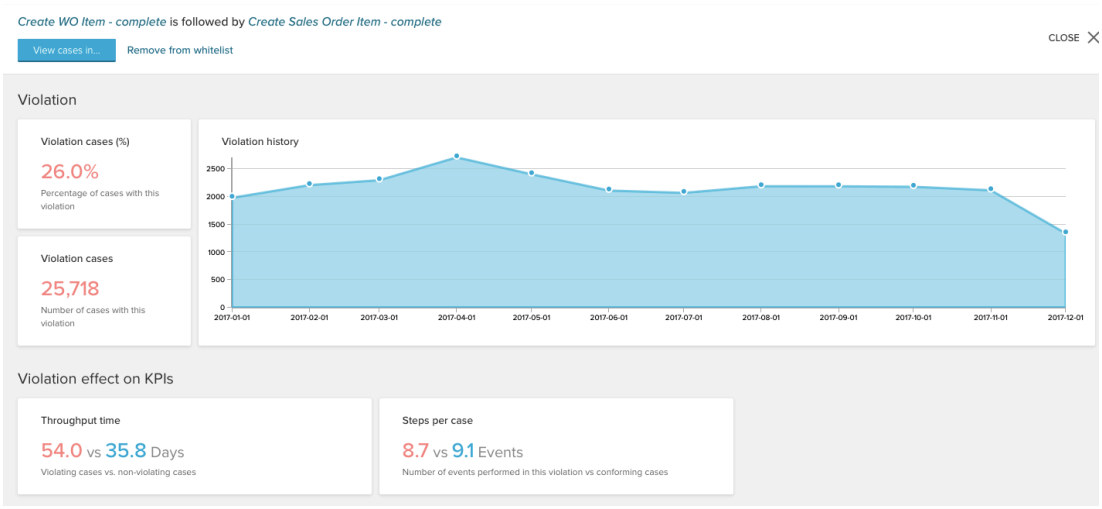


Figure 28. Screenshot of a violation found by Celonis PI

5.4.6 Enhancement

Enhancement stage includes identifying improvement ideas by using the findings of previously executed process mining types, process discovery and conformance checking. Our purpose is to present possible improvement areas for the SSP. Here, we evaluate the outputs of both Celonis and plugins of ProM together to obtain improvement ideas.

In this section findings of previous sections are discussed with the process expert to extract ideas for process enhancement. The facts that are defined in previous sections are evaluated (See Appendix C).

5.4.7 Process analytics

Process analytics include creating visual analytics that may help to improve the business model. To obtain information on SSP, both ProM and Celonis is used. Below, we first express the overview analytics from ProM and then from Celonis.

ProM

ProM has the visualization options to create dotted charts to obtain a general impression on the event log. In Figure 29, the chart shows the dates of the events that are executed to complete the cases. Y-axis is the case ids whereas X-axis is the execution date. The color of the shapes changes according to the activity. The shapes on the same horizontal line indicates the events of the same case. The chart provides insights on the density level of the case throughout the year with the assistance of the line graphs at the bottom of Figure 29. It can be seen that the demand is stable, and the fluctuation is low throughout the year. The blue circles that represent the activities for “Create WO Item” and “Create Sales Order Item” show a smooth flow at the beginning of cases. However, most of the other colors are spread entire graph area. This shows that the sojourn time of these activities deviate. Moreover, it can be emphasized that some of the cases that start at the beginning of the year and the cases that start at the end of the year are completed at the same time. Normally, it is expected to have a parallel line to the start line. In this case, end line does not stand out which implies deviations in the throughput time of cases. In an enhanced model, a narrow band should be observed with the obvious start and end lines and a few exceptions outside the band.

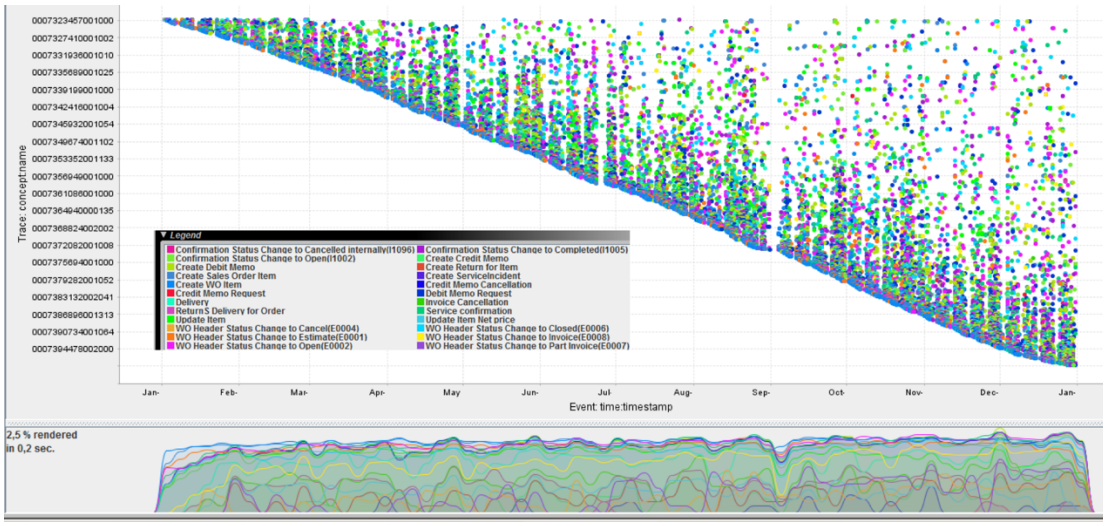


Figure 29. Dotted chart for cases with respect to timestamp by ProM

Moreover, the axis can be changed in the chart to obtain different information from the event log. For instance, Figure 30 shows the time passed since the week started for the events in the event log. In the chart, x-axis displays the time passed since the week started and y-axis displays the activities. The shapes indicate each case. This chart expresses that many activities are executed within the working hours of the weekday. In addition, there are activities that are frequently executed outside the working hours even in the weekends. For instance, the activity “Service Confirmation” is executed in every day of the week for more than 8 hours. Moreover, there are also activities such as “Credit Memo Request” and “Update Net Item price” that never occur on weekends. Furthermore, the chart provides insights on the density level of the case for the time of the week with the assistance of the line graphs at the bottom of Figure 30. These lines reveal that the patterns are similar within the week and the frequency level is higher in the weekdays that the weekends. However, it should be noted that the frequency level on Saturdays is close to the weekdays whereas the frequency level on Sundays is as much as half of the weekdays.

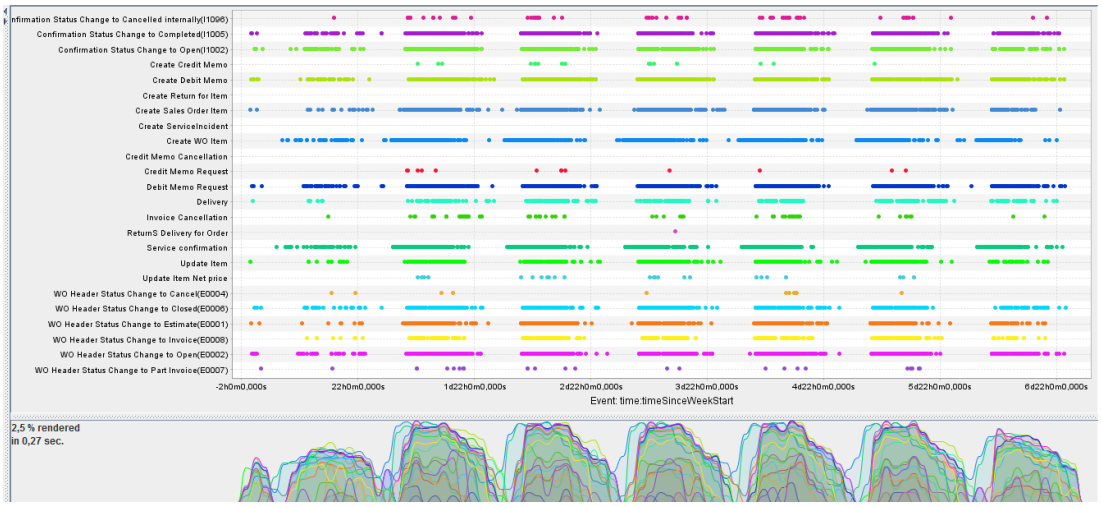


Figure 30. Dotted chart for activities with respect to time since week started by ProM (week starts on sunday)

Celonis

Celonis provides process metrics, throughput time distribution and activities within the process overview section. Figure 31 displays the key statistics about the process.

There are 276 cases and 2,535 events per day. The average throughput time is 40 days. In the Figure 31, the graph the bottom displays development of cases with respect to date. Graph shows the change in the number of cases per day within the year. The minimum number of cases started per case is 179 which is observed in January whereas the maximum number of cases started per day is 706 which is observed in October.

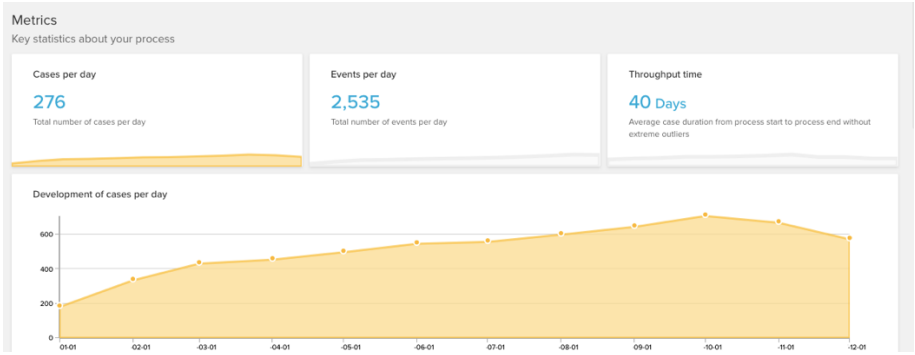


Figure 31. Screenshot of process metrics by Celonis

Moreover, the distribution of throughput times is displayed in Figure 32. Here, it is seen that more that 30,000 cases have a throughput time between 0 and 17 days. As the event log consists of 100,000 cases, it can be said that more than half of the cases has an average time less than 34 for days that is the sum of first two bars in the Figure XX. On the other end of the graph, it is seen that more than 5,000 cases have a throughput time more than 153 days. They have a duration between 5 and 12 months. These cases have a big impact on the average duration of the SSP.

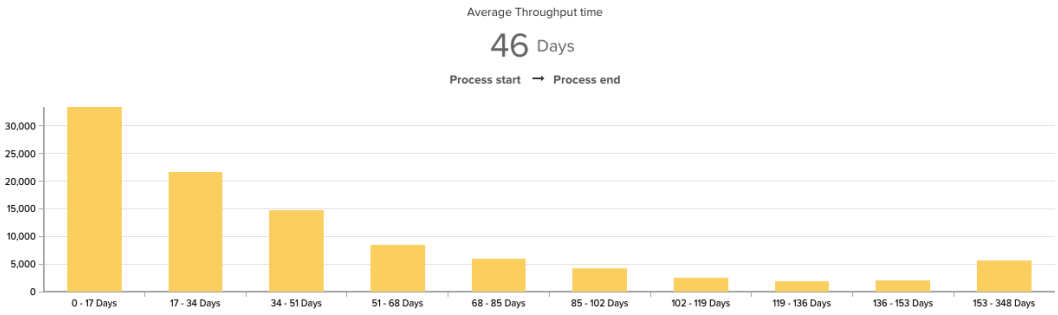


Figure 32. Screenshot of throughput time vs number of cases chart by Celonis

As seen in Figure 33, the final analytics information that is provided by Celonis is activities diagram i.e. activities that are observed in the event log. The size and the color of the activities indicate the frequency of occurrences. For example, the activity “Create WO Item” that is the most frequently observed activity in the event log is displayed as pink and as the largest activity among others. The smaller and blue activities are the least frequent activities. In this diagram, the required activities of the process should be displayed as the biggest activities. For instance, the activities “Create WO Item”, “Service Confirmation”, “Create Sales Order Item” are some of the required activities in the process. They are displayed with the biggest circles at the center of the diagram.

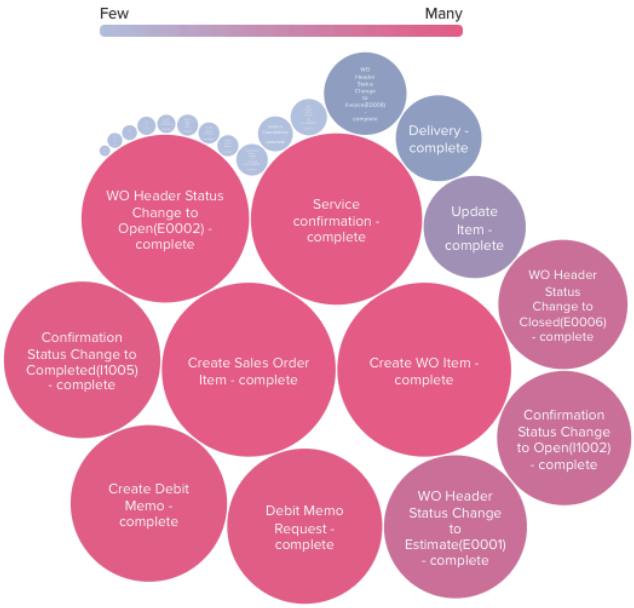


Figure 33. Screenshot of the activity frequency diagram by Celonis

5.5 Evaluation

In evaluation stage, we summarize the findings of previous sections to finalize the improvement ideas to improve process according to the questions defined in planning stage.

By using the event log composed in the data processing stage, both IvM plugin and Celonis provide valuable insights and process information on SSP with the process discovery. These findings are noted for further evaluation in the enhancement stage. After the process discovery, conformance checking is performed to explore the differences between the process model and the event log. Therefore, Celonis and "Replay the event log for performance/conformance" plugin in ProM are utilized. To perform this mining type, company as-is model is needed. As it is not available, company as-is model is created using proper trails in the event log. ProM plugin can successfully check conformance level of the event log with the company as-is model. With this application, trace fitness is found as 0.727. Celonis has failed

in conformance checking with company as-is model. Nevertheless, conformance checking is performed in Celonis by using its proposed model. The output shows that only 45% of the cases fully comply with the improved version of proposed model.

Next, in enhancement stage, the facts that are noted during the analysis and mining stages are evaluated and further analyzed using Celonis. As a result, many improvement ideas are developed for business process reengineering.

In conclusion, the improvement ideas will be assessed as process improvement projects within the organization.

5.6 Process improvement and support

This stage includes, implementing the improvement and supporting daily operations with process mining. As the implementation of the improvement ideas should be considered as separate IT projects within the organization and requires deep know-how covering IT and ERP systems, this part is not included to the scope of this thesis study. Moreover, supporting daily operations is also not in the scope, as this study is one-off case study in order to examine the aims of this thesis research.

CHAPTER 6

COMPARISON OF THE PROCESS MINING TOOLS

In this section, the process mining tools ProM and Celonis, that have been used for this thesis research on SSP are compared. There are studies in the literature that include frameworks for the comparison of process mining tools (Leemans, Fahland, and van der Aalst, 2014, Batyuk and Voityshyn, 2018, and Kebede, 2015). Leemans, Fahland and van der Aalst (2014) compare four tools namely Celonis, Fluxicon, and Perceptive Process Mining and ProM according to the features on semantics, zoomability, evaluation, and speed. Next, Kebede (2015) introduce a framework to compare process mining tools. This framework consists of many features including input and output file formats, mining types, analysis and visualization capabilities. In this study, a case study is also performed to compare three tools, ProM, Fluxicon and Celonis using this framework. Moreover, Batyuk and Voityshyn (2018) state the features with respect to the non-functional requirements that may be used to guide the architectural design of a process mining tool. These features are listed in terms of quality attributes such as compatibility, maintainability, performance efficiency, reliability, and usability. In this thesis study, an evaluation framework is generated as in Table 10 with the features that are selected considering the studies of Leemans, Fahland, and van der Aalst (2014), Batyuk and Voityshyn (2018), and Kebede (2015). Table 10 presents the evaluation results for the comparison of the tools in terms of the features.

Table 10. Comparison of Process Mining Tools ProM and Celonis

Feature	ProM 6.8	Celonis
License	Open source	Academic, commercial, and trial
Supported platform	On-premise	On cloud or on-premise (if purchased)
Setup and Configuration	Procedure should be followed	Instant access via web browser No need for setup on cloud
Application Usage	Offline	Online / Offline
Input Event Log Files	CSV, XES	CSV, XES, XLSX, DBF
Output Files	Process Models, Images and Revised Event Logs	Process Models, Images, Revised Event Logs, and Reports
Capability (# of events)	1 Million (on a computer with 2.20 GHz Intel Core i5 processor and 8 GB memory)	More than 3 Million
Speed	Slow	Fast
Reliability	Low	High
Process Discovery	Easy to understand general behavior	Easy to filter an interested path and explore further
Petri Net Model Output	Yes	No
Filtering Options	Yes	Yes
Statistical Output	Limited	Extensive
Conformance Checking	Easy to understand general conformance	Easy to explore further nonconforming behavior
Base Model	Company as-is model, Proposed model of the tool	Proposed model of the tool
Output	Graphical	Statistical
Analytics	Descriptive Predefined charts	Descriptive Manually prepared charts
Ad-hoc Dashboards	Not possible	Can be created
Mining and Analysis Results	Not preserved	Preserved
User Interface	Basic	Advanced
Perceived ease-of-use	Low	High
Perceived usefulness	High	High

First, the authorization of the tools and the preparation for usage is explained. ProM is an open source framework that allows developers to create new plugins that can be operated with ProM. On the other hand, Celonis is a commercial tool that has predefined applications for process mining and allows users to create customized applications to meet the needs of the users. Besides commercial purchase, there are academic and trial license options for Celonis. These licenses can be obtained online,

and Celonis applications can be used instantly by uploading the event log on cloud. However, ProM is needed to be installed and configured that requires some steps to download both the framework and the plugins. To start with the application is much more easer with Celonis than with ProM. At this point, the constant need for the internet access can be counted as an important drawback for Celonis cloud version. On the other hand, ProM does not require internet access after the plugin package setup. It is important to remark that Celonis has an on-premise commercial version which also does not require internet access but needs to be installed.

Secondly, the inputs and outputs of the tools are mentioned and compared. Both ProM and Celonis use event logs as input file for process mining applications. ProM allows XES file format which is explained in Chapter 5 and CSV file format. ProM requires the conversion of CVS formatted files to XES to use in any plugin. On the other hand, Celonis allows XLSX, and DBF file formats in addition to CSV and XES file formats. Moreover, Celonis also provides an option to upload more than one table and to consolidate them for event log creation. As output, both tools provide process models, images and revised event logs. Additionally, Celonis also supports to create and export various reports.

As it is mentioned in Chapter 5, ProM has performance issues with large event logs. In this thesis study, the event log of 348,712 cases with more than 3 million events could not be used for process mining analysis in ProM. Although ProM can handle to display general information of this large event log and allows execution of some plugins, IvM plugin of ProM could not discover the process for this large event log on a computer with 2.20 GHz Intel Core i5 processor and 8 GB memory. With some iterations, it is found that the size that IvM can handle is an event log of 100,000 cases with almost 1 million events. On the other hand, any

performance problem is not experienced in Celonis with both large and small event logs. Moreover, Celonis can display results in any stage within seconds whereas IvM plugin of ProM may compute up to one hour to display the results. As the performance of ProM depends also on the capability of the computer, it is noticed that ProM may crush which results in the loss of the analyses. During this thesis study, no reliability problem is observed with Celonis.

Next, the outputs of the process mining types are compared. Here, the comparison includes the findings of process discovery and conformance checking. IvM plugin of ProM allows to understand the general behavior of the process. It provides petri nets that enable to differentiate parallel activities and choices in the output model. On the other hand, Celonis provides the most common path as the output that is easy to understand but not comprehensive at all. Although Celonis allows to filter an interested path and explore further, it is not possible to obtain a petri net with process discovery application. It should be emphasized to state that both ProM and Celonis have filtering options to further analyze the business process. Finally, both tools provide statistical outputs. However, the statistical outputs of Celonis are extensive whereas the statistical outputs of ProM are limited. For instance, Celonis provides mean, median and trimmed mean, maximum and minimum durations of activities and process in addition to the case and activities frequencies. However, ProM provides only average durations of activities in addition to the activities and case frequencies.

For conformance checking, it should be noted that ProM cannot provide conformance checking of the as-is model with IvM plugin. Instead it needs to use another plugin called “Replay the event log for performance/conformance plugin”. Yet, IvM can be used for conformance checking with the proposed model of the tool.

On the other hand, Celonis has the option to use the as-is model for conformance analysis. However, it encountered a failure in the thesis study for this claim. There was a failure in conformance checking application of Celonis as it could not confirm any cases and thus, showed 100% violations. The reason for tool failure is left unclear and reported to Celonis. In addition to the company as-is model, Celonis is capable to present conformance checking results by using the process model suggested by itself. When the outputs of conformance checking are analyzed in both tools, it can be expressed that the output of ProM plugin is a graphical representation that displays frequencies of conforming and nonconforming cases, whereas Celonis provides quantitative and statistical process information with a list that allows further investigation of model violations. The results of Celonis include confirming cases ratio and violation list with their effects on the process. The conformance checking features of Celonis are more informative in finding process improvement areas and enhancement. However, conformance checking of ProM may be very beneficial for auditing. For instance, audit applications usually require as-is process models and this as-is model and the sample cases are compared as a part of auditing. Thus, ProM can be preferred for audit applications as it offers better conformance checking results with as-is model.

The next feature in comparing these tools is their capability in descriptive analytics are provided by both tools. ProM and Celonis provide similar performance with predefined descriptive charts. However, Celonis also allows to generate ad hoc analyses with tables, graphs, and lists that cannot be obtained with ProM. It is important to emphasize that Celonis enables users to create dashboards that can include both predefined analyses, filters as well as the ad hoc analysis results that are created manually by the users according to their needs.

Finally, the general evaluation on the appearance and usage of the tools can be made. ProM has a basic user interface whereas Celonis has an advanced user interface. Both tools are evaluated according to vast variety of Technology Acceptance Models: “Perceived ease-of-use” and “Perceived usefulness” (Sumak et al., 2011, p.2068). “Perceived ease-of-use” of ProM is rated low. There are various reasons for this. First, it requires guidance to start using and it is difficult to continue with the analysis without directions. Secondly, ProM does not save the findings of the analysis automatically. Users should save everything from the revised event logs to the outputs of mining and analysis to be able to use or to access them after exiting the software. On the other hand, “Perceived ease-of-use” of Celonis is rated high. Celonis offers an application that is easy to use and enables continuing mining and analysis anytime by preserving them available after exiting the software. Secondly, the latest findings of the analysis are always available in Celonis.

Next, considering the evaluation of the tools with respect to process discovery, conformance checking, analysis, and ad hoc dashboard features, it can be confidently stated that both tools increase the performance of users for improving the business processes. Thus, the perceived usefulness of both ProM and Celonis are high.

To sum up briefly, both tools ProM and Celonis are very useful and beneficial for discovering and improving business processes. ProM is accessible by everyone while Celonis is for commercial usage and has a limited academic license. In the end, Celonis is a practical tool whereas ProM is not very practical and requires completing trainings before using it.

CHAPTER 7

CONCLUSION

In continuously changing and growing competitive business world, companies focus on maintaining or increasing their market shares. To be ahead of their competitors, they aim to acquire the knowledge using the data they have. As part of this goal, the process execution logs may be used to gain knowledge for the business process reengineering. Process mining, which was studied by Maruster, Weijters, and van der Aalst (2002) to create valuable process information from the so-called event logs for process improvement is now proposed as a promising research area by the well-known research workgroups (Adriansyah et al., 2012). This topic is recognized by IEEE Task Force on Process Mining that has members representing software vendors, consultancy firms/end users, and research institutes.

The aims of this thesis study are to observe the role of process mining in process improvement, and to examine and compare process mining tools. To realize these goals, the real data of sales process from an organization is used. To perform the process mining with this data, ProM and Celonis are selected as the tools among many process mining tools for their common usage. In this study, all three process mining types; process discovery, conformance checking, and enhancement are applied using PM^2 methodology initiated by van Eck et al. (2015).

The first aim of this study is to discover the role of process mining in process improvement. Process mining can have competitive advantage among other process improvement methodologies when it can be efficient, effective and flexible. Our experimentation on real data analysis shows that process mining can be performed efficiently on a big data set with 24 activities in a short time and without requiring

many human or technical resources. Moreover, the quality of the to-be process model is increased by the availability of deeper analysis in process discovery and as-is model conformance steps. This provides the BPM expert to make better diagnosis on the occurring problems, identify the bottleneck sub processes and generate better ideas for enhancement. Hence, process mining is a more effective way of process improvement when compared to the traditional process improvement methodologies. Finally, many analyses can be made by process mining in seamless variety of perspectives which makes it a very flexible way of process improvement. Thus, process mining has become essential and inevitable for process improvement in the digital era.

Second aim of this thesis study is to examine and compare process mining tools. First, a framework is generated to evaluate the process mining tools by considering the existing evaluation frameworks in the literature. Then the tools that are used in the current study are compared in terms of the features of the generated framework. This framework includes the aspects for tool installation, technical capacity, process mining analysis, user interface as well as the basic criteria for technology acceptance model. Accordingly, both tools ProM and Celonis are found to be very useful and beneficial for discovering and improving business processes. ProM is accessible by everyone while Celonis is for commercial usage and has a limited academic license. In the end, Celonis is a practical tool whereas ProM is not very practical and requires completing trainings before using it.

In the future, the process mining case study can be extended by adding new attributes to the event log such as resource, start time, vendor, price, and location. Doing these would contribute to both aims of the current thesis research. Resource attribute would enable to make social network analysis such as handover of work,

reassignment, and subcontracting. The remaining attributes will enhance the discovery results and root cause analysis. For example, start time with the completion time would enable to compute both waiting time and processing time. Substituting sojourn time with these durations would ensure healthier analysis results on the process efficiency and queuing time, and provide more precise process improvement suggestions. The remaining attributes such as vendor, price, and location would enhance the analysis results by enabling variant comparisons and displaying differences with respect to these attributes. Furthermore, extending the event log with more attributes such as resource, start time, vendor, price and location would enable us to explore in depth further features of the tools and compare them in more detail.

APPENDIX A

EVENT LOG SUMMARY

Table A1. Log Summary Table of Unfiltered Event Log from ProM

Log Summary
Total number of process instances: **471320**
Total number of events: **3991503**

All events
Total number of classes: **25**

Class	**Occurrences (absolute)**	**Occurrences (relative)**
Create WO Item	471320	11,808%
Create Sales Order Item	471320	11,808%
WO Header Status Change to Open(E0002)	427455	10,709%
Service confirmation	407179	10,201%
Create Debit Memo	327928	8,216%
Confirmation Status Change to Completed(I1005)	324856	8,139%
Debit Memo Request	318231	7,973%
WO Header Status Change to Estimate(E0001)	313345	7,85%
Confirmation Status Change to Open(I1002)	281427	7,051%
Update Item	220046	5,513%
WO Header Status Change to Closed(E0006)	215036	5,387%
WO Header Status Change to Invoice(E0008)	81330	2,038%
Delivery	75239	1,885%
Invoice Cancellation	20121	0,504%
Confirmation Status Change to Cancelled internally(I1096)	10576	0,265%
WO Header Status Change to Cancel(E0004)	8915	0,223%
WO Header Status Change to Part Invoice(E0007)	6623	0,166%
Update Item Net price	4868	0,122%
Credit Memo Request	2964	0,074%
Create Credit Memo	2438	0,061%
Credit Memo Cancellation	210	0,005%
Return Delivery for Order	29	0,001%
Create Return for Item	29	0,001%
Create Service Incident	12	0,0%
Confirmation Status Change to Open(E0002)	1	0,0%

Start events
Total number of classes: **9**

Class	**Occurrences (absolute)**	**Occurrences (relative)**
Create WO Item	404649	85,854%
WO Header Status Change to Estimate(E0001)	58889	12,494%
WO Header Status Change to Open(E0002)	7162	1,52%
Confirmation Status Change to Open(I1002)	537	0,114%
Service confirmation	58	0,012%
WO Header Status Change to Closed(E0006)	15	0,003%
WO Header Status Change to Invoice(E0008)	8	0,002%
WO Header Status Change to Part Invoice(E0007)	1	0,0%
Create Service Incident	1	0,0%

End events
Total number of classes: **20**

Class	**Occurrences (absolute)**	**Occurrences (relative)**
Create Debit Memo	190427	40,403%
WO Header Status Change to Closed(E0006)	109206	23,17%
Create Sales Order Item	45260	9,603%
Update Item	43288	9,184%
WO Header Status Change to Invoice(E0008)	39424	8,365%
Confirmation Status Change to Open(I1002)	23053	4,891%
Service confirmation	5835	1,238%
WO Header Status Change to Open(E0002)	4469	0,948%
WO Header Status Change to Part Invoice(E0007)	2347	0,498%
Delivery	2035	0,432%
WO Header Status Change to Cancel(E0004)	2013	0,427%
Confirmation Status Change to Cancelled internally(I1096)	1150	0,244%

Table A1. Continued

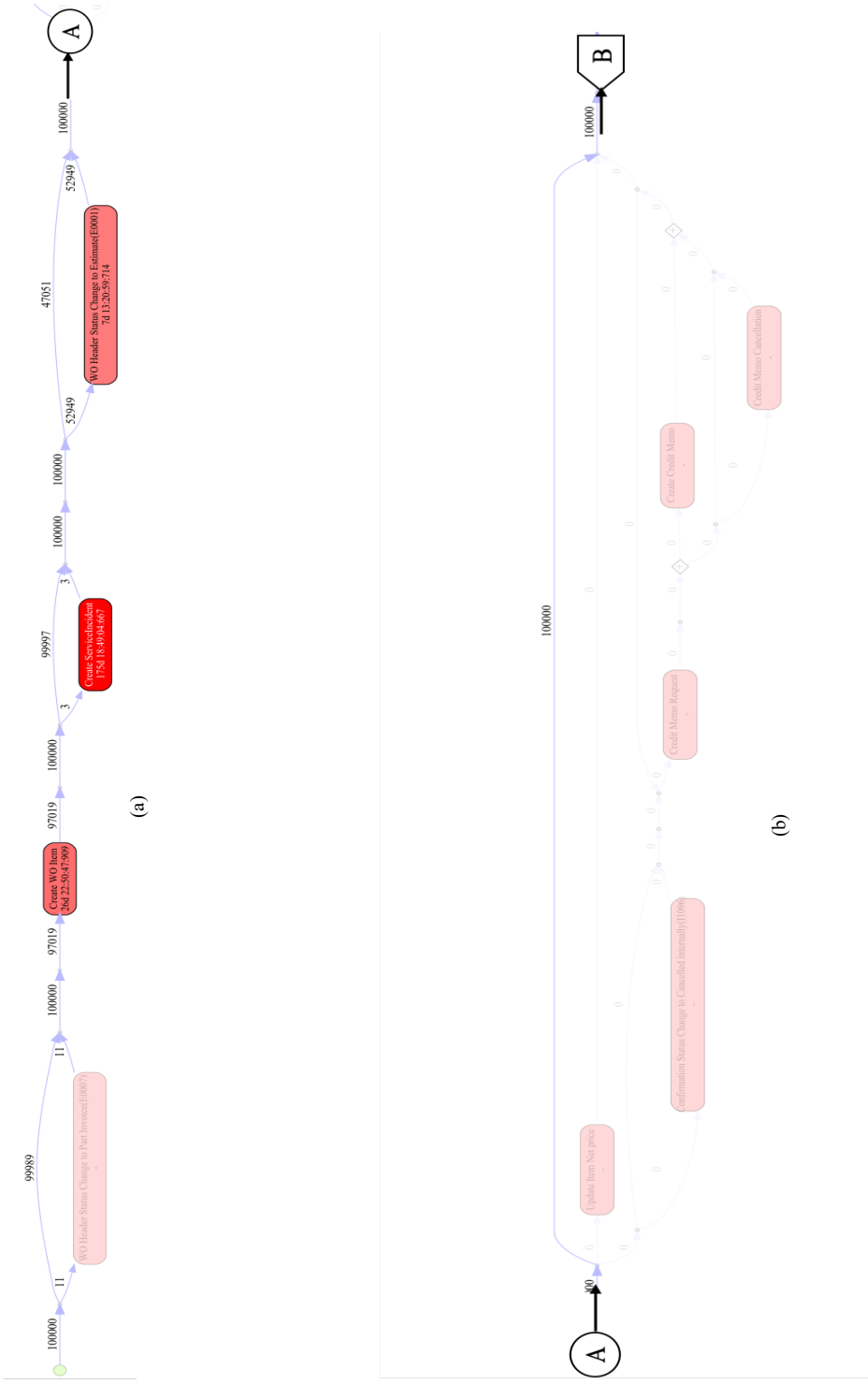
Create Credit Memo	627	0,133%
Confirmation Status Change to Completed(I1005)	589	0,125%
WO Header Status Change to Estimate(E0001)	517	0,11%
Invoice Cancellation	509	0,108%
Debit Memo Request	313	0,066%
Update Item Net price	186	0,039%
Credit Memo Cancellation	40	0,008%
Credit Memo Request	32	0,007%

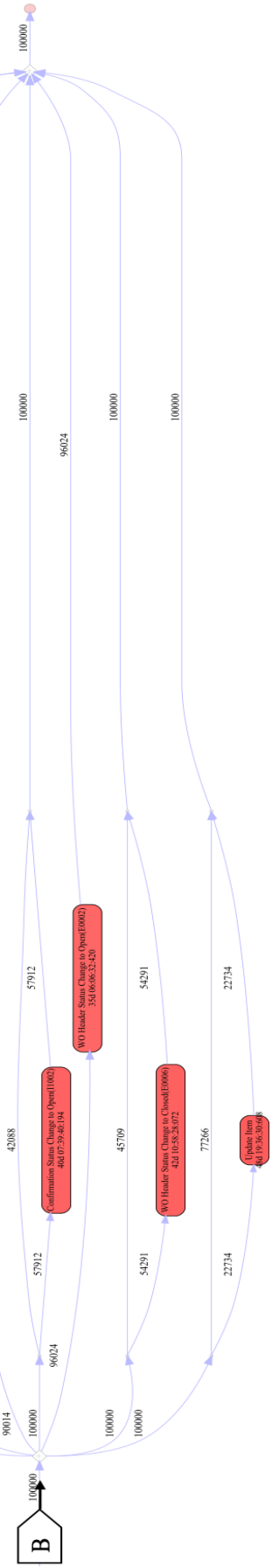
Table A2. Log Summary Table of Final Event Log from ProM

Log Summary Total number of process instances: **100000** Total number of events: **918026**		
All events Total number of classes: **24**		
Class	**Occurrences (absolute)**	**Occurrences (relative)**
Create WO Item	100000	10,893%
Create Sales Order Item	100000	10,893%
Service confirmation	98112	10,687%
WO Header Status Change to Open(E0002)	96078	10,466%
Create Debit Memo	85767	9,343%
Confirmation Status Change to Completed(I1005)	84480	9,202%
Debit Memo Request	82995	9,041%
WO Header Status Change to Estimate(E0001)	65837	7,172%
Confirmation Status Change to Open(I1002)	62544	6,813%
WO Header Status Change to Closed(E0006)	54304	5,915%
Update Item	35656	3,884%
WO Header Status Change to Invoice(E0008)	20273	2,208%
Delivery	20266	2,208%
Invoice Cancellation	5145	0,56%
Confirmation Status Change to Cancelled internally(I1096)	2327	0,253%
WO Header Status Change to Part Invoice(E0007)	1466	0,16%
Update Item Net price	862	0,094%
Credit Memo Request	676	0,074%
WO Header Status Change to Cancel(E0004)	618	0,067%
Create Credit Memo	559	0,061%
Credit Memo Cancellation	43	0,005%
Return Delivery for Order	7	0,001%
Create Return for Item	7	0,001%
Create Service Incident	3	0,0%
Start events Total number of classes: **5**		
Class	**Occurrences (absolute)**	**Occurrences (relative)**
Create WO Item	87563	87,563%
WO Header Status Change to Estimate(E0001)	11453	11,453%
WO Header Status Change to Open(E0002)	856	0,856%
Confirmation Status Change to Open(I1002)	113	0,113%
Service confirmation	15	0,015%
End events Total number of classes: **13**		
Class	**Occurrences (absolute)**	**Occurrences (relative)**
Create Debit Memo	54579	54,579%
WO Header Status Change to Closed(E0006)	31278	31,278%
WO Header Status Change to Invoice(E0008)	11357	11,357%
WO Header Status Change to Part Invoice(E0007)	655	0,655%
WO Header Status Change to Cancel(E0004)	592	0,592%
Delivery	583	0,583%
Confirmation Status Change to Cancelled internally(I1096)	343	0,343%
Create Credit Memo	184	0,184%
Confirmation Status Change to Completed(I1005)	178	0,178%
Invoice Cancellation	151	0,151%
Debit Memo Request	87	0,087%
Credit Memo Request	7	0,007%
Credit Memo Cancellation	6	0,006%

APPENDIX B

THE PROCESS MODEL WITH PATHS AND SOJOURN TIME CREATED BY IVM WITH 1.0 AND 0.8 ACTIVITY AND PATH FILTER LEVELS




B
100000
Create Sales Order Item
1d 10:16:12.479
Service confirmation
28d 00:02:54.294
WO Header Status Change to Invoiced(E0008)
5Sd 19:55:37.602
Confirmation Status Change to Opened(I1002)
40d 07:39:40.194
WO Header Status Change to Open(E0002)
35d 06:06:32.420
WO Header Status Change to Closed(E0006)
42d 10:58:28.072
Update Item
46d 19:36:39.048
Confirmation Status Change to Completed(I1005)
28d 11:07:09.905
Debit Memo Request
0d 02:15:12.800
Delivery
3d 07:20:56.830
Create Debit Memo
0d 20:58:08.115

(c)

APPENDIX C

EVALUATION OF THE FACTS FOR ENHANCEMENT

Order	Fact	Solution
FACT1	In the report, it is seen that there are start activities other than the expected start activity "Create WO Item". As the event log is prepared for the cases that have already passes work order creation step, it can be understood that there are some cases in which other activities are executed before the activity "Create WO Item". This is noted as FACT1 for further analysis in enhancement and evaluation stages.	It is confirmed that the other start activities are observed in the log since new items may be added to the existing work order. This is performed when the existing items do not cover the tasks that are needed to complete the maintenance or repair service.
FACT2	So, the cases with end activities namely "Create Sales Order Item", "Update Item", "Confirmation Status Change to Open(I1002)", "Service confirmation", "WO Header Status Change to Open(E0002)", "WO Header Status Change to Estimate(E0001)", "Update Item Net price" are excluded from further analysis. The high number of incomplete cases is also noted for further analysis as FACT2.	As the cases should be complete by the time of the extraction, the high percentage of incomplete cases was not expected. Since activities on incomplete cases cause problems for the organization, it is suggested to use a new system status that will be triggered after a predefined time to expire work orders.
FACT3	The recurrence of the activities within a case is noted as FACT3. These will be discovered in enhancement phase.	The activities that occur more than once in the event log are analyzed (See Appendix D). rework is usually observed due to the item updates and repetition of confirmations and changes in the debit memos. As these reworks cause delays, process owner will further investigate the reasons for updates and changes in the documents.
FACT4	The darkest red colored activity is "Create Service Incident" as seen in Figure 16. The average sojourn time of this activity is almost 176 days. Although this duration is very high, the number of cases that are exposed to this duration is only three. These extreme cases are noted as a fact in the process flow to be discussed with the process owners in the enhancement phase as FACT4.	These cases are accepted as problematic cases by the business experts. It will be further analyzed by the business experts and business unit collaboratively.
FACT5	Another activity "Create WO Item" has also almost 26 days of an average sojourn time. As the expected first activity is "Create WO Item", this loss of time will be discussed as FACT5.	This fact is evaluated with the FACT1. The reasons that causes this delay is explained above. As this fact affects the average time of 12% of the cases, this item addition will be reviewed for process improvement by the process experts.
FACT6	The average sojourn times of remaining activities that increase the throughput time of the cases are noted and will also be discussed for process improvement as FACT6.	The activities with high sojourn time are listed in Appendix E. These activities will be evaluated by process experts and business unit to decrease the overall throughput time.
FACT7	The transparent activities in Table 9 are noted as FACT7 to evaluate the reasons for executing them at different positions than the expected positions in the model which causes longer throughput times.	This fact is associated with the FACT3. Most of the cases that includes transparent activities includes repetitive activities. Another reason is the order of the activity that are executed after these transparent activities. When the whole of the model is analyzed, it can be suggested that some restrictions should be implemented to prevent the sequence changes of the activities.

Order	Fact	Solution
FACT8	One of parallel activities WO Header Status Change to Open (E0002) has a longer average sojourn time than many other parallel activities. As it is expected that this status should be observed before activities related to Service Confirmation, this high sojourn time is noted as FACT8 to be discussed with the process owners in the enhancement phase	Further investigation shows that open statuses are mainly used to activate the work orders as closed, invoice, and part invoice and status changes are executed at the exact same time with the following status. Due to parallel modelling of this activity, it has a long sojourn time even though it only lasts a second. So, this status can be merged with respective statuses and removed from the event log to reduce the complexity of the process model.
FACT9	The choices may allow to skip some activities that are identified as mandatory by the process experts such as WO Header Status Change to Closed (E0006), WO Header Status Change to Estimate (E0001). These activities are noted as FACT9 to further analyze the reasons that cause to avoid them.	These statuses are expected to be observed in every case. However, this information is not supported by the event log. Thus, existing cases should be analyzed to understand why these statuses are missing. According to the findings, some refinements and restrictions can be added to ERP system.
FACT10	Moreover, the bottleneck between the activities “Create Sales Order Item” and “Service Confirmation” comes second with the highest percentage (69%) of cases affected. This bottleneck is noted as FACT10 for further analysis in enhancement.	Further analysis shows that although the median duration is 14 days, the average throughput time is calculated as 26 days. It shows that there are cases with much higher throughput times that affect overall performance of the process. This is also investigated further, and the distribution of throughput time is found (See Appendix F). It is identified that the duration of 4% of the cases is between 99 and 344 days. Thus, these cases that last much longer than the average should be analyzed by the process owner and service employees to improve the throughput time.
FACT11	There is another bottleneck between Create WO Item and WO Header Status Change to Estimate (E0001) that is highlighted by the tool. The proportion of the cases that are affected is 51%. This means that almost half of the cases include this activity. This is also noted as FACT11 for further analysis in enhancement.	Further analysis shows that the mean throughput time is 8 days whereas median throughput time is only 1 day. This implies that the duration is much higher for the half of the cases than the other half of the cases. So, in overall almost a quarter of the cases lasts 7 days longer. As this status is used to imply that the work order is in evaluation and there isn’t any other status prior to this one, this shouldn’t be seen often. These cases will be evaluated further by process owners to extract root causes.
FACT12	For example, 18377 cases do not include the activity “Create Debit Memo” as shown in the model. It is noted that although 81623 cases include this activity, only 18251 cases include the activity “WO Header Status Change to Invoice (E0008)”. This gap is noted as FACT 12 for further analysis with the process experts.	It is observed that the work order status Invoice (E0008) is not used properly. This creates inconvenience while checking the orders, so automatic status update may be implemented to prevent this issue.
FACT13	The three activities “WO Header Status Change to Open (E0002)”, “Service Confirmation”, and “Confirmation Status Change to Open (I1002)” has ratio of synchronous moves lower than 80%. As they are expected to be performed on the exact position in the model, the existence of nonconforming cases is noted as FACT13 for further analysis in enhancement stage.	It is seen that this issue is related with FACT1 and FACT5. The cases that do not include these activities should be analyzed with IT department and the IS should be controlled to prevent missing records.
FACT14	The violations that are not eligible for the white list are noted as FACT14 will be evaluated for the enhancement stage.	This list will be used as a guide for continuous process improvement. As the first improvement iteration, the activity “Update Item” as an undesired activity and the activity “WO Header Status Change to Estimate (E0001)” as start activity are chosen. These violations will be used to decrease the average throughput time and increase conformance level.

APPENDIX D

ACTIVITY LIST THAT OCCUR MORE THAN ONCE IN THE EVENT LOG

Activity	Activities Count	Case Count	Repeat
Update Item	35656	22734	12922
Service confirmation	98112	90014	8098
Confirmation Status Change to Open (I1002)	62544	58080	4464
Create Debit Memo	85767	81623	4144
Invoice Cancellation	5145	3024	2121
Debit Memo Request	82995	82930	65
Create Credit Memo	559	521	38
Update Item Net price	862	832	30
Credit Memo Request	676	654	22
Credit Memo Cancellation	43	29	14

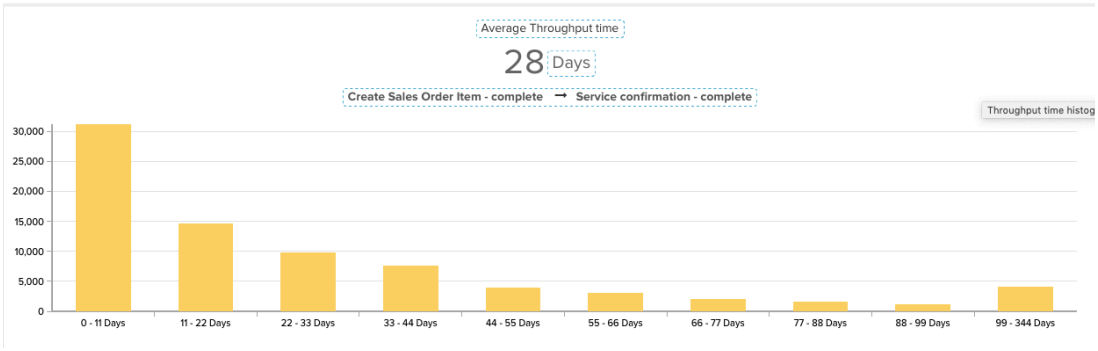
APPENDIX E

ACTIVITIES WITH HIGHER AVERAGE SOJOURN TIME BY IVM

Activity	Average Sojourn Time (days)
WO Header Status Change to Invoice(E0008)	56
Update Item	48
WO Header Status Change to Closed(E0006)	42
Confirmation Status Change to Open(I1002)	40
WO Header Status Change to Open(E0002)	35
Confirmation Status Change to Completed(I1005)	28
Service confirmation	28

APPENDIX F

THROUGHPUT TIME DISTRIBUTION FOR FACT10



REFERENCES

Adriansyah, A., Alves de Medeiros, A. K., Arcieri, F., Blickle, T., Bose, J. C., van den Brand, P., . . . Wynn, M. (2012). Process mining manifesto. (pp. 169-194). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-28108-2_19

Alves de Medeiros, A. K., Van Dongen, B. F., Van der Aalst, W. M. P., & Weijters, A. J. M. M. (2004). *Process mining: Extending the alpha-algorithm to mine short loops* (BETA working paper series, WP 113). Eindhoven: Eindhoven University of Technology.

Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F. M., Marrella, A., ... & Soo, A. (2019). Automated discovery of process models from event logs: Review and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, *31*(4), 686-705.

Batyuk, A. Y., & Voityshyn, V. V. (2018). Process mining: Applied discipline and software implementations. *Research Bulletin of the National Technical University of Ukraine" Kyiv Politechnic Institute"*, (5), 22-36.

Buijs, J. C. A. M. (2014). *Flexible evolutionary algorithms for mining structured process models* (Unpublished PhD Thesis). Eindhoven University of Technology, Eindhoven, The Netherlands.

Chinces, D., & Salomie, I. (2013, September). Business process mining algorithms. In *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 271-277). IEEE.

Dakic, D., Stefanovic, D., Cosic, I., Lolic, T., & Medojevic, M. (2018). Business process mining application: A literature review. In *Proceedings of the 29th DAAAM International Symposium* (pp.0866-0875). DAAAM International.

Fauvet, M. C., La Rosa, M., Sadegh, M., Alshareef, A., Dijkman, R. M., Garcıa-Banuelos, L., ... & Mendling, J. (2010, December). Managing process model collections with AProMoRe. In *International Conference on Service Oriented Computing (ICSOC 2010)* (Vol. 7, p. 10).

Feyyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert, 11*(5), 20-25. doi:10.1109/64.539013

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, *13*(3), 57-57.

Ganesha, K., Soundarya, M., & Supriya, K. V. (2017, March). The best fit process model for the utilization of the physical resources in hospitals by applying inductive visual miner. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 318-322). IEEE.

Günther, C. W., & Van der Aalst, W. M. (2007, September). Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *International Conference on Business Process Management* (pp. 328-343). Berlin, Heidelberg: Springer.

Hull, R., & Nezhad, H.R. (2016). Rethinking BPM in a cognitive world: Transforming how we learn and perform business processes. In *International Conference on Business Process Management* (pp. 3-19). Cham: Springer.

Hwang, S. Y., Wei, C. P., & Yang, W. S. (2004). Discovery of temporal patterns from process instances. *Computers in industry*, *53*(3), 345-364.

Kebede, M., & Dumas, M. (2015). *Comparative evaluation of process mining tools* (Master's Thesis, University of Tartu, Tartu, Estonia). Retrieved from https://pdfs.semanticscholar.org/bb40/63305540e49644f08dd06f6c50f5d0266630.pdf

Krumeich, J., Werth, D., & Loos, P. (2016). Prescriptive control of business processes: New potentials through predictive analytics of big data in the process manufacturing industry. *Business & Information Systems Engineering, 58*(4), 261-280. doi:10.1007/s12599-015-0412-2

Lederer, M., Betz, S., Kurz, M., & Schmidt, W. (2017). Some say digitalization - others say IT-enabled process management thought through to the end. Paper presented at the 1-10. doi:10.1145/3040565.3040574

Leemans, S. J., Fahland, D., & Van der Aalst, W. M. (2014, September). Exploring processes and deviations. In *International Conference on Business Process Management* (pp. 304-316). Cham: Springer.

Leemans, S. J., Fahland, D., & Van der Aalst, W. M. (2016, September). Using life cycle information in process discovery. In *International Conference on Business Process Management* (pp. 204-217). Cham: Springer.

Mans, R. S., Schonenberg, M. H., Song, M., Van der Aalst, W. M., & Bakker, P. J. (2008, January). Application of process mining in healthcare–a case study in a dutch hospital. In *International joint conference on biomedical engineering systems and technologies* (pp. 425-438). Berlin, Heidelberg: Springer.

Maruster, L., Weijters, A. T., Van der Aalst, W. M., & Van den Bosch, A. (2002, November). Process mining: Discovering direct successors in process logs. In *International Conference on Discovery Science* (pp. 364-373). Berlin, Heidelberg: Springer.

Maruster, L, & Van Beest, N. R. (2009). Redesigning business processes: a methodology based on simulation and process mining techniques. *Knowledge and Information Systems*, *21*(3), 267.

Muñoz-Gama, J., & Carmona, J. (2010). A fresh look at precision in process conformance. (pp. 211-226). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-15618-2_16

Eindhoven University of Technology - ProM 6.8 Plug-ins (2019) Integrated SCM & project management. Retrieved from https://svn.win.tue.nl/trac/prom/wiki/ProM68/Plugins

Page, S. (2015). *The power of business process improvement: 10 simple steps to increase effectiveness, efficiency, and adaptability*. New York: American Management Association.

Rozinat, A., Alves de Medeiros, A. K., Günther, C. W., Weijters, A. J. M. M., & Van der Aalst, W. M. (2007). *Towards an evaluation framework for process mining algorithms* (BPM reports; Vol. 0706). Eindhoven: BPMcenter.org.

Rozinat, A., & Van der Aalst, W. M. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, *33*(1), 64-95.

Rozinat, A., de Jong, I. S., Günther, C. W., & Van der Aalst, W. M. (2009). Process mining applied to the test process of wafer scanners in ASML. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *39*(4), 474-479.

Rubin, V. A., Mitsyuk, A. A., Lomazova, I. A., & van der Aalst, W. M. (2014, September). Process mining can be applied to software too! In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (p. 57). Association of Computing Machinery.

Sumak, B., Hericko, M., & Pusnik, M. (2011). A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types. *Computers in Human Behavior*, *27*(6), 2067-2077.

The Cloud Awards (2019). Retrieved from https://www.cloud-awards.com/2019-shortlist/

Van der Aalst, W. M., Weijters, A. J. M. M., & Maruster, L. (2002). *Workflow mining: Which processes can be rediscovered* (BETA Working Paper Series, WP 74). Eindhoven: Eindhoven University of Technology.

Van der Aalst, W. M., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128-1142.

Van der Aalst, W. M., & Alves de Medeiros, A. K. (2005). Process mining and security: Detecting anomalous process executions and checking process conformance. *Electronic Notes in Theoretical Computer Science*, *121*, 3-21.

Van der Aalst, W. M., De Medeiros, A. A., & Weijters, A. J. M. M. (2005, June). Genetic process mining. In *International conference on application and theory of petri nets* (pp. 48-69). Berlin, Heidelberg: Springer.

Van der Aalst, W. M., Reijers, H. A., Weijters, A. J., Van Dongen, B. F., Alves de Medeiros, A. K., Song, M., & Verbeek, H. M. W. (2007). Business process mining: An industrial application. *Information Systems*, *32*(5), 713-732.

Van der Aalst, W. M. (2011). *Process mining: Discovery, conformance and enhancement of business processes* (Vol. 2). Berlin, Heidelberg: Springer.

Van der Aalst, W. M. (2012). Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems (TMIS)*, *3*(2), 7. DOI = 10.1145/2229156.2229157 http://doi.acm.org/10.1145/2229156.2229157

Van der Aalst, W.M., Adriansyah, A., & Dongen, B.F. (2012). Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*, 182-192. doi:10.1002/widm.1045

Van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & Van Der Aalst, W. M. (2005, June). The ProM framework: A new era in process mining tool support. In *International conference on application and theory of petri nets* (pp. 444-454). Berlin, Heidelberg: Springer.

Van Eck, M. L., Lu, X., Leemans, S. J., & van der Aalst, W. M. (2015, June). PM^2: A Process Mining Project Methodology. In *International Conference on Advanced Information Systems Engineering* (pp. 297-313). Cham: Springer.

Weijters, A. J. M. M., & Van der Aalst, W. M. (2001, January). Rediscovering workflow models from event-based data. In *Proceedings of the 11th Dutch-Belgian Conference on Machine Learning (Benelearn 2001)* (pp. 93-100).

Weijters, A. J. M. M., Van der Aalst, W. M., & Alves de Medeiros, A. K. (2006). *Process mining with the heuristics miner-algorithm* (Technical Report WP, 166). Eindhoven: Eindhoven University of Technology.

Weijters, A. J. M. M., & Ribeiro, J. T. S. (2011, April). Flexible heuristics miner (FHM). In *2011 IEEE symposium on computational intelligence and data mining (CIDM)* (pp. 310-317). IEEE.

Wen, L., Van der Aalst, W. M., Wang, J., & Sun, J. (2007). Mining process models with non-free-choice constructs. *Data Mining and Knowledge Discovery*, *15*(2), 145-180.