FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

ASSIGNMENT OF BACHELOR'S THESIS

Title:	Process Mining in Finance Domain
Student:	Katarína Krbilová
Supervisor:	Ing. Marek Skotnica
Study Programme:	Informatics
Study Branch:	Information Systems and Management
Department:	Department of Software Engineering
Validity:	Until the end of winter semester 2020/21

Instructions

In the domain of finance, there are very strict regulations on internal business processes. Moreover, the companies in this industry are subjects to audits by the central authorities. Process mining (PM) is a technique to extract business processes from event logs already existing in the organization. A goal of this thesis is to apply process mining tools on company IT systems event logs and help them to make sure, that they are compliant with regulations.
Steps to take:
1. Explore the state-of-the-art process mining in the finance domain.
2. Explore what are the challenges related to governance, risk and compliance in business processes in this domain.
3. Propose a way how to use process mining to help solve the problems.
4. Create a case study to demonstrate the benefits of process mining applied in the finance domain.

References

Will be provided by the supervisor.

Ing. Michal Valenta, Ph.D.
Head of Department

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
Dean

Prague February 28, 2019

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Bachelor's thesis

Process Mining in Finance Domain

Katarína Krbilová

Department of Software Engineering
Supervisor: Ing. Marek Skotnica

December 15, 2019

Acknowledgements

Firstly, I would like to thank my supervisor for his support, always positive attitude, and constructive feedback. Secondly, my gratitude belongs to my family and friends for creating an optimal environment for writing and for their support. Namely, I would like to thank my fellow colleague Jirka for his negative motivation, and valuable recommendations, and to my friend Kuba for providing proper nourishment for my brain cells. Lastly, I appreciate each living entity willing to sacrifice their time to read this thesis.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as school work under the provisions of Article 60(1) of the Act.

In Prague on December 15, 2019 . . . . . .

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Krbilová, Katarína. *Process Mining in Finance Domain.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2019.

Abstract

Rapidly growing volume of data automatically collected in large institutions offers interesting possibilities to analyze them. Financial institutions depend on many systems that are logging activities of its users, and saved data often contains complex processes. Process mining offers a way to extract processes from large data-sets and perform various types of analysis over them. Since financial institutions are subjects to rules established by different authorities, they must govern their processes in a compliant way and identify risks rigorously. The main purpose of this thesis is to introduce possible use of process mining in the domain of finance to solve issues related to governance, risk, and compliance of a process. The base for the field of process mining is provided in the theoretical part, to understand the analysis in the empirical part. The practical case study demonstrates the application of process mining to mine a model from data-set originating in the financial domain. Mined model is further investigated to obtain a general overview of the process and to disclose possible performance or conformance issues.

Keywords proces model analysis, process mining, compliance checking, GRC, finance domain, ProM

Abstrakt

Rýchlo rastúci objem automatizovane zbieraných dát vo veľkých inšitúciach ponúka zaujímavé možnosti ich analýzy. Finančné inštitúcie sú závislé od mnohých systémov, ktoré evidujú aktivitu svojich používateľov. Tieto dáta často obsahujú zložité procesy. Vďaka vyťažovaniu procesov z dát je možné extrahovať procesy z veľkých súborov dát a analyzovať ich. Finančné inštitúcie sa musia riadiť pravidlami vyšších orgánov, preto je dôležité, aby svoje procesy spravovali v súlade s týmito pravidlami a dôsledne identifikovali hrozby. Hlavným cieľom práce je predostrieť možné využitie vyťažovania procesov z dát v oblasti finančníctva na riešenie problémov súvisiacich s riadením, evidenciou risku a kontrolou súladu procesov s pravidlami. V teoretickej časti predstavujeme základné koncepty vyťažovania procesov z dát pre lepšie pochopenie analýzy v praktickej časti. Praktická časť demonštruje extrahovanie modelu zo súboru dát z oblasti finančníctva. Vyťažený procesný model je ďalej skúmaný s cieľom nadobudnúť všeobecný prehľad o procese a odhaliť možné výkonnostné problémy či nesúlad modelu a dát.

Klíčová slova analýza procesných modelov, vyťažovanie procesov z dát, kontrola compliance, GRC, finančný sektor, ProM

Contents

List of Figures

Introduction

Motivation and Objectives

Over the past ten years, significant growth in the usage of various types of information systems can be observed. [1] Most of the information systems have a potential to collect big amounts of data in the form of event logs that contain valuable information about the processes of the company. When processed, this information can provide insightful reports about the company's current activities that can be later used for future planning or risk management. Nowadays, companies need to make sure that their processes are compliant with current laws and regulations to avoid excessive fines. Laws and regulations are changing very quickly. It is estimated that companies spend over 7 percent of their total operating costs on compliance strategy. [2] This cost could be possibly decreased by monitoring compliance of processes using process mining.

Process mining is still a relatively new discipline that arose as a reaction to the fast growth of the collected data volume. It is one of the most efficient ways of giving sense to data gathered from information systems. Unlike data mining, process mining is mostly working with processes discovered in the event logs. It allows to analyze them in detail to identify bottlenecks, make predictions, or check compliance of these processes.

This thesis aims to explore the possible use of process mining for checking compliance with the law and regulations in the domain of finance. Apart from checking the compliance, the benefits of process mining are very broad. Process mining has the potential to reduce costs, help in naming increased risks in the organization, predict process deviations and violations, or even facilitate reacting appropriately to market dynamics. These aspects of the method are analyzed as well. The thesis also provides a base for further research in this field since possible applications of the process mining in finance haven't been thoroughly analyzed yet.

Problem Statements

Companies using many information systems possess big amounts of collected data. One of the important outputs of information systems is data related to the financial activities of the company. A great value lies in them, and it can be revealed by applying process mining.

Each country has its regulations and rules that have to be obeyed by the companies. Most of them concern the financial domain. Therefore, governance, risk management, and compliance checking become a necessary duty for each company. One of the possible methods to cover these needs is process mining. It helps to discover possible risks in processes of the organization, evaluate problematic processes of the company, check the reliability of financial statements, and provide other insightful analysis.

Process mining is sometimes considered to be a kind of statistical analysis with a strong theoretical base but, examples of its application are lacking. [3] Surprisingly, even after multiple proves of its efficiency when applied to bigger businesses, it's practical application in the financial domain is rather exceptional. Therefore the final goal of this thesis is to demonstrate use of the process mining in the domain of finance on a practical case study.

Structure

This thesis consists of two main parts, structured as follows:

- Theoretical research
  - Chapter 1 - presents governance, risk and compliance in the domain of finance.
  - Chapter 2 - introduces a brief overview of process mining techniques and explains the discovery and evaluation of a process model.
  - Chapter 3 - takes a closer look at possibilities of application of process mining in the financial domain.
- Empirical research
  - Chapter 4 - proposes a case study to demonstrate the use of process mining on a relevant event log. Detailed analysis of the process model is conducted, and its outcome is evaluated in the context of governance, risk and compliance.
  - Conclusion - summarizes the results of the theoretical research and the presented case study.

Chapter 1

Governance, Risk and Compliance (GRC) in Financial Domain

GRC standing for governance, risk, and compliance is a term for activities leading to the achievement of the company's objectives while addressing uncertainty, acting with integrity and proving reliability. [4] Beginnings of this collective name date to 2002, but companies have been facing GRC related challenges before the existence of this term. GRC has nowadays taken the form of "a continuous process that is embedded into the culture of an organization. It governs how management identifies and protects against relevant risks, monitors and evaluates internal controls, and improves operations based on learned insights." [5]

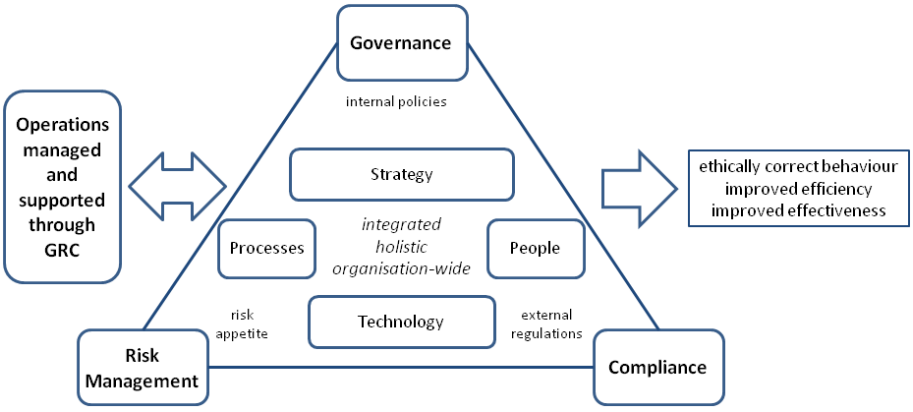


Figure 1.1: Frame of reference for integrated GRC [5]

Technology used to solve these challenges developed over the years as well. Many GRC supporting products are available on the market. GRC is no

longer represented by non-uniform operations in different areas of the company. Nowadays, it is rather a strategy allowing organizations to take full control over their processes, make them more effective, and compliant with all kinds of laws and regulations. GRC can also be a set of integrated concepts that should provide a competitive advantage for its user. [5] It can be seen from Figure 1.1 that GRC as an integrated concept binds more than just governance, risk management, and compliance. It allows building a strategy based on available resources, managing possible risks in a manner fitting the company's objectives, reacting to external regulations with the help of technology, building processes respecting internal policies, and much more. This chapter aims to provide a close-up on each term forming the abbreviation GRC.

1.1 Governance

Governance is a term covering a variety of responsibilities and actions. It represents the fusion of business requirements and technological resources of the firm. Governance should present a plan of allocation for all available resources so that the best possible performance is reached. Term *resources* combines both the technological and human resources of the company. Governance ensures that essential relationships between these resources are developed and accurately managed. Governance also provides guidelines for handling extraordinary situations that can emerge during a process. Following these guidelines should ensure that taken measures are aligned with the company's politics. Aside from previously mentioned activities, governance can also supervise innovations and regulate the expansion of the firm. The main focus of governance is the most efficient and effective achievement of given business objectives. The desired side-effect is a subsequent growth of the profit. According to Tarantino [6], three requirements for successful governance are present:

1. *Current state.* It is crucial to know the current state of the company. It can only be obtained from frank, unbiased reports of internal and/or external auditors.

2. *Vision.* Governance should keep in mind the objectives of the company combined with the goals of stakeholders and based on both propose a balanced vision.

3. *Monitoring of the progress.* Each firm needs to establish regular controls of the progress towards the achievement of its goals.

The last requirement clearly implies the need for monitoring and reporting the state of the company and tracking the fulfillment of its milestones.

1.2 Risk Management

Risks represent unfulfilled threats for each business. Operating a larger company requires proper risk management. In the domain of finance, it is possible to recognize three categories of risks:

- Risks naturally emerging from the use of technology - e.g., security breach, private data leak, unavailability of services or systems,
- Risk of violating external regulations - when processes are not being regularly reviewed for compliance with current laws and regulations, neither updated, firms can be held legally responsible and receive considerable fines,
- Risk of not adhering to internal regulations and policies - lack of control over internal processes can lead to inefficiency of the company and endanger its objectives.

The previous categorization can be supported by the explanation of operational risk in Basel II, which "is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events." [7] Misconduct in a process can be predicted or even avoided if all processes are being monitored.

It is recommended to keep a regularly updated table of identified risks with the following items:

- Description of the risk - information about the source of the risk,
- Probability of the situation - how likely it is for the situation to emerge and what time horizon it endangers,
- Plan for mitigation of the risk - strategy to minimize impacts of the risk or, in the ideal case, eliminate it,
- Crisis plan - set of operations aligned with the company's strategy that can be used in the worst-case scenario to minimize damages of the feared situation,
- Owner of the risk - a person responsible for monitoring the possible risk and for implementing measures leading to its mitigation,
- Possible impacts of the risk - quantified damages that the risk represents, including material damages as well as moral ones. [8]

This table should be regularly reviewed and prioritized. Subsequently, high as well as low-level management has a good overview of possible threats and can react briskly to unexpected changes.

1.3 Compliance

According to Tarantino: [6] "Compliance is a concept of acting in accordance with established laws, regulations, protocols, standards and specifications". On the one hand, as mentioned in the introduction, remaining compliant with all laws can represent a financial burden for the company. On the other hand, consequences of noncompliance are varying from civil and financial ones up to reputational ones therefore companies prefer to invest in high-quality compliance solutions.

Larger firms and especially financial institutions like banks, also have many internal regulations and policies that their employees need to follow. Assuring compliance with internal policies and external laws can become a challenge. To achieve this, companies adopt regular internal controls that are either manually performed or automated. Automation of controls is designed to reach better objectivity and lower error rate. These controls should ensure not only overall compliance but also the reliability of financial reports. Internal auditors can use the results of controls to alert management to inaccuracies and mistakes in their reports.

1.4 GRC Related Challenges

Companies are investing in GRC technologies and solutions for many reasons. They often hope to obtain improvements in administration processes, reduce costs of separate process controls by adopting reasonable automated continuous controls, and boost both efficiency and effectiveness in order to reach their business goals faster. Process mining side by side to the technological GRC solution can help the organization in many areas. It can find the smallest particular problems to be addressed, up to general issues related to GRC.

Deployment of technological solution is often not enough to meet previously mentioned expectations and business needs. As presented by Hunt [9], many companies fail to meet their objectives. Firstly, sufficient time has to be spent on finding the right GRC solution. Vendors offer similar products, but small differences between them can make a big change for the buyer. These differences are usually documented in materials from Forester Wave like [10]. Secondly, organizations focus on the technical side of deployment of a GRC solution but often forget about business changes that need to be done and neglect a proper introduction of new systems to employees. Both are crucial for the adoption of a GRC solution. Otherwise, faults in the usage of the technology can endanger its performance and asset to the company. Sporadic faults are hard to find. This is the opportunity to use process mining.

The previous example was very specific, but process mining can be used to solve more general issues too. Opportunities for its application appear in risk management and also in compliance checking. The developed risk strategy of

a firm is not sufficient if it's not reviewed regularly. Process mining can offer guidance to the improvement of risk management processes and help with their adjustment by identifying their weaknesses. Fundamentals of process mining and more examples of its usage can be found in Chapter 2 and Chapter 3.

Chapter 2

Process Mining

Introducing process mining is not entirely possible without first explaining its relation to process modeling and data mining. It could be described as the missing connection between these two disciplines. Process mining takes the best of process modeling and data mining to deliver reliable and detailed data analysis. Why is data mining or process modeling not enough on its own? When analyzing data, data mining does not take into consideration any internal processes that can be hidden in the collected data. Data mining is purely data-centered and therefore limited by the quality of analyzed data. Four main factors also called 4Vs of big data determine the quality of the data:

- Volume - the quantity of the data,
- Velocity - the source of the data,
- Variety - multiple types of collected data,
- Veracity - the uncertainty of the data. [11]

Since data mining is unable to discover processes, a big part of the information hidden in the data remains unexplored. Process modeling, on the other hand, is oriented mainly on processes. Process models are often viewed as a way to capture the transition between inputs and outputs. Another definition describes them as networks in which a number of roles collaborate and interact to achieve a business goal. Both of these views forget about the primary added value of process models that lies in their ability to represent the human factor. [12] This potential of process models has been overlooked for a long time. Nowadays, process models also aim to discover how people behave. Still, the representation is not always genuinely accurate because people tend to break the rules or find new ways of performing certain acts.

Process mining is an interconnection of multiple disciplines as visible from Figure 2.1. It binds algorithms from data science and game theory with con-

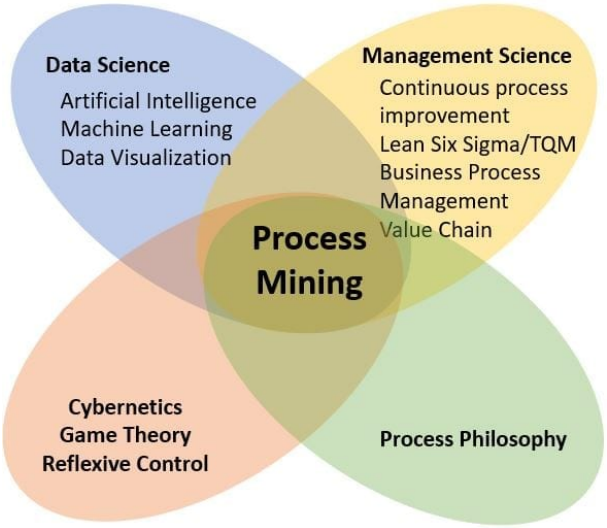


Figure 2.1: Interconnecting disciplines forming process mining [13]

cepts of management sciences, mainly those related to process management. It is a well-balanced method of analyzing business processes. It shows its strengths in extracting the most useful from mined data and giving them informational value by creating a well-founded process model for further research. In relation to process modeling, big advantage of process mining is also its compatibility with different process model notations such as BPMN, UML, Petri Nets. . .If needed, process mining can be used for various types of analysis from general ones, e.g., tracking the processes performance statistics or identifying bottlenecks to very detailed ones serving for audit purposes or predictions. Based on a detailed analysis, it is easier to improve processes.

The following sections contain a brief overview of the most commonly used process mining techniques, such as the alpha algorithm or heuristic miner. They also aim to introduce the practice of discovering the process model from mining the data up to the evaluation of the model's quality.

2.1 Process Mining Practices

The discipline of process mining does not solely consist of creating the process model based on collected data. The preliminary condition for conducting a successful process mining project is the right data selection as well as forming questions relevant to the analysis of critical processes in the company. Based on the presented questions, an appropriate process mining technique needs to be chosen. The following sections aim to introduce the correct form of event logs, the topic of data filtering, and types of questions that can be answered by process mining.

2.1.1 Data Selection and Preparation

At the beginning of each process mining project is a collection of raw data. Data usually comes from an IT system that regularly logs the activities of its users. Data can be stored in a database, CSV file, or a log. Information about one entire process can be spread across multiple systems that are being used in the company, and its form can vary from transaction logs to mail archives. Hence it can be challenging to extract all necessary data of sufficient quality. To obtain maximal value from the data, it is recommended to incorporate various sources in the analysis.

Efforts to standardize logging of data were recognized, yet rapidly augmenting quantity of software using process mining algorithms is opposing them. Attempt to unite the form of logs used by numerous software vendors would be more than demanding. Nonetheless, academia adopted XES event log format. This format is based on the fact that each case, e.g., customer refund in an electronics shop, has multiple traces. Traces document what happened during one particular handling of the process. For example, the refund must have been additionally reviewed by a manager, or it was denied, but customer filed a request for the refund again. Each trace is composed of events that should refer to activity. "Events represent atomic states of an activity that have been observed during the execution of a process." [14] Event logs that are extracted from systems should minimally contain:

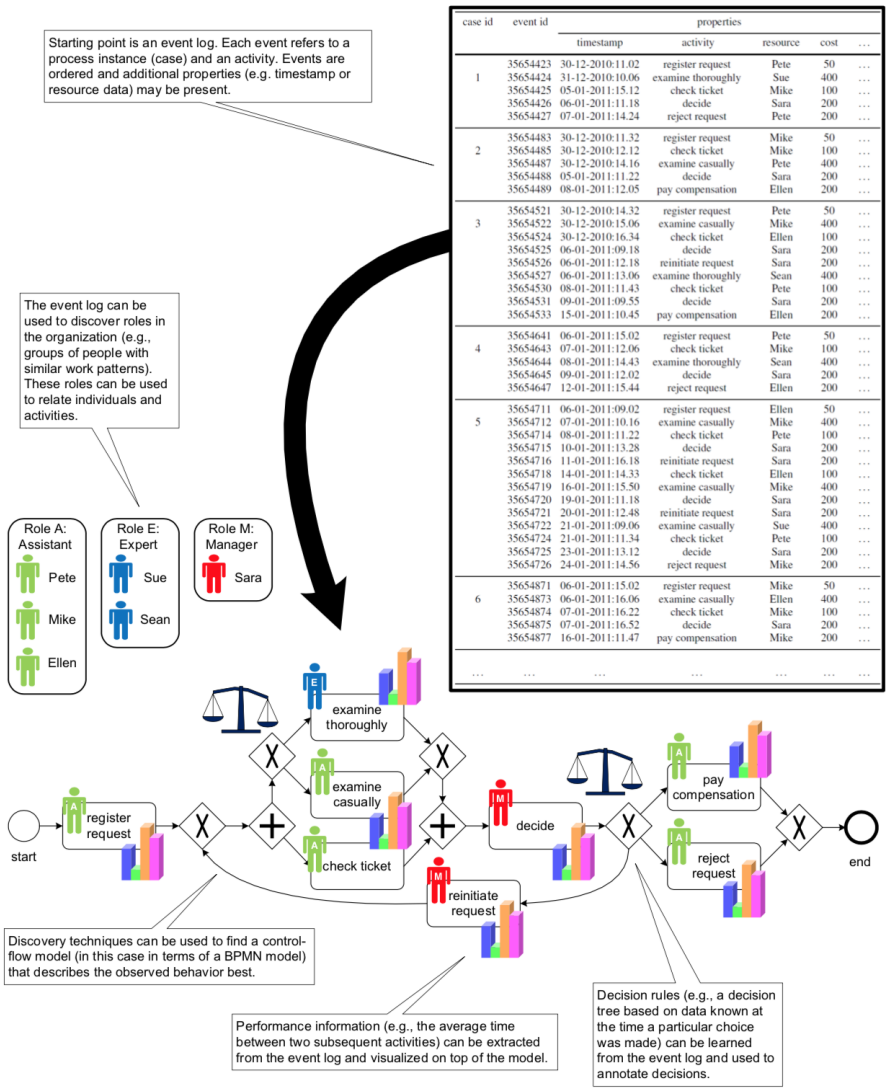
Case ID - unique identification of the activity corresponding to one specific flow through the process,

Activity name - clear characterization of the activity that represents one step in the process, this step can occur in multiple traces,

Timestamp - date and/or time when the activity was performed, allows to position different activities in time and determine their successions. [15]

This is the minimum of information required for the extraction of a process model. Whenever present, process mining also uses information about the resource, meaning the person or the system that executed the activity. Logs can include much more information, such as the duration of the activity, start, delay, or completion of the activity and other case attributes and data elements that may have been recorded with the event. Figure 2.2 shows an example of a log and explains what kinds of information can be retrieved from it.

It may seem that the more data available, the more precise model can be extracted, but that is usually a false assumption. In process mining, one of the crucial first steps before the extraction of the model is data filtering and preprocessing. Too much data can lead to unnecessarily complicated model or even hide unexpected and suspicious behavior that was meant to be observed. Furthermore, obtained data can be incomplete, imprecise, or even incorrect. In such cases it is recommended to separate incorrect events and explicitly



case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	decide	Sara	200	...
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	check ticket	Pete	100	...
	35654531	09-01-2011:09.55	decide	Sara	200	...
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	register request	Pete	50	...
	35654643	07-01-2011:12.06	check ticket	Mike	100	...
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	decide	Sara	200	...
	35654647	12-01-2011:15.44	reject request	Ellen	200	...
5	35654711	06-01-2011:09.02	register request	Ellen	50	...
	35654712	07-01-2011:10.16	examine casually	Mike	400	...
	35654714	08-01-2011:11.22	check ticket	Pete	100	...
	35654715	10-01-2011:13.28	decide	Sara	200	...
	35654716	11-01-2011:16.18	reinitiate request	Sara	200	...
	35654718	14-01-2011:14.33	check ticket	Ellen	100	...
	35654719	16-01-2011:15.50	examine casually	Mike	400	...
	35654720	19-01-2011:11.18	decide	Sara	200	...
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	...
	35654722	21-01-2011:09.06	examine casually	Sue	400	...
	35654724	21-01-2011:11.34	check ticket	Pete	100	...
	35654725	23-01-2011:13.12	decide	Sara	200	...
	35654726	24-01-2011:14.56	reject request	Mike	200	...
6	35654871	06-01-2011:15.02	register request	Mike	50	...
	35654873	06-01-2011:16.06	examine casually	Ellen	400	...
	35654874	07-01-2011:16.22	check ticket	Mike	100	...
	35654875	07-01-2011:16.52	decide	Sara	200	...
	35654877	16-01-2011:11.47	pay compensation	Mike	200	...
...	...	...	...	...	...	...

Figure 2.2: Process mining extracts knowledge from event logs in order to discover, monitor and improve processes. [16]

indicate uncertain data. In the final data-set used for mining the process, only relevant attributes related to the asked questions should be present. Process Mining Manifesto defines event log maturity levels that rank logs from 5 as best to 1 as worst considering how they were created. Highest quality logs are those recorded in "automatic, systematic, reliable and safe manner". [16]

2.1.2 Posing Relevant Questions

Questions that are expected to be answered by process mining need to be well defined. On the one hand, it is useful for the early data selection, before extraction of the model and for proper choice of process mining technique, on the other hand, it plays a big role in choosing correct visualization for the end-user. Some questions may emerge naturally based on the character of available data. Other questions can also be posed by the management of the company, but a consultant shall be present to transform them into a form suitable for process mining. Most frequently asked process mining questions are divided into two main categories by van der Aalst [11]:

Performance questions, addressing the effectiveness and smooth flow of the process. E.g.:

- Where do delays happen the most often?
- Where does the delay start?
- Which resources receive tasks that are piling up?
- What is the fastest trace in the process?

Conformance questions, addressing the reliability of the process and it's alignment to the company's politics. E.g.:

- Which steps of the process are often omitted?
- Which principles are violated?
- Which person is present the most often when violations happen?
- How many traces represent a non-compliant behavior?

Moreover, many other questions can be answered by process mining. In the obtained process model, deviations and unusual traces are observed. Repetitive delays or forbidden shortcuts are examined in order to detect their origin. Knowing the origin of deviations allows forming predictions. This topic will be further covered in Chapter 3.

Filtering the data and posing questions for the analysis are activities closely tied together preceding the construction of the model. We need to pose such questions that can be answered by available data and only choose data that are related to posed questions. Both of these activities are interconnected and should be performed with regard to each other. The next section is going to cover different process mining techniques and algorithms for the creation of a model.

2.2 Mining the Process Model

The previous section presented what data needs to stand at the input of a process mining algorithm and which problems can be revealed by process mining. To interpret answers correctly, a fitting method and algorithm need to be applied. The following sections discuss differences between the most frequently used algorithms and between utilization of three fundamental process mining methods.

2.2.1 Three Types of Process Mining

Process mining offers an ample range of applicability, though three basic techniques can be recognized:

Discovery - *play-in* - process model is inferred from raw data, it is purely based on available examples, no a priori information is used. This technique provides the first insight into the logged behavior.

Conformance checking - *replay* - compares the original process model with real-life data and evaluates if it corresponds to reality. Violence and negligence of the existing process can be unveiled together with specific deviations.

Enhancement - *replay* - aims to extend or improve the original process model by using the additional information available in the recorded log. Aside from extending the model with new perspectives, it also allows its reparation and adjustment to real-life needs. [11]

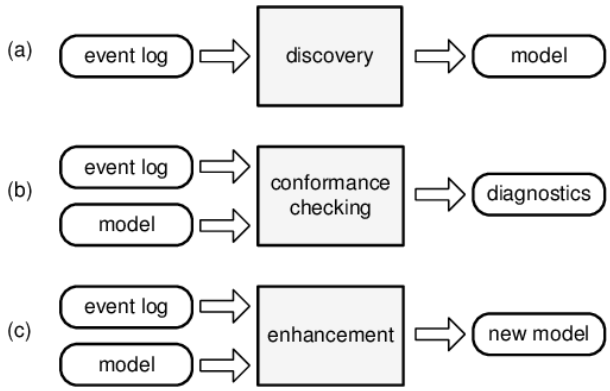


Figure 2.3: The three basic types of process mining explained in terms of input and output: (a) discovery, (b) conformance checking, (c) enhancement [16]

Figure 2.3 shows the comparison of inputs and outputs of the different methods. Process mining always works with available logs, but it is important

to stress that they do not cover the complete representation of reality and only contain a fragment of all possible behaviors.

These techniques can be sorted into three categories, play-in, replay, and play-out. [17] The last one that hasn't been mentioned yet serves only to generate behavior from an already existing model. It can be used to verify that incorrect traces cannot take place. The replay technique can also be used to simulate real behavior on the model and conduct performance analysis.

2.2.2 Process Mining Algorithms

The previous section described different applications of process mining. After choosing the technique fitting posed questions, an algorithm needs to be selected too. The detailed description of different algorithms is beyond the scope of this thesis. Therefore this section offers just a brief overview of the most frequently used algorithms.

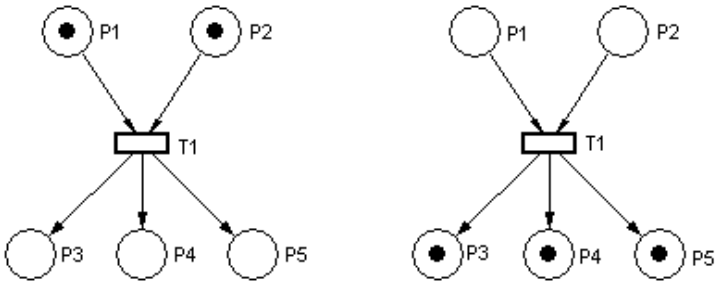


Figure 2.4: Transition in a Petri net before and after execution [18]

Most of the process mining algorithms work with Petri nets. As explained by Peterson [19], Petri nets allow to modeling informational flow in the process. They are constructed from so-called places (circles) and transitions (rectangles), as can be seen in Figure 2.4. Transitions are the activities that are executed during the process. Often some other activities precede them obligatorily. This is represented by tokens 'traveling' in the net. To execute a transition, tokens need to be present in all places that stand right ahead of it. When the transition takes place, new tokens are produced on all outputs places of the transition.

Many categories of process mining algorithms were researched, e.g., deterministic algorithms, evolutionary approaches, region-based mining, or inductive mining. Deterministic algorithms producing process models directly are considered the base of the process mining. Some of the well known are α-algorithm, heuristic miner, or fuzzy miner. To complete the list of most important algorithms, the inductive miner is shortly presented as well, since it is going to be used later in the practical case study.

α-algorithm discovers relations between events and mostly works with the order of activities. Based on these relations, the algorithm creates a footprint of the log. The footprint is a table noting down relations between all events, such as succession, parallel execution, etc. It is possible to create a footprint from an event log as well as from an existing model. From the footprint, the algorithm is able to recreate traces in the process that are visualized in the final model. Downsides of this basic algorithm are mostly its inability to work with short loops or non-local dependencies between events, and creating too wide Petri nets with unnecessary, rare behavior, meaning the algorithm is not performing well when dealing with noise.

Heuristic miner works with direct successions found in the log as well, but it is possible to specify how many times activity A should be followed by activity B to be incorporated in the final dependency graph. This approach ensures that unimportant behavior is excluded from the final model, and only meaningful causalities between actions are explored. Setting thresholds for dependencies manually helps to deal with noise and incompleteness. On top of it, this approach allows creating and comparing multiple process models with a different concentration on details.

Fuzzy miner deals with noise and incompleteness at the beginning. It offers a wider range of setting thresholds and even allows a precise selection of behavior that should be included in the resulting model. Consequently, it is possible to construct multiple levels of process model containing either only general process flow or also sub-processes with infrequent behavior.

Inductive miner works with process trees instead of Petri nets. Their considerable advantage compared to other notations, is that their construction doesn't allow deadlocks, livelocks, or other anomalies by default. The inductive miner can deal very well with noise in huge logs while ensuring the precision of the model. [11]

Many modifications of the mentioned algorithms are available, and it can be challenging to choose the right one. For the general overview, the heuristic miner is often used, since it is capable of mining different models thanks to the possibility of setting thresholds.

2.3 Evaluating the Discovered Model

Evaluating the quality of the discovered model correctly is valuable for the final outcome of the process mining project. It helps to recognize the need to change the algorithm or the need to extend the model with more perspectives

if required data is available. This chapter presents the manners of grading the quality of the mined model, together with different perspectives and visualizations that can be employed to reach the final model.

2.3.1 Analytical Biases

When mining a model, it is necessary to be aware of possible biases that can influence the result. Three biases influencing the analysis of the model can be distinguished:

Inductive bias is a preference for a certain solution based on external factors.

Learning bias results from using a particular algorithm that may prefer or accentuate certain solutions.

Representational bias is limiting the solution because the expressive power of a used language may be restricted. [11]

Most of the notations used to represent processes contain the same constructs, such as splits, joins, choices. Therefore it is possible to transform almost any model into a different notation. Nevertheless, process mining algorithms still struggle with some constructs that may appear in process models, specifically with loops or unbalanced splits and joins.

A mined model can be disturbed by more factors than just bias. Problematic can be unrelated traces in the log, that do not represent typical behavior in the process. Such traces are called noise. Certain algorithms are able to deal with noise, e.g., heuristic mining, genetic mining, or fuzzy mining. [11] Challenge is also posed by the incompleteness of the data, as mentioned in earlier sections.

2.3.2 Quality Criteria and Validation of the Model

To estimate the quality of the model, multiple criteria can be used. The following list merges criteria mentioned in [11] and in the free introduction into the suggested evaluation framework for process mining algorithms by Rozinat [20].

- *fitness* - shows how much of the behavior captured in the log is also represented by the mined model,
- *simplicity* - indicates the complexity of the model, how many nodes or arcs in the graph are present,
- *precision* - the precise model should not contain forbidden or unobserved behavior,

- *generalization* - the general model should not be too precise, matching only examples from the log but also allow different possible traces,

- *structure* - represents how the behavior is captured and depends on the chosen modeling formalism,

- *entropy* - addresses the uncertainty and variability of the model and its ability to represent various behaviors.

Some of these criteria can be transformed into measurable variables, just like metrics for quality of process models were presented in [21]. Figure 2.5 represents that certain criteria are opposites of each other, like precision and generalization, and a good balance needs to be found between them. With precision and generalization are associated terms overfitting and underfitting. Underfitting happens when the model allows behavior too different from the one captured by the log, meaning the model is too general. Overfitting is the exact opposite when the mined model is way too specific and captures very limited behavior.

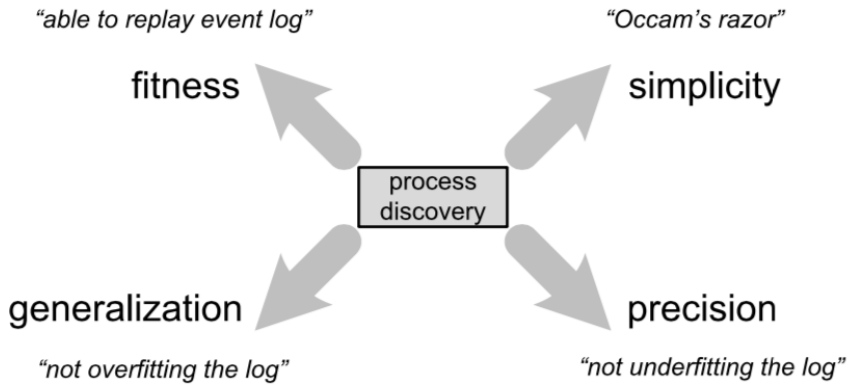


Figure 2.5: Quality criteria influencing the mined process model [11]

To find the correct ratio between precision and generalization, a concept often used in data mining proves its strength. Cross-validation splits data (event log in this case) into a training log and test log. The training log serves for the creation of the model, and test log allows evaluating how well the mined model captures unseen cases. The advanced method is k-fold cross-validation when the log is split into k equal parts, and multiple tests are performed to obtain more reliable result. This method provides a comparison of mined models and helps to choose the best one.

2.3.3 Model Perspectives and Final Representation

Benefits of the process mining come from its complexity. Firstly, process mining is able to work with many existing model notations allowing consequently to present its results to the wide public, varying from data scientists to top management of companies. Secondly, the model can reveal many different perspectives. Event logs containing additional process data logged with each event can enrich the representation. Process mining works most productively with these perspectives:

Control-flow perspective - simplest view on the process model, displaying order of activities. The model tries to capture possible paths in the process.

Organizational perspective - offers additional information about resources, from particular roles or systems to departments up to relations and interactions between them.

Case perspective - uses data elements recorded with particular events to extend the model with values that are changing in the process.

Time perspective - enables to analyze timing and frequency of events that further show bottlenecks, utilization of resources, or predictions. One can broaden the model only if timestamps are properly recorded. [11]

All of these perspectives grant access to a multitude of insights into the functioning of the process. Those will be further discussed in Chapter 3. Despite the incredible versatility of the process mining, it is important to stress that "suitability of the model ultimately depends on the questions one would like to answer". [11]

2.4 Conclusion

Conducting a process mining project is a complex task where a variety of inputs needs to be considered. This chapter provided insight on possible steps of the process mining project from the preparation of the data and its correct form, up to model refinement and extension, through the choice of correct process mining technique and algorithm. The project does not end with the discovery of the model. Modeling language to display the model has to be considered, and analysis of facts captured by the model has to be performed. Possibilities of analytics over the mined process model and its benefits for managing governance, risk, and compliance in the company are introduced in the later chapter.

Chapter 3

Process Mining in Finance

Process mining as a discipline firstly appeared at the threshold of a new millennium when IBM Research Center and Software Engineering Research Laboratory at the University of Colorado explored mining process models from workflow logs and event-based data in [22] and [23]. Since that time, process mining has undergone large theoretical development, but its use is still hardly noticeable, especially in the domain of finance. The year 2018 can be considered a breakthrough in the process mining since banks started adopting this new way of analyzing processes. In the following chapters, possible applications of process mining in finance will be introduced as well as examples and benefits of its use.

3.1 Motivations for Applying Process Mining in Finance

The use of process mining in the domain of finance has lately proven it's tremendous impact when implemented in VTB24 bank and Piraeus bank as described in [24] and [25]. VTB24 estimated that the opening of an account would take around three hours. However, process mining showed that only 13 percent of new accounts were opened within 3 hours. [26] Department of process efficiency was not aware of this fact before, because they lacked a tool that could offer an insight into this process. Thanks to process mining, critical delays in the process could have been identified. Analysis of processes also helped to discover employees who missed training for a new IT system and consequently were not using it correctly. Data from 15 different IT systems running in the bank were used for the analysis.

Another example is the Piraeus bank, where the IT department estimated that their newly automated process of consumer loans was working well, yet they were receiving complaints from customers as stated by QPR Software company in [27]. Traditional analysis of the process did not show any problem,

but process mining did. Bank was able to identify the bottlenecks and optimize the process in order to make their customers satisfied.

These two examples demonstrate only a fragment of possible applications of process mining and of the positives it can bring. Process mining can be beneficial across all three areas of the GRC, as shown in the following chapters.

3.2 Analysis of Processes

Process mining can be used not only for an in-depth analysis of the company's processes visualized with a model but also for its improvement or extension. Furthermore, it is possible to perform a simulation on the modified model and predict its positive or negative impacts. To fully understand where do the outcomes of the process mining project originate from, it is needed to introduce the topic of conformance checking.

3.2.1 Conformance Checking

Conformance checking is a mean of evaluating the fitness of the mined model. Fitness determines how well the constructed process model captures real-life processes. In other words, one is able to see true events that are happening, not just events allowed by a process model that was designed as optimal by the management. Various methods of evaluating the fitness of the model are available:

Causal footprints - footprints were explained in Section 2.2.2. Since both models and event logs have footprints, this technique offers multiple utilizations. It allows a comparison of different models or event logs or also a model with an event log. It is often applied to verify that the mined model represents a good match to the reality captured by the event log. This technique doesn't take into consideration frequencies of events, and it tries to capture various metrics in a single one, which results in compact but not detailed analysis.

Token-based replay - in order to use this technique, the model needs to be first converted into a Petri net. Traces are replayed on the Petri net and missing, or remaining tokens are counted. Counting tokens provides detailed diagnostics of places in the Petri net where deviations take place, and it is possible to determine how often they happen. Traces can be separated into groups of fitting and non-fitting ones. Unfortunately, by restricting the input of the technique, it's flexibility may be insufficient, yet it's outcome remains easily understandable.

Alignment-based checking - checking the conformance should not be limited by the technique used. The previous technique worked only with Petri nets. Alignment is universal because it compares logged traces

with traces enabled by the model and counts in how many steps this trace doesn't correspond to the model. Based on this knowledge, the fitness of the model is estimated. Alignment is a very flexible technique providing a detailed analysis of the process. [11]

All mentioned techniques allow tracing deviations. The characteristics of the deviation and additional data available in the logs offer better insight into specific problems in the process. Different examples are mentioned in the following sections.

3.3 Governance

The application of process mining in governance is very broad. All companies want to achieve optimal governance settings. The following sections explain what process mining can offer to improve processes or even organizational structures.

3.3.1 Performance Analysis

As shown in Section 3.2 process mining is able to identify bottlenecks in the processes by observing waiting times or completion times if timestamps are available in the log. Observations can be turned into performance statistics. Causes of bottlenecks can be revealed, e.g., locating incompetent resource or too complicated condition to perform a step in the process. Consequently, it is possible to address the discovered problems individually.

3.3.2 Resource Management

Using process mining, it is possible to discover unexpected connections between human or technological resources. On the one hand, process mining can uncover inefficiencies in the communication of certain people or groups of people or even drawbacks of systems that are being used. On the other hand, it can also point out groups that are functioning very well together. The resource-activity matrix is often used, as it displays responsibilities and competencies of different actors in the process. The resource-activity matrix is also covered by the term *staff assignment rules*. "Staff assignment rules define, to a certain extent, the profile of agents capable of or eligible for performing an activity." [28] Roles of resources can be determined, and important resources having multiple responsibilities can be located. Furthermore, management can ease the workload posed on burdened resources to improve the performance of the process.

Resource analysis provides grounds for reallocating resources to suit process needs better. Imbalance of resources can be unveiled, and individuals

who are struggling with some tasks can be assigned to the team with more experienced colleagues who can help them. Skillful newcomers can be identified, and management can make them progress in the company faster to ensure their personal growth and motivation.

3.3.3 Improvement of the Process

Process models are either *descriptive*, meaning they do not pose restrictions on the process and only try to capture existing situation, or they are *prescriptive* and suggest the way that process should be executed. [29] Descriptive models can be modified based on the outcome of conformance checking. The original model can be extended with behavior that was not captured so far but represents a valid trace in the process.

In many companies, IT and business are working independently of each other. Processes are changing in time, and soon it can be found that they are not aligned well anymore. Updating processes regularly can eliminate a lot of errors and confusion. Before putting a new process model into practice, it is extremely useful to test it. Process mining offers the possibility to simulate the behavior of a new process model. Multiple alternatives of a model can be tested, and optimal modifications can be adopted, e.g., new quantities or allocations of resources. Besides, when stakeholders are planning the future development of their product or company, they are often considering various options. Options usually represent possible optimization of processes or their modification. With process mining options can be explored and simulated, and it is easier to predict the possible benefit they could have for the company.

3.3.4 Operational Support

Process mining is mostly used for the analysis of traces in the process that already finished. The capability of process mining to work with running cases is often neglected. Based on previously collected data, predictions about the current execution of the process can be made. Three main actions can be taken when exploring the process at run-time:

- Detection - when deviation happens and is observed in the log, an alert is generated,
- Prediction - based on the context predictions are made, e.g., about the completion time of a running case,
- Recommendation - multiple future traces are compared, and the best one can be chosen. [11]

Thanks to predictions, companies can learn from their past mistakes, and thanks to recommendations, they can further optimize internal processes.

3.4 Risk Management

Examples of the use of process mining to tackle risks in the company are often missing in the literature. Nevertheless, suggestions for practical applications of process mining for risk management can be derived from its assets in governance and compliance.

As stated in Chapter 1, most of the risks arise from internal or external non-compliance of processes. If processes are monitored, regularly reviewed, and modifications suggested in Section 3.3 are applied early enough, risks of non-compliance can be mitigated. Multiple entities in the company can benefit from the process mining due to activities they are responsible for:

- Management - sets the risk management process of the organization and designs internal control system,
- Internal Audit - objectively evaluates quality and adequacy of the risk management and control measures, provides recommendations,
- External Audit - provides a completely independent opinion on the reliability of statements capturing the current status of the processes in the company. [30]

Only management uses process mining directly to design and maintain risk management processes. As shown in the previous section, risks in the form of bottlenecks or deviations can be identified and relationships between risks can be uncovered, and should be properly documented. Furthermore, traces of future risks can be identified in historical data. Internal audit can use process mining to provide recommendations. Contrarily external audit uses it rather for checking compliance, which will be covered in the next section.

3.5 Compliance Checking

Compliance checking is associated with internal and external audits. The main purpose of an audit is to control that processes of the company respect certain boundaries. Business rules can have defined restrictions, e.g., duration restriction, parallel execution restriction.

Internal auditors can benefit from using process mining for faster and more objective control of these constraints. Additionally, it enables them to visualize flaws of processes in a way easily comprehensible for the management. They use process mining to "investigate sets of transactions to determine whether adequate risk responses were formulated in the first place". [30] Auditors can monitor performance and effectiveness of internal process controls but also observe progress in optimization of the process model and provide specific recommendations. For example, to eliminate violations in an internal process,

auditors can suggest motivating resources with rewards for establishing the correct process.

Business processes are "directly influenced by the existing legislation, business policies, and external directives, which pose restrictions on the process model designed by the business analysts." [30] External auditor's responsibility is to control the alignment of processes with directives and present a completely independent opinion on the reliability of the provided financial statements. [31] Moreover, they identify risks and evaluate the technical maturity of solutions used by the firm. Companies or employees may sometimes try to hide their insufficiency, but process mining can offer unbiased insight into the internal control system of the company. It can uncover attempts to change data of the processes or discover non-compliant behavior falsely.

3.6 Conclusion

Process mining has a broad use as it can offer a big-picture view on the end-to-end process as well as a detailed observation on a particular part of it. The advantages of analyzing processes with process mining are unnegotiable. As seen in this chapter, different available perspectives support effortless detection of diverse problems that range from bottleneck analysis up to conformance checking or resource management. Many of the previously mentioned techniques are going to be demonstrated on a practical case study in Chapter 4.

Chapter 4

Case Study

The empirical part of this thesis, in the form of a case study, aims to explore different applications of process mining in the domain of finance. For this purpose, we chose a data set from BPI Challenge 2014 [32], gathering real-life event logs from Rabobank Netherlands Group ICT. This data set captures change requests on the IT department from the Service Desk of the company after a new software release or update. The main goal is to uncover the process in the data and interesting information that can be beneficial for the management. The mined model should capture the flow of events in the process and their frequencies. For processing the data, we chose the open-source Java framework ProM 6.9. [33] A variety of commercial tools is available on the market, but not all of them are platform-independent, and their offer of tools and algorithms is limited. ProM, as a tool developed by academics, offers a wide variety of algorithms and plugins for processing the data. All visualizations used in this section originate from ProM 6.9.

4.1 Data Source and Preparation of the Data

Primary information about the process is available from the description in Quick Reference Guide [32] that is part of the examined dataset. Users of the IT system in Rabobank are reporting their issues to the Service Desk that records them as *interaction*. Some of the interactions can be solved directly by the Service Desk, more complicated ones are transformed into *incidents*. Multiple interactions can be transformed into one single incident if the problem is identical. Incidents are further solved by the IT department, and we can observe activities executed in order to fix the problem in `Detail_Incident_Activity.csv` event log. The process hidden in the log should capture the handling of incidents by the IT department.

Data for the BPI challenge was in the form of a CSV file, not supported by ProM. Therefore, it was necessary first to convert the file into XLog file using the ProM converter.

Incident ID	DateStamp	IncidentActivity_Number	IncidentActivity_Type	Assignment Group	KM number	Interaction ID
IM0000004	07/01/2013 08:17	001A3689763	Reassignment	TEAM0001	KM0000553	SD0000007
IM0000004	04/11/2013 13:41	001A5852941	Reassignment	TEAM0002	KM0000553	SD0000007
IM0000004	04/11/2013 13:41	001A5852943	Update from customer	TEAM0002	KM0000553	SD0000007
IM0000004	04/11/2013 12:09	001A5849980	Operator Update	TEAM0003	KM0000553	SD0000007
IM0000004	04/11/2013 12:09	001A5849979	Assignment	TEAM0003	KM0000553	SD0000007
IM0000004	04/11/2013 13:41	001A5852942	Assignment	TEAM0002	KM0000553	SD0000007
IM0000004	04/11/2013 13:51	001A5852172	Closed	TEAM0003	KM0000553	SD0000007
IM0000004	04/11/2013 13:51	001A5852173	Caused By CI	TEAM0003	KM0000553	SD0000007
IM0000004	04/11/2013 12:09	001A5849978	Reassignment	TEAM0003	KM0000553	SD0000007
IM0000004	25/09/2013 08:27	001A5544096	Operator Update	TEAM0003	KM0000553	SD0000007
IM0000005	03/06/2013 11:15	001A4725475	Update	TEAM9999	KM0000611	SD0000011
IM0000005	03/04/2013 11:29	001A4327777	Operator Update	TEAM0003	KM0000611	SD0000011

Figure 4.1: Sample of the `Detail_Incident_Activity.csv` event log

All columns of the raw event log can be seen in Figure 4.1. Column `Incident ID` was chosen to be the case identifier, column `IncidentActivity Type` represents event in the process and column `Assignment Group` serves to determine the resource that performed an event. Column `DateStamp` was used as a timestamp of the start of the event. Events not containing all necessary data were omitted. Missing data could be completed by additional data from other systems or generated by machine learning algorithms.

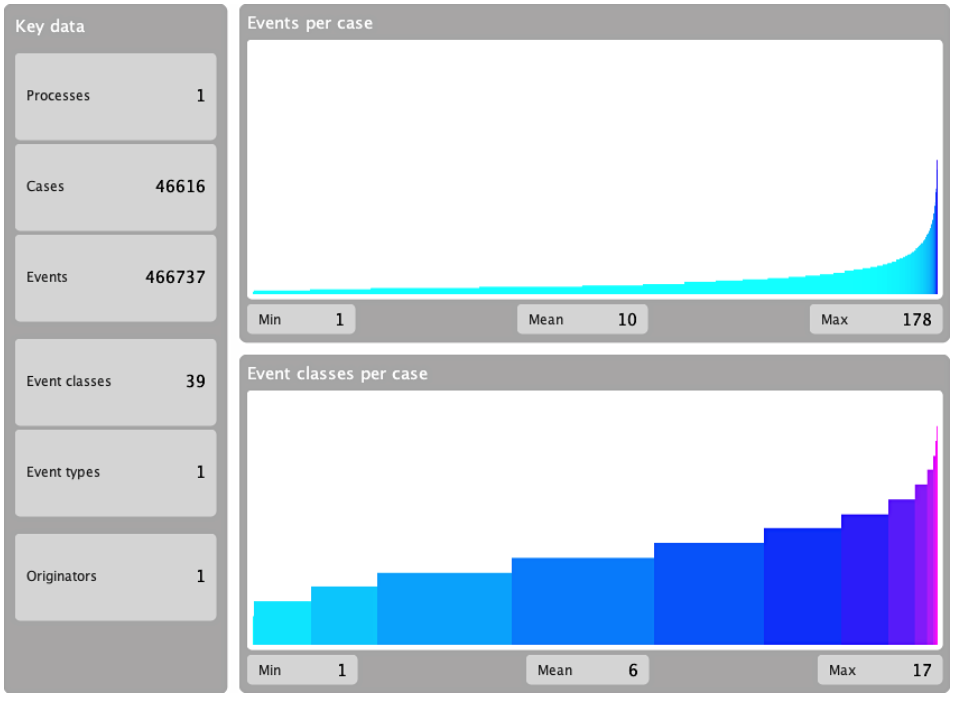


Figure 4.2: First insight about the event log from Log Visualizer

4.1.1 First Look on the Data

The converted event log can be visualized in ProM in many different ways. Solely the visualization of the log may, in many cases, provide meaningful information about the process. For the first insight about the raw data, we used the basic Log Visualizer. From Figure 4.2 can be seen that event log consists of around half a million events grouped in 46 616 cases.

Event Name

Event classes defined by Event Name
All events

Total number of classes: **39**

Class	Occurrences (absolute)	Occurrences (relative)
Assignment	88502	18.962%
Operator Update	56292	12.061%
Reassignment	51961	11.133%
Status Change	50914	10.908%
Closed	50145	10.744%
Open	46607	9.986%
Update	35969	7.706%
Caused By CI	34382	7.366%
Quality Indicator Fixed	7791	1.669%
Communication with customer	6148	1.317%

Figure 4.3: Most frequent event classes

Thirty-nine event classes were recognized in the log, with an average of six event classes per case. A relatively high number of event classes can signal the complexity of the process, as observed in the latter sections. Ten most frequent event classes, together with the percentage of their occurrences, can be seen in Figure 4.3.

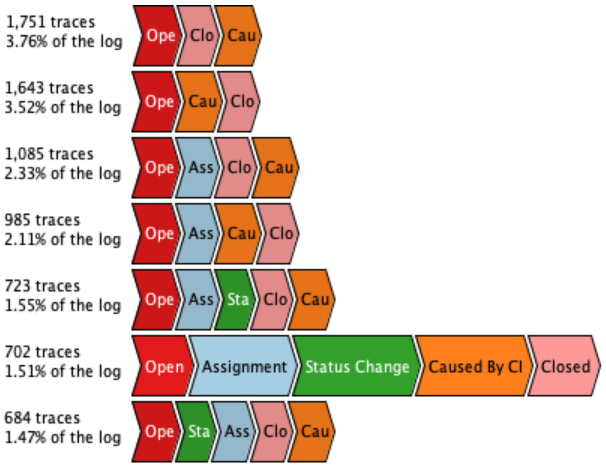


Figure 4.4: Most frequent traces

After switching to Explore Event Log visualization, specific traces of the process can be observed. Seven most frequent traces can be seen in Figure 4.4. We can observe that two end events are altering.

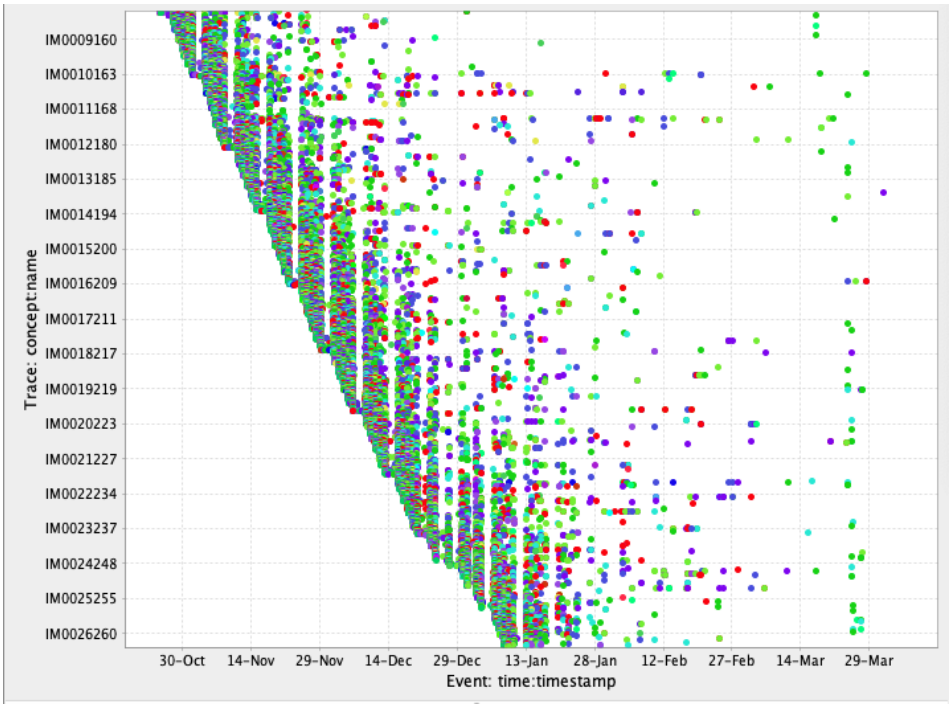


Figure 4.5: Visualization of event log using dotted chart

The last visualization we use is the Dotted Chart. The first thing that can be noticed from Figure 4.5 is the periodicity of events. It is clear that most of the events take place during workdays and gaps signal weekends or holidays during which incidents are rarely submitted. Chart is colored by different activities listed in Figure 4.6.

After regrouping the dotted chart to be sorted by length of different traces, we obtain Figure 4.7. Shortest traces are at the top and longest at the bottom. Periodicity and holidays are visible again, but what is more interesting is the coloring of traces. The top of the graph is rather green, and the bottom slowly becomes more purple and red. Red and purple activities are *reassignments* and *updates*. Their high frequency in longer running traces can signify that they tend to cause delays. In this particular case, activities reassignment and update are expected to cause delays. Nevertheless, in a different process, this type of visualization could help uncover problematic event class causing bottlenecks.

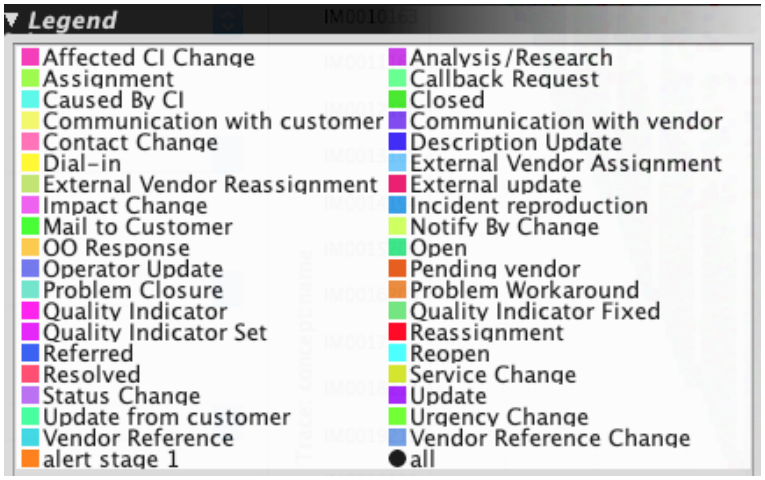


Figure 4.6: List of events found in dotted chart

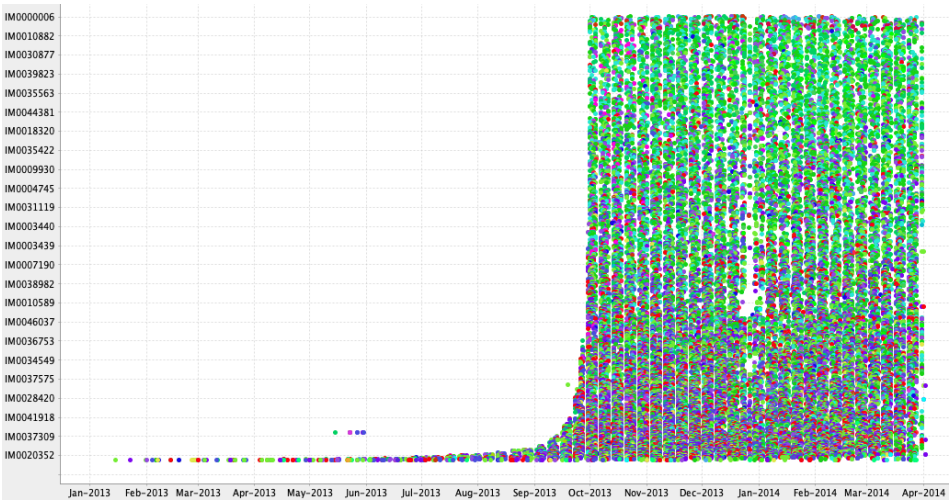


Figure 4.7: Traces of event log sorted by duration

4.1.2 Filtration of the Data

To obtain a reasonably sized process model, we decided to filter the data using plug-in `Filter Log using Simple Heuristics`. [33] Infrequent events were removed from the log, meaning noise in the log decreased. Only most frequent start and end activities were kept, precisely those used in ninety percent of traces. Only ten most frequent events from Figure 4.3 were retained. New log summary can be found in Appendix A as Figure A.1. The result contains ten event classes, yet the log shrank only to 85% of the original size. Important traces forming the core model have been preserved, and the noise has been removed. The upcoming chapter covering the creation of the model is going to

clarify the importance of log filtering and demonstrate the difference between models based on filtered or unfiltered event logs.

4.2 Creation of the Model

A multitude of process mining algorithms is available in ProM 6.9 in the form of plug-ins with a variety of different outputs such as Petri Nets, BPMN, process trees, or causal graphs. To mine the model, we chose plug-ins `Mine with Inductive Visual Miner` and `Interactive Data-aware Heuristic Miner`. [33] Both algorithms used in these plug-ins were briefly introduced in Section 2.2.2 and compared to other frequently used algorithms. Heuristic miner seems to be the ideal choice to gain a general overview of the process model and of connections between steps in the model. Inductive miner, on the other hand, produces an easily understandable clear model that is very functional for performance or conformance checking.

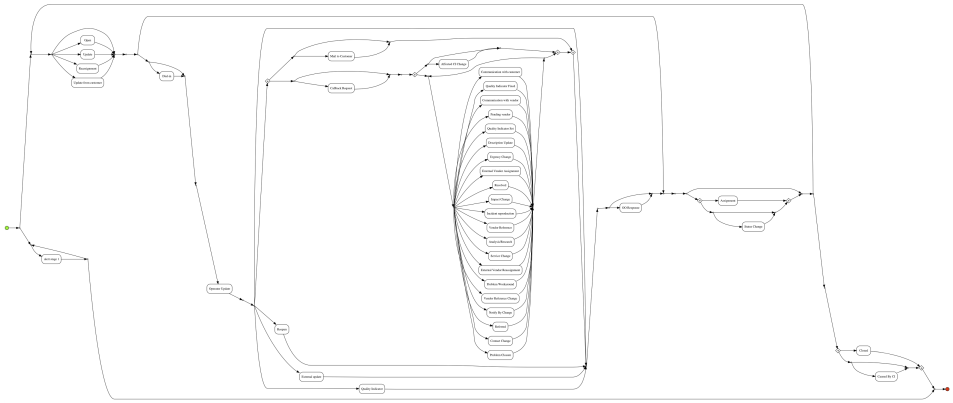


Figure 4.8: Process model from unfiltered event log created using inductive miner

Initially, we tried to obtain the model from unfiltered data using the inductive miner. Not even the names of different activities are legible in Figure 4.8 as the model is very chaotic and large. We can see that certain activities could be grouped into clusters or even omitted as they are not very usual. Figure 4.9 represents the process model mined from filtered data that contained only ten different event classes. It is already easier to imagine the process flow as it clearly consists of *Open* event, two parts where multiple alternatives can be chosen and of the last part where either *Closed* or *Caused by CI* can be executed as the close event. We use a Petri net version of this model for further analysis, and its BPMN version can be found in Appendix A as Figure A.2.

Interactive heuristic miner allowed us to see the process from a different perspective. Figure 4.10 shows the number of direct successions between events and the frequency of specific events as well. Darker the event, the more

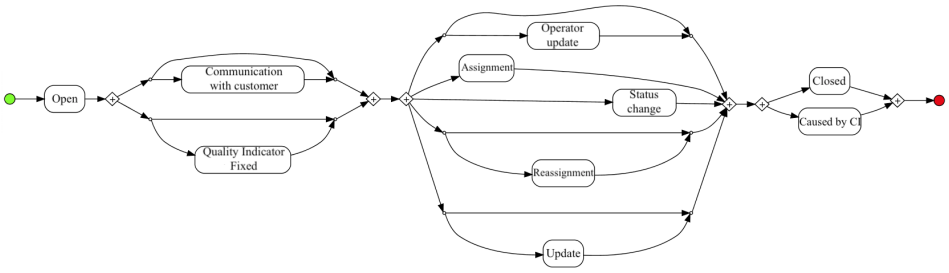


Figure 4.9: Process model mined from filtered event log created using inductive miner

frequent it is. Since *Assignment* is the most frequent activity, we can further inspect how long it takes, what resources are executing it, and whether it could be optimized. The same goes for *Operator Update* and *Reassignment*. In general, it is practical to look into events executed the most often because of their potential for the most significant savings in case there is a way to handle them better.

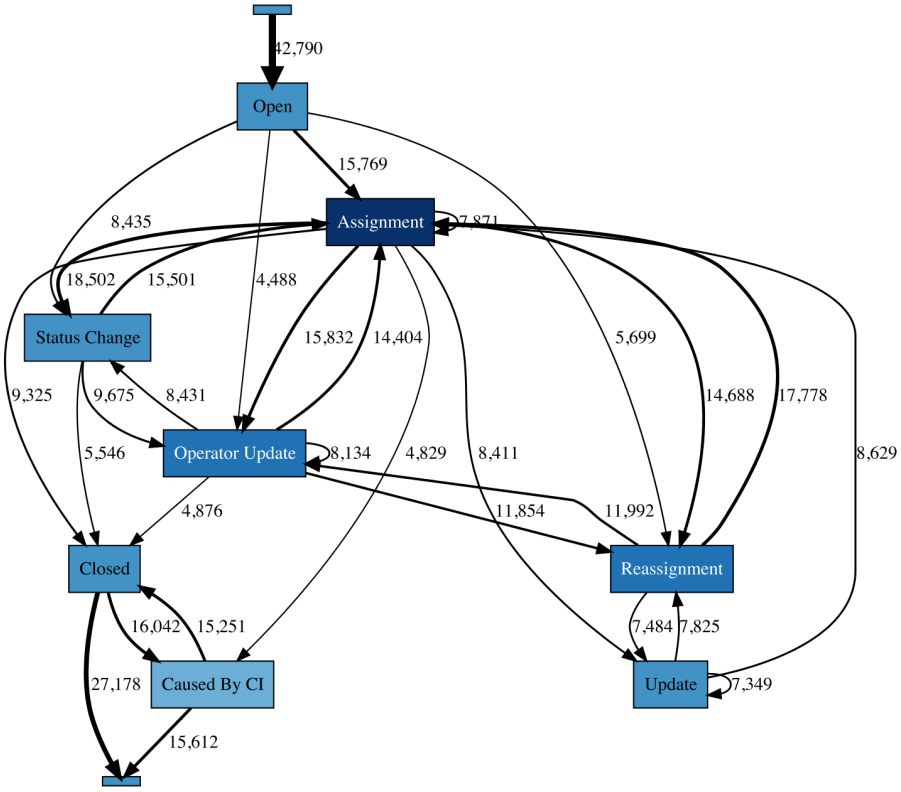


Figure 4.10: Process model mined from filtered event log created using heuristic miner

Many loops are noticeable in the causal graph, and the order of different activities varies a lot. We can deduce that the process of handling incidents does not have a clearly established structure, either regulated flow. A more detailed analysis of the mined model follows in the next section.

4.3 Analysis of the Model

ProM offers many tools to investigate different aspects of the process. The choice of tools needs to be appropriated to the type of the examined process. In case the process is well structured, it is easy to perform almost any kind of analysis over it. Otherwise, the ability to use certain tools depends also on accessible data. Data at our disposal allowed us to do regular performance and conformance checking. Moreover, knowing the resources of different events, we were able to create social networks.

4.3.1 Performance Checking

Performance analysis was carried out using ProM plug-in `Replay a Log on Flexible model for Performance Analysis`. [33] Inputs of the plug-in are aPetri net and a log. The output is the Petri net showing average time that events take represented by Figure 4.12. Red color marks events that take the longest. It is clear that *Update* is the most critical, with an average duration of 2.47 days. *Reassignment* takes more than two days as well.

Performance overview of the entire process listed in Figure 4.11 shows that the average time to solve an incident is 5.12 days. On the other hand, extreme cases with a duration of 21 seconds up to almost 13 months were logged. Extremely short case duration can signify a non-compliant behavior skipping many events in the process and should be investigated further by the management.

Case Property	Value
#Cases	42790
#Perfectly-fitting cases	21682
#Non-fitting cases	21108
#Properly started cases	42781
Case Throughput time (avg)	5.12 days
Case Throughput time (min)	21.00 seconds
Case Throughput time (max)	12.66 months
Case Throughput time (std. dev)	16.62 days
Observation period	14.73 months

Figure 4.11: Performance checking - process overview

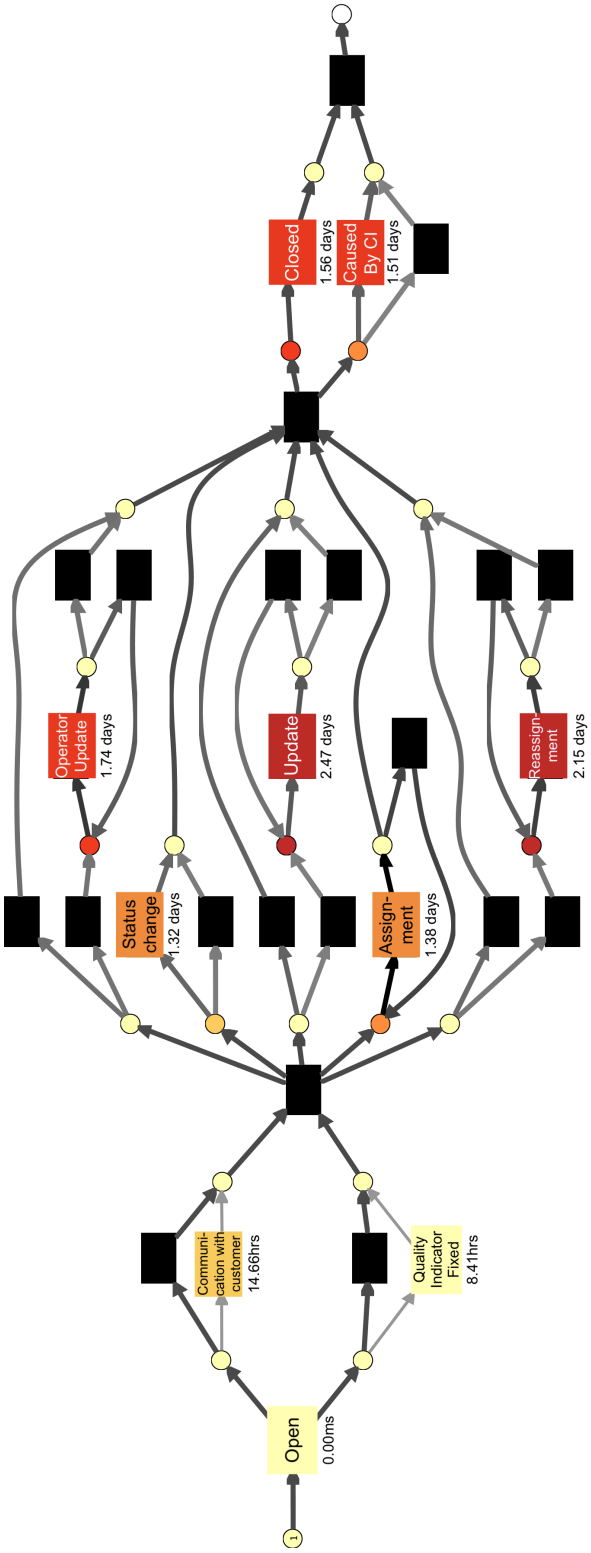


Figure 4.12: Performance checking - Petri net displaying average duration of events

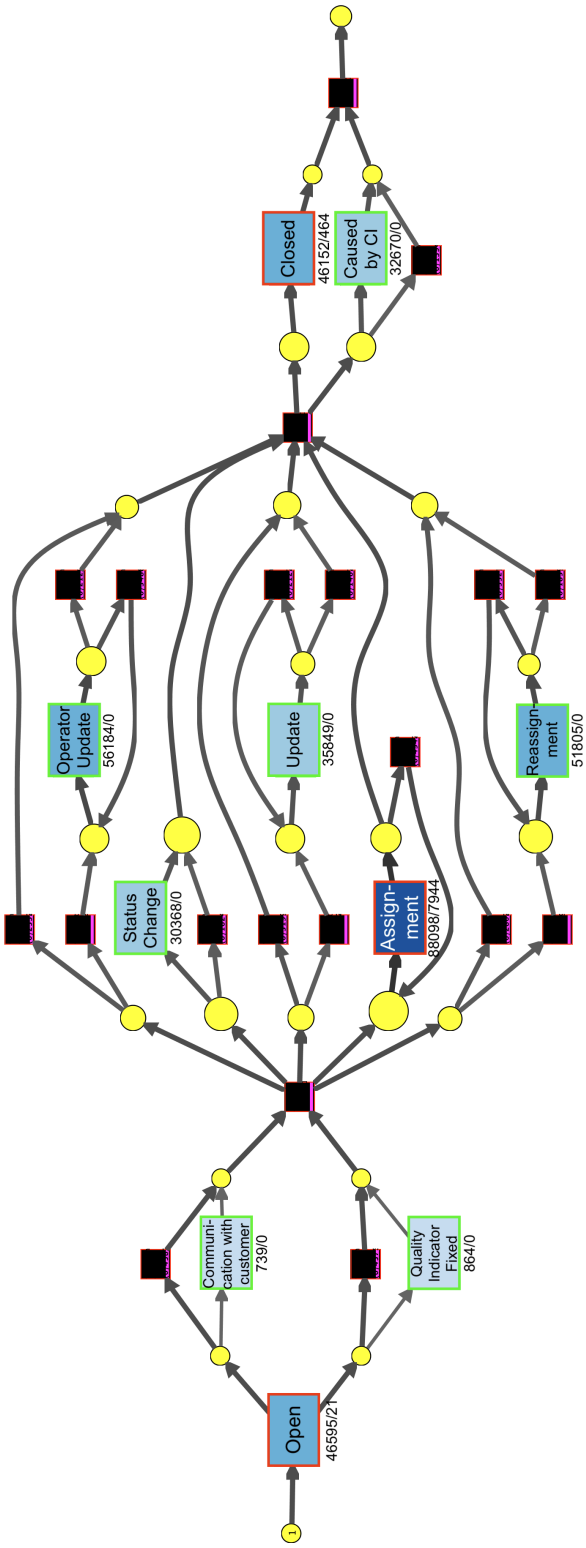


Figure 4.13: Conformance checking

4.3.2 Conformance Checking

Conformance checking measures how well the model captures real-life events. In our case, we can only evaluate the fitness of the model in terms of event data from which it was mined. We do not possess a reference model for the observed process. In case the reference model or model designed by the management is available, it is possible to replay real-life event data on this model and inspect deviations and unexpected paths that were taken in this model.

Conformance checking was applied mostly to confirm that our model represents the given data well. For doing so, we used `Replay a Log on Petri Net for Performance/Conformance Analysis` plug-in. [33] Figure 4.13 displays analysed process with frequencies of different events. The overall fitness of the model was 87.38 %. Deviations that were found are illustrated with a red border. Minor non-conformance can be noticed at *Open* and *Closed* activities but its less than 1% of events. *Assignment*, on the other hand, has around 9% of non-conforming events that can imply the need to adjust this part of the model.

4.3.3 Social Network Mining

The reviewed data-set contains information about the originators of each activity with 242 different resources. This allowed us to mine social networks and obtain useful information for resource management. Multiple ways of creating a social network are available. We used following ProM plug-ins: `Mine for a Handover-of-Work Social Network`, `Mine for a Subcontracting Social Network`, `Mine for a Working-Together Social Network`. [33]

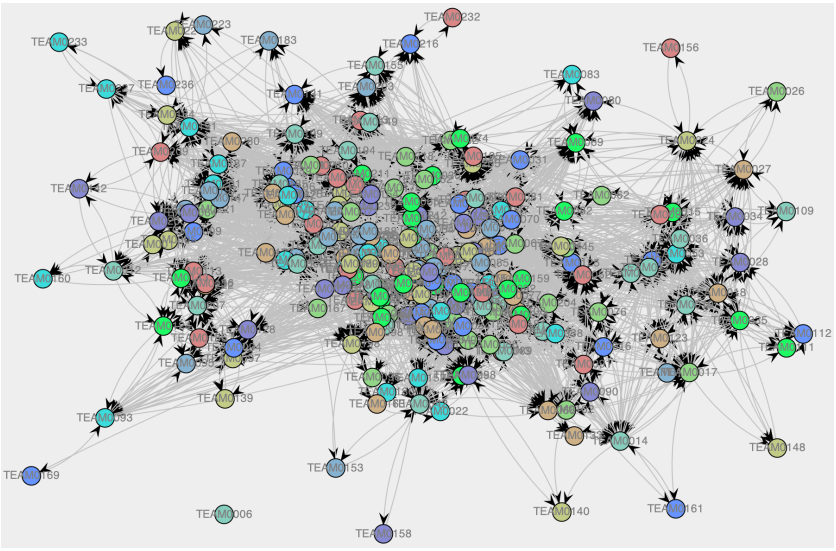


Figure 4.14: Handover of work social network

Figure 4.14 shows how resources handover work between themselves. One bigger cluster is visible in the middle, yet multiple entities are standing outside, handing work over just occasionally. It is visible that *Team0006* is not handing over any work.

ProM allows displaying social networks with many different settings. When we visualized the subcontracting social network using `Ranking view`, we obtained interesting fig. 4.15. Subcontracting means that resource A performs an activity, and then waits for resource B that executes some activity, so it (A) can perform another activity. It is clear that *Team9999* and *Team0008* (in the middle) are subcontracted to most of the surrounding resources. Very often, such engaged resources represent automatic steps in some system. Contrarily, blue *Team0066* on the right or green *Team0015* and red *Team0007* at the top are most probably important teams receiving a disproportionate amount of work and management should review their workload. Different visualization of the subcontracting network is available in Appendix A as Figure A.3, revealing some more teams with an unbalanced workload, and supporting previous claims.

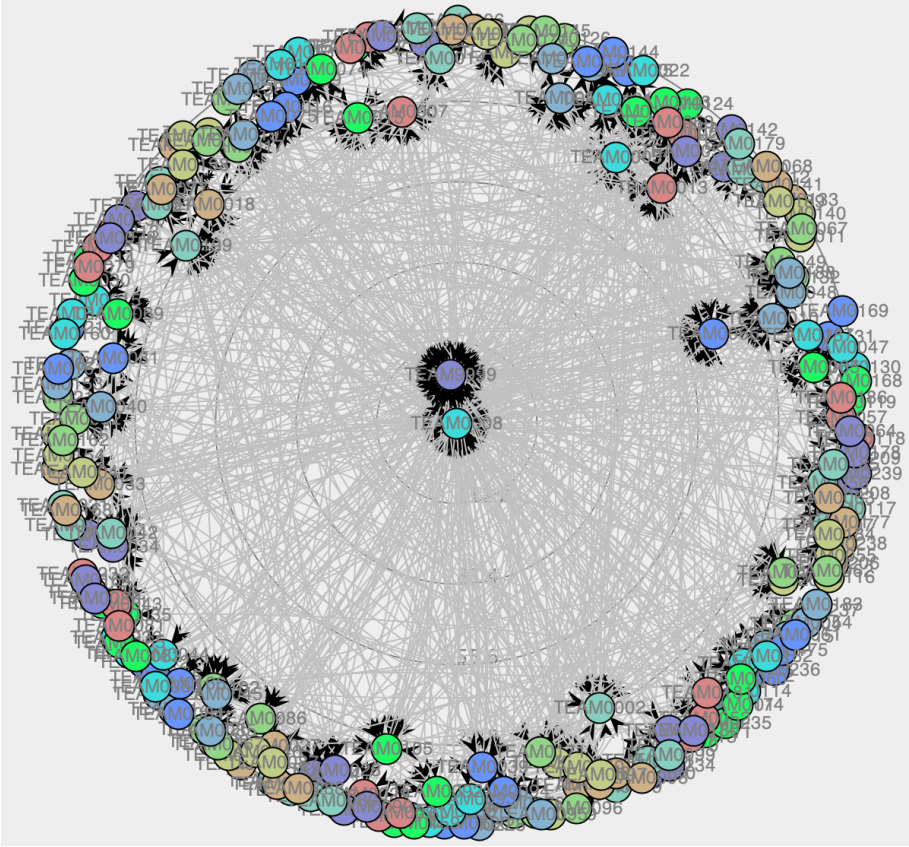


Figure 4.15: Subcontracting social network

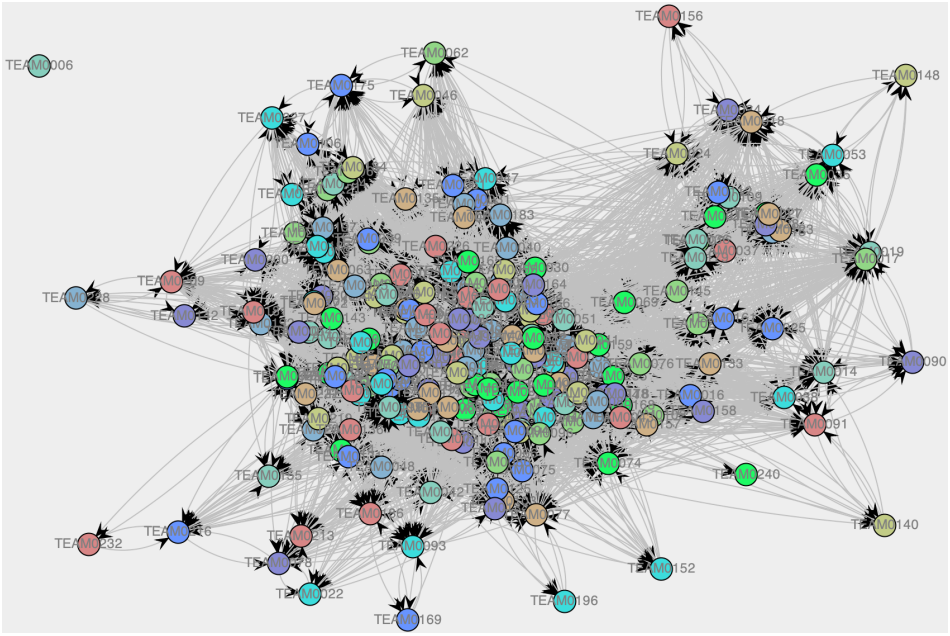


Figure 4.16: Working-together social network

Figure 4.16 shows teams working together on the same case. Similarly to Handover-of-work social network, we can see the main cluster in the middle and then multiple smaller clusters around. Some teams are not related to other teams and do not play a big role in the cooperation. *Team0006* is standing aside again. Based on the number of connections, we can sometimes identify the role or importance of the resource.

Dotted chart visualization can be used for resource analysis as well. We used resource as Y-axis to obtain Figure 4.17. The chart is colored by teams executing tasks. It is visible that at the threshold of the new year, new teams were created and started executing some work. We colored Figure 4.18 by different activities, and this allowed us to see which resources usually perform which activities. This is only a sample from all the teams, but it already unveiled that *Team0123* tends to reassign the work very often. We already suspect that reassignment is the cause of delays. Therefore this resource should be investigated. Using the dotted chart, we can also explore the habits of the resources as shown in

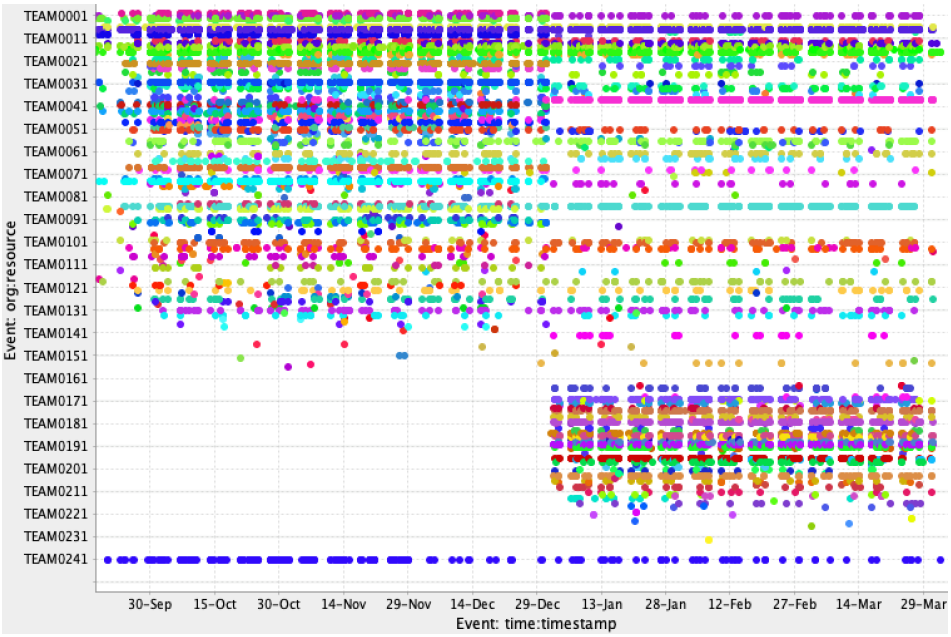


Figure 4.17: Creation of new resources showed on a dotted chart

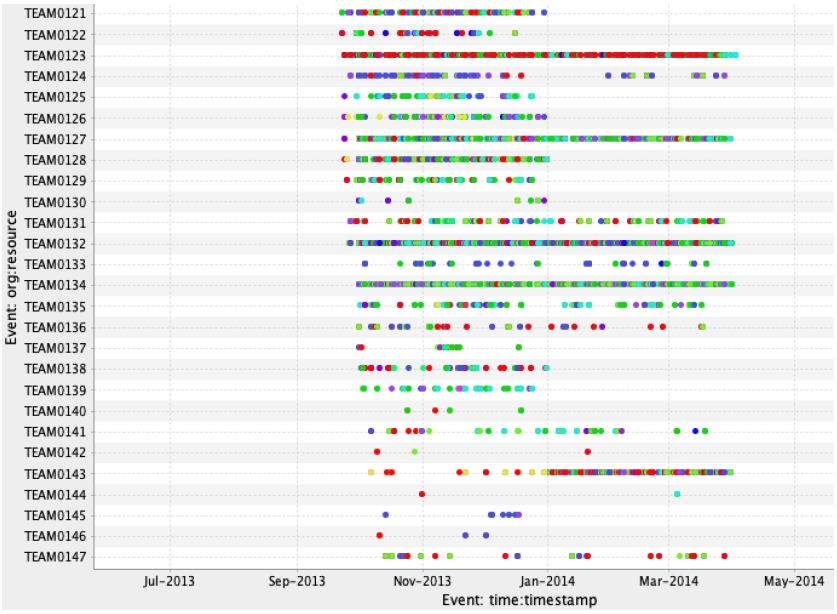


Figure 4.18: Activities executed by different resources demonstrated on a dotted chart

4.4 Outcome of the Process Mining Project

In this practical case study, we demonstrated the benefit of process mining on a real-life data-set from the financial domain. Automated extraction of the process from an event log is a non-negotiable advantage, as it offers savings on expensive consulting services. The average salary of a senior business process analyst is 35âĆň per day. [34] For analysis of complex processes, teams of analysts are working during multiple weeks or even months to obtain detailed reports. Process mining allows processing big amounts of data in a very short time and even from various perspectives. Types of analysis used in this process mining project and its outcomes are summed up in Figure 4.19.

Type of Conducted Analysis	**Results**
General overview	
	≈ 0.5 mil. of events were used for the analysis
	39 event classes were found
	10 events compose the usual length of the case
Frequency analysis	
	7 most frequent traces were determined
	10 most frequent events were identified
	15% of events with the lowest frequency were removed
Performance analysis	
	5.12 days is the average duration of a case
	2 activities with critical duration > 2 days were diagnosed
	3 activities with duration > 1.5 day were identified
Conformance analysis	
	87.38% fitness was estimated for the process model
	3 non-conforming activities were found in the model
	9% of non-conformant events happen at *Assignment*
Social network analysis	
	242 different resources were found
	3 overloaded teams were pinpointed
	creation of new teams was noted at the beginning of the new year

Figure 4.19: Results of the process mining project

Firstly, the general overview of the data was provided, allowing us to see basic patterns of the process. Problematic activities that could possibly be a source of delays were identified. Data-set was filtered in order to create a reasonably sized process model.

Secondly, two obtained process models were compared, and the need for filtration of the data was justified. The simplified process model was further used for analysis. We also created a causal graph showing dependencies of different events and their occurrence, that unveiled the most frequent activity.

Lastly, a thorough analysis of the process was performed, covering different perspectives. In performance checking, the average time to solve an incident

was obtained together with events that last the longest and perhaps could be optimized. We disclosed possible bottlenecks of the process. Conformance checking allowed us to discover irregularities in the process and diagnosed place where adjustment of the model is needed. Non-conforming events also suggest a place with a high risk of deviations. Social network mining offered insights into relations between resources and pinpointed important teams with big workloads. It also offered insight into the working habits of resources and activities they usually perform. We were able to recognize the critical team executing many delaying activities.

Overall, insights obtained by this process mining project could be very important for the management, and they have the potential to improve governance of the process. The analysis calls attention to places with risks of deviations and to problematic resources that risk management should take into account.

Conclusion

This thesis aimed to explore challenges in the domain of finance that could be solved by process mining and demonstrate its use on a practical case study. Theoretical research provided a foundation for the topic of process mining and introduced a wide range of process mining based tools and algorithms dedicated to the analysis of specific parts of processes. Algorithms for mining and analyzing processes were put in relation to GRC problems to illustrate their possible use. In the practical case study, process mining was applied to a data-set from a financial domain. Process model was extracted from the data using the ProM framework, and 87.38% fitness was achieved. The process model analysis was performed on the mined model, and its performance and compliance were examined. We estimated the average duration of a case to be 5.12 days, and we located two critical activities with an average duration longer than two days. We were able to mine social networks from the data and identify problems related to a specific resource. The benefits of process mining to tackle GRC related problems were explored, and specific examples of use were thoroughly described.

A new perspective of using process mining in finance was documented and shall serve as an inspiration to use process mining techniques in the environment of a bank or financial institution. Compared to gathering process data in an excel sheet and constructing the process manually, process mining proposes a more effective alternative. Additionally, process mining can be used to monitor compliance of running processes and report misconduct immediately. This thesis used the most known process mining algorithms, but more could be achieved by combining it with machine learning or business intelligence tools. Using a wider scope of available disciplines to analyze new perspectives of process models and relating them to reference models could represent the future challenge.

Bibliography

[1] OECD. ICT Access and Usage by Businesses. [Online], accessed on 03 March 2019. Available from: `https://stats.oecd.org/Index.aspx?DataSetCode=ICT_BUS`

[2] KPMG International. The cost of compliance. [Online], 2014, accessed on 03 March 2019. Available from: `https://home.kpmg/content/dam/kpmg/pdf/2014/07/Cost-of-Compliance.pdf`

[3] Jans, M.; Alles, M.; et al. The case for process mining in auditing: Sources of value added and areas of application. *International Journal of Accounting Information Systems*, volume 14, no. 1, 2013: pp. 1–20.

[4] Rasmussen, M. GRC 3.0, A History of GRC. [Online], 2013, accessed on 24 April 2019. Available from: `https://grc2020.com/2013/04/16/11grc-3-0-a-history-of-grc/`

[5] Racz, N.; Weippl, E.; et al. A frame of reference for research of integrated governance, risk and compliance (GRC). In *IFIP International Conference on Communications and Multimedia Security*, Springer, 2010, pp. 106–117.

[6] Tarantino, A. *Governance, risk, and compliance handbook: technology, finance, environmental, and international guidance and best practices*. John Wiley & Sons, 2008.

[7] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards: a revised framework. *Basel Committee on Banking Supervision, Bank for International Settlements*, 2004.

[8] Náplava, P. Procesní pohled, projekt, implementace IS. [presentation from lecture], 11 2018, accessible from moodle.fit.cvut.cz.

[9] Hunt, R. Why governance, risk and compliance projects fail–and how to prevent it. *Computer Fraud & Security*, volume 2014, no. 6, 2014: pp. 5–7.

[10] Murphy, R.; O'Malley, C. The Forrester Wave: Governance, Risk, And Compliance Platforms, Q1 2018. *Forrester Res*, 2018.

[11] van der Aalst, W. *Data Science in Action*. Springer Berlin Heidelberg, 2016, ISBN 978-3-662-49851-4, 25–54 pp., doi:10.1007/978-3-662-49851-4_1.

[12] Melão, N.; Pidd, M. A conceptual framework for understanding business processes and business process modelling. *Information Systems Journal*, volume 10, no. 2, 2000: pp. 105–129, doi:10.1046/j.1365-2575.2000.00075.x.

[13] More Cowbell Unlimited. National Security Education Augmentation. [Online], 2019, accessed on 4 November 2019. Available from: `https://morecowbellunlimited.com/national-security-education/`

[14] Buijs, J. Mapping data sources to xes in a generic way. *Department of Mathematics and Computer Science, Eindhoven University of Technology*, 2010.

[15] van der Aalst, W. M. P.; van Dongen, B. F. Discovering Workflow Performance Models from Timed Logs. In *Engineering and Deployment of Cooperative Information Systems*, edited by Y. Han; S. Tai; D. Wikarski, Springer Berlin Heidelberg, 2002, ISBN 978-3-540-45785-5, pp. 45–63.

[16] van der Aalst, W.; Adriansyah, A.; et al. Process mining manifesto. In *International Conference on Business Process Management*, Springer, 2011, pp. 169–194.

[17] van der Aalst, W. M. What makes a good process model? *Software & Systems Modeling*, volume 11, no. 4, 2012: pp. 557–569.

[18] Universität Bielefeld: Technischen Fakultät. Petri Nets. [Online], 2002, accessed on 31 October 2019. Available from: `https://www.techfak.uni-bielefeld.de/~mchen/BioPNML/Intro/pnfaq.html`

[19] Peterson, J. L. Petri Nets. *ACM Comput. Surv.*, volume 9, no. 3, 1977: pp. 223–252, ISSN 0360-0300, doi:10.1145/356698.356702.

[20] Rozinat, A.; de Medeiros, A. A.; et al. Towards an evaluation framework for process mining algorithms. *BPM Center Report BPM-07-06, BPMcenter. org*, volume 123, 2007: p. 142.

[21] Mendling, J.; Neumann, G.; et al. Understanding the occurrence of errors in process models based on metrics. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2007, pp. 113–130.

[22] Agrawal, R.; Gunopulos, D.; et al. Mining process models from workflow logs. In *International Conference on Extending Database Technology*, Springer, 1998, pp. 467–483.

[23] Cook, J. E.; Wolf, A. L. Discovering models of software processes from event-based data. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, volume 7, no. 3, 1998: pp. 215–249.

[24] Press-center, R. Efficiency of Process Mining implementation in a bank. [Online], 2018, accessed on 07 April 2019. Available from: `http://www.ramax.ru/en/o-kompanii/press-tsentr/media-publications/229/`

[25] QPR. Piraeus bank. [Online], 2018, accessed on 07 April 2019. Available from: `https://www.qpr.com/customers/piraeus-bank`

[26] Celonis. Process Mining Story VTB: Banking in the 21st Century. [Online], 2018, accessed on 23 April 2019. Available from: `https://www.youtube.com/watch?v=QfN-iTaeDG0`

[27] QPR Software. Process Mining Interview Piraeus Bank: Locate Bottlenecks in 5 Minutes. [Online], 2018, accessed on 23 April 2019. Available from: `https://www.youtube.com/watch?v=Nxz-pTv3EPw`

[28] Ly, L. T.; Rinderle, S.; et al. Mining staff assignment rules from event-based data. In *International Conference on Business Process Management*, Springer, 2005, pp. 177–190.

[29] Rozinat, A.; van der Aalst, W. M. Conformance checking of processes based on monitoring real behavior. *Information Systems*, volume 33, no. 1, 2008: pp. 64–95.

[30] Caron, F.; Vanthienen, J.; et al. A comprehensive framework for the application of process mining in risk management and compliance checking. *KU Leuven Faculty of Business and Economics KBI*, volume 1226, 2012.

[31] Caron, F.; Vanthienen, J.; et al. A comprehensive investigation of the applicability of process mining techniques for enterprise risk management. *Computers in Industry*, volume 64, no. 4, 2013: pp. 464–475.

[32] van Dongen, B.F. (Boudewijn). BPI Challenge 2014. [Online], 2014, accessed on 3 December 2019. Available from: `https://data.4tu.nl/repository/uuid:c3e5d162-0cfd-4bb0-bd82-af5268819c35`

[33] van der Aalst, W. e. a. ProM 6.9 - Process mining workbench. [Online], accessed on 03 December 2019. Available from: http://www.promtools.org/doku.php?id=prom69

[34] Payscale Inc. Average Late-Career Business Process Analyst Salary. [Online], 2019, accessed on 10 December 2019. Available from: https://www.payscale.com/research/US/Job=Business_Process_Analyst/Salary/7ff75497/Late-Career

Appendix A

Graphs and Visualizations

Figure A.1: Summary of filtered event log

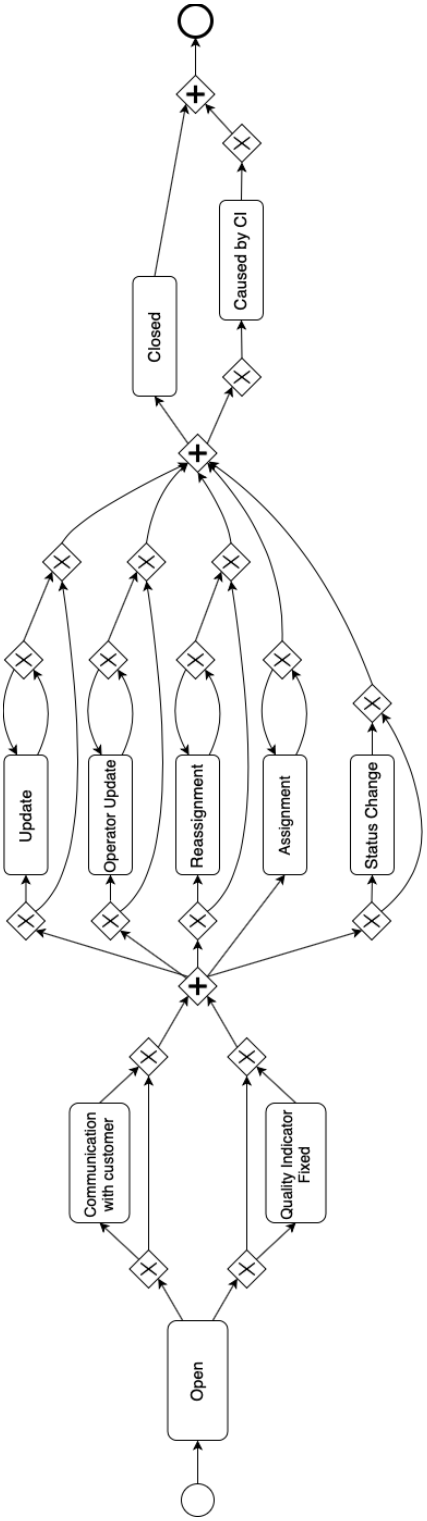


Figure A.2: Process model mined from filtered event log, created using the inductive miner in BPMN

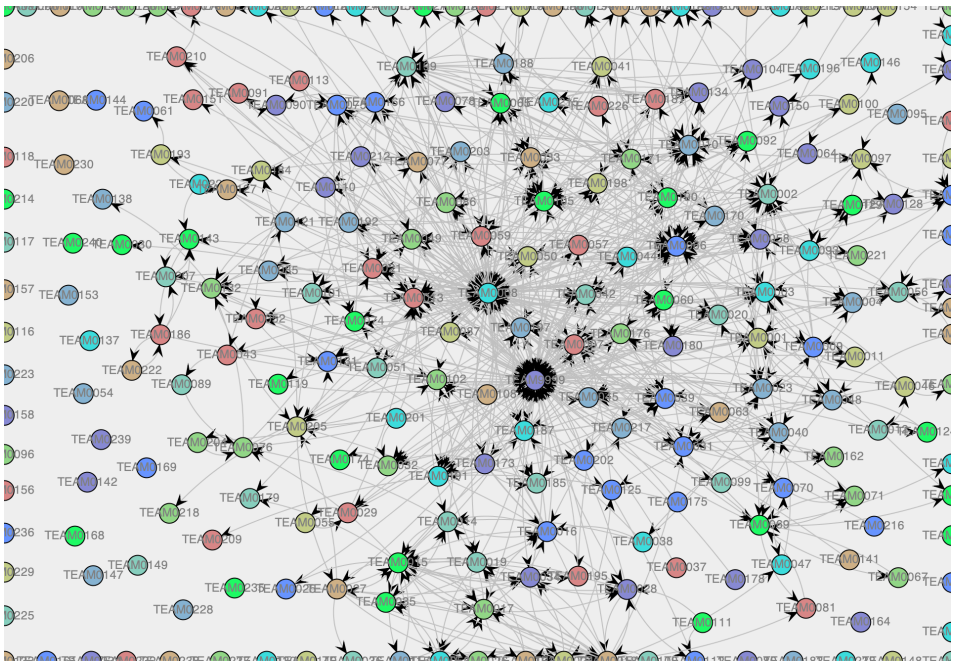


Figure A.3: Subcontracting social network II

Figure A.4: Dotted chart demonstrating working habits of the resources

Appendix B

Contents of Enclosed SD Card

- `src` ..... the directory of source codes
  - `thesis` ..... the directory of LaTeX source codes of the thesis
  - `images` ..... the directory of images from case study
  - `data` ..... the directory with dataset used in this thesis
- `text` ..... the thesis text directory
  - `BP_Krbilova_Katarina_2019.pdf` ..... the thesis text in PDF format