

Wil van der Aalst

Process Mining

Data Science in Action

Second Edition



Springer

Process Mining

Wil van der Aalst

Process Mining

Data Science in Action

Second Edition



Springer

Wil van der Aalst
Department of Mathematics and Computer Science
Eindhoven University of Technology
Eindhoven, The Netherlands
url: <http://www.vdaalst.com>

ISBN 978-3-662-49850-7
DOI 10.1007/978-3-662-49851-4

ISBN 978-3-662-49851-4 (eBook)

Library of Congress Control Number: 2016938641

Springer Heidelberg New York Dordrecht London
© Springer-Verlag Berlin Heidelberg 2011, 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*Thanks to Karin for understanding that
science is more rewarding than running
errands*

*Thanks to all people that contributed to
ProM; the fruits of their efforts demonstrate
that sharing a common goal is more
meaningful than “cashing in the next
publon”¹*

*In remembrance of Gerry Straatman-Beelen
(1932–2010)*

¹Publon = smallest publishable unit.

Preface

The interest in *data science* is rapidly growing. Many consider data science as *the* profession of the future. Just like computer science emerged as a discipline in the 1970s, we now witness the rapid creation of research centers and bachelor/master programs in data science. The hype related to Big Data and predictive analytics illustrates this. Data (“Big” or “small”) are essential for people and organizations and their importance will only increase. However, it is not sufficient to focus on data storage and data analysis. A data scientist also needs to relate data to operational processes and be able to ask the right questions. This requires an understanding of end-to-end processes. *Process mining* bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining. Process mining provides a new means to improve processes in a variety of application domains. The omnipresence of event data combined with process mining allows organizations to diagnose problems based on facts rather than fiction.

Although traditional Business Process Management (BPM) and Business Intelligence (BI) technologies received lots of attention, they did not live up to the expectations raised by academics, consultants, and software vendors. Probably, the same will happen to most of the Big Data technologies vigorously promoted today. The goal should be to improve the operational processes themselves rather than the artifacts (models, data, and systems) they use. As will be demonstrated in this book, there are novel ways to put “data science in action” and improve processes based on the data they generate.

Process mining is an emerging discipline providing comprehensive sets of tools to provide fact-based insights and to support process improvements. This new discipline builds on process model-driven approaches and data mining. However, process mining is much more than an amalgamation of existing approaches. For example, existing data mining techniques are too data-centric to provide a comprehensive understanding of the end-to-end processes in an organization. BI tools focus on simple dashboards and reporting rather than clear-cut business process insights. BPM suites heavily rely on experts modeling idealized to-be processes and do not help the stakeholders to understand the as-is processes.

This book presents a range of process mining techniques that help organizations to uncover their actual business processes. Process mining is not limited to process discovery. By tightly coupling event data and process models, it is possible to check conformance, detect deviations, predict delays, support decision making, and recommend process redesigns. Process mining breathes life into otherwise static process models and puts today’s massive data volumes in a process context. Hence, managements trends related to process improvement (e.g., Six Sigma, TQM, CPI, and CPM) and compliance (SOX, Basel II, etc.) can benefit from process mining.

Process mining, as described in this book, emerged in the last decade [156, 160]. However, the roots date back about half a century. For example, Anil Nerode presented an approach to synthesize finite-state machines from example traces in 1958 [108], Carl Adam Petri introduced the first modeling language adequately capturing concurrency in 1962 [111], and Mark Gold was the first to systematically explore different notions of learnability in 1967 [61]. When data mining started to flourish in the 1990s, little attention was given to processes. Moreover, only recently event logs have become omnipresent thus enabling end-to-end process discovery. Since the first survey on process mining in 2003 [156], progress has been spectacular. Process mining techniques have become mature and supported by various tools. Moreover, whereas initially the primary focus was on process discovery, the process mining spectrum has broadened markedly. For instance, conformance checking, multi-perspective process mining, and operational support have become integral parts of ProM, one of the leading process mining tools.

The book provides a comprehensive overview of the state-of-the-art in process mining. It is intended as an introduction to the topic for practitioners, students, and academics. On the one hand, the book is accessible for people that are new to the topic. On the other hand, the book does not avoid explaining important concepts on a rigorous manner. The book aims to be self-contained while covering the entire process mining spectrum from process discovery to operational support. Therefore, it also serves as a reference handbook for people dealing with BPM or BI on a day-to-day basis.

The first edition of this book appeared in 2011 under the title “Process Mining: Discovery, Conformance and Enhancement of Business Processes” [140]. Given the rapid developments in process mining, there was a clear need for an updated version. The original book has been extended in several ways. First of all, process mining has been put into the broader context of *data science* (see the new Chap. 1). This explains the new subtitle “Data Science in Action”. There is an urgent need for data scientists able to help organizations improve their operational processes. Therefore, the new edition of the book positions process mining in this broader context and relates it to statistics, data mining, Big Data, etc. Second, there has been significant progress in process discovery in recent years. This is exemplified by the family of *inductive mining* techniques that can handle large incomplete event logs with infrequent behavior, but still provide formal guarantees. The basic elements of inductive mining (Sect. 7.5) and the notion of process trees (Sect. 3.2.8) have been added to this book. Third, the notion of *alignments* has become a key concept to relate observed behavior and modeled behavior. The chapter on conformance checking has been extended to carefully introduce alignments (Sect. 8.3). Moreover, next

to fitness, also quality dimensions like *precision* are now defined. Fourth, a chapter on “process mining in the large” (Chap. 12) has been added to illustrate that process mining can exploit modern infrastructures and that process discovery and conformance checking can be decomposed and distributed. Since the first edition of the book, many new process mining products emerged (often inspired by the open source platform ProM and the previous edition of this book). The chapter on tools (Chap. 11) has been completely rewritten and discusses commercial tools like Celonis Process Mining, Disco, Enterprise Discovery Suite, Interstage Business Process Manager Analytics, Minit, myInvenio, Perceptive Process Mining, QPR ProcessAnalyzer, Rialto Process, SNP Business Process Analysis, and webMethods Process Performance Manager (next to open-source initiatives like ProM and RapidProM). Finally, pointers to recent literature have been added and a new section of data quality has been added (Sect. 5.4). These changes justify a revised edition of the book.

The reader can immediately put process mining into practice due to the applicability of the techniques, the availability of (open-source) process mining software, and the abundance of event data in today’s information systems. I sincerely hope that you enjoy reading this book and start using some of the amazing process mining techniques available today.

Eindhoven, The Netherlands
January 2016

Wil van der Aalst

Acknowledgements

Many individuals and organizations contributed to the techniques and tools described in this book. Therefore, it is a pleasure to acknowledge their support, efforts, and contributions.

All of this started in 1999 with a research project named “Process Design by Discovery: Harvesting Workflow Knowledge from Ad-hoc Executions” initiated by Ton Weijters and myself. At that time, I was still working as a visiting professor at the University of Colorado in Boulder. However, the research school BETA had encouraged me to start collaborating with existing staff in my new research group at TU/e (Eindhoven University of Technology). After talking to Ton it was clear that we could benefit from combining his knowledge of machine learning with my knowledge of workflow management and Petri nets. Process mining (at that time we called it workflow mining) was the obvious topic for which we could combine our expertise. This was the start of a very successful collaboration. Thanks Ton!

Since the turn of the century, many PhD students have been working on the topic: Arya Adriansyah, Ana Karla Alves de Medeiros, Alfredo Bolt Iriondo, R.P. Jagadeesh Chandra (JC) Bose, Carmen Bratosin, Joos Buijs, Alok Dixit, Boudewijn van Dongen, Maikel van Eck, Robert Engel, Eduardo González Lopéz de Murillas, Christian Günther, Bart Hompes, Anna Kalenkova, Marie Koornneef, Maikel Leemans, Sander Leemans, Guangming Li, Cong Liu, Xixi Lu, Felix Mannhardt, Ronny Mans, Laura Maruster, Alexey Mitsyuk, Richard Müller, Jorge Munoz-Gama, Joyce Nakatumba, Maja Pesic, Elham Ramezani, Anne Rozinat, Alifah Syamsiyah, Helen Schonenberg, Dennis Schunselaar, Minseok Song, Niek Tax, and Bas van Zelst. I am extremely grateful for their efforts.

Ana Karla Alves de Medeiros was the first PhD student to work on the topic under my supervision (genetic process mining). She did a wonderful job; her thesis on genetic process mining was awarded with the prestigious ASML 2007 Promotion Prize and was selected as the best thesis by the KNAW research school BETA. Also Boudewijn van Dongen has been involved in the development of ProM right from the start. As a Master student he already developed the process mining tool EMiT, i.e., the predecessor of ProM. He turned out to be a brilliant PhD student and developed a variety of process mining techniques. Eric Verbeek did a PhD on work-

flow verification, but over time he got more and more involved in process mining research and the development of ProM. Many people underestimate the importance of a scientific programmer like Eric. Tool development and continuity are essential for scientific progress! Boudewijn and Eric have been the driving force behind ProM and their contributions have been crucial for process mining research at TU/e. Moreover, they are always willing to help others. Thanks guys!

Christian Günther and Anne Rozinat joined the team in 2005. Their contributions have been of crucial importance for extending the scope of process mining and lifting the ambition level. Christian managed to make ProM look beautiful while significantly improving its performance. Moreover, his Fuzzy miner facilitated dealing with Spaghetti processes. Anne managed to widen the process mining spectrum by adding conformance checking and multi-perspective process mining to ProM. It is great that they succeeded in founding a process mining company (Fluxicon). Anne and Christian are great process mining ambassadors and build software that people can and also want to use. Another person crucial for the development of ProM is Peter van den Brand. He set up the initial framework and played an important role in the development of the architecture of ProM 6. Based on his experience with ProM, he set up a process mining company (Futura Process Intelligence) that joined forces with Pallas Athena which, in turn, was taken over by Lexmark's Perceptive Software. It is great to work with people like Peter, Christian, and Anne; they are essential for turning research results into commercial products (although I am still waiting for the sports cars they promised ...).

Next to Boudewijn, Eric, and the PhDs mentioned, the current “process mining team” at TU/e consists of Joos Buijs, Dirk Fahland, Massimiliano de Leoni, Hajo Reijers, Natalia Sidorova, Patrick Mukala, Nour Assy, Farideh Heidari, and—of course—Ine van der Ligt, our secretary.

Academics from various universities contributed to ProM and supported our process mining research. We are grateful to the Technical University of Lisbon, Katholieke Universiteit Leuven, Universitat Politècnica de Catalunya, Universität Paderborn, University of Rostock, Humboldt-Universität zu Berlin, University of Calabria, Queensland University of Technology, Tsinghua University, Universität Innsbruck, Ulsan National Institute of Science and Technology, Università di Bologna, Zhejiang University, Vienna University of Technology, Universität Ulm, Open University, Jilin University, National Research University Higher School of Economics, Free University of Bozen-Bolzano, University of Tartu, Pontificia Universidad Católica de Chile, University of Vienna, Pontifícia Universidade Católica do Paraná, Technion, VU University Amsterdam, Hasso-Plattner-Institut, University of Freiburg, Vienna University of Economics and Business, University of Haifa, University of Naples Federico II, University of Padua, and University of Nancy for their help. I would also like to thank the members of the IEEE Task Force on Process Mining for promoting the topic. We are grateful to all other organizations that supported process mining research at TU/e: NWO, STW, EU, IOP, LOIS, BETA, SIKS, Stichting EIT Informatica Onderwijs, Pallas Athena, IBM, LaQuSo, Philips Healthcare, Philips Research, Vanderlande, BrandLoyalty, ESI, Jacquard, Nuffic, BPM Usergroup, and WWTF. Special thanks go to Pallas Athena and Fluxicon

for promoting the topic of process mining and their collaboration in a variety of projects. Over 150 organizations provided event logs that helped us to improve our process mining techniques. Here, I would like to explicitly mention the AMC hospital, Philips Healthcare, ASML, Ricoh, Vestia, Catharina hospital, Thales, Océ, Rijkswaterstaat, Heusden, Harderwijk, Deloitte, and all organizations involved in the SUPER, ACSI, PoSecCo, and CoSeLoG projects. We are grateful for allowing us to use their data and for providing feedback.

Since 2013 I am also serving as the scientific director of the *Data Science Center Eindhoven* (DSC/e). This is a great initiative. Research and education in data science are of growing importance, and there is a natural fit with process mining. I am grateful for the support from people like Emile Aarts, Maurice Groten, Joos Buijs, Jack van Wijk, and many others. They helped to create and develop DSC/e.

It is impossible to name all of the individuals that contributed to ProM or helped to advance process mining. Peter van den Brand, Boudewijn van Dongen, Dirk Fahland, Christian Günther, Sander Leemans, Xixi Lu, Massimiliano de Leoni, Felix Mannhardt, Eric Verbeek, Michael Westergaard, and many others contributed to the current version of ProM. Moreover, besides the people mentioned earlier, I would like to thank Han van der Aa, Rafael Accorsi, Michael Adams, Piet Bakker, Huub de Beer, Tobias Blickle, Seppe vanden Broucke, Andrea Burattin, Riet van Buul, Toon Calders, Diego Calvanese, Jorge Cardoso, Josep Carmona, Alina Chipaila, Jan Claes, Raffaele Conforti, Francisco Curbela, Ernesto Damiani, Marcus Dees, Benoît Depaire, Jörg Desel, John Domingue, Marlon Dumas, Schahram Dustdar, Skip Ellis, Paul Eertink, Dyon Egberts, Dirk Fahland, Diogo Ferreira, Colin Fidge, Walid Gaaloul, Frank van Geffen, Stijn Goedertier, Adela Grando, Gianluigi Greco, Vladimir Gromov, Dolf Grünbauer, Shengnan Guo, Antonella Guzzo, Kees van Hee, Joachim Herbst, Sergio Hernández, Rastislav Hlavac, Arthur ter Hofstede, John Hoogland, Mieke Jans, Theo Janssen, Urszula Jessen, Georgi Jojgov, Ivo de Jong, Ivan Khodyrev, Albert Kisjes, Martin Klenk, Pieter de Kok, Angelique Koopman, Rudolf Kuhn, Thom Langerwerf, Teemu Lehto, Giorgio Leonardi, Jiafei Li, Zheng Liu, Niels Lohmann, Irina Lomazova, Wei Zhe Low, Peter Hornix, Fabrizio Maggi, Paola Mello, Jan Mendling, Frits Minderhoud, Arnold Moleman, Marco Montali, Michael zur Muehlen, Mariska Netjes, Andriy Nikolov, Rudi Niks, Bastian Nominacher, Chun Ouyang, Zbigniew Paszkiewicz, Mykola Pechenizkiy, Carlos Pedrinaci, Anastasiia Pika, Viara Popova, Silvana Quaglini, Manfred Reichert, Remmert Remmerts de Vries, Joel Ribeiro, Jaap Rigter, Stefanie Rinderle-Ma, Alexander Rinke, Andreas Rogge-Solti, Marcello La Rosa, Michael Rosemann, Marcella Rovani, Vladimir Rubin, Nick Russell, Stefania Rusu, Eduardo Portela Santos, Marcos Sepúlveda, Shiva Shabaninejad, Pnina Soffer, Alessandro Sperduti, Christian Stahl, Suriadi Suriadi, Keith Swenson, Nikola Trcka, Kenny van Uden, Irene Vanderfeesten, Jan Vanthienen, Rob Vanwersch, George Varvaressos, Marc Verdonk, Sicco Verwer, Jan Vogelaar, Hans Vrins, Jianmin Wang, Teun Wagelmakers, Barbara Weber, Jochen De Weerdt, Lijie Wen, Jan Martijn van der Werf, Mathias Weske, Jack van Wijk, Moe Wynn, Bart Ydo, Marco Zapletal, Reng Zeng, and Indre Zliobaite for their support.

Thanks to Springer-Verlag for publishing this book. Ralf Gerstner encouraged me to write this book and handled things in a truly excellent manner. He also repeatedly triggered me to make a new edition of this book. Thanks Ralf!

More than 95% of the original book was written in beautiful Schleiden. Despite my sabbatical, there were many other tasks competing for attention. Thanks to my weekly visits to Schleiden (without Internet access!), it was possible to write the first edition of this book in a three month period. The excellent coffee of Serafin helped when proofreading the individual chapters, the scenery did the rest.

As always, acknowledgements end with thanking the people most precious. Lion's share of credits should go to Karin, Anne, Willem, Sjaak, and Loes. They often had to manage without me under difficult circumstances. Without their continuing support, this book would have taken ages.

Eindhoven, The Netherlands
January 2016

Wil van der Aalst

Contents

Part I Introduction

1	Data Science in Action	3
1.1	Internet of Events	3
1.2	Data Scientist	10
1.3	Bridging the Gap Between Process Science and Data Science	15
1.4	Outlook	20
2	Process Mining: The Missing Link	25
2.1	Limitations of Modeling	25
2.2	Process Mining	30
2.3	Analyzing an Example Log	35
2.4	Play-In, Play-Out, and Replay	41
2.5	Positioning Process Mining	44
2.5.1	How Process Mining Compares to BPM	44
2.5.2	How Process Mining Compares to Data Mining	46
2.5.3	How Process Mining Compares to Lean Six Sigma	46
2.5.4	How Process Mining Compares to BPR	49
2.5.5	How Process Mining Compares to Business Intelligence	49
2.5.6	How Process Mining Compares to CEP	50
2.5.7	How Process Mining Compares to GRC	50
2.5.8	How Process Mining Compares to ABPD, BPI, WM,	51
2.5.9	How Process Mining Compares to Big Data	52
3	Process Modeling and Analysis	55

Part II Preliminaries

3	Process Modeling and Analysis	55
3.1	The Art of Modeling	55
3.2	Process Models	57
3.2.1	Transition Systems	58
3.2.2	Petri Nets	59
3.2.3	Workflow Nets	65
3.2.4	YAWL	66

3.2.5	Business Process Modeling Notation (BPMN)	68
3.2.6	Event-Driven Process Chains (EPCs)	70
3.2.7	Causal Nets	72
3.2.8	Process Trees	78
3.3	Model-Based Process Analysis	83
3.3.1	Verification	83
3.3.2	Performance Analysis	85
3.3.3	Limitations of Model-Based Analysis	88
4	Data Mining	89
4.1	Classification of Data Mining Techniques	89
4.1.1	Data Sets: Instances and Variables	90
4.1.2	Supervised Learning: Classification and Regression	92
4.1.3	Unsupervised Learning: Clustering and Pattern Discovery	94
4.2	Decision Tree Learning	94
4.3	k -Means Clustering	100
4.4	Association Rule Learning	104
4.5	Sequence and Episode Mining	107
4.5.1	Sequence Mining	107
4.5.2	Episode Mining	109
4.5.3	Other Approaches	111
4.6	Quality of Resulting Models	112
4.6.1	Measuring the Performance of a Classifier	113
4.6.2	Cross-Validation	115
4.6.3	Occam's Razor	118
Part III From Event Logs to Process Models		
5	Getting the Data	125
5.1	Data Sources	125
5.2	Event Logs	128
5.3	XES	138
5.4	Data Quality	144
5.4.1	Conceptualizing Event Logs	145
5.4.2	Classification of Data Quality Issues	148
5.4.3	Guidelines for Logging	151
5.5	Flattening Reality into Event Logs	153
6	Process Discovery: An Introduction	163
6.1	Problem Statement	163
6.2	A Simple Algorithm for Process Discovery	167
6.2.1	Basic Idea	167
6.2.2	Algorithm	171
6.2.3	Limitations of the α -Algorithm	174
6.2.4	Taking the Transactional Life-Cycle into Account	177
6.3	Rediscovering Process Models	178
6.4	Challenges	182

6.4.1	Representational Bias	183
6.4.2	Noise and Incompleteness	185
6.4.3	Four Competing Quality Criteria	188
6.4.4	Taking the Right 2-D Slice of a 3-D Reality	192
7	Advanced Process Discovery Techniques	195
7.1	Overview	195
7.1.1	Characteristic 1: Representational Bias	197
7.1.2	Characteristic 2: Ability to Deal With Noise	198
7.1.3	Characteristic 3: Completeness Notion Assumed	199
7.1.4	Characteristic 4: Approach Used	199
7.2	Heuristic Mining	201
7.2.1	Causal Nets Revisited	201
7.2.2	Learning the Dependency Graph	202
7.2.3	Learning Splits and Joins	205
7.3	Genetic Process Mining	207
7.4	Region-Based Mining	212
7.4.1	Learning Transition Systems	212
7.4.2	Process Discovery Using State-Based Regions	216
7.4.3	Process Discovery Using Language-Based Regions	218
7.5	Inductive Mining	222
7.5.1	Inductive Miner Based on Event Log Splitting	222
7.5.2	Characteristics of the Inductive Miner	229
7.5.3	Extensions and Scalability	233
7.6	Historical Perspective	236
Part IV Beyond Process Discovery		
8	Conformance Checking	243
8.1	Business Alignment and Auditing	243
8.2	Token Replay	246
8.3	Alignments	256
8.4	Comparing Footprints	263
8.5	Other Applications of Conformance Checking	268
8.5.1	Repairing Models	268
8.5.2	Evaluating Process Discovery Algorithms	269
8.5.3	Connecting Event Log and Process Model	272
9	Mining Additional Perspectives	275
9.1	Perspectives	275
9.2	Attributes: A Helicopter View	277
9.3	Organizational Mining	281
9.3.1	Social Network Analysis	282
9.3.2	Discovering Organizational Structures	287
9.3.3	Analyzing Resource Behavior	288
9.4	Time and Probabilities	290
9.5	Decision Mining	294
9.6	Bringing It All Together	297

10 Operational Support	301
10.1 Refined Process Mining Framework	301
10.1.1 Cartography	303
10.1.2 Auditing	304
10.1.3 Navigation	305
10.2 Online Process Mining	305
10.3 Detect	307
10.4 Predict	311
10.5 Recommend	316
10.6 Processes Are Not in Steady State!	318
10.6.1 Daily, Weekly and Seasonal Patterns in Processes	318
10.6.2 Contextual Factors	318
10.6.3 Concept Drift in Processes	320
10.7 Process Mining Spectrum	321
Part V Putting Process Mining to Work	
11 Process Mining Software	325
11.1 Process Mining Not Included!	325
11.2 Different Types of Process Mining Tools	327
11.3 ProM: An Open-Source Process Mining Platform	331
11.3.1 Historical Context	331
11.3.2 Example ProM Plug-Ins	333
11.3.3 Other Non-commercial Tools	337
11.4 Commercial Software	339
11.4.1 Available Products	339
11.4.2 Strengths and Weaknesses	345
11.5 Outlook	352
12 Process Mining in the Large	353
12.1 Big Event Data	353
12.1.1 $N = \text{All}$	354
12.1.2 Hardware and Software Developments	356
12.1.3 Characterizing Event Logs	364
12.2 Case-Based Decomposition	368
12.2.1 Conformance Checking Using Case-Based Decomposition	369
12.2.2 Process Discovery Using Case-Based Decomposition	370
12.3 Activity-Based Decomposition	373
12.3.1 Conformance Checking Using Activity-Based Decomposition	374
12.3.2 Process Discovery Using Activity-Based Decomposition	376
12.4 Process Cubes	378
12.5 Streaming Process Mining	381
12.6 Beyond the Hype	384

13	Analyzing “Lasagna Processes”	387
13.1	Characterization of “Lasagna Processes”	387
13.2	Use Cases	391
13.3	Approach	392
13.3.1	Stage 0: Plan and Justify	393
13.3.2	Stage 1: Extract	395
13.3.3	Stage 2: Create Control-Flow Model and Connect Event Log	395
13.3.4	Stage 3: Create Integrated Process Model	396
13.3.5	Stage 4: Operational Support	396
13.4	Applications	397
13.4.1	Process Mining Opportunities per Functional Area	397
13.4.2	Process Mining Opportunities per Sector	398
13.4.3	Two Lasagna Processes	402
14	Analyzing “Spaghetti Processes”	411
14.1	Characterization of “Spaghetti Processes”	411
14.2	Approach	415
14.3	Applications	418
14.3.1	Process Mining Opportunities for Spaghetti Processes	418
14.3.2	Examples of Spaghetti Processes	420
Part VI Reflection		
15	Cartography and Navigation	431
15.1	Business Process Maps	431
15.1.1	Map Quality	432
15.1.2	Aggregation and Abstraction	432
15.1.3	Seamless Zoom	434
15.1.4	Size, Color, and Layout	438
15.1.5	Customization	440
15.2	Process Mining: TomTom for Business Processes?	441
15.2.1	Projecting Dynamic Information on Business Process Maps	441
15.2.2	Arrival Time Prediction	444
15.2.3	Guidance Rather than Control	444
16	Epilogue	447
16.1	Process Mining as a Bridge Between Data Mining and Business Process Management	447
16.2	Challenges	449
16.3	Start Today!	451
References		453
Index		463

Part I

Introduction

Part I: Introduction**Chapter 1**
Data Science in Action**Chapter 2**
Process Mining:
The Missing Link

Part II: Preliminaries**Chapter 3**
Process Modeling
and Analysis**Chapter 4**
Data Mining

Part III: From Event Logs to Process Models**Chapter 5**
Getting the Data**Chapter 6**
Process Discovery:
An Introduction**Chapter 7**
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery**Chapter 8**
Conformance
Checking**Chapter 9**
Mining Additional
Perspectives**Chapter 10**
Operational Support

Part V: Putting Process Mining to Work**Chapter 11**
Process Mining
Software**Chapter 12**
Process Mining in the
Large**Chapter 13**
Analyzing “Lasagna
Processes”**Chapter 14**
Analyzing “Spaghetti
Processes”

Part VI: Reflection**Chapter 15**
Cartography and
Navigation**Chapter 16**
Epilogue

The goal of process mining is to turn event data into insights and actions. Process mining is an integral part of data science, fueled by the availability of data and the desire to improve processes. Part I sets the scene for the more technical chapters on process modeling, data mining, data extraction, process discovery, conformance checking, performance analysis, and operational support. Chapter 1 starts with an overview of the data science discipline and is used to position process mining. Chapter 2 introduces the basic concepts of process mining.

Chapter 1

Data Science in Action

In recent years, *data science* emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. Existing approaches need to be combined to turn abundantly available data into value for individuals, organizations, and society. Moreover, new challenges have emerged, not just in terms of size (“Big Data”) but also in terms of the questions to be answered. This book focuses on the *analysis of behavior based on event data*. *Process mining* techniques use event data to discover processes, check compliance, analyze bottlenecks, compare process variants, and suggest improvements. In later chapters, we will show that process mining provides powerful tools for today’s data scientist. However, before introducing the main topic of the book, we provide an overview of the data science discipline.

1.1 Internet of Events

As described in [73], society shifted from being predominantly “analog” to “digital” in just a few years. This has had an incredible impact on the way we do business and communicate [99]. Society, organizations, and people are “Always On”. Data are collected *about anything, at any time, and at any place*. Nowadays, the term “Big Data” is often used to refer the expanding capabilities of information systems and other systems that depend on computing. These developments are well characterized by *Moore’s law*. Gordon Moore, the co-founder of Intel, predicted in 1965 that the number of components in integrated circuits would double every year. During the last 50 years the growth has indeed been exponential, albeit at a slightly slower pace. For example, the number of transistors on integrated circuits has been doubling every two years. Disk capacity, performance of computers per unit cost, the number of pixels per dollar, etc. have been growing at a similar pace. Besides these incredible technological advances, people and organizations depend more and more on computerized devices and information sources on the Internet. The IDC Digital Universe Study of April 2014 confirms again the spectacular growth of data [134].

This study estimates that the amount of digital information (cf. personal computers, digital cameras, servers, sensors) stored in 2014 already exceeded 4 Zettabytes and predicts that the “digital universe” will grow to 44 Zettabytes in 2020. The IDC study characterizes 44 Zettabytes as “6.6 stacks of iPads from Earth to the Moon”. This illustrates that the long anticipated *data explosion* has become an undeniable reality.

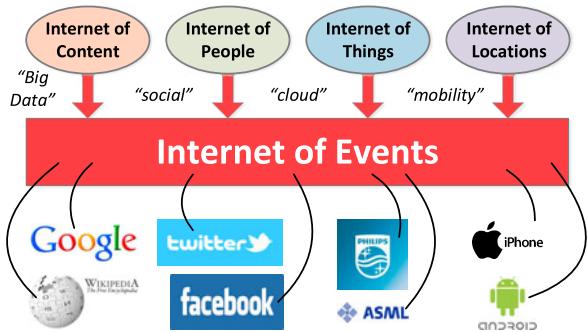
From Bits to Zettabytes

A “bit” is the smallest unit of information possible. One bit has two possible values: 1 (on) and 0 (off). A “byte” is composed of 8 bits and can represent $2^8 = 256$ values. To talk about larger amounts of data, multiples of 1000 are used: 1 Kilobyte (KB) equals 1000 bytes, 1 Megabyte (MB) equals 1000 KB, 1 Gigabyte (GB) equals 1000 MB, 1 Terabyte (TB) equals 1000 GB, 1 Petabyte (PB) equals 1000 TB, 1 Exabyte (EB) equals 1000 PB, and 1 Zettabyte (ZB) equals 1000 EB. Hence, 1 Zettabyte is $10^{21} = 1,000,000,000,000,000,000$ bytes. Note that here we used the International System of Units (SI) set of unit prefixes, also known as SI prefixes, rather than binary prefixes. If we assume binary prefixes, then 1 Kilobyte is $2^{10} = 1024$ bytes, 1 Megabyte is $2^{20} = 1048576$ bytes, and 1 Zettabyte is $2^{70} \approx 1.18 \times 10^{21}$ bytes.

Most of the data stored in the digital universe is unstructured, and organizations have problems dealing with such large quantities of data. One of the main challenges of today’s organizations is to *extract information and value from data* stored in their information systems.

The importance of information systems is not only reflected by the spectacular growth of data, but also by the role that these systems play in today’s business processes as the digital universe and the physical universe are becoming more and more aligned. For example, the “state of a bank” is mainly determined by the data stored in the bank’s information system. Money has become a predominantly digital entity. When booking a flight over the Internet, a customer is interacting with many organizations (airline, travel agency, bank, and various brokers), often without being aware of it. If the booking is successful, the customer receives an e-ticket. Note that an e-ticket is basically a number, thus illustrating the alignment between the digital and physical universe. When the SAP system of a large manufacturer indicates that a particular product is out of stock, it is impossible to sell or ship the product even when it is available in physical form. Technologies such as RFID (Radio Frequency Identification), GPS (Global Positioning System), and sensor networks will stimulate a further alignment of the digital universe and the physical universe. RFID tags make it possible to track and trace individual items. Also note that more and more devices are being monitored. Already 14 billion devices are connected to the Internet [134]. For example, Philips Healthcare is monitoring its medical equipment (e.g., X-ray machines and CT scanners) all over the world. This helps Philips to

Fig. 1.1 Internet of Events (IoE): Event data are generated from a variety of sources connected to the Internet



understand the needs of customers, test their systems under realistic circumstances, anticipate problems, service systems remotely, and learn from recurring problems. The success of the “App Store” of Apple illustrates that location-awareness combined with a continuous Internet connection enables new ways to pervasively intertwine the digital universe and the physical universe.

The spectacular growth of the digital universe, summarized by the overhyped term “Big Data”, makes it possible to record, derive, and analyze *events*. Events may take place inside a machine (e.g., an X-ray machine, an ATM, or baggage handling system), inside an enterprise information system (e.g., an order placed by a customer or the submission of a tax declaration), inside a hospital (e.g., the analysis of a blood sample), inside a social network (e.g., exchanging e-mails or Twitter messages), inside a transportation system (e.g., checking in, buying a ticket, or passing through a toll booth), etc. Events may be “life events”, “machine events”, or “organization events”. The term *Internet of Events* (IoE), coined in [146], refers to all event data available. The IoE is composed of:

- The *Internet of Content* (IoC), i.e., all information created by humans to increase knowledge on particular subjects. The IoC includes traditional web pages, articles, encyclopedia like Wikipedia, YouTube, e-books, newsfeeds, etc.
- The *Internet of People* (IoP), i.e., all data related to social interaction. The IoP includes e-mail, Facebook, Twitter, forums, LinkedIn, etc.
- The *Internet of Things* (IoT), i.e., all physical objects connected to the network. The IoT includes all things that have a unique id and a presence in an Internet-like structure.
- The *Internet of Locations* (IoL) which refers to all data that have a geographical or geospatial dimension. With the uptake of mobile devices (e.g., smartphones) more and more events have location or movement attributes.

Note that the IoC, the IoP, the IoT, and the IoL are overlapping. For example, a place name on a webpage or the location from which a tweet was sent. *Process mining aims to exploit event data in a meaningful way*, for example, to provide insights, identify bottlenecks, anticipate problems, record policy violations, recommend countermeasures, and streamline processes. This explains our focus on event data.

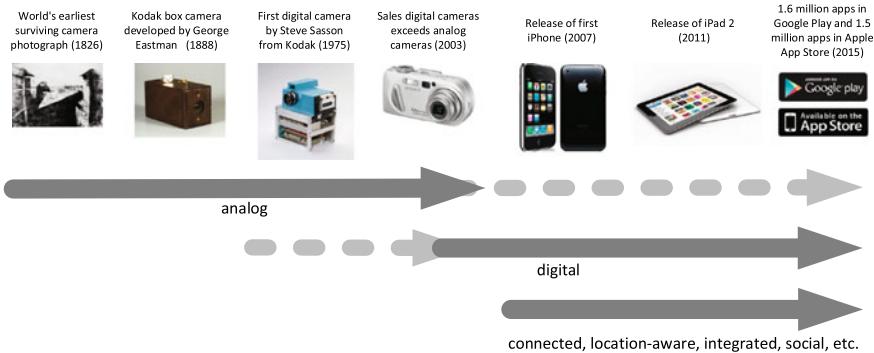


Fig. 1.2 The transition from analog to digital dramatically changed the way we create and share photos. This is one of the factors contributing to the rapid expansion of the Internet of Events (IoE)

To illustrate the above developments, let us consider the development of photography over time (see Fig. 1.2). Photography emerged at the beginning of the 19th century. Around 1800, Thomas Wedgwood attempted to capture the image in a camera obscura by means of a light-sensitive substance. The earliest remaining photo dates from 1826. Towards the end of the 19th century, photographic techniques matured. George Eastman founded Kodak around 1890 and produced “The Kodak” box camera that was sold for \$25, thus making photography accessible for a larger group of people. The company witnessed the rapid growth of photography while competing with companies like Fujifilm. In 1976, Kodak was responsible for 90% of film sales and 85% of camera sales in the United States [57]. Kodak developed the first digital camera in 1975, i.e., at the peak of its success. The Kodak digital camera had the size of a toaster and a CCD image sensor that only allowed for 0.01 megapixel black and white pictures. It marked the beginning of digital photography, but also the decline of Kodak. Kodak was unable to adapt to the market of digital photography. Competitors like Sony, Canon, and Nikon better adapted to the rapid transition from analog to digital. In 2003, the sales of digital cameras exceeded the sales of traditional cameras for the first time. Today, the market for analog photography is virtually non-existent. Soon after their introduction, smartphones with built-in cameras overtook dedicated cameras. The first iPad having a camera (iPad 2) was presented on March 2nd, 2011 by Steve Jobs. Today, the sales of tablet-like devices like the iPad exceed the sales of traditional PCs (desktops and laptops). As a result of these developments, most photos are made using mobile phones and tablets. The remarkable transition from analog to digital photography has had an impact that goes far beyond the photos themselves. Today, photos have GPS coordinates allowing for localization. Photos can be shared online (e.g., Flickr, Instagram, Facebook, and Twitter) and changed the way we communicate and socialize (see the uptake of the term “selfie”). Smartphone apps can detect eye cancer, melanoma, and other diseases by analyzing photos. A photo created using a smartphone may generate a wide range of events (e.g., sharing) having data attributes (e.g., location) that reach far beyond the actual image. As illustrated by

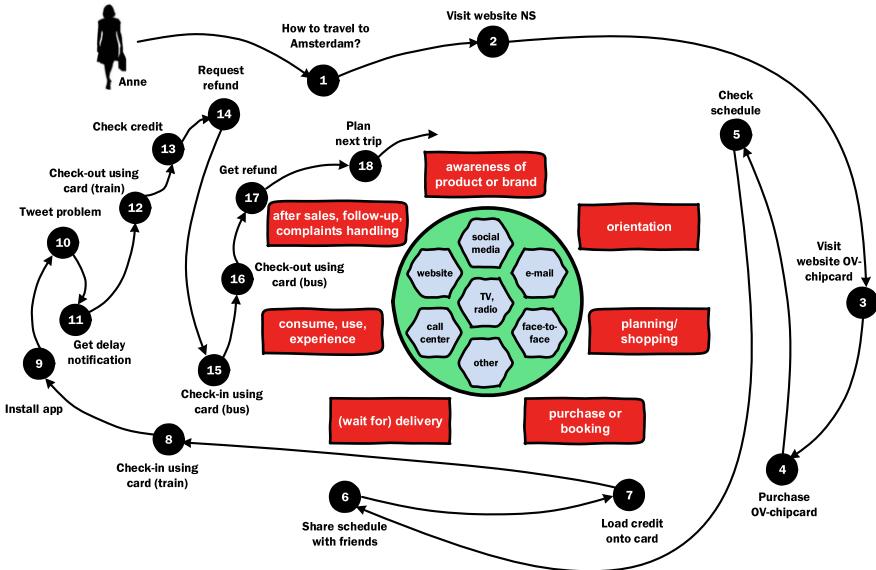


Fig. 1.3 An example of a customer journey illustrating the many (digital) touchpoints generating events that allow us to understand and serve customers better

Fig. 1.2, developments in photography accelerated since the first digital camera, and the transition from analog to digital photography contributed significantly to the growth of the Internet of Events (IoE). Digitalization resulted not just in content (e.g., photos) but also in new ways to capture “events” showing what is really happening.

Let us now consider another development: the digitization¹ of the *customer journey* where customers interact in multiple novel ways with organizations [36]. In the digital era, there are many *touchpoints* using different media. The center of Fig. 1.3 shows the different media: social media, e-mails, websites, face-to-face contacts, call-centers, etc. Although there are significant differences between the wide variety communication channels, “content” tends to become less dependent on the media used (phone, laptop, etc.). Smartphones and iPads can make photographs, computers can be used to make phone calls (through Skype or FaceTime), and customer complaints can be expressed via a website or call-center. Different devices and services co-exist in an integrated ecosystem. Consider, for example, the capturing, managing, publication, viewing, and sharing of photos using digital cameras, mobile phones, computers, websites, media-players, printers, televisions, interactive whiteboards, etc.

¹ Some distinguish between the terms digitization and digitalization where *digitization* is the process of converting an analogue reality into digital bits and *digitalization* refers to the way in which social life and businesses are impacted by digital communication infrastructures. In this book, we do not use this distinction as both are too much intertwined.

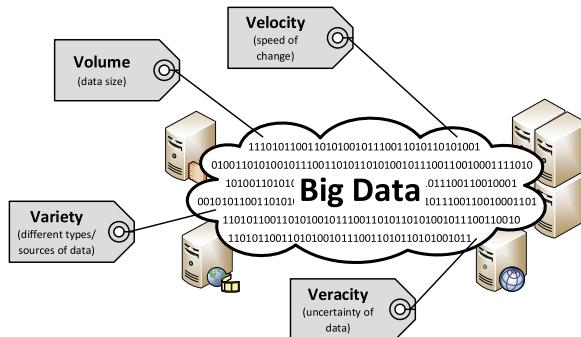
Figure 1.3 distinguishes seven stages for an archetypal customer journey:

1. *Awareness of product or brand.* The customer needs to be aware of the product and/or brand to start a customer journey. For example, a customer that does not know about the existence of air purifiers will not consider buying one. (An air purifier removes contaminants from the air in a room to fight allergies, asthmatics, or tobacco smoke.)
2. *Orientation.* During the second stage, the customer is interested in a product, possibly of a particular brand. For example, the customer searches for the differences between air purifiers, e.g., there are devices that use thermodynamic sterilization, ultraviolet germicidal irradiation, HEPA filters, etc.
3. *Planning/shopping.* After the orientation phase the customer may decide to purchase a product or service. This requires planning and/or shopping, e.g., browsing websites for the best offer.
4. *Purchase or booking.* If the customer is satisfied with a particular offering, the product is bought or the service (e.g., flight or hotel) is booked.
5. *(Wait for) delivery.* This is the stage after purchasing the product or booking the service, but before the actual delivery. For example, the air purifier that was purchased is unexpectedly out of stock, resulting in a long delivery time and an unhappy customer. Events like this are an integral part of the customer journey.
6. *Consume, use, experience.* At the sixth stage, the product or service is used. For example, the air purifier arrived and is used on a daily basis. While using the product or service, a multitude of events may be generated. For example, some air purifiers are connected to the Internet measuring the air quality. The user can control the purifier via an app and monitor the air quality remotely. The recorded event data can be used to understand the actual use of the product by the customer.
7. *After sales, follow-up, complaints handling.* This is the stage that follows the actual use of the product or service. For example, the customer may want to return the air purifier because it is broken or does not deliver the performance expected. At this seventh stage, new add-on products may be offered (e.g., air filters).

Given a particular product or organization, many customer journeys are possible. The customer journey is definitely *not* a linear process. Stages may be skipped and revisited. Moreover, customers may use many products of the same brand leading to an overall customer experience influencing future purchase decisions.

Figure 1.3 shows one particular customer journey to illustrate the different touchpoints potentially providing lots of event data for analysis. Consider a teenager (let us call her Anne) that wants to make a trip from Eindhoven Central Station to Amsterdam to visit the Van Gogh Museum. Anne first explores different ways to travel to Amsterdam (1) followed by a visit to the website of NS (Dutch railroad company) (2). Anne finds out that she needs to buy a so-called “OV-chipcard”. Such a card gives access to a contactless smartcard system used for all public transport in the Netherlands. Using the card Anne can check-in at the start of a trip and check-out at the end of trip. After visiting the OV-chipcard website (3), Anne purchases

Fig. 1.4 The four V's of Big Data: Volume, Velocity, Variety, and Veracity



the OV-chipcard from a machine in the train station (4), and checks the schedule (5) using her mobile phone. She shares the selected schedule with her friends (6). Before checking in using the card (8), she first loads 100 euro credit onto her OV-chipcard (7). While traveling she installs the NS app obtained from iTunes (9). Due to a broken cable, the train gets a 90 minute delay. Anne tweets about the problem while mentioning @NS_online to express her disappointment (10). A bit later, she gets a push message from her newly installed app (11). Customers build expectations based on experiences, and Anne is clearly not happy. Due to the digitization of the customer journey, such negative sentiments can be detected and acted upon. Finally, Anne reaches Amsterdam Central Station and checks out (12). Anne checks her credit on the card using a machine (13) and requests a refund using the app on her mobile phone (14). She takes the bus to the Van Gogh Museum. When entering the bus she checks in (15) and checks out (16) when exiting. A few days later she gets the requested refund (17) and starts planning her next trip (18).

During Anne's journey many events were recorded. It is easy to relate all events involving the OV-chipcard. However, some of the other events may be difficult to relate to Anne. This complicates analysis. *Event correlation*, i.e., establishing relationships between events, is one of the key challenges in data science.

The seven customer journey stages in Fig. 1.3 illustrate that the journey does not end after the 4th stage (purchase or booking). The classical “funnel-oriented” view towards purchasing a product is too restrictive. The availability of customer data from all seven stages helps shifting attention from sales to loyalty.

The development of photography and the many digital touchpoints in today's customer journey exemplify the growing availability of event data. Although data science is definitely not limited to Big Data, the dimensions of data are rapidly changing resulting in new challenges. It is fashionable to list challenges starting with the letter 'V'. Figure 1.4 lists the "four V's of Big Data": *Volume*, *Velocity*, *Variety*, and *Veracity*. The first 'V' (Volume) refers to the incredible scale of some data sources. For example, Facebook has over 1 billion active users and stores hundreds of petabytes of user data. The second 'V' (Velocity) refers to the frequency of incoming data that need to be processed. It may be impossible to store all data or the data may change so quickly that traditional batch processing approaches cannot cope with high-velocity streams of data. The third 'V' (Variety) refers to the differ-

ent types of data coming from multiple sources. Structured data may be augmented by unstructured data (e.g., free text, audio, and video). Moreover, to derive maximal value, data from different sources needs combined. As mentioned before, the correlation of data is often a major challenge. The fourth ‘V’ (Veracity) refers to the trustworthiness of the data. Sensor data may be uncertain, multiple users may use the same account, tweets may be generated by software rather than people, etc.

Already in 2001, Doug Laney wrote a report introducing the first three V’s [87]. Later the fourth ‘V’ (Veracity) was added. Next to the basic four V’s of Big Data shown in Fig. 1.4, many authors and organizations proposed additional V’s: Variability, Visualization, Value, Venue, Validity, etc. However, there seems to be a consensus that *Volume*, *Velocity*, *Variety*, and *Veracity* are the key characteristics.

Later in this book we will focus exclusively on event data. However, these are an integral part of any Big Data discussion. Input for process mining is an event log which can be seen as a particular view on the event data available. For example, an event log may contain all events related to a subset of customers and used to build a customer journey map.

1.2 Data Scientist

Fueled by the developments just described, *Data science* emerged as a new discipline in recent years. Many definitions have been suggested [48, 112]. For this book, we propose the following definition:

Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.

The above definition implies that data science is broader than applied statistics and data mining. *Data scientists* assist organizations in turning data into value. A data scientist can answer a variety of data-driven questions. These can be grouped into the following four main categories [146]:

- (Reporting) *What happened?*
- (Diagnosis) *Why did it happen?*
- (Prediction) *What will happen?*
- (Recommendation) *What is the best that can happen?*

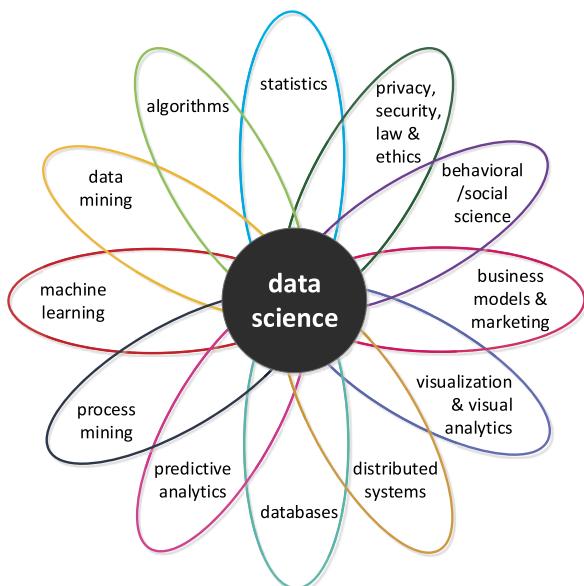
The definition of data science given is quite broad. Some consider data science as just a fancy term for *statistics*. Clearly, data science has its roots in statistics,

a discipline that developed over four centuries. John Graunt (1620–1674) started to study London’s death records around 1660. Based on this he was able to predict the life expectancy of a person at a particular age. Francis Galton (1822–1911) introduced statistical concepts like regression and correlation at the end of the 19th century. Although data science can be seen as a continuation of statistics, the majority of statisticians did not contribute much to recent progress in data science. Most statisticians focused on theoretical results rather than real-world analysis problems. The computational aspects, which are critical for larger data sets, are typically ignored by statisticians. The focus is on generative modeling rather than prediction and dealing with practical challenges related to data quality and size. When the data mining community realized major breakthroughs in the discovery of patterns and relationships (e.g., efficiently learning decision trees and association rules), most statisticians referred to these discovery practices as “data fishing”, “data snooping”, and “data dredging” to express their dismay.

A few well-known statisticians criticized their colleagues for ignoring the actual needs and challenges in data analysis. John Tukey (1915–2000), known for his fast Fourier transform algorithm and the box plots, wrote in 1962: “For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. . . . I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.” [133]. This text was written over 50 years ago. Also Leo Breiman (1928–2005), another distinguished statistician, wrote in 2001 “This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics.” [25]. David Donoho adequately summarizes the 50 year old struggle between old-school statistics and real-life data analysis in [48].

Data science is also closely related to data processing. Turing award winner Peter Naur (1928–2016) used the term “data science” long before it was in vogue. In 1974, Naur wrote: “A basic principle of *data science*, perhaps the most fundamental that may be formulated, can now be stated: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available” [107]. Earlier, Peter Naur also defined *datalogy* as the “science of the nature and use of data” and suggested to use this term rather than “computer science”. The book from 1974 also has two parts considering “large data”: “Part 5—Processes with Large Amounts of Data” and “Part 6—Large Data Systems” [107]. In the book, “large amounts of data” are all data sets that cannot be stored in working memory. The maximum capacity of magnetic disk stores considered in [107] ranges between 1.25 and 250 megabytes. Not only the disks are orders of magnitude smaller than today’s disks, also the notion of what is “large/big” has changed dramatically since the early 1970s. Nevertheless, many of the core principles of data processing have remained invariant.

Fig. 1.5 The ingredients contributing to data science



Like data science, computer science had its roots in a number of related areas, including mathematics. Computer science emerged because of the availability of computing resources and the need for computer scientists. Data science is now emerging because of the omnipresence and abundance of data and the need for data scientists that can turn data into value.

Data science is an amalgamation of different partially overlapping (sub)disciplines. Figure 1.5 shows the main ingredients of data science. The diagram should be taken with a grain of salt. The (sub)disciplines are overlapping and varying in size. Moreover, the boundaries are not clear-cut and seem to change over time. Consider, for example, the difference between data mining and machine learning or statistics. Their roots are very different: data mining emerged from the database community, and machine learning emerged from the Artificial Intelligence (AI) community, both quite disconnected from the statistics community. Despite the different roots, the three (sub)disciplines are definitely overlapping.

- *Statistics* can be viewed as the origin of data science. The discipline is typically split into *descriptive* statistics (to summarize sample data using notions like mean, standard deviation, and frequency) and *inferential* statistics (using sample data to estimate characteristics of all data or to test a hypothesis).
- *Algorithms* are crucial in any approach analyzing data. When data sets get larger, the complexity of the algorithms becomes a primary concern. Consider, for example, the Apriori algorithm for finding frequent items sets, the MapReduce approach for parallelizing algorithms, and the PageRank algorithm used by Google search.
- *Data mining* can be defined as “the analysis of (often large) data sets to find unsuspected relationships and to summarize the data in novel ways that are both

understandable and useful to the data owner” [69]. The input data are typically given as a table and the output may be rules, clusters, tree structures, graphs, equations, patterns, etc. Clearly, data mining builds on statistics, databases, and algorithms. Compared to statistics, the focus is on scalability and practical applications.

- *Machine learning* is concerned with the question of how to construct computer programs that automatically improve with experience [102]. The difference between data mining and machine learning is equivocal. The field of machine learning emerged from within Artificial Intelligence (AI) with techniques such as neural networks. Here, we use the term machine learning to refer to algorithms that give computers the capability to learn *without* being explicitly programmed (“learning from experience”). To learn and adapt, a model is built from input data (rather than using fixed routines). The evolving model is used to make data-driven predictions or decisions.
- *Process mining* adds the process perspective to machine learning and data mining. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). Event data are related to explicit process models, e.g., Petri nets or BPMN models. For example, process models are discovered from event data or event data are replayed on models to analyze compliance and performance.
- *Predictive analytics* is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. To generate predictions, existing mining and learning approaches are applied in a business context. Predictive analytics is related to business analytics and business intelligence.
- *Databases* are used to store data. The database discipline forms one of the cornerstones of data science. Database Management (DBM) systems serve two purposes: (i) structuring data so that they can be managed easily and (ii) providing scalability and reliable performance. Using database technology, application programmers do not need to worry about data storage. Until recently, relational databases and SQL (Structured Query Language) were the norm. Due to the growing volume of data, massively distributed databases and so-called NoSQL databases emerged. Moreover, in-memory computing (cf. SAP HANA) can be used to answer questions in real-time. Related is OLAP (Online Analytical Processing) where data are stored in multidimensional cubes facilitating analysis from different points of view.
- *Distributed systems* provide the infrastructure to conduct analysis. A distributed system is composed of interacting components that coordinate their actions to achieve a common goal. Cloud, grid, and utility computing rely on distributed systems. Some analysis tasks are too large or too complex to be performed on a single computer. Such tasks can be split into many smaller tasks that can be performed concurrently on different computing nodes. Scalability may be realized by sharing and/or extending the set of computing nodes.
- *Visualization & visual analytics* are key elements of data science. In the end people need to interpret the results and guide analysis. Automated learning and

mining techniques can be used to extract knowledge from data. However, if there are many “unknown unknowns” (things we don’t know we don’t know),² analysis heavily relies on human judgment and direct interaction with the data. The perception capabilities of the human cognitive system can be exploited by using the right visualizations [178]. Visual analytics, a term coined by Jim Thomas (1946–2010), combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [83].

- *Business models & marketing* also appear in Fig. 1.5 because data science is about turning data into value, including business value. The market capitalization of Facebook in November 2015 was approximately US \$300 billion while having approximately 1500 million monthly active users. Hence, the average value of a Facebook user was US \$200. At the same time, the average value of a Twitter user was US \$55 (market capitalization of approximately US \$17 billion with 307 million users). Via the website www.tvalue.com one can even compute the value of a particular Twitter account. In November 2015, the author’s Twitter account (@wvdaalst) was estimated to have a value of US \$1002.98. These numbers illustrate the economic value of data and the success of young companies based on new business models. Airbnb (helping people to list, find and rent lodging), Uber (connecting travelers and drivers who use their own cars), and Alibaba (an online business-to-business trading platform) are examples of data-driven companies that are radically changing the hotel, taxi, and trading business. Marketing is also becoming more data-driven (see Sect. 1.1 describing the increase in digital touchpoints during a customer journey). Data scientists should understand how business considerations are driving the analysis of new types of data.
- *Behavioral/social science* appears in Fig. 1.5 because most data are (indirectly) generated by people and analysis results are often used to influence people (e.g., guiding the customer to a product or encouraging a manager to eliminate waste). Behavioral science is the systematic analysis and investigation of human behavior. Social sciences study the processes of a social system and the relationships among individuals within a society. To interpret the results of various types of analytics, it is important to understand human behavior and the social context in which humans and organizations operate. Moreover, analysis results often trigger questions related to coaching and positively influencing people.
- *Privacy, security, law, and ethics* are key ingredients to protect individuals and organizations from “bad” data science practices. Privacy refers to the ability to seclude sensitive information. Privacy often depends on security mechanisms which aim to ensure the confidentiality, integrity and availability of data. Data should be accurate and stored safely, not allowing for unauthorized access. Privacy and security need to be considered carefully in all data science applications. Individuals

²On February 12, 2002, when talking about weapons of mass destruction in Iraq, United States Secretary of Defense Donald Rumsfeld used the following classification: (i) “known knowns” (things we know we know), (ii) “known unknowns” (things we know we don’t know), and (iii) “unknown unknowns” (things we don’t know we don’t know).

need to be able to trust the way data are stored and transmitted. Next to concrete privacy and security breaches, there may be ethical notions related to “good” and “bad” conduct. Not all types of analysis possible are morally defendable. For example, mining techniques may favor particular groups (e.g., a decision tree may reveal that it is better to give insurance to middle-aged white males rather than other groups). Moreover, due to a lack of sufficient data, minority groups may be wrongly classified. A data scientist should be aware of such problems and provide safeguards for “irresponsible” forms of data science.

Figure 1.5 shows that data science is quite broad and located at the intersection of existing disciplines. It is difficult to combine all the different skills needed in a single person. Josh Wills, former director of data science at Cloudera, defined a data scientist as “a person who is better at statistics than any software engineer and better at software engineering than any statistician”. It will be a challenge to find and/or educate “unicorn” data scientists able to cover the full spectrum depicted in Fig. 1.5. As a result, ‘unicorn’ data scientists are in high demand and extremely valuable for data-driven organizations. As an alternative it is also possible to form multi-disciplinary teams covering the “flower” of Fig. 1.5. In the latter case, it is vital that the team members are able to see the bigger picture and complement each other in terms of skills.

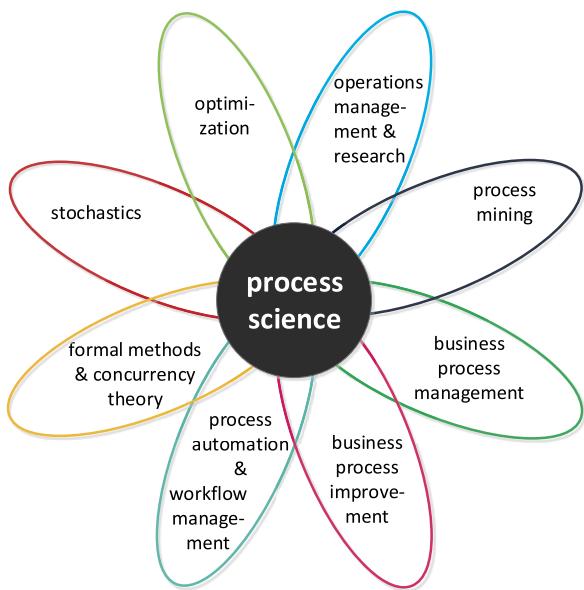
1.3 Bridging the Gap Between Process Science and Data Science

In Fig. 1.5, we listed process mining, the topic of this book, as one of the essential ingredients of data science. Unfortunately, this is not a common view. The process perspective is absent in many Big Data initiatives and data science curricula. We argue that *event data should be used to improve end-to-end processes*. Process mining can be seen as a means to *bridge the gap between data science and process science*. Data science approaches tend to be process agnostic whereas process science approaches tend to be model-driven without considering the “evidence” hidden in the data.

We use the umbrella term “*process science*” to refer to the *broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes*. Figure 1.6 shows the ingredients of process science. Just like Fig. 1.5, the diagram should be taken with a grain of salt. The (sub)disciplines mentioned in Fig. 1.6 are also overlapping and varying in size.

- *Stochastics* provides a repertoire of techniques to analyze random processes. The behavior of a process or system is modeled using random variables in order to allow for analysis. Well-known approaches include Markov models, queueing networks/systems, and simulation. These can be used to analyze waiting times, reliability, utilization, etc. in the context stochastic processes.

Fig. 1.6 Process science is an umbrella term for the broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes



- *Optimization* techniques aim to provide a “best” alternative (e.g., cheapest or fastest) from a large or even infinite set of alternatives. Consider, for example, the following question: Given a list of cities and the distances between each pair of cities, what is a shortest possible route that visits each city exactly once and returns to the origin city? Numerous optimization techniques have been developed to answer such questions as efficient as possible. Well-known approaches include Linear Programming (LP), Integer Linear Programming (ILP), constraint satisfaction, and dynamic programming.
- *Operations management & research* deals with the design, control and management of products, processes, services and supply chains. Operations Research (OR) tends to focus on the analysis of mathematical models. Operations Management (OM) is closer to industrial engineering and business administration.
- *Business process management* is the discipline that combines approaches for the design, execution, control, measurement and optimization of business processes. Business Process Management (BPM) efforts tend to put emphasis on explicit process models (e.g., Petri nets or BPMN models) that describe the control-flow and, optionally, other perspectives (organization, resources, data, functions, etc.) [50, 143, 187].
- *Process mining* is also part of process science. For example, process mining techniques can be used to discover process models from event data. By replaying these data, bottlenecks and the effects of non-compliance can be unveiled. Compared to mainstream BPM approaches the focus is not on process modeling, but on exploiting event data. Sometimes the terms Workflow Mining (WM), Business Process Intelligence (BPI), and Automated Business Process Discovery (ABPD) are used to refer to process-centric data-driven approaches.

- *Business process improvement* is an umbrella term for a variety of approaches aiming at process improvement. Examples are Total Quality Management (TQM), Kaizen, (Lean) Six Sigma, Theory of Constraints (TOC), and Business Process Reengineering (BPR). Note that most of the ingredients in Fig. 1.6 ultimately aim at process improvement, thus making the term business process improvement rather unspecific. One could argue that the whole of process science aims to improve processes.
- *Process automation & workflow management* focuses on the development of information systems supporting operational business processes including the routing and distribution of work. Workflow Management (WFM) systems are model-driven, i.e., a process model suffices to configure the information system and run the process. As a result, a process can be changed by modifying the corresponding process model.
- *Formal methods & concurrency theory* build on the foundations of theoretical computer science, in particular logic calculi, formal languages, automata theory, and program semantics. Formal methods use a range of languages to describe processes. Examples are transition systems, Petri nets, process calculi such as CSP, CCS and π -calculus, temporal logics such as LTL and CTL, and statecharts. Model checkers such as SPIN can be used to verify logical properties such as the absence of deadlocks. Concurrency complicates analysis, but is also essential: In reality parts of a process or system may be executing simultaneously and potentially interacting with each other. Petri nets were the first formalism to model and analyze concurrent processes. Many BPM, WFM, and process mining approaches build upon such formalisms.

As mentioned earlier, Fig. 1.6 should not be taken too seriously. It is merely a characterization of process science and its main ingredients. Note, for example, that stochastics and optimization are partly overlapping (e.g., solving Markov decision processes) and that BPM can be viewed as a continuation or extension of WFM with less emphasis on automation.

Process mining brings together traditional model-based process analysis and data-centric analysis techniques. As shown in Fig. 1.7, *process mining can be viewed as the link between data science and process science*. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). Mainstream data science approaches tend to be process agnostic. Data mining, statistics and machine learning techniques do not consider end-to-end process models. Process science approaches are process-centric, but often focus on modeling rather than learning from event data. The unique positioning of process mining, as sketched in Fig. 1.7, makes it a powerful tool to exploit the growing availability of data for improving end-to-end processes.

Process mining only recently emerged as a subdiscipline of both data science and process science, but the corresponding techniques can be applied to any type of operational processes (organizations and systems). Example applications include: analyzing treatment processes in hospitals, improving customer service processes in a multinational corporation, understanding the browsing behavior of customers

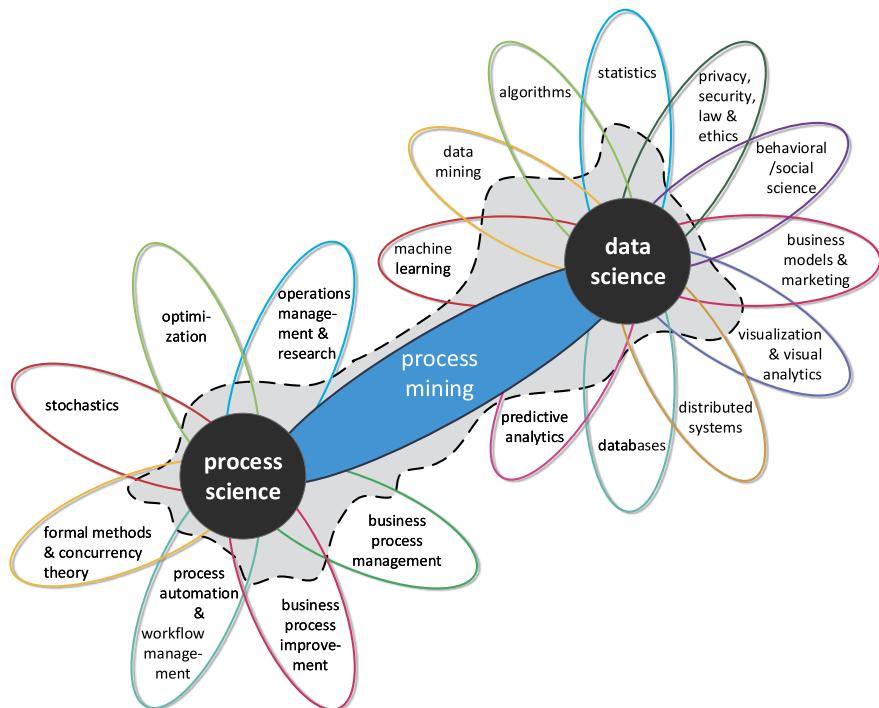


Fig. 1.7 Process mining as the bridge between data science and process science

using a booking site, analyzing failures of a baggage handling system, and improving the user interface of an X-ray machine. What all of these applications have in common is that dynamic behavior needs to be related to process models. Hence, we refer to this as “data science in action”.

Spreadsheets: Dealing with numbers rather than dynamic behavior

Spreadsheet software can be found on most computers, and over the last 25 years many computers have been purchased just to be able to create and use spreadsheets. A spreadsheet is composed of cells organized in rows and columns. Some cells serve as input, other cells have values computed over a collection of other cells (e.g., taking the sum over an array of cells). The expression associated to a cell may use a range of arithmetic operations (add, subtract, multiply, etc.) and predefined functions. For example, Microsoft’s Excel provides hundreds of functions including statistical functions, math and trigonometry functions, financial functions, and logical functions. Most organizations use spreadsheets in financial planning, budgeting, work distribution, etc. Hence, it is interesting to view process mining against the backdrop of this widely used technology.

The first widely used spreadsheet program was VisiCalc (“Visible Calculator”) developed by Dan Bricklin and Bob Frankston, founders of Software Arts (later named VisiCorp). VisiCalc was released in 1979 for the Apple II computer. It is generally considered as Apple II’s “killer application” because numerous organizations purchased the Apple II computer just to be able to use VisiCalc. When Lotus 1-2-3 was launched in 1983, VisiCalc sales dropped dramatically. Lotus 1-2-3 took full advantage of the IBM PC’s capabilities and better supported data handling and charting. What VisiCalc was for the Apple II, Lotus 1-2-3 was for the IBM PC. For the second time, a spreadsheet program generated a tremendous growth in computer sales. People were buying computers in order to run spreadsheet software: A nice example of the “tail” (VisiCalc/Lotus 1-2-3) wagging the “dog” (Apple-II/IBM PC). Lotus 1-2-3 dominated the spreadsheet market until 1992. The dominance ended with the uptake of Microsoft Windows. After decades of spectacular IT-developments, spreadsheet software can still be found on most computers (e.g., Excel is part of Microsoft’s Office) and can be accessed online (e.g., Google Sheets as part of Google Docs).

The situations in which spreadsheets can be used in a meaningful way are almost endless. In short, *spreadsheets can be used to do anything with numbers*. However, spreadsheets are *not* suitable for analyzing event data. One can count frequencies, sums, and the number of events per case using a so-called pivot table, but spreadsheets cannot be used to analyze bottlenecks and deviations (see Fig. 1.8). Consider questions like:

- What are the most frequent paths in my process? Do they change over time?
- What do the cases that take longer than 3 months have in common? Where are the bottlenecks causing these delays?
- Which cases deviate from the reference process? Do these deviations also cause delays?

Obviously, these questions cannot be answered using spreadsheets because the process perspective is completely absent in spreadsheets. Processes *cannot* be captured in numerical data and operations like summation. Process models and concepts such as cases, events, activities, timestamps, and resources need to be treated as first-class citizens during analysis. Data mining tools and spreadsheet programs take as input any tabular data without distinguishing these key concepts. As a result, such tools tend to be process-agnostic. Nevertheless, there is an obvious need for spreadsheet-like technology tailored towards processes and event data.

Where spreadsheets work with *numbers*, process mining starts from *event data* with the aim to analyze processes. Instead of pie charts, bar charts, and tables, results include end-to-end process models, conformance diagnostics, and bottlenecks.

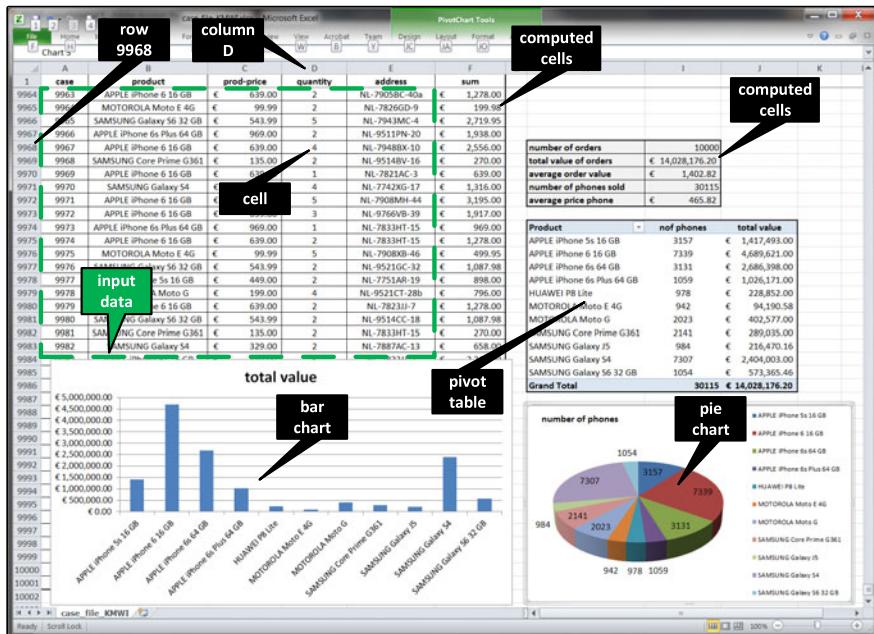


Fig. 1.8 Spreadsheets can be used to do anything with numbers, but have difficulties adequately capturing dynamic behavior

As will be demonstrated in later chapters, the process mining spectrum is quite *broad*. It is not limited to automated process discovery: Process mining can also be used to check compliance, diagnose deviations, pinpoint bottlenecks, improve performance, predict flow times, and recommend actions. Process mining techniques are also *generic*: just like spreadsheet software. Event logs and operational processes can be found everywhere and the analysis techniques are not limited to specific application domains. Just like Excel can be used in finance, production, sales, education, and sports, process mining software can be used in a variety of application domains.

1.4 Outlook

Process mining provides an important bridge between data mining and business process modeling and analysis. Process mining research at TU/e (Eindhoven University of Technology) started in 1999. At that time there was little event data available and the initial process mining techniques were extremely naïve and hence unusable in practice. Over the last decade event data have become readily available and process mining techniques have matured. Moreover, process mining algorithms have been implemented in various academic and commercial systems. Today, there is an active group of researchers working on process mining, and it has become one of the

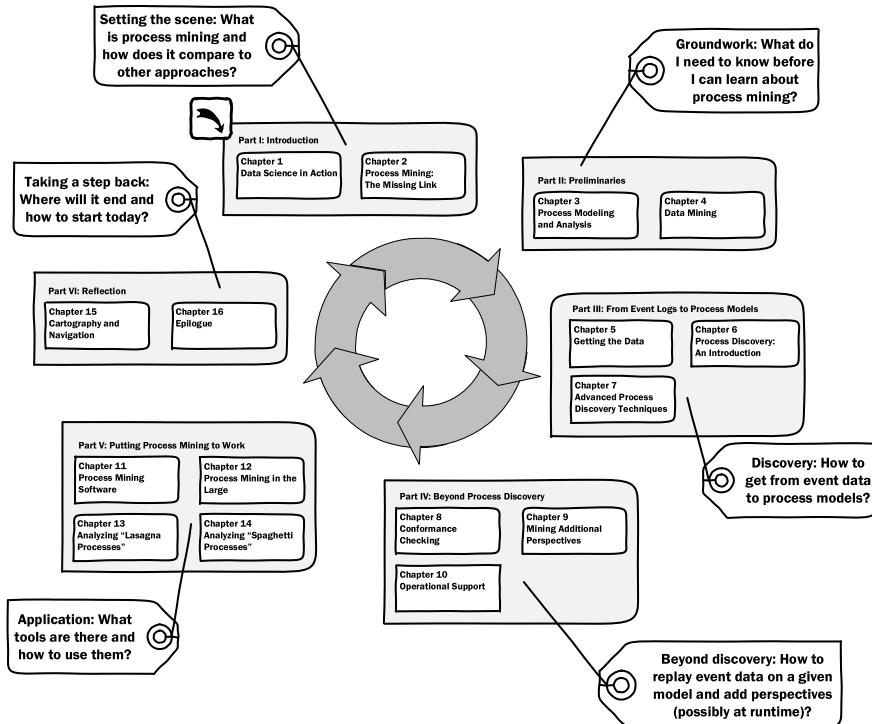


Fig. 1.9 Outline of the book

“hot topics” in BPM research. Moreover, there is a rapidly growing interest from industry in process mining. More and more software vendors started adding process mining functionality to their tools. Our open-source process mining tool ProM is widely used all over the globe and provides an easy starting point for practitioners, students, and academics. These developments are the main motivation for writing this book. There are many books on data mining, business intelligence, process reengineering, and BPM, but these rarely address process mining.

This book aims to provide a comprehensive overview of process mining. The book is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers. On the one hand, the book avoids delving into unnecessary details. On the other hand, the book does not shy away from formal definitions and technical issues needed to fully understand the essence of process mining. As Einstein said: “Everything should be made as simple as possible, but not one bit simpler”.

Figure 1.9 provides an overview of the book. *Part I* introduces process mining and positions this emerging discipline in the context of data science and process science. *Chap. 2* discusses the role of process models, introduces the notion of event logs, and illustrates the main process mining tasks using a small example.

Part II provides the preliminaries necessary for reading the remainder of the book. *Chap. 3* introduces different process modeling languages and provides an overview of model-based analysis techniques. *Chap. 4* introduces standard data mining techniques such as decision tree learning and association rule learning. Process mining can be seen as a bridge between the preliminaries presented in both chapters.

Part III focuses on one particular process mining task: process discovery. *Chap. 5* discusses the input needed for process mining. The chapter discusses different input formats and issues related to the extraction of event logs from heterogeneous data sources. *Chap. 6* presents the α -algorithm step-by-step in such a way that the reader can understand how it works and see its limitations. This simple algorithm has problems dealing with less structured processes. Nevertheless, it provides a basic introduction into the topic and serves as a “hook” for discussing more advanced algorithms and general issues related to process mining. *Chap. 7* introduces more advanced process discovery approaches. This way the reader gets a good understanding of the state-of-the-art and is guided in selecting suitable techniques.

Part IV moves beyond process discovery, i.e., the focus is no longer on discovering the control-flow. *Chap. 8* presents conformance checking approaches, i.e., techniques to compare and relate event logs and process models. It is shown that conformance can be quantified and that deviations can be diagnosed. *Chap. 9* focuses on other perspectives: the organizational perspective, the case perspective, and the time perspective. *Chap. 10* shows that process mining can also be used to support operational processes at runtime, i.e., while cases are running it is possible to detect violations, make predictions, and provide recommendations.

Part V guides the reader in successfully applying process mining in practice. *Chap. 11* provides an overview of the different process mining tools. Data science is often related to Big Data. The “four V’s of Big Data” (Fig. 1.4) are obviously also relevant for event data and their analysis. *Chap. 12* shows that process mining problems can be decomposed in various ways and many of the techniques can be adapted to provide scalability. The next two chapters are based on the observation that there are essentially two types of processes: “Lasagna processes” and “Spaghetti processes”. Lasagna processes are well-structured and relatively simple. Therefore, process discovery is less interesting, but the techniques presented in Part IV are highly relevant for Lasagna processes. The added value of process mining can be found in conformance checking, detailed performance analysis, and operational support. *Chap. 13* explains how process mining can be applied in such circumstances and provides various real-life examples. Spaghetti processes are less structured. Therefore, the added value of process mining shifts to providing insights and generating ideas for better controlled processes, but advanced techniques such as prediction are less relevant for Spaghetti processes. *Chap. 14* shows how to apply process mining in such less-structured environments.

Part VI takes a step back and reflects on the material presented in the preceding parts. *Chap. 15* provides a broader vision on the topic by comparing process modeling with cartography, and relating BPM systems to navigation systems provided by vendors such as TomTom, Garmin, and Navigon. The goal of this chapter is to provide a refreshing view on process management and reveal the limitations of existing

information systems. *Chap. 16* concludes the book by summarizing improvement opportunities provided by process mining. The chapter also discusses some of the key challenges and provides concrete pointers to start applying the material presented in this book.

Chapter 2

Process Mining: The Missing Link

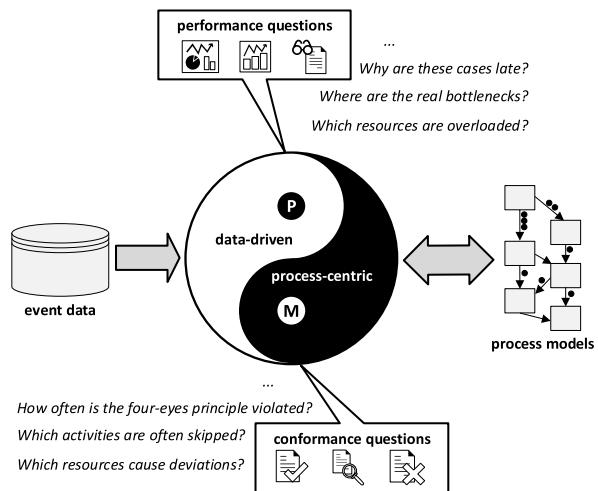
Information systems are becoming more and more intertwined with the operational processes they support. As discussed in the previous chapter, multitudes of events are recorded by today's information systems. Nevertheless, organizations have problems extracting value from these data. The goal of *process mining* is to use event data to extract process-related information, e.g., to automatically *discover* a process model by observing events recorded by some enterprise system. A small example is used to explain the basic concepts. These concepts will be elaborated in later chapters.

2.1 Limitations of Modeling

Process mining can be viewed as the missing link between *data science* and *process science*, as demonstrated in the previous chapter using Fig. 1.7. Another way to characterize process mining is shown in Fig. 2.1. The diagram shows that process mining starts from *event data* and uses *process models* in various ways, e.g., process models are discovered from event data, serve as reference models, or are used to project bottlenecks on. Figure 2.1 shows that event data and process models can be viewed as “yin and yang” in process mining. Like in Chinese philosophy, we aim for a duality (yin and yang). *Data-driven forces* and *process-centric forces* are viewed as complementary, interconnected, and interdependent in process mining. Figure 2.1 also provides examples of questions that can be answered using process mining. These questions can be grouped into *performance* and *conformance* related questions. Clearly, such questions cannot be answered using a spreadsheet program. We later show concrete examples. However, before introducing process mining using a concrete event log, we first discuss the limitations of modeling when it comes to process analysis and process improvement.

To discuss the limitations of modeling, in particular the use of hand-made models, we briefly introduce *Petri nets* as an example language. A plethora of notations exists to model operational (business) processes (next to Petri nets there are languages like BPMN, UML, and EPCs), some of which will be discussed in the next

Fig. 2.1 Process mining is both data-driven and process-centric: Using a combination of event data and process models a wide range of conformance and performance questions can be answered



chapter. We refer to all of these as *process models*. The notations mentioned have in common that processes are described in terms of activities (and possibly subprocesses). The ordering of these activities is modeled by describing causal dependencies. Moreover, the process model may also describe temporal properties, specify the creation and use of data, e.g., to model decisions, and stipulate the way that resources interact with the process (e.g., roles, allocation rules, and priorities).

Figure 2.2 shows a process model expressed in terms of a *Petri net* [46]. The model describes the handling of a request for compensation within an airline. Customers may request compensation for various reasons, e.g., a delayed or canceled flight. As Fig. 2.2 shows the process starts by registering the request. This activity is modeled by transition *register request*. Each *transition* is represented by a square. Transitions are connected through *places* that model possible states of the process. Each place is represented by a circle. In a Petri net a transition is *enabled*, i.e., the corresponding activity can occur, if all input places hold a *token*. Transition *register request* has only one input place (*start*) and this place initially contains a token to represent the request for compensation. Hence, the corresponding activity is enabled and can occur. This is also referred to as *firing*. When firing, the transition consumes one token from each of its input places and produces one token for each of its output places. Hence, the firing of transition *register request* results in the removal of the token from input place *start* and the production of two tokens: one for output place *c1* and one for output place *c2*. Tokens are shown as black dots. The configuration of tokens over places—in this case the state of the request—is referred to as *marking*. Figure 2.2 shows the initial marking consisting of one token in place *start*. The marking after firing transition *register request* has two tokens: one in place *c1* and one in place *c2*. After firing transition *register request*, three transitions are enabled. The token in place *c2* enables transition *check ticket*. This transition models an administrative check to see whether the customer is eligible to issue a request. For example, while executing *check ticket* it is verified whether the customer indeed has

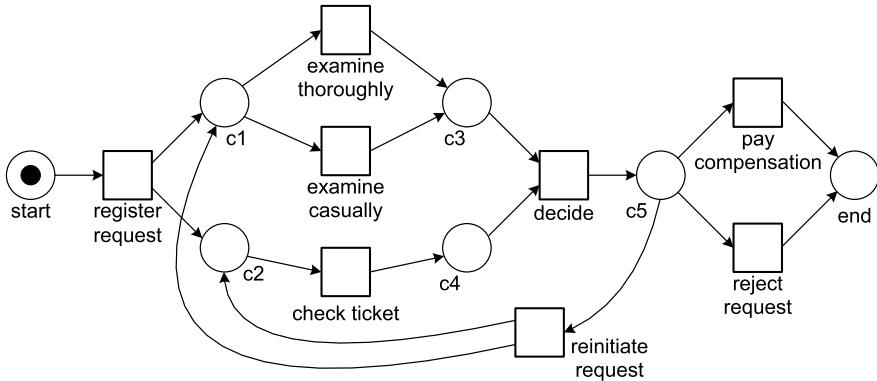


Fig. 2.2 A Petri net modeling the handling of compensation requests

a ticket issued by the airline. In parallel, the token in $c1$ enables both *examine thoroughly* and *examine casually*. Firing *examine thoroughly* will remove the token from $c1$, thus disabling *examine casually*. Similarly, the occurrence of *examine casually* will disable *examine thoroughly*. In other words, there is a choice between these two activities. Transition *examine thoroughly* is executed for requests that are suspicious or complex. Straightforward requests only need a casual examination. Firing *check ticket* does not disable any other transition, i.e., it can occur concurrently with *examine thoroughly* or *examine casually*. Transition *decide* is only enabled if both input places contain a token. The ticket needs to be checked (token in place $c4$) and the casual or thorough examination of the request has been conducted (token in place $c3$). Hence, the process synchronizes before making a decision. Transition *decide* consumes two tokens and produces one token for $c5$. Three transitions share $c5$ as an input place, thus modeling the three possible outcomes of the decision. The requested compensation is paid (transition *pay compensation* fires), the request is declined (transition *reject request* fires), or further processing is needed (transition *reinitiate request* fires). In the latter case the process returns to the state marking places $c1$ and $c2$: transition *reinitiate request* consumes a token from $c5$ and produces a token for each of its output places. This was the marking directly following the occurrence of *register request*. In principle, several iterations are possible. The process ends after paying the compensation or rejecting the request.

Process-Aware Information Systems

Process-Aware Information Systems (PAISs) include all software systems that support processes and not just isolated activities [49]. For example, ERP (Enterprise Resource Planning) systems (SAP, Oracle, etc.), BPM (Business Process Management) systems (Pegasystems, Bizagi, Appian, IBM BPM, etc.), WFM (Workflow Management) systems, CRM (Customer Relationship Management) systems, rule-based systems, call center software, high-end middle-

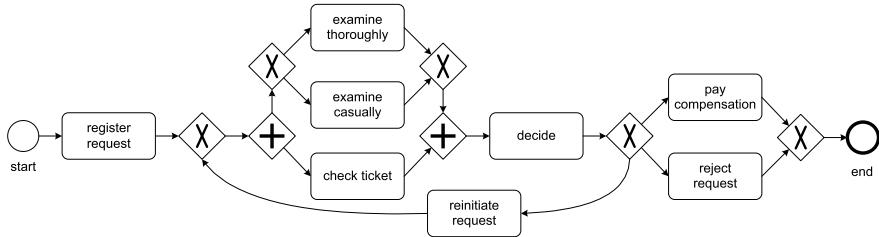


Fig. 2.3 The same process modeled in terms of BPMN

ware (WebSphere), etc. can be seen as process-aware. What these systems have in common is that there is a process notion present in the software (e.g., the completion of one activity triggers another activity) and that the information system is aware of the processes it supports (e.g., collecting information about flow times). This is very different from a database system, e-mail program, text editor, spreadsheet program, or currency transfer application. The latter set of example tools may be used to execute steps in some business process. However, these tools are not “aware” of the processes they are used in. Therefore, they cannot be actively involved in the management and orchestration of the processes they are used for.

A particular class of PAISs is formed by generic systems that are *driven by explicit process models*. Examples are BPM and WFM systems. WFM primarily focuses on the automation of business processes [76, 92, 151], whereas BPM has a broader scope: from process automation and process analysis to process management and the organization of work [50, 143, 187]. However, both BPM and WFM systems start from process models in a Petri-net or BPMN-like language. These models can be executed for any number of process instances. Changing the model corresponds (in theory) to automatically changing the process. This way flexibility and adaptability are supported.

Note that ERP systems are often hybrid. They provide a WFM subsystem, but also support processes driven by (configurable) code rather than process models. What all PAISs have in common is that they can be configured in some way (through an explicit process model, via predefined settings, or using customization).

Figure 2.2 models the process as a Petri net. There exist many different notations for process models. Figure 2.3 models the same process in terms of a so-called BPMN diagram [110, 187]. The *Business Process Modeling Notation* (BPMN) uses explicit *gateways* rather than places to model the control-flow logic. The diamonds with a “X” sign denote XOR split/join gateways, whereas diamonds with a “+” sign denote AND split/join gateways. The diamond directly following activity *register request* is an XOR-split gateway. This gateway is used to be able to “jump back”

after making the decision to reinitiate the request. After this XOR-join gateway there is an AND-split gateway to model that the checking of the ticket can be done in parallel with the selected examination type (thorough or casual). The remainder of the BPMN diagram is self explanatory as the behavior is identical to the Petri net described before.

Figures 2.2 and 2.3 show only the *control-flow*, i.e., the ordering of activities for the process described earlier. This is a rather limited view on business processes. Therefore, most modeling languages offer notations for modeling other perspectives such as the organizational or resource perspective (“The decision needs to be made by a manager”), the data perspective (“After four iteration always a decision is made unless more than 1 million Euro is claimed”), and the time perspective (“After two weeks the problem is escalated”). Although there are important differences between the various process modeling languages, we do not elaborate one these in this book. Instead, we refer to the systematic comparisons in the context of the *Workflow Patterns Initiative* [155, 191]. This allows us to focus on the role that process models play in process science, rather than worrying about notation. Although process mining can be used in a variety of applications domains, we often assume a BPM context for clarity. However, the techniques in this book can be used for *all* types of (discrete) events (e.g., in healthcare logistics, luggage handling systems, software analysis, smart maintenance, website analytics, and customer journey analysis).

What are process models used for?

- *insight*: while making a model, the modeler is triggered to view the process from various angles;
- *discussion*: the stakeholders use models to structure discussions;
- *documentation*: processes are documented for instructing people or certification purposes (cf. ISO 9000 quality management);
- *verification*: process models are analyzed to find errors in systems or procedures (e.g., potential deadlocks);
- *performance analysis*: techniques like simulation can be used to understand the factors influencing response times, service levels, etc.;
- *animation*: models enable end users to “play out” different scenarios and thus provide feedback to the designer;
- *specification*: models can be used to describe a PAIS before it is implemented and can hence serve as a “contract” between the developer and the end user/management; and
- *configuration*: models can be used to configure a system.

Clearly, process models play an important role in larger organizations. When redesigning processes and introducing new information systems, process models are used for a variety of reasons. Typically, two types of models are used: (a) *informal models* and (b) *formal models* (also referred to as “executable” models). Informal models are used for discussion and documentation whereas formal models

are used for analysis or enactment (i.e., the actual execution of process). On the one end of the spectrum there are “PowerPoint diagrams” showing high-level processes whereas on the other end of the spectrum there are process models captured in executable code. Whereas informal models are typically ambiguous and vague, formal models tend to have a rather narrow focus or are too detailed to be understandable by the stakeholders. The lack of alignment between both types of models has been discussed extensively in BPM literature [49, 70, 132, 137, 154, 187, 193]. Here, we would like to provide another view on the matter. Independent of the kind of model—informal or formal—one can reflect on the alignment between model and reality. A process model used to configure a workflow management system is probably well-aligned with reality as the model is used to force people to work in a particular way. Unfortunately, most hand-made models are disconnected from reality and provide only an idealized view on the processes at hand. Moreover, also formal models that allow for rigorous analysis techniques may have little to do with the actual process.

The value of models is limited if too little attention is paid to the alignment of model and reality. Process models become “paper tigers” when the people involved cannot trust them. For example, it makes no sense to conduct simulation experiments while using a model that assumes an idealized version of the real process. It is likely that—based on such an idealized model—incorrect redesign decisions are made. It is also precarious to start a new implementation project guided by process models that hide reality. A system implemented on the basis of idealized models is likely to be disruptive and unacceptable for end users. A nice illustration is the limited quality of most *reference models*. Reference models are used in the context of large enterprise systems such as SAP [37] but also to document processes for particular branches, cf. the NVVB (Nederlandse Vereniging Voor Burgerzaken) models describing the core processes in Dutch municipalities. The idea is that “best practices” are shared among different organizations. Unfortunately, the quality of such models leaves much to be desired. For example, the SAP reference model has very little to do with the processes actually supported by SAP. In fact, more than 20 percent of the SAP models contain serious flaws (deadlocks, livelocks, etc.) [101]. Such models are not aligned with reality and, thus, have little value for end users.

Given (a) the interest in process models, (b) the abundance of event data, and (c) the limited quality of hand-made models, it seems worthwhile to relate event data to process models. This way the actual processes can be discovered and existing process models can be evaluated and enhanced. This is precisely what process mining aims to achieve.

2.2 Process Mining

To position process mining, we first describe the so-called *BPM life-cycle* using Fig. 2.4. The life-cycle describes the different phases of managing a particular business process. In the *design* phase, a process is designed. This model is transformed

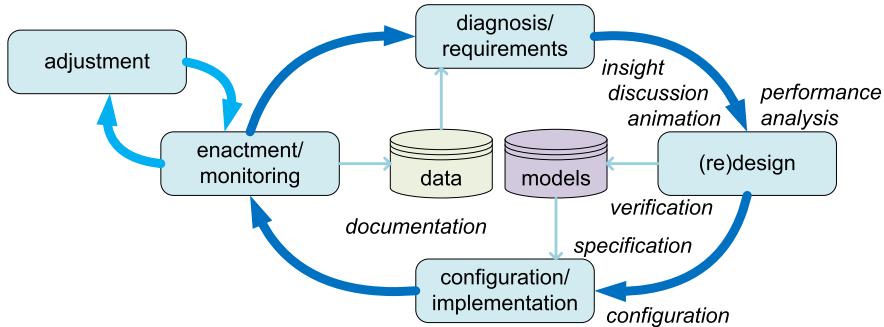


Fig. 2.4 The BPM life-cycle showing the different uses of process models

into a running system in the *configuration/implementation* phase. If the model is already in executable form and a WFM or BPM system is already running, this phase may be very short. However, if the model is informal and needs to be hardcoded in conventional software, this phase may take substantial time. After the system supports the designed processes, the *enactment/monitoring* phase starts. In this phase, the processes are running while being monitored by management to see if any changes are needed. Some of these changes are handled in the *adjustment* phase shown in Fig. 2.4. In this phase, the process is not redesigned and no new software is created; only predefined controls are used to adapt or reconfigure the process. The *diagnosis/requirements* phase evaluates the process and monitors emerging requirements due to changes in the environment of the process (e.g., changing policies, laws, competition). Poor performance (e.g., inability to meet service levels) or new demands imposed by the environment may trigger a new iteration of the BPM life-cycle starting with the *redesign* phase.

As Fig. 2.4 shows, process models play a dominant role in the (re)design and configuration/implementation phases, whereas data plays a dominant role in the enactment/monitoring and diagnosis/requirements phases. The figure also lists the different ways in which process models are used (as identified in Sect. 2.1). Until recently, there were few connections between the data produced while executing the process and the actual process design. In fact, in most organizations the diagnosis/requirements phase is not supported in a systematic and continuous manner. Only severe problems or major external changes will trigger another iteration of the life-cycle, and factual information about the current process is not actively used in redesign decisions. Process mining offers the possibility to truly “close” the BPM life-cycle. Data recorded by information systems can be used to provide a better view on the actual processes, i.e., deviations can be analyzed and the quality of models can be improved.

Process mining is a relative young research discipline that sits between machine learning and data mining on the one hand and process modeling and analysis on the other hand. The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today’s systems.

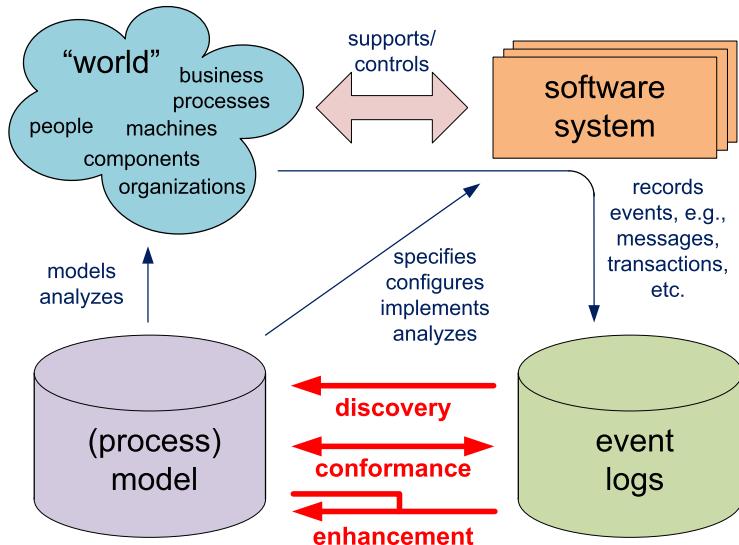


Fig. 2.5 Positioning of the three main types of process mining: *discovery*, *conformance*, and *enhancement*

Figure 2.5 shows that process mining establishes links between the actual processes and their data on the one hand and process models on the other hand. As explained in the previous chapter, the digital universe and the physical universe become more and more aligned. Today's information systems log enormous amounts of events. Classical WFM systems (e.g., Staffware and COSA), BPM systems (e.g., BPM|one by Pallas Athena, SmartBPM by Pegasystems, FileNet, Global 360, and Teamwork by Lombardi Software), ERP systems (e.g., SAP Business Suite, Oracle E-Business Suite, and Microsoft Dynamics NAV), PDM systems (e.g., Windchill), CRM systems (e.g., Microsoft Dynamics CRM and SalesForce), middleware (e.g., IBM's WebSphere and Cordys Business Operations Platform), and hospital information systems (e.g., Chipsoft and Siemens Soarian) provide detailed information about the activities that have been executed. Figure 2.5 refers to such data as *event logs*. All of the PAISs just mentioned directly provide such event logs. However, most information systems store such information in unstructured form, e.g., event data is scattered over many tables or needs to be tapped off from subsystems exchanging messages. In such cases, event data exist but some efforts are needed to extract them. Data extraction is an integral part of any process mining effort.

Let us assume that it is possible to *sequentially record events* such that each event refers to an *activity* (i.e., a well-defined step in the process) and is related to a particular *case* (i.e., a process instance). Consider, for example, the handling of requests for compensation modeled in Fig. 2.2. The cases are individual requests and per case a *trace* of events can be recorded. An example of a possible trace is *(register request, examine casually, check ticket, decide, reinitiate request, check ticket, examine thoroughly, decide, pay compensation)*. Here activity names are used

to identify events. However, there are two *decide* events that occurred at different times (the fourth and eighth event of the trace), produced different results, and may have been conducted by different people. Obviously, it is important to distinguish these two decisions. Therefore, most event logs store additional information about events. In fact, whenever possible, process mining techniques use extra information such as the *resource* (i.e., person or device) executing or initiating the activity, the *timestamp* of the event, or *data elements* recorded with the event (e.g., the size of an order).

Event logs can be used to conduct three types of process mining as shown in Fig. 2.5.

The first type of process mining is *discovery*. A discovery technique takes an event log and produces a model without using any a-priori information. An example is the α -algorithm [157] that will be described in Chap. 6. This algorithm takes an event log and produces a Petri net explaining the behavior recorded in the log. For example, given sufficient example executions of the process shown in Fig. 2.2, the α -algorithm is able to automatically construct the Petri net without using any additional knowledge. If the event log contains information about resources, one can also discover resource-related models, e.g., a social network showing how people work together in an organization.

The second type of process mining is *conformance*. Here, an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. For instance, there may be a process model indicating that purchase orders of more than one million Euro require two checks. Analysis of the event log will show whether this rule is followed or not. Another example is the checking of the so-called “four-eyes” principle stating that particular activities should not be executed by one and the same person. By scanning the event log using a model specifying these requirements, one can discover potential cases of fraud. Hence, conformance checking may be used to detect, locate and explain deviations, and to measure the severity of these deviations. An example is the conformance checking algorithm described in [121]. Given the model shown in Fig. 2.2 and a corresponding event log, this algorithm can quantify and diagnose deviations.

The third type of process mining is *enhancement*. Here, the idea is to extend or improve an existing process model using information about the actual process recorded in some event log. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model. One type of enhancement is *repair*, i.e., modifying the model to better reflect reality. For example, if two activities are modeled sequentially but in reality can happen in any order, then the model may be corrected to reflect this. Another type of enhancement is *extension*, i.e., adding a new perspective to the process model by cross-correlating it with the log. An example is the extension of a process model with performance data. For instance, by using timestamps in the event log of the “request for compensation” process, one can extend Fig. 2.2 to show bottlenecks, service levels, throughput times, and frequencies. Similarly, Fig. 2.2 can be extended with information about resources, decision rules, quality metrics, etc.

As indicated earlier, process models such as depicted in Figs. 2.2 and 2.3 show only the control-flow. However, when extending process models, additional perspectives are added. Moreover, discovery and conformance techniques are not limited to control-flow. For example, one can discover a social network and check the validity of some organizational model using an event log. Hence, orthogonal to the three types of mining (discovery, conformance, and enhancement), different perspectives can be identified.

In the remainder, we consider the following *perspectives*.

- The *control-flow perspective* focuses on the control-flow, i.e., the ordering of activities. The goal of mining this perspective is to find a good characterization of all possible paths, e.g., expressed in terms of a Petri net or some other notation (e.g., EPCs, BPMN, and UML ADs).
- The *organizational perspective* focuses on information about resources hidden in the log, i.e., which actors (e.g., people, systems, roles, and departments) are involved and how are they related. The goal is to either structure the organization by classifying people in terms of roles and organizational units or to show the social network.
- The *case perspective* focuses on properties of cases. Obviously, a case can be characterized by its path in the process or by the originators working on it. However, cases can also be characterized by the values of the corresponding data elements. For example, if a case represents a replenishment order, it may be interesting to know the supplier or the number of products ordered.
- The *time perspective* is concerned with the timing and frequency of events. When events bear timestamps it is possible to discover bottlenecks, measure service levels, monitor the utilization of resources, and predict the remaining processing time of running cases.

Note that the different perspectives are partially overlapping and non-exhaustive. Nevertheless, they provide a good characterization of the aspects that process mining aims to analyze.

In most examples given thus far it is assumed that process mining is done *off-line*, i.e., processes are analyzed afterward to see how they can be improved or better understood. However, more and more process mining techniques can also be used in an *online* setting. We refer to this as *operational support*. An example is the detection of non-conformance at the moment the deviation actually takes place. Another example is time prediction for running cases, i.e., given a partially executed case the remaining processing time is estimated based on historic information of similar cases. This illustrates that the “process mining spectrum” is broad and not limited to process discovery. In fact, today’s process mining techniques are indeed able to support the whole BPM life-cycle shown in Fig. 2.4. Process mining is not only relevant for the design and diagnosis/requirements phases, but also for the enactment/monitoring and adjustment phases.

2.3 Analyzing an Example Log

After providing an overview of process mining and positioning it in the broader BPM discipline, we use the event log shown in Table 2.1 to clarify some of the foundational concepts. The table shows just a fragment of a possible log corresponding to the handling of requests for compensation. Each line presents one event. Note that events are already grouped per case. Case 1 has five associated events. The first event of Case 1 is the execution of activity *register request* by Pete on December 30th 2010. Table 2.1 also shows a unique id for this event: 35654423. This is merely used for the identification of the event, e.g., to distinguish it from event 35654483 that also corresponds to the execution of activity *register request* (first event of second case). Table 2.1 shows a date and a timestamp for each event. In some event logs this information is more coarse-grained and only a date or partial ordering of events is given. In other logs there may be more elaborate timing information also showing when the activity was started, when it was completed, and sometimes even when it was offered to the resource. The times shown in Table 2.1 should be interpreted as completion times. In this particular event log, activities are considered to be atomic and the table does not reveal the duration of activities. In the table, each event is associated to a resource. In some event logs this information will be missing. In other logs more detailed information about resources may be stored, e.g., the role a resource has or elaborate authorization data. The table also shows the costs associated to events. This is an example of a data attribute. There may be many other data attributes. For example, in this particular example it would be interesting to record the outcome of the different types of examinations and checks. Another data element that could be useful for analysis is the amount of compensation requested. This could be an attribute of the whole case or stored as an attribute of the *register request* event.

Table 2.1 illustrates the typical information present in an event log. Depending on the process mining technique used and the questions at hand, only part of this information is used. The minimal requirements for process mining are that any event can be related to both a case and an activity and that events within a case are ordered. Hence, the “case id” and “activity” columns in Table 2.1 represent the bare minimum for process mining. By projecting the information in these two columns we obtain the more compact representation shown in Table 2.2. In this table, each case is represented by a sequence of activities also referred to as *trace*. For clarity the activity names have been transformed into single-letter labels, e.g., *a* denotes activity *register request*.

Process mining algorithms for process discovery can transform the information shown in Table 2.2 into process models. For instance, the basic α -algorithm [157] discovers the Petri net described earlier when providing it with the input data in Table 2.2. Figure 2.6 shows the resulting model with the compact labels just introduced. It is easy to check that all six traces in Table 2.2 are possible in the model. Let us replay the trace of the first case— $\langle a, b, d, e, h \rangle$ —to show that the trace “fits” (i.e., conforms to) the model. In the initial marking shown in Fig. 2.6, *a* is indeed enabled because of the token in *start*. After firing *a* places *c1* and *c2* are marked,

Table 2.1 A fragment of some event log: each line corresponds to an event

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	...	

Table 2.1 (Continued)

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
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	...
...

Table 2.2 A more compact representation of log shown in Table 2.1: $a = \text{register request}$, $b = \text{examine thoroughly}$, $c = \text{examine casually}$, $d = \text{check ticket}$, $e = \text{decide}$, $f = \text{reinitiate request}$, $g = \text{pay compensation}$, and $h = \text{reject request}$

Case id	Trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
...	...

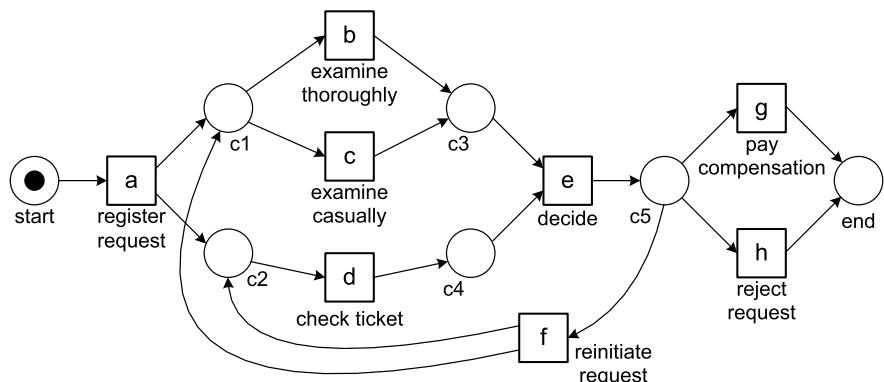


Fig. 2.6 The process model discovered by the α -algorithm [157] based on the set of traces $\{\langle a, b, d, e, h \rangle, \langle a, d, c, e, g \rangle, \langle a, c, d, e, f, b, d, e, g \rangle, \langle a, d, b, e, h \rangle, \langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle, \langle a, c, d, e, g \rangle\}$

i.e., both places contain a token. b is enabled at this marking and its execution results in the marking with tokens in c_2 and c_3 . Now we have executed $\langle a, b \rangle$ and the sequence $\langle d, e, h \rangle$ remains. The next event d is indeed enabled and its execution results in the marking enabling e (tokens in places c_3 and c_4). Firing e results in the marking with one token in c_5 . This marking enables the final event h in the trace. After executing h , the case ends in the desired final marking with just a token in

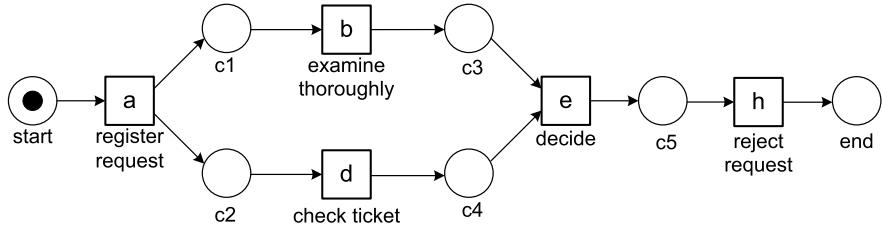


Fig. 2.7 The process model discovered by the α -algorithm based on cases 1 and 4, i.e., the set of traces $\{\langle a, b, d, e, h \rangle, \langle a, d, b, e, h \rangle\}$

place *end*. Similarly, it can be checked that the other five traces shown in Table 2.2 are also possible in the model and that all of these traces result in the marking with just a token in place *end*.

The Petri net shown in Fig. 2.6 also allows for traces not present in Table 2.2. For example, the traces $\langle a, d, c, e, f, b, d, e, g \rangle$ and $\langle a, c, d, e, f, c, d, e, f, c, d, e, f, c, d, e, f, b, d, e, g \rangle$ are also possible. This is a desired phenomenon as the goal is *not* to represent just the *particular set of example traces* in the event log. Process mining algorithms need to generalize the behavior contained in the log to show the most likely underlying model that is not invalidated by the next set of observations. One of the challenges of process mining is to balance between “overfitting” (the model is too specific and only allows for the “accidental behavior” observed) and “underfitting” (the model is too general and allows for behavior unrelated to the behavior observed).

When comparing the event log and the model, there seems to be a good balance between “overfitting” and “underfitting”. All cases start with *a* and end with either *g* or *h*. Every *e* is preceded by *d* and one of the examination activities (*b* or *c*). Moreover, *e* is followed by *f*, *g*, or *h*. The repeated execution of *b* or *c*, *d*, and *e* suggests the presence of a loop. These characteristics are adequately captured by the net of Fig. 2.6.

Let us now consider an event log consisting of only two traces $\langle a, b, d, e, h \rangle$ and $\langle a, d, b, e, h \rangle$, i.e., cases 1 and 4 of the original log. For this log, the α -algorithm constructs the Petri net shown in Fig. 2.7. This model only allows for two traces and these are exactly the ones in the small event log. *b* and *d* are modeled as being concurrent because they can be executed in any order. For larger and more complex models it is important to discover concurrency. Not modeling concurrency typically results in large “Spaghetti-like” models in which the same activity needs to be duplicated.¹

The α -algorithm is just one of many possible process discovery algorithms. For real-life logs more advanced algorithms are needed to better balance between “overfitting” and “underfitting” and to deal with “incompleteness” (i.e., logs containing

¹See, for example, Figs. 14.1 and 14.10 to understand why we use the term “Spaghetti” to refer to models that are difficult to comprehend.

Table 2.3 Another event log: cases 7, 8, and 10 are not possible according to Fig. 2.6

Case id	Trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
7	$\langle \mathbf{a}, \mathbf{b}, \mathbf{e}, \mathbf{g} \rangle$
8	$\langle \mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e} \rangle$
9	$\langle a, d, c, e, f, d, c, e, f, b, d, e, h \rangle$
10	$\langle \mathbf{a}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{f}, \mathbf{b}, \mathbf{d}, \mathbf{g} \rangle$

only a small fraction of the possible behavior due to the large number of alternatives) and “noise” (i.e., logs containing exceptional/infrequent behavior that should not automatically be incorporated in the model). This book will describe several of such algorithms and guide the reader in selecting one. In this section, we used Petri nets to represent the discovered process models, because Petri nets are a succinct way of representing processes and have unambiguous and simple semantics. However, most mining techniques are independent of the desired representation. For instance, the discovered Petri net model shown in Fig. 2.6 can be (automatically) transformed into the BPMN model shown in Fig. 2.3.

As explained in Sect. 2.2, process mining is not limited to process discovery. Event logs can be used to check conformance and enhance existing models. Moreover, different perspectives may be taken into account. To illustrate this, let us first consider the event log shown in Table 2.3. The first six cases are as before. It is easy to see that Case 7 with trace $\langle a, b, e, g \rangle$ is not possible according to the model in Fig. 2.6. The model requires the execution of d before e , but d did not occur. This means that the ticket was not checked at all before making a decision and paying compensation. Conformance checking techniques aim at discovering such discrepancies [121]. When checking the conformance of the remainder of the event log it can also be noted that cases 8 and 10 do not conform either. Case 9 conforms although it is not identical to one of the earlier traces. Trace $\langle a, b, d, e \rangle$ (i.e., Case 8) has the problem that no concluding action was taken (rejection or payment). Trace $\langle a, c, d, e, f, b, d, g \rangle$ (Case 10) has the problem that the airline paid compensation without making a final decision. Note that conformance can be viewed from two angles: (a) the model does not capture the real behavior (“the model is wrong”) and (b) reality deviates from the desired model (“the event log is wrong”). The first viewpoint is taken when the model is supposed to be *descriptive*, i.e., capture or predict reality. The second viewpoint is taken when the model is *normative*, i.e., used to influence or control reality.

The original event log shown in Table 2.1 also contains information about resources, timestamps and costs. Such information can be used to discover other perspectives, check the conformance of models that are not pure control-flow models, and to extend models with additional information. For example, one could derive

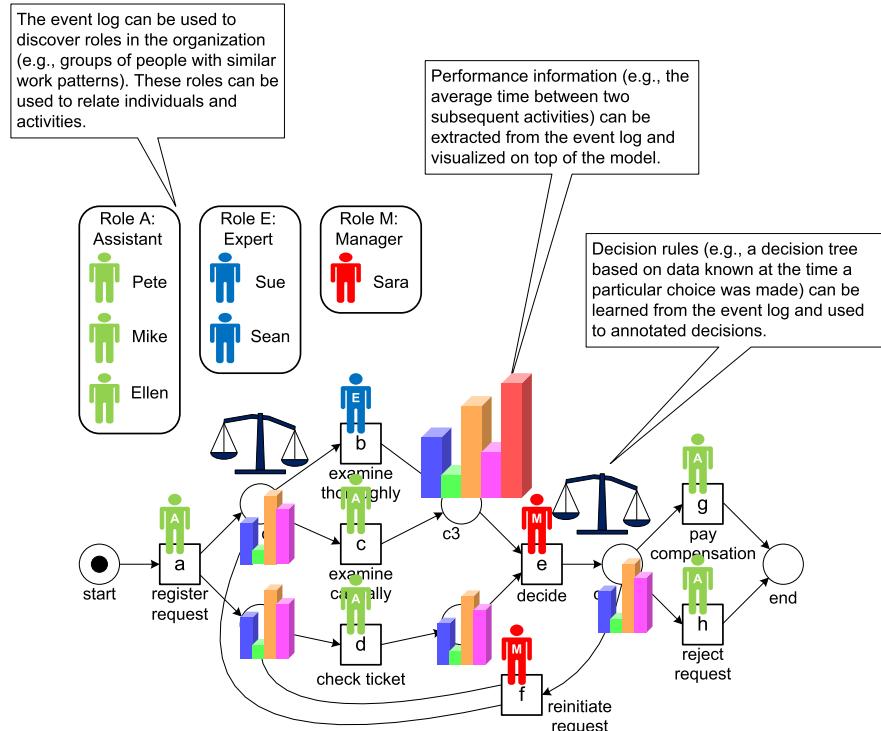


Fig. 2.8 The process model extended with additional perspectives: the organizational perspective (“What are the organizational roles and which resources are performing particular activities?”), the case perspective (“Which characteristics of a case influence a particular decision?”), and the time perspective (“Where are the bottlenecks in my process?”)

a social network based on the interaction patterns between individuals. The social network can be based on the “handover of work” metric, i.e., the more frequent individual x performed an activity that is causally followed by an activity performed by individual y , the stronger the relation between x and y is [159].

Figure 2.8 illustrates the way in which a control-flow oriented model can be extended with the three other main perspectives mentioned in Sect. 2.2. Analysis of the event log shown in Table 2.1 may reveal that Sara is the only one performing the activities *decide* and *reinitiate request*. This suggests that there is a “manager role” and that Sara is the only one having this role. Activity *examine thoroughly* is performed only by Sue and Sean. This suggests some “expert role” associated to this activity. The remaining activities are performed by Pete, Mike and Ellen. This suggests some “assistant role” as shown in Fig. 2.8. Techniques for organizational process mining [130] will discover such organizational structures and relate activities to resources through roles. By exploiting resource information in the log, the organizational perspective can be added to the process model. Similarly, information on timestamps and frequencies can be used to add performance related information

to the model. Figure 2.8 sketches that it is possible to measure the time that passes between an examination (activities b or c) and the actual decision (activity e). If this time is remarkably long, process mining can be used to identify the problem and discover possible causes. If the event log contains case-related information, this can be used to further analyze the decision points in the process. For instance, through decision point analysis it may be learned that requests for compensation of more than € 800 tend to be rejected.

Using process mining, the different perspectives can be cross-correlated to find surprising insights. Examples of such findings could be: “requests examined by Sean tend to be rejected more frequently”, “requests for which the ticket is checked after examination tend to take much longer”, “requests of less than € 500 tend to be completed without any additional iterations”. Moreover, these perspectives can also be linked to conformance questions. For example, it may be shown that Pete is involved in relatively many incorrectly handled requests. These examples show that privacy issues need to be considered when analyzing event logs with information about individuals (see Sect. 9.3.3).

2.4 Play-In, Play-Out, and Replay

One of the key elements of process mining is the emphasis on establishing a strong relation between a process model and “reality” captured in the form of an event log. Inspired by the terminology used by David Harel in the context of Live Sequence Charts [70], we use the terms *Play-In*, *Play-Out*, and *Replay* to reflect on this relation. Figure 2.9 illustrates these three notions.

Play-Out refers to the classical use of process models. Given a Petri net, it is possible to generate behavior. The traces in Table 2.2 could have been obtained by repeatedly “playing the token game” using the Petri net of Figure 2.6. Play-Out can be used both for the analysis and the enactment of business processes. A workflow engine can be seen as a “Play-Out engine” that controls cases by only allowing the “moves” allowed according to the model. Hence, Play-Out can be used to enact operational processes using some executable model. Simulation tools also use a Play-Out engine to conduct experiments. The main idea of simulation is to repeatedly run a model and thus collect statistics and confidence intervals. Note that a simulation engine is similar to a workflow engine. The main difference is that the simulation engine interacts with a modeled environment whereas the workflow engine interacts with the real environment (workers, customers, etc.). Also classical verification approaches using exhaustive state-space analysis—often referred to as model checking [30]—can be seen as Play-Out methods.

Play-In is the opposite of Play-Out, i.e., example behavior is taken as input and the goal is to construct a model. Play-In is often referred to as *inference*. The α -algorithm and other process discovery approaches are examples of Play-In techniques. Note that the Petri net of Fig. 2.6 can be derived automatically given an event log like the one in Table 2.2. Most data mining techniques use Play-In, i.e.,

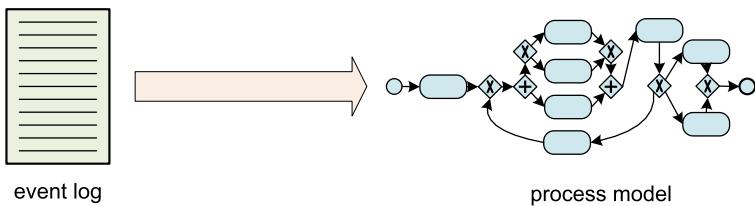
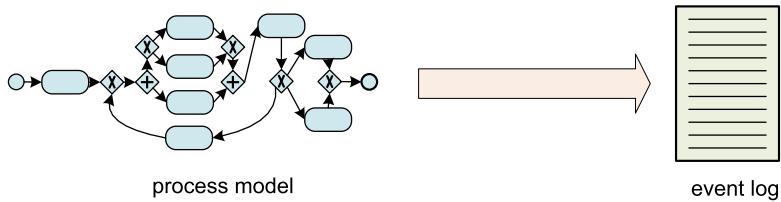
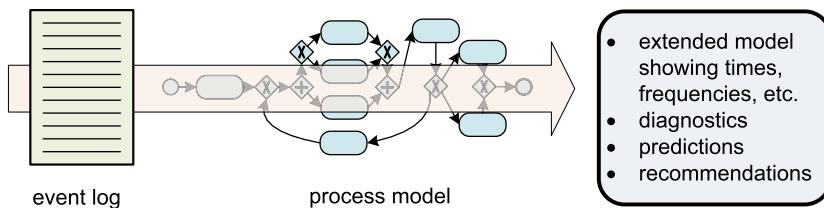
Play-In**Play-Out****Replay**

Fig. 2.9 Three ways of relating event logs (or other sources of information containing example behavior) and process models: *Play-In*, *Play-Out*, and *Replay*

a model is learned on the basis of examples. However, traditionally, data mining has not been concerned with process models. Typical examples of models are decision trees (“people that drink more than five glasses of alcohol and smoke more than 56 cigarettes tend to die young”) and association rules (“people that buy diapers also buy beer”). Unfortunately, it is not possible to use conventional data mining techniques to Play-In process models. Only recently, process mining techniques have become readily available to discover process models based on event logs.

Replay uses an event log *and* a process model as input. The event log is “replayed” on top of the process model. As shown earlier it is possible to replay trace $\langle a, b, d, e, h \rangle$ on the Petri net in Fig. 2.6; simply “play the token game” by forcing the transitions to fire (if possible) in the order indicated. An event log may be replayed for different purposes:

- *Conformance checking*: discrepancies between the log and the model can be detected and quantified by replaying the log. For instance, replaying trace

$\langle a, b, e, h \rangle$ on the Petri net in Fig. 2.6 will show that d should have happened but did not.

- *Extending the model with frequencies and temporal information.* By replaying the log one can see which parts of the model are visited frequently. Replay can also be used to detect bottlenecks. Consider, for example, the trace $\langle a^8, b^9, d^{20}, e^{21}, h^{21} \rangle$ in which the superscripts denote timestamps. By replaying the trace on top of Fig. 2.6 one can see that e was enabled at time 20 and occurred at time 21. The enabling of e was delayed by the time it took to complete d ; although d was enabled already at time 8, it occurred only at time 20.
- *Constructing predictive models.* By replaying event logs one can build predictive models, i.e., for the different states of the model particular predictions can be made. For example, a predictive model learned by replaying many cases could show that the expected time until completion after enabling e is eight hours.
- *Operational support.* Replay is not limited to historic event data. One can also replay partial traces of cases still running. This can be used for detecting deviations at run-time, e.g., the partial trace $\langle a^8, e^{11} \rangle$ of a case that is still running will never fit into Fig. 2.6. Hence, an alert can be generated before the case completes. Similarly, it is possible to predict the remaining processing time or the likelihood of being rejected of a case having a partial trace, e.g., a partial executed case $\langle a^8, b^9 \rangle$ has an expected remaining processing time of 3.5 days and a 40 percent probability of being rejected. Such predictions can also be used to recommend suitable next steps to progress the case.

Desire lines in process models

A desire line—also known as the social trail—is a path that emerges through erosion caused by footsteps of humans (or animals). The width and amount of erosion of the path indicates how frequently the path is used. Typically, the desire line follows the shortest or most convenient path between two points. Moreover, as the path emerges more people are encouraged to use it, thus stimulating further erosion. Dwight Eisenhower is often mentioned as one of the persons using this emerging group behavior. Before becoming the 34th president of the United States, he was the president of Columbia University. When he was asked how the university should arrange the sidewalks to best interconnect the campus buildings, he suggested letting the grass grow between buildings and delay the creation of sidewalks. After some time the desire lines revealed themselves. The places where the grass was most worn by people's footsteps were turned into sidewalks. In the same vein, replay can be used to show the *desire lines in processes*. The paths in the process model traveled most can be highlighted by using brighter colors or thicker arcs (cf. ProM's Fuzzy Miner [66]).

An interesting question is how desire lines can be used to better manage business processes. Operational support, e.g., predictions and recommendations derived from historic information, can be used to reinforce successful behavior and thus create suitable “sidewalks” in processes.

2.5 Positioning Process Mining

The process mining spectrum is quite broad and extends far beyond process discovery and conformance checking. Process mining also connects data science and process science (see Fig. 1.7). As a result, it is inevitable that process mining objectives are overlapping with those of other approaches, methodologies, principles, methods, tools, and paradigms. For example, some will argue that “process mining is part of data mining”, but discussions on such inclusion relations are seldom useful and are often politically motivated. Most data mining tools do *not* provide process mining capabilities, most data mining books do *not* describe process mining techniques, and it seems that process mining techniques like conformance checking do *not* fit in any of the common definitions of data mining. It is comparable to claiming that “data mining is part of statistics”. Taking the transitive closure of both statements, we would even be able to conclude that process mining is part of statistics. Obviously, this does not make any sense. Making definitions all-encompassing does not help to provide actual analysis capabilities. Nevertheless, it is important to position process mining in the context of existing technologies and management approaches.

2.5.1 How Process Mining Compares to BPM

Business Process Management (BPM) is the discipline that combines approaches for the design, execution, control, measurement and optimization of business processes. Process mining can be best related to BPM by looking at the so-called BPM life-cycle in Fig. 2.4. Initially, the main focus of BPM was on process design and implementation [143]. Process modeling plays a key role in the (re)design phase and directly contributes to the configuration/implementation phase. Originally, BPM approaches had a tendency to be model-driven without considering the “evidence” hidden in the data.

There is now a clear trend in the BPM community to focus more on the enactment/monitoring, adjustment, and diagnosis/requirements phases. These phases are more data-driven and process mining techniques are frequently used in this part of the BPM life-cycle. Hence, process mining can easily be positioned in Fig. 2.4. However, process mining is *not* limited to BPM. Any process for which events can be recorded, is a candidate for process mining.

Learning more about Business Process Management (BPM)

Although process mining is not limited to BPM, both are clearly related. Hence, the interested reader may want to read more on BPM. Developments in BPM have resulted in a well-established set of principles, methods and tools that combine knowledge from information technology, management sciences and industrial engineering for the purpose of improving business processes. BPM can be viewed as a continuation of the Workflow Management (WFM) wave in the 1990s. The survey paper [143] structures the BPM field using 20 *BPM Use Cases* and describes the development of the field since the late 1970s. For more details, we refer to the following BPM/WFM books that served as milestones in the evolution of the field:

- *Workflow Management: Modeling Concepts, Architecture, and Implementation* [76]: first comprehensive WFM book focusing on the different workflow perspectives and the MOBILE language,
- *Production Workflow: Concepts and Techniques* [92]: book on production WFM systems closely related to IBM's workflow products,
- *Business Process Management: Models, Techniques, and Empirical Studies* [152]: edited book that served as the basis for the BPM conference series,
- *Workflow Management: Models, Methods, and Systems* [151]: most cited WFM book using a Petri net-based approach to model, analyze and enact workflow processes,
- *Workflow-based Process Controlling: Foundation, Design and Application of workflow-driven Process Information Systems* [192]: book relating WFM systems to operational performance,
- *Process-Aware Information Systems: Bridging People and Software through Process Technology* [49]: edited book on process-aware information systems,
- *Business Process Management: The Third Wave* [127]: visionary book linking management perspectives to the π -calculus,
- *Business Process Management: Concepts, Languages, Architectures* [187]: book presenting the foundations of BPM, including different languages and architectures,
- *Modern Business Process Automation: YAWL and its Support Environment* [132]: book based on YAWL and the workflow patterns,
- *Handbooks on Business Process Management* [180, 181]: edited handbooks covering the broader BPM discipline,
- *Process Management: A Guide for the Design of Business Processes* [18]: book on the design of process-oriented organizations,
- *Enabling Flexibility in Process-Aware Information Systems: Challenges, Methods, Technologies* [115]: book on supporting flexibility in process-aware information systems, and
- *Fundamentals of Business Process Management* [50]: tutorial-style book covering the whole BPM life-cycle.

2.5.2 How Process Mining Compares to Data Mining

Data mining techniques aim to analyze (often large) data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [69]. Like process mining, data mining is *data-driven*. However, unlike process mining, mainstream data mining techniques are typically *not process-centric*. Process models expressed in terms of Petri nets or BPMN diagrams cannot be discovered or analyzed in any way by the main data mining tools.

There are a few data mining techniques that come close to process mining. Examples are sequence and episode mining. However, these techniques do not consider end-to-end processes. Through process mining, it becomes easier to apply data mining techniques to event data. For example, decision rules can be learned using standard data mining tools after the control-flow backbone (e.g., a Petri net) has been learned using a process mining tool. *RapidProM*, available through the RapidMiner Marketplace, shows that process mining and data mining can be combined in various ways. Chapter 4 discusses the relation in more detail.

2.5.3 How Process Mining Compares to Lean Six Sigma

Lean Six Sigma is a methodology that combines ideas from *lean manufacturing* and *Six Sigma*. The idea is to improve performance by systematically removing waste. Lean principles originate from the Japanese manufacturing industry. The *Toyota Production System* (TPS) is a well-known example of a lean manufacturing approach developed by Taiichi Ohno and Eiji Toyoda between 1948 and 1975. The main objectives of the TPS are to eliminate “muri” (overburdening of people and equipment), “mura” (unevenness in operations), and “muda” (waste). The emphasis is on waste (“muda”) reduction. Typically, seven types of waste are mentioned in this context [109]:

- *Transportation waste*: Each time a product is moved, it encounters the risk of being damaged, lost, delayed, etc. Transportation does not make any transformation to the product that the consumer is willing to pay for (except for the final delivery).
- *Inventory waste*: Inventory may exist in the form of raw materials, work-in-progress, or finished goods. Inventory that is not being actively processed can be considered as waste because it consumes capital and space.
- *Motion waste*: Resources (equipment and people) that are used in the production processes suffer from “wear and tear”. Unnecessary activities (e.g., transformation and double work) result in additional degradation of resources and increase the risk of incidents (e.g., accidents).
- *Unnecessary waiting*: Whenever goods are not in transport or being processed, they are waiting. In traditional processes, a large part of an individual product’s life is spent waiting to be worked on. The total flow time of a case is often orders of magnitude larger than the sum of all service times.

- *Over-processing waste*: All additional efforts done for a product not directly required by the customer are considered as waste. This includes using components that are more precise, complex, of higher quality, and thus more expensive than absolutely required.
- *Overproduction waste*: Producing more than what is required by the customers at a particular time is a potential form of waste. Overproduction may lead to excess inventory and the customer's preferences may change over time making products outdated or less valuable.
- *Defects*: Rework, scrap, missing parts, poor work instructions, and correction activities are defects that can increase the costs of a product drastically.

Various additional types of waste have been identified. Although the terminology is oriented towards production processes and physical products, the same principles can be used for information/financial services, administrative work, and other BPM-like processes. The above examples illustrate that the focus of lean manufacturing is on *eliminating all non-value added activities*. Six Sigma focuses on *improving the quality of value added activities*. Both complement each other and are combined in Lean Six Sigma.

What does “Six Sigma” mean?

Today the term “Six Sigma” refers to a broad set of tools, techniques and methods to improve the quality of processes [113]. Six Sigma was originally developed by Motorola in the early 1980s and extended by many others. The term “Six Sigma” refers to the initial goal set by Motorola to minimize defects. In fact, the σ in “Six Sigma” refers to the standard deviation of a normal distribution. Given a normal distribution, 68.3% of the values lie within 1 standard deviation of the mean, i.e., a random draw from normal distribution with a mean value of μ and a standard deviation of σ has a probability of 0.683 to be in the interval $[\mu - \sigma, \mu + \sigma]$. Given the same normal distribution, 95.45% of randomly sampled values lie within two standard deviations of the mean, i.e., $[\mu - 2\sigma, \mu + 2\sigma]$, and 99.73% of the values lie within three standard deviations of the mean, i.e., $[\mu - 3\sigma, \mu + 3\sigma]$. The traditional quality paradigm in manufacturing defines a process as “capable” if the process’s natural spread, plus and minus three σ , was less than the engineering tolerance. So, if deviations of up to three times the standard deviations are allowed, then on average 2700 out of one million cases will have a defect (i.e., samples outside the $[\mu - 3\sigma, \mu + 3\sigma]$ interval). Six Sigma aims to create processes where the standard deviation is so small that any value within 6 standard deviations of the mean can be considered as non-defective. In the literature, often a 1.5 sigma shift (to accommodate for long term variations and decreasing quality) is taken into account [113]. This results in the following table:

Quality level	Defective Parts per Million Opportunities (DPMO)	Percentage passed
One Sigma	690,000 DPMO	31%
Two Sigma	308,000 DPMO	69.2%
Three Sigma	66,800 DPMO	93.32%
Four Sigma	6,210 DPMO	99.379%
Five Sigma	230 DPMO	99.977%
Six Sigma	3.4 DPMO	99.9997%

A process that “runs at One Sigma” has less than 690,000 defective cases per million cases, i.e., at least 31% of the cases are handled properly. A process that “runs at Six Sigma” has only 3.4 defective cases per million cases, i.e., on average 99.9997% of the cases are handled properly.

A typical Lean Six Sigma project follows the so-called *DMAIC* approach consisting of five steps:

- *Define* the problem and set targets,
- *Measure* key performance indicators and collect data,
- *Analyze* the data to investigate and verify cause-and-effect relationships,
- *Improve* the current process based on this analysis, and
- *Control* the process to minimize deviations from the target.

Numerous organizations heavily invested in (Lean) Six Sigma training over the past decade. Based on Karate-like skill levels (green belt, black belt, etc.), certification programs were implemented. Unfortunately, the actual techniques are typically very basic (from a data science point of view). As a result, many consider Lean Six Sigma training as a management fad. Fortunately, process mining can be used as a tool to add more substance to the methodology. For example, process discovery can be used to eliminate all non-value added activities and reduce waste. If the relevant events are being recorded, we can visualize unnecessary waiting and rework. Conformance checking can also improve the quality of value added activities. Deviations can be found and diagnosed easily, provided that the event data and normative process models are present.

Related to Lean Six Sigma are management approaches such as: *Continuous Process Improvement* (CPI), *Total Quality Management* (TQM), *5S* (workplace organization method characterized by the terms Sort, Straighten, Shine, Standardize, and Sustain), *Kaizen* (another continuous improvement method), and *Theory of Constraints* (management paradigm by Eliyahu Goldratt based in the idea that “a chain is no stronger than its weakest link”, thus focusing on the constraints limiting performance). What these approaches have in common is that processes are “put under a microscope” to see whether further improvements are possible. Clearly, process mining can help to analyze deviations and inefficiencies.

2.5.4 How Process Mining Compares to BPR

Business Process Reengineering (BPR) is a management approach developed by people like Michael Hammer [68]. BPR is characterized by four key words: *fundamental*, *radical*, *dramatic* and *process* [151]. The keyword *fundamental* indicates that, when revitalizing a business process, it is of great importance always to ask the basic question: Why are we doing this, and why are we doing it in this way? *Radical* means that the reengineered process must represent a complete break from the current way of working. BPR does not advocate to gradually improve existing processes: It aims at finding by completely new ones. The third keyword also refers to the fact that BPR does not aim at marginal or superficial changes. Changes must be *dramatic* in terms of costs, service and quality. In order to achieve dramatic improvements, it is necessary to focus on the *processes* and not start from data or systems.

BPR is *process-centric*, i.e., the focus is on the process just like in process mining. However, BPR is *not data-driven*. It promotes “thinking outside the box” rather than analyzing data in great detail. Process mining helps to identify the problems and assists shareholders in defining improvement actions. However, process mining cannot come up with completely different ways of working (unless event data is enriched with domain knowledge).

2.5.5 How Process Mining Compares to Business Intelligence

Process mining can be positioned under the umbrella of *Business Intelligence* (BI). There is no clear definition for BI. On the one hand, it is a very broad term that includes anything that aims at providing actionable information that can be used to support decision making. On the other hand, vendors and consultants tend to conveniently skew the definition towards a particular tool or methodology. Process mining provides innovations highly relevant for the next generation of BI techniques. However, it is important to note that current BI tools are not really “intelligent” and do not provide any process mining capabilities. The focus is on querying and reporting combined with simple visualization techniques showing dashboards and scorecards. Some systems provide data mining capabilities or support *Online Analytical Processing* (OLAP). OLAP tools are used to view multidimensional data from different angles. On the one hand, it is possible to aggregate and consolidate data to create high-level reports. On the other hand, OLAP tools can drill down into the data to find detailed information. There are approaches that combine process mining with OLAP to create and analyze so-called *process cubes* filled with event data (see Sect. 12.4).

Under the BI umbrella, many fancy terms have been introduced to refer to rather simple reporting and dashboard tools. *Business Activity Monitoring* (BAM) refers to the real-time monitoring of business processes. *Corporate Performance Management* (CPM) is another buzzword for measuring the performance of a process or

organization. Typically, CPM focuses on financial aspects. Recently, more and more software vendors started to use the term “analytics” to refer to advanced BI capabilities. *Visual analytics* focuses on the analysis of large amounts of data while exploiting the remarkable capabilities of humans to visually identify patterns and trends. *Predictive analytics* uses historic data to make forecasts. Clearly, process mining also aims at providing advanced analytics and some process mining techniques also heavily rely on advanced visualization and human interpretation. Moreover, as will be demonstrated in Chapt. 10, process mining is not restricted to analyzing historic data and also includes operational support, i.e., providing predictions and recommendations in an online setting.

2.5.6 How Process Mining Compares to CEP

Process mining complements *Complex Event Processing* (CEP). CEP combines data from multiple sources to infer events or patterns that suggest higher-level events. The goal of CEP is to identify meaningful events (such as opportunities or threats) and respond to them as quickly as possible, e.g., immediately generate an alert when a combination of events occurs. CEP can be used as a preprocessing step for process mining, i.e., low level event data with many (seemingly) meaningless events can be converted into higher-level event streams used by process mining techniques (online or offline). CEP is particularly useful if there are many (low-level) events. By reducing torrents of event data to manageable streams or logs, analysis becomes easier.

2.5.7 How Process Mining Compares to GRC

Whereas management approaches such as Lean Six Sigma and BPR mainly aim at improving operational performance, e.g., reducing flow time and defects, organizations are also putting increased emphasis on *corporate governance, risk, and compliance*. The frequently used acronym *GRC* is composed of the pillars *Governance*, *Risk management* and *Compliance*, and refers to an organization’s capability to reliably achieve its objectives while addressing uncertainty and acting with integrity. *Governance* is the combination of culture, policies, processes, laws, and institutions that define the structure by which the organization is directed and managed. *Risk management* is the process of identifying, assessing, and prioritizing risks, as well as creating a plan for minimizing or eliminating the impact of negative events. *Compliance* is the act of adhering to, and demonstrating adherence to, external laws and regulations as well as corporate policies and procedures.

Major corporate and accounting scandals including those affecting Enron, Tyco, Adelphia, Peregrine, and WorldCom have fueled interest in more rigorous auditing practices. Legislation such as the *Sarbanes–Oxley Act* (SOX) of 2002 and the different *Basel Accords* was enacted in response to such scandals. The financial crisis

a few years ago also underscores the importance of verifying that organizations operate “within their boundaries”. Process mining techniques offer a means to a more rigorous compliance checking and ascertaining the validity and reliability of information about an organization’s core processes. Since conformance checking can be used to reveal deviations, defects, and near incidents, it is a valuable tool to check compliance and manage risks.

2.5.8 How Process Mining Compares to ABPD, BPI, WM, ...

As described in the *Process Mining Manifesto* [75] by the IEEE Task Force on Process Mining, there are several alternative terms for process mining. Sometimes these terms are synonyms and at other times they refer to particular process mining tasks.

Automated Business Process Discovery (ABPD) is an example of such a term. ABPD was introduced by Gartner introduced in 2008 [84]. Often ABPD is used to refer to just process discovery (i.e., discovering process models from event data). This does not include bottleneck analysis, conformance checking, prediction, social network analysis, etc. Sometimes ABPD is used as a synonym for process mining. However, Fig. 2.5 clearly shows that process mining includes, for example, conformance checking. Moreover, later other forms of process mining will be described (e.g., prediction). Vendors that claim to support ABPD typically only uphold a fraction of the process mining spectrum covered in this book.

The term *Business Process Intelligence* (BPI) is used in different ways. It is often used as a synonym for process mining (perhaps with less emphasis on process models). See, for example, the annual International Workshop on Business Process Intelligence that has been running since 2005. Almost all papers presented at these BPI workshops use or propose process mining techniques.

BPI can also be understood as BI with a focus on analyzing operational processes. Some of the products positioned as BPI tools do not support discovery, i.e., performance data are mapped onto hand-made models. These tools assume a stable and known process model. Terms comparable to BPI are used by a range of vendors. For example, IBM’s Business Process Manager refers to the latter BPI-like functionality (i.e., without process discovery) as *Business Process Analytics* (BPA).

The term *Workflow Mining* (WM) was a precursor for process mining. It dates from a time where the main aim of process mining was the automatic configuration of a WFM system. Ideally, one would like to observe an existing process and automatically generate the corresponding executable workflow model. This view turned out to be too narrow and often unrealistic. Process mining has a much wider applicability, also in areas unrelated to WFM/BPM systems. Moreover, through discovery one can indeed find a skeleton of the workflow model. However, the model needs to be enriched with technical details to obtain an executable workflow. This makes the automated generation of workflow models less practical. Today, process mining is predominantly an approach for performance and conformance analysis (see Fig. 2.1). Therefore, the term WM is no longer actively used.

2.5.9 How Process Mining Compares to Big Data

In Chap. 1, we listed the “four V’s of Big Data”: Volume, Velocity, Variety, and Veracity (Fig. 1.4). These reflect the typical characteristics of some of the exciting new data sources interesting for analysis. Big Data does not focus on a particular type of analysis and is not limited to process-related data. However, Big Data infrastructures enable us to collect, store, and process huge event logs. Process mining tools can exploit such infrastructures to distribute large analysis tasks over multiple computing resources. For example, the MapReduce programming model can be used for discovery algorithms and the Hadoop Distributed File System (HDFS) can be used to store event data in a distributed fashion. In principle, one can use thousands of compute nodes to perform process mining analyses. Chapter 12 will elaborate on this.

The many acronyms in this section—BPM, BI, OLAP, TPS, BAM, CEP, CPM, CPI, TQM, SOX, etc.—are just a subset of the jargon used by business consultants and vendors. Some are just variations on the same theme, others emphasize a particular aspect. What can be distilled from the above is that there is a clear trend towards actually using the data available in today’s systems. The data is used to *reason about the process* and for *decision making within the process*. Moreover, the acronyms express a clear desire to get more *insight* into the actual processes, to *improve* them, and to make sure that they are *compliant*. Unfortunately, buzzwords are often used when the actual analysis capabilities are weak. When listening to a product presentation of conference talk, one is often tempted to play “buzzword bingo” (also known as bullshit bingo) illustrating that the foundational issues are not addressed. This book aims to provide a clear and refreshing view on the matter. Using recent breakthroughs in process mining, we will show that it is possible to simplify and unify the analysis of business processes based on facts. Moreover, the techniques and insights presented are directly applicable and are supported by process mining tools such as *ProM* (www.processmining.org).

Part II

Preliminaries

Part I: Introduction**Chapter 1**
Data Science in Action**Chapter 2**
Process Mining:
The Missing Link

Part II: Preliminaries**Chapter 3**
Process Modeling
and Analysis**Chapter 4**
Data Mining

Part III: From Event Logs to Process Models**Chapter 5**
Getting the Data**Chapter 6**
Process Discovery:
An Introduction**Chapter 7**
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery**Chapter 8**
Conformance
Checking**Chapter 9**
Mining Additional
Perspectives**Chapter 10**
Operational Support

Part V: Putting Process Mining to Work**Chapter 11**
Process Mining
Software**Chapter 12**
Process Mining in the
Large**Chapter 13**
Analyzing “Lasagna
Processes”**Chapter 14**
Analyzing “Spaghetti
Processes”

Part VI: Reflection**Chapter 15**
Cartography and
Navigation**Chapter 16**
Epilogue

Process mining provides a bridge between data mining and process modeling and analysis. Therefore, we provide an introduction to both fields. Chapter 3 reviews various process modeling notations and their analysis. Chapter 4 explains the main data mining techniques.

Chapter 3

Process Modeling and Analysis

The plethora of process modeling notations available today illustrates the relevance of process modeling. Some organizations may use only informal process models to structure discussions and to document procedures. However, organizations that operate at a higher BPM maturity level use models that can be analyzed and used to enact operational processes. Today, most process models are made by hand and are not based on a rigorous analysis of existing process data. This chapter serves two purposes. On the one hand, preliminaries are presented that will be used in later chapters. For example, various process modeling notations are introduced and some analysis techniques are reviewed. On the other hand, the chapter reveals the limitations of classical approaches, thus motivating the need for process mining.

3.1 The Art of Modeling

In Sect. 1.3, we introduced the umbrella term “*process science*” to refer to the broader discipline that combines knowledge from information technology and knowledge from management sciences to improve and run operational processes. Many of the (sub)disciplines mentioned in Fig. 1.6 heavily rely on *modeling* using a variety of formalisms and notations. In this book, we will use transition systems, Petri nets, BPMN, C-nets, EPCs, YAWL, and process trees as example representations. Before providing a “crash course” in these process representations, we briefly reflect on the role of models and the limitations of modeling in process science.

Since the industrial revolution, productivity has been increasing because of technical innovations, improvements in the organization of work, and the use of information technology. Adam Smith (1723–1790) showed the advantages of the division of labor. Frederick Taylor (1856–1915) introduced the initial principles of scientific management. Henry Ford (1863–1947) introduced the production line for the mass production of “black T-Fords”. Around 1950 computers and digital communication infrastructures started to influence business processes. This resulted in dramatic changes in the organization of work and enabled new ways of doing business. Today, innovations in computing and communication are still the main drivers behind

change in business processes. So, business processes have become more complex, heavily rely on information systems, and may span multiple organizations. Therefore, process modeling has become of the utmost importance. Process models assist in managing complexity by providing insight and documenting procedures. Information systems need to be configured and driven by precise instructions. Cross-organizational processes can only function properly if there is a common agreement on the required interactions. As a result, process models are widely used in today's organizations.

Operations management, and in particular *operation research*, is a branch of management science heavily relying on modeling. Here a variety of mathematical models ranging from linear programming and project planning to queueing models, Markov chains, and simulation are used. For example, the location of a warehouse is determined using linear programming, server capacity is added on the basis of queueing models, and an optimal route in a container terminal is determined using integer programming. Models are used to reason *about processes* (redesign) and to make decisions *inside processes* (planning and control). The models used in operations management are typically tailored towards a particular analysis technique and only used for answering a specific question. In contrast, process models in BPM typically serve *multiple* purposes. A process model expressed in BPMN may be used to discuss responsibilities, analyze compliance, predict performance using simulation, and configure a WFM system. However, BPM and operations management have in common that making a good model is “an art rather than a science”. Creating models is therefore a difficult and error-prone task. Typical errors include:

- *The model describes an idealized version of reality.* When modeling processes the designer tends to concentrate on the “normal” or “desirable” behavior. For example, the model may only cover 80% of the cases assuming that these are representative. Typically this is not the case as the other 20% may cause 80% of the problems. The reasons for such oversimplifications are manifold. The designer and management may not be aware of the many deviations that take place. Moreover, the perception of people may be biased, depending on their role in the organization. Hand-made models tend to be subjective, and often there is a tendency to make things too simple just for the sake of understandability.
- *Inability to adequately capture human behavior.* Although simple mathematical models may suffice to model machines or people working in an assembly line, they are inadequate when modeling people involved in multiple processes and exposed to multiple priorities [139, 163]. A worker who is involved in multiple processes needs to distribute his attention over multiple processes. This makes it difficult to model one process in isolation. Workers also do not work at constant speed. A well-known illustration of this is the so-called *Yerkes–Dodson law* that describes the relation between workload and performance of people [139]. In most processes one can easily observe that people will take more time to complete a task and effectively work fewer hours per day if there is hardly any work to do. Nevertheless, most simulation models sample service times from a fixed probability distribution and use fixed time windows for resource availability.

- *The model is at the wrong abstraction level.* Depending on the input data and the questions that need to be answered, a suitable abstraction level needs to be chosen. The model may be too abstract and thus unable to answer relevant questions. The model may also be too detailed, e.g., the required input cannot be obtained or the model becomes too complex to be fully understood. Consider, for example, a car manufacturer that has a warehouse containing thousands of spare parts. It may be tempting to model all of them in a simulation study to compare different inventory policies. However, if one is not aiming at making statements about a specific spare part, this is not wise. Typically it is very time consuming to change the abstraction level of an existing model. Unfortunately, questions may emerge at different levels of granularity.

These are just some of the problems organizations face when making models by hand. Only experienced designers and analysts can make models that have a good predictive value and can be used as a starting point for a (re)implementation or redesign. An inadequate model can lead to wrong conclusions. Therefore, we advocate the use of event data. Process mining allows for the extraction of models based on *facts*. Moreover, process mining does not aim at creating a single model of the process. Instead, it provides *various views on the same reality at different abstraction levels*. For example, users can decide to look at the most frequent behavior to get a simple model (“80% model”). However, they can also inspect the full behavior by deriving the “100% model” covering all cases observed. Similarly, abstraction levels can be varied to create different views. Process mining can also reveal that people in organizations do not function as “machines”. On the one hand, it may be shown that all kinds of inefficiencies take place. On the other hand, process mining can also visualize the remarkable flexibility of some workers to deal with problems and varying workloads.

3.2 Process Models

It is not easy to make good process models. Yet, they are important. Fortunately, process mining can facilitate the construction of better models in less time. Process discovery algorithms like the α -algorithm can automatically generate a process model. As indicated in Chap. 2, various process modeling notations exist. Sometimes the plethora of notations is referred to as the new “tower of Babel”. Therefore, we describe only some basic notations. This section does not aim to provide a complete overview of existing process modeling notations. We just introduce the notations that we will use in the remainder. We would like to stress that it is relatively easy to automatically translate process mining results into the desired notation. For example, although the α -algorithm produces a Petri net, it is easy to convert the result into a BPMN model, BPEL model, or UML Activity Diagram. Again we refer to the systematic comparisons in the context of the Workflow Patterns Initiative [155, 191] for details.

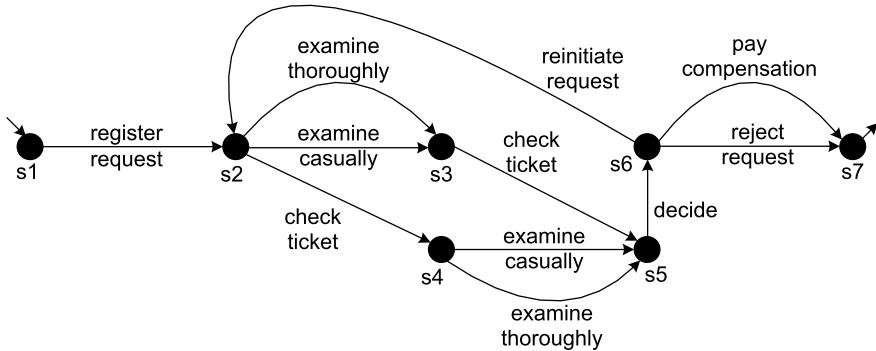


Fig. 3.1 A transition system having one initial state and one final state

In this section we focus on the control-flow perspective of processes. We assume that there is a set of *activity labels* \mathcal{A} . The goal of a process model is to decide which *activities* need to be executed and in *what order*. Activities can be executed sequentially, activities can be optional or concurrent, and the repeated execution of the same activity may be possible.

3.2.1 Transition Systems

The most basic process modeling notation is a *transition system*. A transition system consists of *states* and *transitions*. Figure 3.1 shows a transition system consisting of seven states. It models the handling of a request for compensation within an airline as described in Sect. 2.1. The states are represented by black circles. There is one initial state labeled s_1 and one final state labeled s_7 . Each state has a unique label. This label is merely an identifier and has no meaning. Transitions are represented by arcs. Each transition connects two states and is labeled with the name of an activity. Multiple arcs can bear the same label. For example, *check ticket* appears twice.

Definition 3.1 (Transition system) A *transition system* is a triplet $TS = (S, A, T)$ where S is the set of *states*, $A \subseteq \mathcal{A}$ is the set of *activities* (often referred to as *actions*), and $T \subseteq S \times A \times S$ is the set of *transitions*. $S^{start} \subseteq S$ is the set of *initial states* (sometimes referred to as “start” states), and $S^{end} \subseteq S$ is the set of *final states* (sometimes referred to as “accept” states).

The sets S^{start} and S^{end} are defined implicitly. In principle, S can be infinite. However, for most practical applications the state space is finite. In this case the transition system is also referred to as a Finite-State Machine (FSM) or a finite-state automaton.

The transition system depicted in Fig. 3.1 can be formalized as follows: $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$, $S^{start} = \{s_1\}$, $S^{end} = \{s_7\}$, $A = \{\text{register request}, \text{examine thoroughly}, \text{examine casually}, \text{check ticket}, \text{decide}, \text{reinvoke request}, \text{reject}\}$

request, pay compensation}, and $T = \{(s1, \text{register request}, s2), (s2, \text{examine casually}, s3), (s2, \text{examine thoroughly}, s3), (s2, \text{check ticket}, s4), (s3, \text{check ticket}, s5), (s4, \text{examine casually}, s5), (s4, \text{examine thoroughly}, s5), (s5, \text{decide}, s6), (s6, \text{reinitiate request}, s2), (s6, \text{pay compensation}, s7), (s6, \text{reject request}, s7)\}$.

Given a transition system one can reason about its behavior. The transition starts in one of the initial states. Any path in the graph starting in such a state corresponds to a possible *execution sequence*. For example, the path *register request, examine casually, check ticket* in Fig. 3.1 is an example of an execution sequence starting in state $s1$ and ending in $s5$. There are infinitely many execution sequences for this transition system. A path *terminates successfully* if it ends in one of the final states. A path *deadlocks* if it reaches a non-final state without any outgoing transitions. Note that the absence of deadlocks does not guarantee successful termination. The transition system may *livelock*, i.e., some transitions are still enabled but it is impossible to reach one of the final states.

Any process model with executable semantics can be mapped onto a transition system. Therefore, many notions defined for transition systems can easily be translated to higher-level languages such as Petri nets, BPMN, and UML activity diagrams. Consider, for example, the seemingly simple question: “When are two processes the same from a behavioral point of view”. As shown in [176], many equivalence notions can be defined. *Trace equivalence* considers two transition systems to be equivalent if their execution sequences are the same. More refined notions like *branching bisimilarity* also take the moment of choice into account. These notions defined for transition systems can be used for any pair of process models as long as the models are expressed in a language with executable semantics (see also Sect. 6.3).

Transition systems are simple but have problems expressing concurrency succinctly. Suppose that there are n parallel activities, i.e., all n activities need to be executed but any order is allowed. There are $n!$ possible execution sequences. The transition system requires 2^n states and $n \times 2^{n-1}$ transitions. This is an example of the well-known “state explosion” problem [135]. Consider for example 10 parallel activities. The number of possible execution sequences is $10! = 3,628,800$, the number of reachable states is $2^{10} = 1024$, and the number of transitions is $10 \times 2^{10-1} = 5120$. The corresponding Petri net is much more compact and needs only 10 transitions and 10 places to model the 10 parallel activities. Given the concurrent nature of business processes, more expressive models like Petri nets are needed to adequately represent process mining results.

3.2.2 Petri Nets

Petri nets are the oldest and best investigated process modeling language allowing for the modeling of concurrency. Although the graphical notation is intuitive and simple, Petri nets are executable and many analysis techniques can be used to analyze them [82, 117, 149]. In the introduction we already showed an example Petri

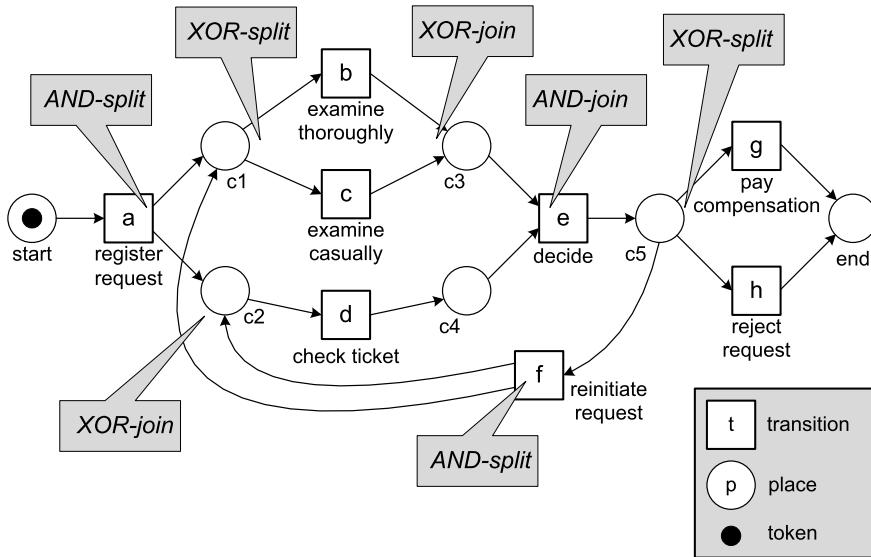


Fig. 3.2 A marked Petri net

net. Figure 3.2 shows the Petri net again with the various constructs highlighted. A Petri net is a bipartite graph consisting of *places* and *transitions*. The network structure is static, but, governed by the firing rule, *tokens* can flow through the network. The state of a Petri net is determined by the distribution of tokens over places and is referred to as its *marking*. In the initial marking shown in Fig. 3.2, there is only one token; *start* is the only marked place.

Definition 3.2 (Petri net) A *Petri net* is a triplet $N = (P, T, F)$ where P is a finite set of *places*, T is a finite set of *transitions* such that $P \cap T = \emptyset$, and $F \subseteq (P \times T) \cup (T \times P)$ is a set of directed arcs, called the *flow relation*. A *marked Petri net* is a pair (N, M) , where $N = (P, T, F)$ is a Petri net and where $M \in \mathbb{B}(P)$ is a *multi-set* over P denoting the *marking* of the net. The set of all marked Petri nets is denoted \mathcal{N} .

The Petri net shown Fig. 3.2 can be formalized as follows: $P = \{\text{start}, c1, c2, c3, c4, c5, \text{end}\}$, $T = \{a, b, c, d, e, f, g, h\}$, and $F = \{(\text{start}, a), (a, c1), (a, c2), (c1, b), (c1, c), (c2, d), (b, c3), (c, c3), (d, c4), (c4, e), (c3, e), (e, c5), (c5, f), (f, c1), (f, c2), (c5, g), (c5, h), (g, \text{end}), (h, \text{end})\}$.

Multi-sets

A marking corresponds to a multi-set of tokens. However, multi-sets are not only used to represent markings; later we will use multi-sets to model event logs where the same trace may appear multiple times. Therefore, we provide some basic notations used in the remainder.

A multi-set (also referred to as *bag*) is like a set in which each element may occur multiple times. For example, $[a, b^2, c^3, d^2, e]$ is the multi-set with nine elements: one a , two b 's, three c 's, two d 's, and one e . The following three multi-set are identical: $[a, b, b, c^3, d, d, e]$, $[e, d^2, c^3, b^2, a]$, and $[a, b^2, c^3, d^2, e]$. Only the number of occurrences of each value matters, not the order. Formally, $\mathbb{B}(D) = D \rightarrow \mathbb{N}$ is the set of multi-sets (bags) over a finite domain D , i.e., $X \in \mathbb{B}(D)$ is a multi-set, where for each $d \in D$, $X(d)$ denotes the number of times d is included in the multi-set. For example, if $X = [a, b^2, c^3]$, then $X(b) = 2$ and $X(e) = 0$.

The sum of two multi-sets ($X \uplus Y$), the difference ($X \setminus Y$), the presence of an element in a multi-set ($x \in X$), and the notion of subset ($X \leq Y$) are defined in a straightforward way. For example, $[a, b^2, c^3, d] \uplus [c^3, d, e^2, f^3] = [a, b^2, c^6, d^2, e^2, f^3]$ and $[a, b] \leq [a, b^3, c]$. Moreover, we can also apply these operators to sets, where we assume that a set is a multi-set in which every element occurs exactly once. For example, $[a, b^2] \uplus \{b, c\} = [a, b^3, c]$.

The operators are also robust with respect to the domains of the multi-sets, i.e., even if X and Y are defined on different domains, $X \uplus Y$, $X \setminus Y$, and $X \leq Y$ are defined properly by extending the domain whenever needed.

The marking shown in Fig. 3.2 is $[start]$, i.e., a multi-set containing only one token. The dynamic behavior of such a marked Petri net is defined by the so-called *firing rule*. A transition is *enabled* if each of its input places contains a token. An enabled transition can *fire* thereby consuming one token from each input place and producing one token for each output place. Hence, transition a is enabled at marking $[start]$. Firing a results in the marking $[c1, c2]$. Note that one token is consumed and two tokens are produced. At marking $[c1, c2]$, transition a is no longer enabled. However, transitions b , c , and d have become enabled. From marking $[c1, c2]$, firing b results in marking $[c2, c3]$. Here, d is still enabled, but b and c not anymore. Because of the loop construct involving f there are infinitely many firing sequences starting in $[start]$ and ending in $[end]$.

Assume now that the initial marking is $[start]^5$. Firing a now results in the marking $[start^4, c1, c2]$. At this marking a is still enabled. Firing a again results in marking $[start^3, c1^2, c2^2]$. Transition a can fire five times in a row resulting in marking $[c1^5, c2^5]$. Note that after the first occurrence of a , also b , c , and d are enabled and can fire concurrently.

To formalize the firing rule, we introduce a notation for input (output) places (transitions). Let $N = (P, T, F)$ be a Petri net. Elements of $P \cup T$ are called *nodes*. A node x is an *input node* of another node y if and only if there is a directed arc from x to y (i.e., $(x, y) \in F$). Node x is an *output node* of y if and only if $(y, x) \in F$. For any $x \in P \cup T$, $\bullet x = \{y \mid (y, x) \in F\}$ and $x\bullet = \{y \mid (x, y) \in F\}$. In Fig. 3.2, $\bullet c1 = \{a, f\}$ and $c1\bullet = \{b, c\}$.

Definition 3.3 (Firing rule) Let (N, M) be a marked Petri net with $N = (P, T, F)$ and $M \in \mathbb{B}(P)$. Transition $t \in T$ is *enabled*, denoted $(N, M)[t]$, if and only if

$\bullet t \leq M$. The *firing rule* $\underline{\underline{\cdot}} \subseteq \mathcal{N} \times T \times \mathcal{N}$ is the smallest relation satisfying for any $(N, M) \in \mathcal{N}$ and any $t \in T$, $(N, M)[t] \Rightarrow (N, M)[t] (N, (M \setminus \bullet t) \uplus t\bullet)$.

$(N, M)[t]$ denotes that t is enabled at marking M , e.g., $(N, [start])[a]$ in Fig. 3.2. $(N, M)[t](N, M')$ denotes that firing this enabled transition results in marking M' . For example, $(N, [start])[a](N, [c1, c2])$ and $(N, [c3, c4])[e](N, [c5])$.

Let (N, M_0) with $N = (P, T, F)$ be a marked Petri net. A sequence $\sigma \in T^*$ is called a *firing sequence* of (N, M_0) if and only if, for some natural number $n \in \mathbb{N}$, there exist markings M_1, \dots, M_n and transitions $t_1, \dots, t_n \in T$ such that $\sigma = \langle t_1 \dots t_n \rangle$ and, for all i with $0 \leq i < n$, $(N, M_i)[t_{i+1}]$ and $(N, M_i)[t_{i+1}](N, M_{i+1})$.¹

Let (N, M_0) be the marked Petri net shown in Fig. 3.2, i.e., $M_0 = [start]$. The empty sequence $\sigma = \langle \rangle$ is enabled in (N, M_0) , i.e., $\langle \rangle$ is a firing sequence of (N, M_0) . The sequence $\sigma = \langle a, b \rangle$ is also enabled and firing σ results in marking $[c2, c3]$. Another possible firing sequence is $\sigma = \langle a, c, d, e, f, b, d, e, g \rangle$. A marking M is *reachable* from the initial marking M_0 if and only if there exists a sequence of enabled transitions whose firing leads from M_0 to M . The set of reachable markings of (N, M_0) is denoted $[N, M_0]$. The marked Petri net shown in Fig. 3.2 has seven reachable markings.

In Fig. 3.2, transitions are identified by a single letter, but also have a longer label describing the corresponding activity. Thus far we ignored these labels.

Definition 3.4 (Labeled Petri net) A *labeled Petri net* is a tuple $N = (P, T, F, A, l)$ where (P, T, F) is a Petri net as defined in Definition 3.2, $A \subseteq \mathcal{A}$ is a set of *activity labels*, and $l : T \rightarrow A$ is a *labeling function*.

In principle, multiple transitions may bear the same label. One can think of the transition label as the *observable action*. Sometimes one wants to express that particular transitions are not observable. For this we reserve the label τ . A transition t with $l(t) = \tau$ is unobservable. Such transitions are often referred to as *silent* or *invisible*. It is easy to convert any Petri net into a labeled Petri net; just take $A = T$ and $l(t) = t$ for any $t \in T$. The reverse is not always possible, e.g., when several transitions have the same label. It is also possible to convert a marked (labeled) Petri net into a transition system as is shown next.

Definition 3.5 (Reachability graph) Let (N, M_0) with $N = (P, T, F, A, l)$ be a marked labeled Petri net. (N, M_0) defines a transition system $TS = (S, A', T')$ with $S = [N, M_0]$, $S^{start} = \{M_0\}$, $A' = A$, and $T' = \{(M, l(t), M') \in S \times A \times S \mid \exists_{t \in T} (N, M)[t](N, M')\}$. TS is often referred to as the *reachability graph* of (N, M_0) .

Figure 3.3 shows the transition system generated from the labeled marked Petri net shown in Fig. 3.2. States correspond to reachable markings, i.e., multi-sets of

¹ X^* is the set of sequences containing elements of X , i.e., for any $n \in \mathbb{N}$ and $x_1, x_2, \dots, x_n \in X$: $\langle x_1, x_2, \dots, x_n \rangle \in X^*$. See also Sect. 5.2.

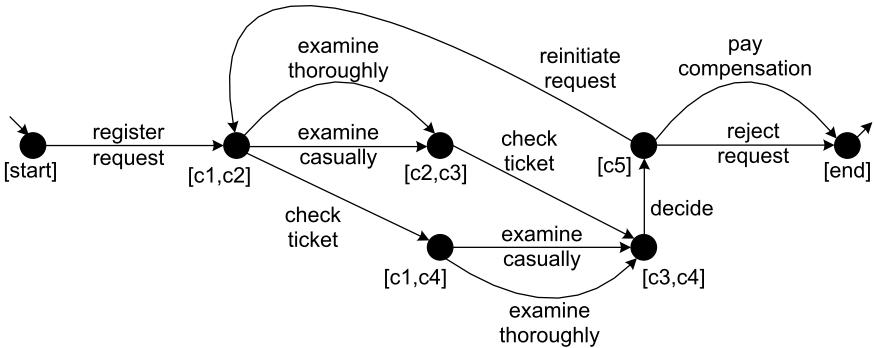
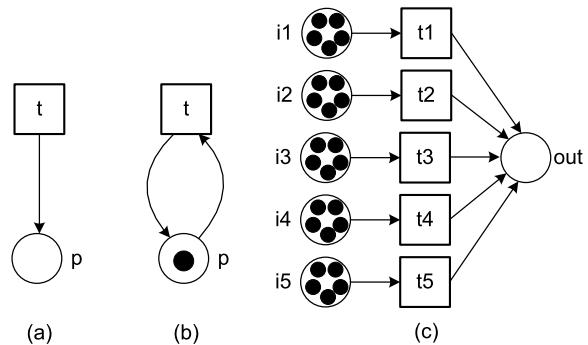


Fig. 3.3 The reachability graph of the marked Petri net shown in Fig. 3.2

Fig. 3.4 Three Petri nets:
(a) a Petri net with an infinite state space,
(b) a Petri net with only one reachable marking,
(c) a Petri net with 7776 reachable markings



tokens. Note that $S^{start} = \{[start]\}$ is a singleton containing the initial marking of the Petri net. The Petri net does not explicitly define a set of final markings S^{end} . However, in this case it is obvious to take $S^{end} = \{[end]\}$. Later, we will see that it is sometimes useful to distinguish deadlocks and livelocks from successful termination.

Note that we are overloading the term “transition”; the term may refer to a “box” in a Petri net or an “arc” in a transition system. In fact, one transition in a Petri net may correspond to many transitions in the corresponding transition system.

The Petri net in Fig. 3.2 and the transition system in Fig. 3.3 are of similar sizes. If the model contains a lot of concurrency or multiple tokens reside in the same place, then the transition system is much bigger than the Petri net. In fact, a marked Petri net may have infinitely many reachable states. The marked Petri net in Fig. 3.4(a) consists of only one place and one transition. Nevertheless, its corresponding transition system has infinitely many states: $S = \{[p^k] \mid k \in \mathbb{N}\}$. In this example, transition t is continuously enabled because it has no input place. Therefore, it can put any number of tokens in p . The Petri net in Fig. 3.4(b) has two arcs rather than one and now the only reachable state is $[p]$. The marked Petri net in Fig. 3.4(c) shows the effect of concurrency. The corresponding transition system has $6^5 = 7776$ states and 32,400 transitions.

Modern computers can easily compute reachability graphs with millions of states and analyze them. If the reachability graph is infinite, one can resort to the so-called *coverability graph* that presents a kind of over-approximation of the state space [117]. By constructing the reachability graph (if possible) or the coverability graph one can answer a variety of questions regarding the behavior of the process modeled. Moreover, dedicated analysis techniques can also answer particular questions without constructing the state space, e.g., using the linear-algebraic representation of the Petri net. It is outside the scope of this book to elaborate on these. However, we list some generic properties typically investigated in the context of a marked Petri net.

- A marked Petri net (N, M_0) is *k-bounded* if no place ever holds more than k tokens. Formally, for any $p \in P$ and any $M \in [N, M_0]$: $M(p) \leq k$. The marked Petri net in Fig. 3.4(c) is 25-bounded because in none of the 7776 reachable markings there is a place with more than 25 tokens. It is not 24-bounded, because in the final marking place *out* contains 25 tokens.
- A marked Petri net is *safe* if and only if it is 1-bounded. The marked Petri net shown in Fig. 3.2 is safe because in each of the seven reachable markings there is no place holding multiple tokens.
- A marked Petri net is *bounded* if and only if there exists a $k \in \mathbb{N}$ such that it is k -bounded. Figure 3.4(a) shows an unbounded net. The two other marked Petri nets in Fig. 3.4 (i.e., (b) and (c)) are bounded.
- A marked Petri net (N, M_0) is *deadlock free* if at every reachable marking at least one transition is enabled. Formally, for any $M \in [N, M_0]$ there exists a transition $t \in T$ such that $(N, M)[t]$. Figure 3.4(c) shows a net that is not deadlock free because at marking $[out^{25}]$ no transition is enabled. The two other marked Petri nets in Fig. 3.4 are deadlock free.
- A transition $t \in T$ in a marked Petri net (N, M_0) is *live* if from every reachable marking it is possible to enable t . Formally, for any $M \in [N, M_0]$ there exists a marking $M' \in [N, M]$ such that $(N, M')[t]$. A marked Petri net is live if each of its transitions is live. Note that a deadlock-free Petri net does not need to be live. For example, merge the nets (b) and (c) in Fig. 3.4 into one marked Petri net. The resulting net is deadlock free, but not live.

Petri nets have a strong theoretical basis and can capture concurrency well. Moreover, a wide range of powerful analysis techniques and tools exists [117]. Obviously, this succinct model has problems capturing data-related and time-related aspects. Therefore, various types of high-level Petri nets have been proposed. *Colored Petri nets* (CPNs) are the most widely used Petri-net based formalism that can deal with data-related and time-related aspects [82, 149]. Tokens in a CPN carry a data value and have a timestamp. The data value, often referred to as “color”, describes the properties of the object modeled by the token. The timestamp indicates the earliest time at which the token may be consumed. Transitions can assign a delay to produced tokens. This way waiting and service times can be modeled. A CPN may be hierarchical, i.e., transitions can be decomposed into subprocesses. This way large models can be structured. CPN Tools is a toolset providing support for the modeling and analysis of CPNs (www.cpntools.org).

3.2.3 Workflow Nets

When modeling business processes in terms of Petri nets, we often consider a subclass of Petri nets known as *Workflow nets* (WF-nets) [136, 168]. A WF-net is a Petri net with a dedicated source place where the process starts and a dedicated sink place where the process ends. Moreover, all nodes are on a path from source to sink.

Definition 3.6 (Workflow net) Let $N = (P, T, F, A, l)$ be a (labeled) Petri net and \bar{t} a fresh identifier not in $P \cup T$. N is a *workflow net* (WF-net) if and only if (a) P contains an input place i (also called source place) such that $\bullet i = \emptyset$, (b) P contains an output place o (also called sink place) such that $o\bullet = \emptyset$, and (c) $\bar{N} = (P, T \cup \{\bar{t}\}, F \cup \{(o, \bar{t}), (\bar{t}, i)\}, A \cup \{\tau\}, l \cup \{(\bar{t}, \tau)\})$ is strongly connected, i.e., there is a directed path between any pair of nodes in \bar{N} .

\bar{N} is referred to as the short-circuited net [136]. The unique sink place o is connected to the unique source place i in the resulting net.

Figure 3.2 shows an example of a WF-net with $i = \text{start}$ and $o = \text{end}$. None of the three Petri nets in Fig. 3.4 is a WF-net.

Why are WF-nets particularly relevant for business process modeling? The reason is that the process models used in the context of BPM describe the *life-cycle of cases* of a given kind. Examples of cases are insurance claims, job applications, customer orders, replenishment orders, patients, and credit applications. The process model is instantiated once for each case. Each of these process instances has a well-defined start (“case creation”) and end (“case completion”). In-between these points, activities are conducted according to a predefined procedure. One model may be instantiated many times. For example, the process of handling insurance claims may be executed for thousands or even millions of claims. These instances can be seen as copies of the same WF-net, i.e., tokens of different cases are not mixed.

WF-nets are also a natural representation for process mining. There is an obvious relation between the firing sequences of a WF-net and the traces found in event logs. Note that one can only learn models based on examples. In the context of market basket analysis, i.e., finding patterns in what customers buy, one needs many examples of customers buying particular collections of products. Similarly, process discovery uses sequences of activities in which each sequence refers to a particular process instance. These can be seen as firing sequences of an unknown WF-net. Therefore, we will often focus on WF-nets. Recall that the α -algorithm discovered the WF-net in Fig. 2.6 using the set of traces shown in Table 2.2. Every trace corresponds to a case executed from begin to end.

Not every WF-net represents a correct process. For example, a process represented by a WF-net may exhibit errors such as deadlocks, activities that can never become active, livelocks, or garbage being left in the process after termination. Therefore, we define the following well-known correctness criterion [136, 168]:

Definition 3.7 (Soundness) Let $N = (P, T, F, A, l)$ be a WF-net with input place i and output place o . N is *sound* if and only if

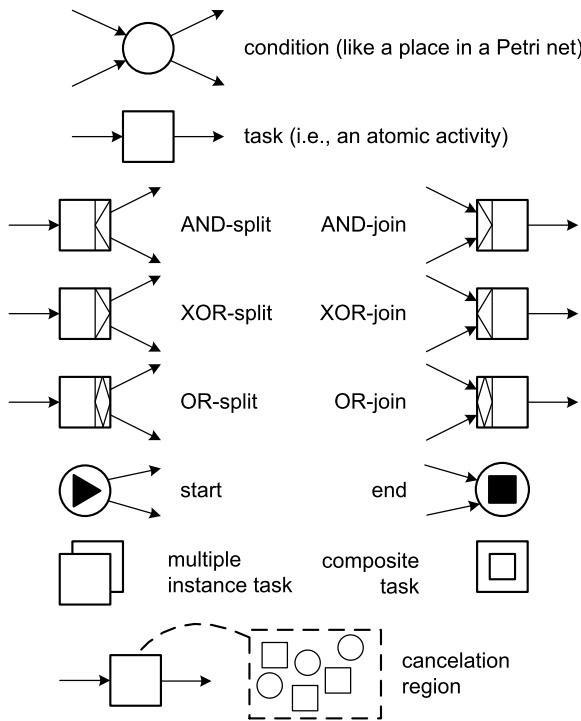
- (*safeness*) $(N, [i])$ is safe, i.e., places cannot hold multiple tokens at the same time;
- (*proper completion*) for any marking $M \in [N, [i]]$, $o \in M$ implies $M = [o]$;
- (*option to complete*) for any marking $M \in [N, [i]]$, $[o] \in [N, M]$; and
- (*absence of dead parts*) $(N, [i])$ contains no dead transitions (i.e., for any $t \in T$, there is a firing sequence enabling t).

Note that the option to complete implies proper completion. The WF-net shown in Fig. 3.2 is sound. Soundness can be verified using standard Petri-net-based analysis techniques. In fact soundness corresponds to liveness and safeness of the corresponding short-circuited net \bar{N} introduced in Definition 3.6 [136]. This way efficient algorithms and tools can be applied. An example of a tool tailored towards the analysis of WF-nets is *Woflan* [179]. This functionality is also embedded in our process mining tool *ProM* described in Sect. 11.3.

3.2.4 YAWL

YAWL is both a workflow modeling *language* and an open-source workflow *system* [132]. The acronym YAWL stands for “Yet Another Workflow Language”. The development of the YAWL language was heavily influenced by the *Workflow Patterns Initiative* [155, 191] mentioned earlier. Based on a systematic analysis of the constructs used by existing process modeling notations and workflow languages, a large collection of patterns was identified. These patterns cover all workflow perspectives, i.e., there are control-flow patterns, data patterns, resource patterns, change patterns, exception patterns, etc. The aim of YAWL is to offer direct support for many patterns while keeping the language simple. It can be seen as a reference implementation of the most important workflow patterns. Over time, the YAWL language and the YAWL system have increasingly become synonymous and have garnered widespread interest from both practitioners and the academic community alike. YAWL is currently one of the most widely used open-source workflow systems.

Here we restrict ourselves to the control-flow perspective. Figure 3.5 shows the main constructs. Each process has a dedicated start and end condition, like in WF-nets. Activities in YAWL are called *tasks*. *Conditions* in YAWL correspond to places in Petri nets. However, it is also possible to directly connect tasks without putting a condition in-between. Tasks have—depending on their type—a well-defined split and join semantics. An *AND-join/AND-split* task behaves like a transition, i.e., it needs to consume one token via each of the incoming arcs and produces a token along each of the outgoing arcs. An *XOR-split* selects precisely one of its outgoing arcs. The selection is based on evaluating data conditions. Only one token is produced and sent along the selected arc. An *XOR-join* is enabled once for every incoming token and does not need to synchronize. An *OR-split* selects one or more of its outgoing arcs. This selection is again based on evaluating data conditions. Note

Fig. 3.5 YAWL notation

that an OR-split may select 2 out of three 3 outgoing arcs. The semantics of the *OR-join* are more involved. The OR-join requires at least one input token, but also synchronizes tokens that are “on their way” to the OR-join. As long as another token may arrive via one of the ingoing arcs, the OR-join waits. YAWL also supports *cancelation regions*. A task may have a cancelation region consisting of conditions, tasks, and arcs. Once the task completes all tokens are removed from this region. Note that tokens for the task’s output conditions are produced after emptying the cancelation region. YAWL’s cancelation regions provide a powerful mechanism to abort work in parallel branches and to reset parts of the workflow. Tasks in a YAWL model can be *atomic* or *composite*. A composite task refers to another YAWL model. This way models can be structured hierarchically. Atomic and composite tasks can be instantiated multiple times in parallel. For example, when handling a customer order, some tasks need to be executed for every order line. These order lines can be processed in any order. Therefore, a loop construct is less suitable. Figure 3.5 shows the icon for such a multiple instance task and all other constructs just mentioned.

Figure 3.6 shows an example YAWL model for the handling of a request for compensation within an airline. To show some of the features of YAWL, we extended the process described in Sect. 2.1 with some more complex behaviors. In the new model it is possible that both examinations are executed. By using an OR-split and an OR-join *examine causally* and/or *examine thoroughly* are executed. The model has also been extended with a cancelation region (see dotted box in Fig. 3.6). As

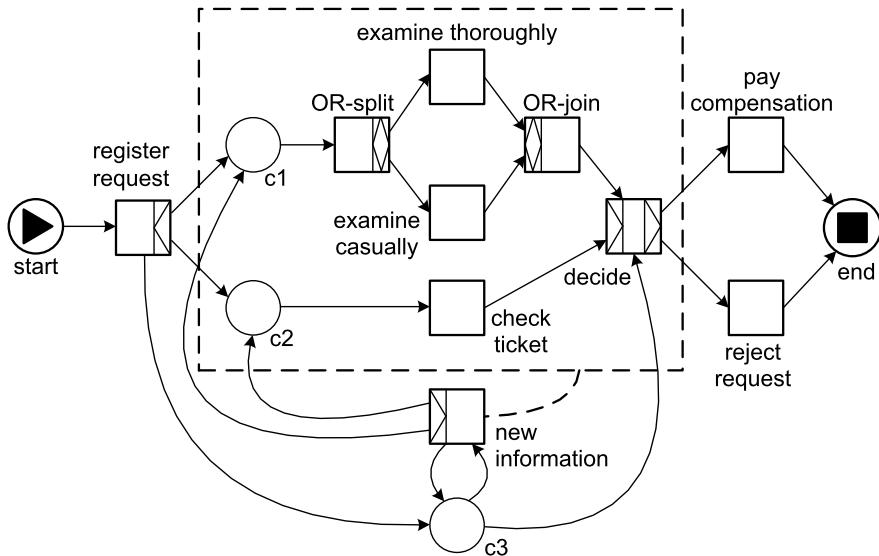


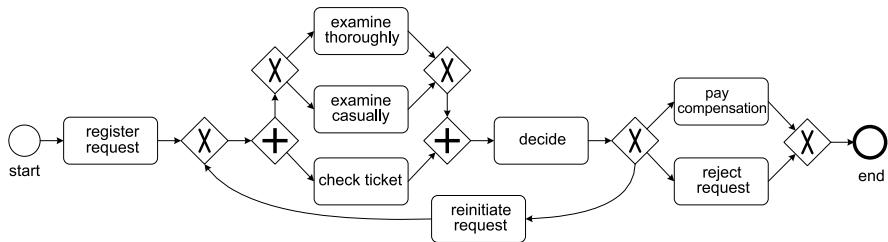
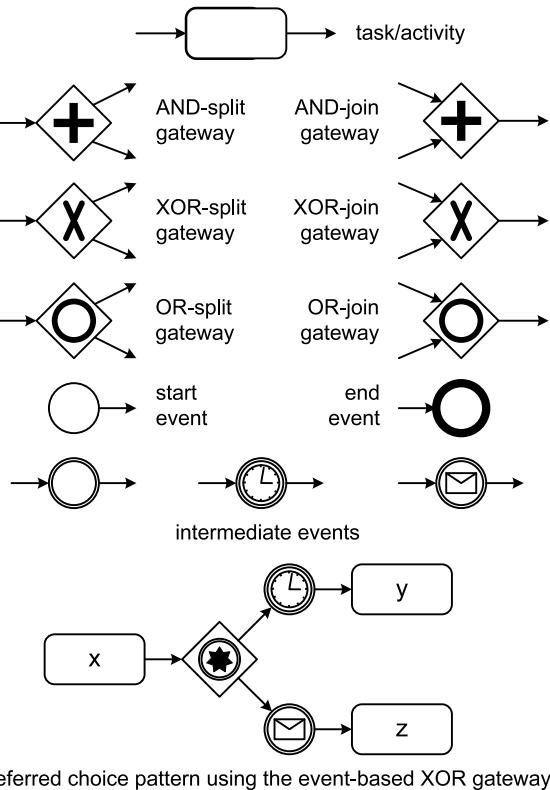
Fig. 3.6 Process model using the YAWL notation

long as there is a token in c_3 , task *new information* may be executed. When this happens, all tokens are removed from the region, i.e., checks and examinations are aborted. Task *new information* does not need to know where all tokens are and after the reset by this task the new state is $[c_1, c_2, c_3]$. Explicit choices in YAWL (i.e., XOR/OR-splits) are driven by data conditions. In the Petri net in Fig. 3.2, all choices were non-deterministic. In the example YAWL model, the decision may be derived from the outcome of the check and the examination(s), i.e., the XOR-split *decide* may be based on data created in earlier tasks. As indicated, both the YAWL language and the YAWL system cover all relevant perspectives (resources, data, exceptions, etc.). For example, it is possible to model that decisions are taken by the manager and that it is not allowed that two examinations for the same request are done by the same person (4-eyes principle) [132].

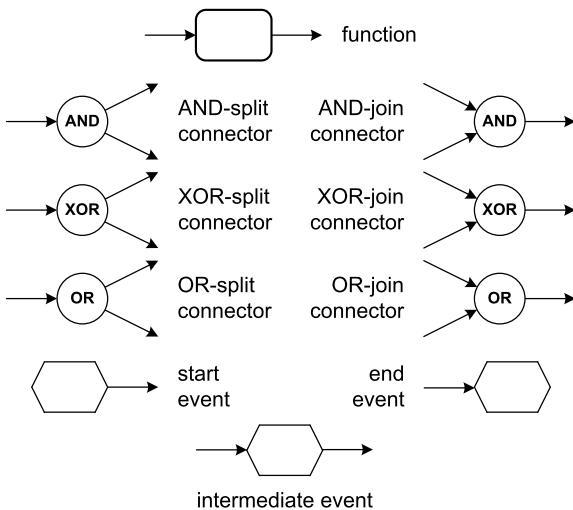
3.2.5 Business Process Modeling Notation (BPMN)

Recently, the *Business Process Modeling Notation* (BPMN) has become one of the most widely used languages to model business processes. BPMN is supported by many tool vendors and has been standardized by the OMG [110]. Figure 3.7 shows the BPMN model already introduced in Sect. 2.1.

Figure 3.8 shows a small subset of all notational elements. Atomic activities are called *tasks*. Like in YAWL activities can be nested. Most of the constructs can be easily understood after the introduction to YAWL. A notable difference is that the

**Fig. 3.7** Process model using the BPMN notation**Fig. 3.8** BPMN notation

routing logic is not associated with tasks but with separate *gateways*. Figure 3.8 shows that there are split and join gateways of different types: AND, XOR, OR. The splits are based on data conditions. An *event* is comparable to a place in a Petri net. However, the semantics of places in Petri nets and events in BPMN are quite different. There is no need to insert events in-between activities and events cannot have multiple input or output arcs. *Start events* have one outgoing arc, *intermediate events* have one incoming and one outgoing arc, and *end events* have one incoming arc. Unlike in YAWL or a Petri net, one cannot have events with multiple

Fig. 3.9 EPC notation

incoming or outgoing arcs; splitting and joining needs to be done using gateways. To model the so-called *deferred choice* workflow pattern [155] one needs to use the event-based XOR gateway shown in Fig. 3.8. This illustrates the use of events. After executing task x there is a race between two events. One of the events is triggered by a timeout. The other event is triggered by an external message. The first event to occur determines the route taken. If the message arrives before the timer goes off, task z is executed. If the timer goes off before the message arrives, task y is executed. Note that this construct can easily be modeled in YAWL using a condition with two output arcs.

Figure 3.8 shows just a tiny subset of all notations provided by BPMN. Most vendors support only a small subset of BPMN in their products. Moreover, users typically use only few BPMN constructs. In [193], it was shown that the average subset of BPMN used in real-life models consists of less than 10 different symbols (despite the more than 50 distinct graphical elements offered to the modeler). For this reason, we will be rather pragmatic when it comes to process models and their notation.

3.2.6 Event-Driven Process Chains (EPCs)

Event-driven Process Chains (EPCs) provide a classical notation to model business processes [126]. The notation is supported by products such as ARIS and SAP R/3. Basically, EPCs cover a limited subset of BPMN and YAWL while using a dedicated graphical notation.

Figure 3.9 provides an overview of the different notational elements. *Functions* correspond to activities. A function has precisely one input arc and one output arc. Therefore, splitting and joining can only be modeled using *connectors*. These are

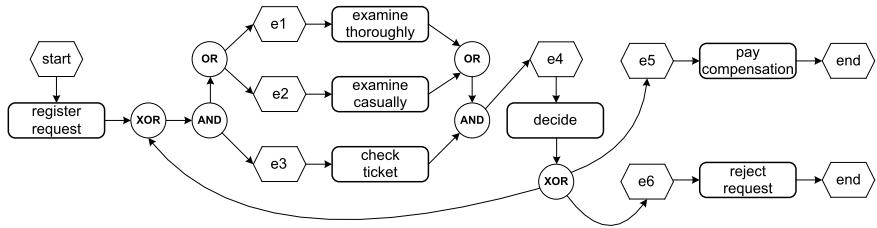
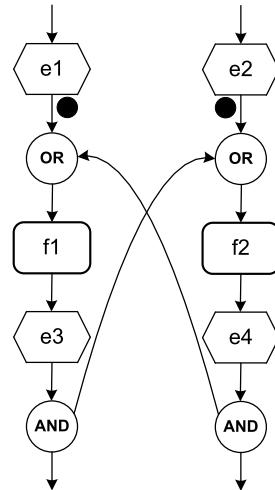


Fig. 3.10 Process model using the EPC notation

Fig. 3.11 The so-called “vicious circle” expressed using the EPC notation



comparable to the gateways in BPMN. Again splits and joins of type AND, XOR, and OR are supported. Like in BPMN there are three types of *events* (start, intermediate, and end). Events and functions need to alternate along any path, i.e., it is not allowed to connect events to events or functions to functions.

Figure 3.10 shows another variation of the process for handling a request for compensation. Note that, because of the two OR connectors, it is possible to do both examinations or just one.

The EPC notation was one of the first notations allowing for OR splits and joins. However, the people who developed and evangelized EPCs did not provide clear semantics nor some reference implementation [154]. This triggered lively debates resulting in various proposals and alternative implementations. Consider, for example, the so-called “vicious circle” shown in Fig. 3.11. The two tokens show the state of this process fragment; events *e1* and *e2* hold a token. It is unclear what could happen next, because both OR-joins depend on one another.

Should the OR-join below *e1* block or not? Suppose that this OR-join blocks, then by symmetry also the other OR-join following *e2* should block and the whole EPC deadlocks in the state shown Fig. 3.11. This seems to be wrong because if it deadlocks, the OR join will never receive an additional token and hence should

not have waited in the first place. Suppose that the OR-join following $e1$ does not block. By symmetry the other OR-join should also not block and both $f1$ and $f2$ are executed and tokens flow towards both OR-joins via the two AND-splits. However, this implies that the OR-joins should both have blocked. Hence, there is a *paradox* because all possible decisions are wrong.

The vicious circle paradox shows that higher-level constructs may introduce all kinds of subtle semantic problems. Despite these problems and the different notations, the core concepts of the various languages are very similar.

3.2.7 Causal Nets

The notations discussed thus far connect activities (i.e., transitions, tasks, functions) through model elements like places (Petri nets), conditions (YAWL), connectors and events (EPC), gateways and events (BPMN). These elements interconnect activities but do not leave any “marks” in the event log, i.e., they need to be inferred by analyzing the behavior. Since the log does not provide concrete information about places, conditions, connectors, gateways and events, some mining algorithms use a representation consisting of just activities and no connecting elements [4, 12, 66, 183, 184].

Causal nets are a representation tailored towards process mining. A causal net is a graph where nodes represent activities and arcs represent causal dependencies. Each activity has a set of possible *input bindings* and a set of possible *output bindings*. Consider, for example, the causal net shown in Fig. 3.12. Activity a has only an empty input binding as this is the start activity. There are two possible output bindings: $\{b, d\}$ and $\{c, d\}$. This means that a is followed by either b and d , or c and d . Activity e has two possible input bindings ($\{b, d\}$ and $\{c, d\}$) and three possible output bindings ($\{g\}$, $\{h\}$, and $\{f\}$). Hence, e is preceded by either b and d , or c and d , and is succeeded by just g , h or f . Activity z is the end activity having two input bindings and one output binding (the empty binding). This activity has been added to create a unique end point. All executions commence with start activity a and finish with end activity z . As will be shown later, the causal net shown in Fig. 3.12 and the Petri net shown in Fig. 3.2 are trace equivalent, i.e., they both allow for the same set of traces. However, there are no places in the causal net; the routing logic is solely represented by the possible input and output bindings.

Definition 3.8 (Causal net) A *Causal net* (C-net) is a tuple $C = (A, a_i, a_o, D, I, O)$ where:

- $A \subseteq \mathcal{A}$ is a finite set of *activities*;
- $a_i \in A$ is the *start activity*;
- $a_o \in A$ is the *end activity*;
- $D \subseteq A \times A$ is the *dependency relation*,
- $AS = \{X \subseteq \mathcal{P}(A) \mid X = \{\emptyset\} \vee \emptyset \notin X\}$;²

² $\mathcal{P}(A) = \{A' \mid A' \subseteq A\}$ is the powerset of A . Hence, elements of AS are *sets of sets* of activities.

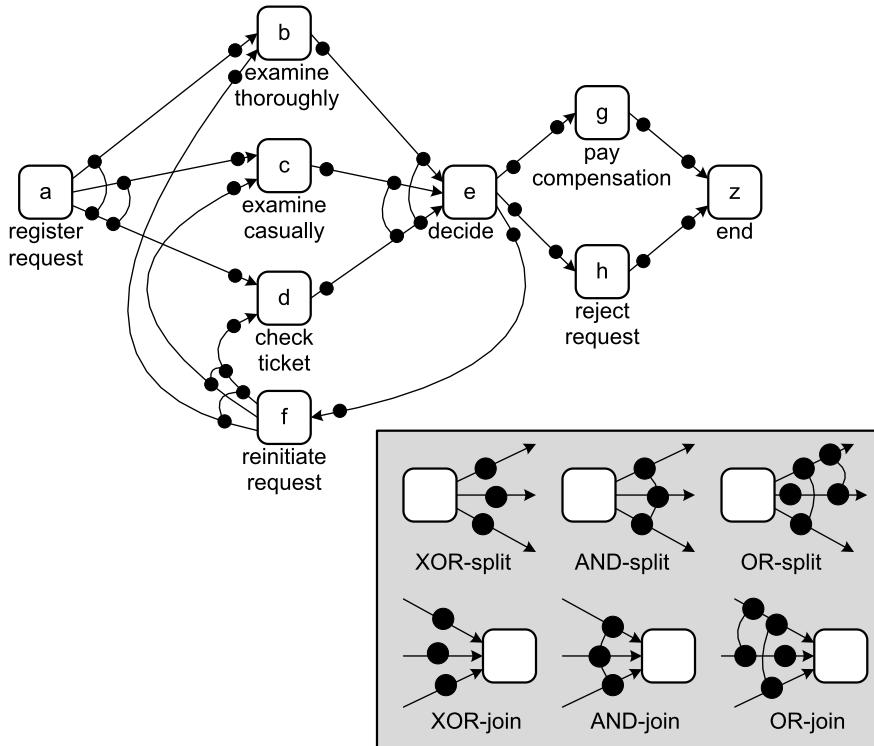


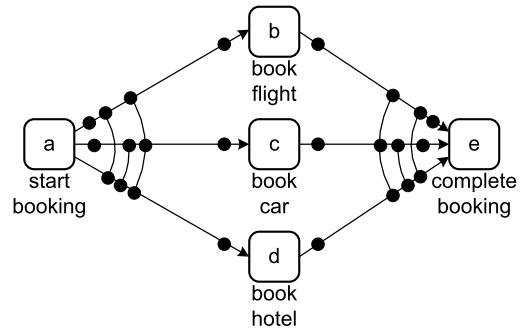
Fig. 3.12 Causal net C_1

- $I \in A \rightarrow AS$ defines the set of possible *input bindings* per activity; and
- $O \in A \rightarrow AS$ defines the set of possible *output bindings* per activity,

such that

- $D = \{(a_1, a_2) \in A \times A \mid a_1 \in \bigcup_{as \in I(a_2)} as\};$
- $D = \{(a_1, a_2) \in A \times A \mid a_2 \in \bigcup_{as \in O(a_1)} as\};$
- $\{a_i\} = \{a \in A \mid I(a) = \emptyset\};$
- $\{a_o\} = \{a \in A \mid O(a) = \emptyset\};$ and
- all activities in the graph (A, D) are on a path from a_i to a_o .

The C-net of Fig. 3.12 can be described as follows. $A = \{a, b, c, d, e, f, g, h, z\}$ is the set of activities, $a = a_i$ is the unique start activity, and $z = a_o$ is the unique end activity. The arcs shown in Fig. 3.12 visualize the dependency relation $D = \{(a, b), (a, c), (a, d), (b, e), \dots, (g, z), (h, z)\}$. Functions I and O describe the sets of possible input and output bindings. $I(a) = \{\emptyset\}$ is the set of possible input bindings of a , i.e., the only input binding is the empty set of activities. $O(a) = \{\{b, d\}, \{c, d\}\}$ is the set of possible output bindings of a , i.e., activity a is followed by d and either b or c . $I(b) = \{\{a\}, \{f\}\}$, $O(b) = \{\{e\}\}, \dots$,

Fig. 3.13 Causal net C_2 

$I(z) = \{\{g\}, \{h\}\}$, $O(z) = \emptyset$. Note that any element of AS is a set of sets of activities, e.g., $\{\{b, d\}, \{c, d\}\} \in AS$. If one of the elements is the empty set, then there cannot be any other elements, i.e., for any $X \in AS$: $X = \{\emptyset\}$ or $\emptyset \notin X$. This implies that only the unique start activity a_i has the empty binding as (only) possible input binding. Similarly, only the unique end activity a_o has the empty binding as (only) possible output binding.

An *activity binding* is a tuple (a, as^I, as^O) denoting the occurrence of activity a with input binding as^I and output binding as^O . For example, $(e, \{b, d\}, \{f\})$ denotes the occurrence of activity e in Fig. 3.12 while being preceded by b and d , and succeeded by f .

Definition 3.9 (Binding) Let $C = (A, a_i, a_o, D, I, O)$ be a C-net. $B = \{(a, as^I, as^O) \in A \times \mathcal{P}(A) \times \mathcal{P}(A) \mid as^I \in I(a) \wedge as^O \in O(a)\}$ is the set of *activity bindings*. A *binding sequence* σ is a sequence of activity bindings, i.e., $\sigma \in B^*$.

A possible binding sequence for the C-net of Fig. 3.12 is $\langle (a, \emptyset, \{b, d\}), (b, \{a\}, \{e\}), (d, \{a\}, \{e\}), (e, \{b, d\}, \{g\}), (g, \{e\}, \{z\}), (z, \{g\}, \emptyset) \rangle$.

Figure 3.13 shows another C-net modeling the booking of a trip. After activity a (*start booking*) there are three possible activities: b (*book flight*), c (*book car*), and d (*book hotel*). The process ends with activity e (*complete booking*). $O(a) = I(e) = \{\{b\}, \{c\}, \{b, d\}, \{c, d\}, \{b, c, d\}\}$, $I(a) = O(e) = \emptyset$, $I(b) = I(c) = I(d) = \{\{a\}\}$, and $O(b) = O(c) = O(d) = \{\{e\}\}$. A possible binding sequence for the C-net of Fig. 3.12 is $\langle (a, \emptyset, \{b, d\}), (d, \{a\}, \{e\}), (b, \{a\}, \{e\}), (e, \{b, d\}, \emptyset) \rangle$, i.e., the scenario in which a flight and a hotel are booked. Note that Fig. 3.13 does not allow for booking just a hotel nor is it possible to just book a flight and a car.

A binding sequence is *valid* if a predecessor activity and successor activity always “agree” on their bindings. For a predecessor activity x and successor activity y we need to see the following “pattern”: $\langle \dots, (x, \{\dots\}, \{y, \dots\}), \dots, (y, \{x, \dots\}, \{\dots\}), \dots \rangle$, i.e., the occurrence of activity x with y in its output binding needs to be followed by the occurrence of activity y and the occurrence of activity y with x in its input binding needs to be preceded by the occurrence of activity x . To formalize the notion of a valid sequence, we first define the notion of *state*.

Definition 3.10 (State) Let $C = (A, a_i, a_o, D, I, O)$ be a C-net. $S = \mathbb{B}(A \times A)$ is the *state space* of C . $s \in S$ is a *state*, i.e., a multi-set of pending *obligations*. Function $\psi \in B^* \rightarrow S$ is defined inductively: $\psi(\langle \rangle) = []$ and $\psi(\sigma \oplus (a, as^I, as^O)) = (\psi(\sigma) \setminus (as^I \times \{a\})) \uplus (\{a\} \times as^O)$ for any binding sequence $\sigma \oplus (a, as^I, as^O) \in B^*$.³ $\psi(\sigma)$ is the state after executing binding sequence σ .

Consider C-net C_1 shown in Fig. 3.12. Initially there are no pending “obligations”, i.e., no output bindings have been enacted without having corresponding input bindings. If activity binding $(a, \emptyset, \{b, d\})$ occurs, then $\psi((a, \emptyset, \{b, d\})) = \psi(\langle \rangle \setminus (\emptyset \times \{a\}) \uplus (\{a\} \times \{b, d\})) = [] \setminus [] \uplus [(a, b), (a, d)] = [(a, b), (a, d)]$. State $[(a, b), (a, d)]$ denotes the obligation to execute both b and d using input bindings involving a . Input bindings remove pending obligations whereas output bindings create new obligations.

A *valid sequence* is a binding sequence that (a) starts with start activity a_i , (b) ends with end activity a_o , (c) only removes obligations that are pending, and (d) ends without any pending obligations. Consider, for example, the valid sequence $\sigma = \langle (a, \emptyset, \{b, d\}), (d, \{a\}, \{e\}), (b, \{a\}, \{e\}), (e, \{b, d\}, \emptyset) \rangle$ for C-net C_2 in Fig. 3.13:

$$\begin{aligned}\psi(\langle \rangle) &= [] \\ \psi((a, \emptyset, \{b, d\})) &= [(a, b), (a, d)] \\ \psi((a, \emptyset, \{b, d\}), (d, \{a\}, \{e\})) &= [(a, b), (d, e)] \\ \psi((a, \emptyset, \{b, d\}), (d, \{a\}, \{e\}), (b, \{a\}, \{e\})) &= [(b, e), (d, e)] \\ \psi((a, \emptyset, \{b, d\}), (d, \{a\}, \{e\}), (b, \{a\}, \{e\}), (e, \{b, d\}, \emptyset)) &= []\end{aligned}$$

Sequence σ indeed starts with start activity a , ends with end activity e , only removes obligations that are pending (i.e., for every input binding there was an earlier output binding), and ends without any pending obligations: $\psi(\sigma) = []$.

Definition 3.11 (Valid) Let $C = (A, a_i, a_o, D, I, O)$ be a C-net and $\sigma = \langle (a_1, as_1^I, as_1^O), (a_2, as_2^I, as_2^O), \dots, (a_n, as_n^I, as_n^O) \rangle \in B^*$ a binding sequence. σ is a *valid sequence* of C if and only if:

- $a_1 = a_i$, $a_n = a_o$, and $a_k \in A \setminus \{a_i, a_o\}$ for $1 < k < n$;
- $\psi(\sigma) = []$; and
- for any prefix $\langle (a_1, as_1^I, as_1^O), (a_2, as_2^I, as_2^O), \dots, (a_k, as_k^I, as_k^O) \rangle = \sigma' \oplus (a_k, as_k^I, as_k^O) \in pref(\sigma)$: $(as_k^I \times \{a_k\}) \leq \psi(\sigma')$.

$V(C)$ is the set of all valid sequences of C .

³ $\sigma_1 \oplus \sigma_2$ is the concatenation of two sequences, e.g., $\langle a, b, c \rangle \oplus \langle d, e \rangle = \langle a, b, c, d, e \rangle$. It is also possible to concatenate a sequence and an element, e.g., $\langle a, b, c \rangle \oplus d = \langle a, b, c, d \rangle$. Recall that X^* is the set of all sequences containing elements of X and $\langle \rangle$ is the empty sequence. See also Sect. 5.2 for more notations for sequences.

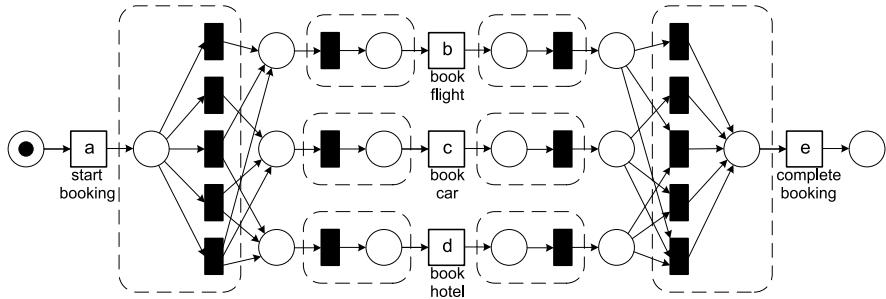


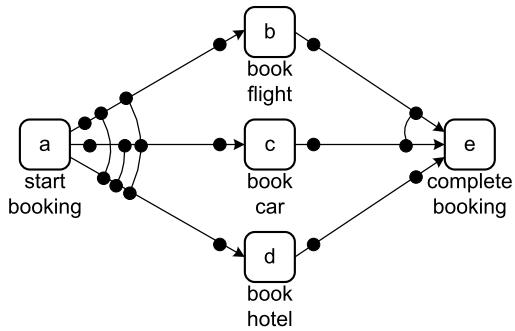
Fig. 3.14 A C-net transformed into a WF-net with silent transitions: every “sound run” of the WF-net corresponds to a valid sequence of the C-net C_2 shown in Fig. 3.13

The first requirement states that valid sequences start with a_i and end with a_o (a_i and a_o cannot appear in the middle of valid sequence). The second requirement states that at the end there should not be any pending obligations. (One can think of this as the constraint that no tokens left in the net.) The last requirement considers all non-empty prefixes of σ , $((a_1, as_1^I, as_1^O), (a_2, as_2^I, as_2^O), \dots, (a_k, as_k^I, as_k^O))$. The last activity binding of the prefix (i.e., (a_k, as_k^I, as_k^O)) should only remove pending obligations, i.e., $(as_k^I \times \{a_k\}) \leq \psi(\sigma')$ where $as_k^I \times \{a_k\}$ are the obligations to be removed and $\psi(\sigma')$ are the pending obligations just before the occurrence of the k -th binding. (One can think of this as the constraint that one cannot consume tokens that have not been produced.)

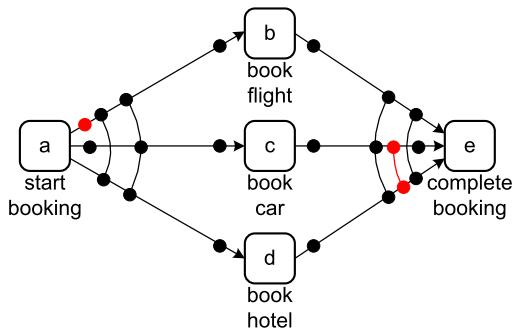
Figure 3.13 has 12 valid sequences: only b is executed $((a, \emptyset, \{b\}), (b, \{a\}, \{e\}), (e, \{b\}, \emptyset))$, only c is executed (besides a and e), b and d are executed (two possibilities), c and d are executed (two possibilities), and b , c and d are executed ($3! = 6$ possibilities). The C-net in Fig. 3.12 has infinitely many valid sequences because of the loop construct involving f . For example, $((a, \emptyset, \{c, d\}), (c, \{a\}, \{e\}), (d, \{a\}, \{e\}), (e, \{c, d\}, \{f\}), (f, \{e\}, \{c, d\}), (c, \{f\}, \{e\}), (d, \{f\}, \{e\}), (e, \{c, d\}, \{g\}), (g, \{e\}, \{z\}), (z, \{g\}, \emptyset))$.

For the semantics of a C-net we only consider valid sequences, i.e., *invalid sequences are not part of the behavior* described by the C-net. This means that C-nets do not use plain “token-game like semantics” as in BPMN, Petri nets, EPCs, and YAWL. The semantics of C-nets are more declarative as they are defined over complete sequences rather than a local firing rule. This is illustrated by the WF-net shown in Fig. 3.14. This WF-net aims to model the semantics of the C-net C_2 in Fig. 3.13. The input and output bindings are modeled by *silent transitions*. In Fig. 3.14, these are denoted by black rectangles without labels. Note that the WF-net also allows for many invalid sequences. For example, it is possible to enable b , c and d . After firing b it is possible to fire e without firing c and d . This firing sequence does not correspond to a valid sequence because there are still pending commitments when executing the end activity e . However, if we only consider firing sequences of the WF-net that start with a token in the source place and end with a token in the sink place, then these match one-to-one with the valid sequences in $V(C_2)$.

Fig. 3.15 Two C-nets that are not sound. The first net does not allow for any valid sequence, i.e., $V(C) = \emptyset$. The second net has valid sequences but also shows input/output bindings that are not realizable



(a) unsound because there are no valid sequences



(b) unsound although there exist valid sequences

The C-net shown in Fig. 3.12 and the WF-net shown in Fig. 3.2 are trace equivalent. Recall that in this comparison we consider all possible firing sequences of the WF-net and only valid sequences for the C-net.

We defined the notion of soundness for WF-nets (Definition 3.7) to avoid process models that have deadlocks, livelocks, and other anomalies. A similar notion can be defined for C-nets.

Definition 3.12 (Soundness of C-nets) A C-net $C = (A, a_i, a_o, D, I, O)$ is *sound* if (a) for all $a \in A$ and $as^I \in I(a)$ there exists a $\sigma \in V(C)$ and $as^O \subseteq A$ such that $(a, as^I, as^O) \in \sigma$, and (b) for all $a \in A$ and $as^O \in O(a)$ there exists a $\sigma \in V(C)$ and $as^I \subseteq A$ such that $(a, as^I, as^O) \in \sigma$.

Since the semantics of C-nets already enforce “proper completion” and the “option to complete”, we only need to make sure that there are valid sequences and that all parts of the C-net can potentially be activated by such a valid sequence. The C-nets C_1 and C_2 in Figs. 3.12 and 3.13 are sound. Figure 3.15 shows two C-nets that are not sound. In Fig. 3.15(a), there are no valid sequences because the output bindings of a and the input bindings of e do not match. For example, consider the binding sequence $\sigma = \langle (a, \emptyset, \{b\}), (b, \{a\}, \{e\}) \rangle$. Sequence σ cannot be extended into a valid sequence because $\psi(\sigma) = [(b, e)]$ and $\{b\} \notin I(e)$, i.e., the input bind-

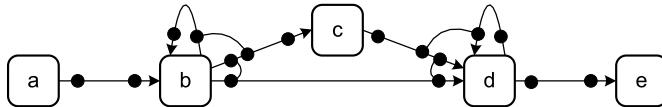


Fig. 3.16 A sound C-net that has no corresponding WF-net

ings of e do not allow for just booking a flight whereas the output bindings of a do. In Fig. 3.15(b), there are valid sequences, e.g., $\langle(a, \emptyset, \{c\}), (c, \{a\}, \{e\}), (e, \{c\}, \emptyset)\rangle$. However, not all bindings appear in one or more valid sequences. For example, the output binding $\{b\} \in O(a)$ does not appear in any valid sequence, i.e., after selecting just a flight the sequence cannot be completed properly. The input binding $\{c, d\} \in I(e)$ also does not appear in any valid sequence, i.e., the C-net suggests that only a car and hotel can be booked but there is no corresponding valid sequence.

Figure 3.16 shows an example of a sound C-net. One of the valid binding sequences for this C-net is $\langle(a, \emptyset, \{b\}), (b, \{a\}, \{b, c\}), (b, \{b\}, \{c, d\}), (c, \{b\}, \{d\}), (c, \{b\}, \{d\}), (d, \{b, c\}, \{d\}), (d, \{c, d\}, \{e\}), (e, \{d\}, \emptyset)\rangle$, i.e., the sequence $\langle a, b, b, c, c, d, d, e \rangle$. This sequence covers all the bindings. Therefore, the C-net is sound. Examples of other valid sequences are $\langle a, b, c, d, e \rangle$, $\langle a, b, c, b, c, d, d, e \rangle$, and $\langle a, b, b, b, c, c, c, d, d, d, e \rangle$. Figure 3.16 illustrates the expressiveness of C-nets. Note that there is no sound WF-net that reproduces exactly the set of valid sequences of this C-net. If we use the construction shown in Fig. 3.14 for the C-net of Fig. 3.16, we get a WF-net that is able to simulate the valid sequences. However, the resulting WF-net also allows for invalid behavior and it is impossible to modify the model such that the set of firing sequences coincides with the set of valid sequences.

Causal nets are particularly suitable for process mining given their declarative nature and expressiveness without introducing all kinds of additional model elements (places, conditions, events, gateways, etc.). Several process discovery and conformance checking approaches use a similar representation [4, 12, 66, 183, 184]. In Chap. 7, we elaborate on this when discussing some of the more advanced process mining algorithms.

3.2.8 Process Trees

Petri nets, WF-nets, BPMN models, EPCs, YAWL models, and UML activity diagrams may suffer from deadlocks, livelocks, and other anomalies. Models having undesirable properties *independent* of the event log are called *unsound*. One does not need to look at the event log to see that an unsound model cannot describe the observed behavior well. Process discovery approaches using any of the graph-based process notations mentioned may produce unsound models. In fact, the majority of models in the search space tend to be unsound. This complicates discovery. C-nets address this problem by using more relaxed semantics. It is also possible to use

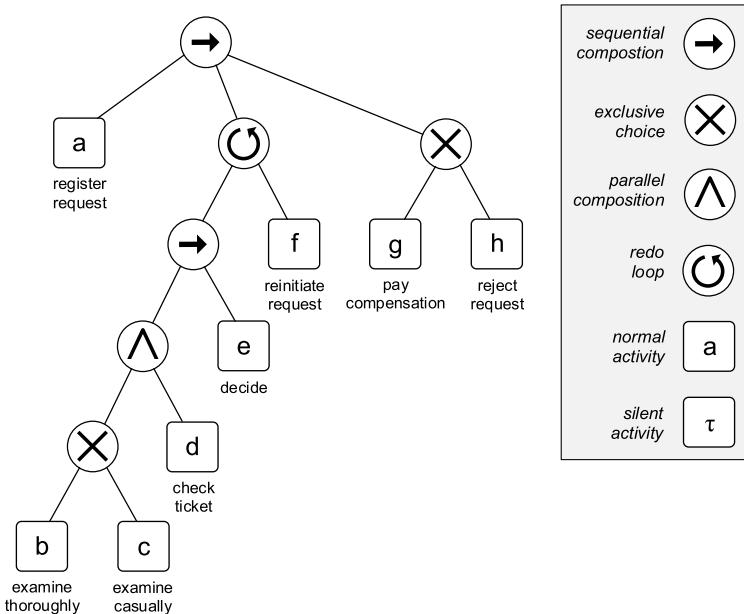


Fig. 3.17 Process tree $\rightarrow(a, \odot(\rightarrow(\wedge(\times(b, c), d), e), f), \times(g, h))$ showing the different process tree operators

block-structured models that are sound by construction. In this section, we introduce *process trees* as a notation to represent such block-structured models. A process tree is a hierarchical process model where the (inner) nodes are operators such as sequence and choice and the leaves are activities.

Process trees are tailored towards process discovery. A range of *inductive process discovery* techniques exists for process trees [88–91]. These techniques benefit from the fact that the representation ensures soundness. The family of inductive mining techniques has variants that can handle infrequent behavior and deal with huge models and logs while ensuring formal correctness criteria such as the ability to rediscover the original model (see Sect. 7.5). Also the ETM (Evolutionary Tree Miner) approach described in [26] exploits the process tree representation. The fact that the search space is limited to sound models is a key ingredient of this highly flexible genetic process mining approach.

Figure 3.17 shows a *process tree* modeling the handling of a request for compensation within an airline. The set of traces that can be generated by this model is identical to the traces generated by the WF-net in Fig. 3.2 (the two models are trace equivalent). The inner nodes of the process tree represent operators. The leaves represent activities. There is one root node. Figure 3.17 shows the four types of operators that can be used in a process tree: \rightarrow (sequential composition), \times (exclusive choice), \wedge (parallel composition), and \odot (redo loop).

A sequence operator executes its children in sequential order. Activity a is the first child of the root node in Fig. 3.17. Since this node is a sequence node, every

process instance starts with activity a followed by the subtree starting with the redo loop (\circlearrowright). After this subtree in the middle, the rightmost subtree is executed. The latter subtree models a choice (\times) between g and h .

The process tree in Fig. 3.17 can also be represented textually:

$$\rightarrow(a, \circlearrowright(\rightarrow(\wedge(\times(b, c), d), e), f), \times(g, h))$$

The rightmost subtree modeling the choice between activities g and h is represented as $\times(g, h)$. The redo loop $\circlearrowright(\rightarrow(\wedge(\times(b, c), d), e), f)$ starts with its leftmost child and may loop back through any of its other children. In the process tree of Fig. 3.17, it is possible to loop back via “redo” activity f . The leftmost child (“do part”) is $\rightarrow(\wedge(\times(b, c), d), e)$, i.e., a sequence that ends with activity e which is preceded by the subtree $\wedge(\times(b, c), d)$ where activity d is executed in parallel with a choice between b and c . The subtree $\wedge(\times(b, c), d)$ has four potential behaviors: $\langle b, d \rangle$, $\langle c, d \rangle$, $\langle d, b \rangle$, and $\langle d, c \rangle$.

The same activity may appear multiple times in the same process tree. For example, process tree $\rightarrow(a, a, a)$ models a sequence of three a activities. From a behavioral point of view, $\rightarrow(a, a, a)$ and $\wedge(a, a, a)$ are indistinguishable. Both have one possible trace, $\langle a, a, a \rangle$.

A silent activity is denoted by τ and cannot be observed. Process tree $\times(a, \tau)$ can be used to model an activity a that can be skipped. Process tree $\circlearrowright(a, \tau)$ can be used to model the process that executes a at least once. The “redo” part is silent, so the process can loop back without executing any activity. Process tree $\circlearrowright(\tau, a)$ models a process that executes a any number of times. The “do” part is now silent and activity a is in the “redo” part. This way it is also possible to not execute a at all. The smallest process tree is a tree consisting of just one activity. In this case the root node is also a leaf node and there are no operator nodes.

Definition 3.13 (Process tree) Let $A \subseteq \mathcal{A}$ be a finite set of activities with $\tau \notin A$. $\oplus = \{\rightarrow, \times, \wedge, \circlearrowright\}$ is the set of *process tree operators*.

- If $a \in A \cup \{\tau\}$, then $Q = a$ is a process tree,
- If $n \geq 1$, Q_1, Q_2, \dots, Q_n are process trees, and $\oplus \in \{\rightarrow, \times, \wedge\}$, then $Q = \oplus(Q_1, Q_2, \dots, Q_n)$ is a process tree, and
- If $n \geq 2$ and Q_1, Q_2, \dots, Q_n are process trees, then $Q = \circlearrowright(Q_1, Q_2, \dots, Q_n)$ is a process tree.

\mathcal{Q}_A is the set of *all process trees* over A .

The redo loop operator \circlearrowright has at least two children. The first child is the “do” part and the other children are “redo” parts. Process tree $\circlearrowright(a, b, c)$ allows for traces $\{\langle a \rangle, \langle a, b, a \rangle, \langle a, c, a \rangle, \langle a, b, a, b, a \rangle, \langle a, c, a, c, a \rangle, \langle a, c, a, b, a \rangle, \langle a, b, a, c, a \rangle, \dots\}$. Activity a is executed at least once and the process always starts and ends with a . The “do” part alternates with the “redo” parts b or c . When looping back either b or c is executed.

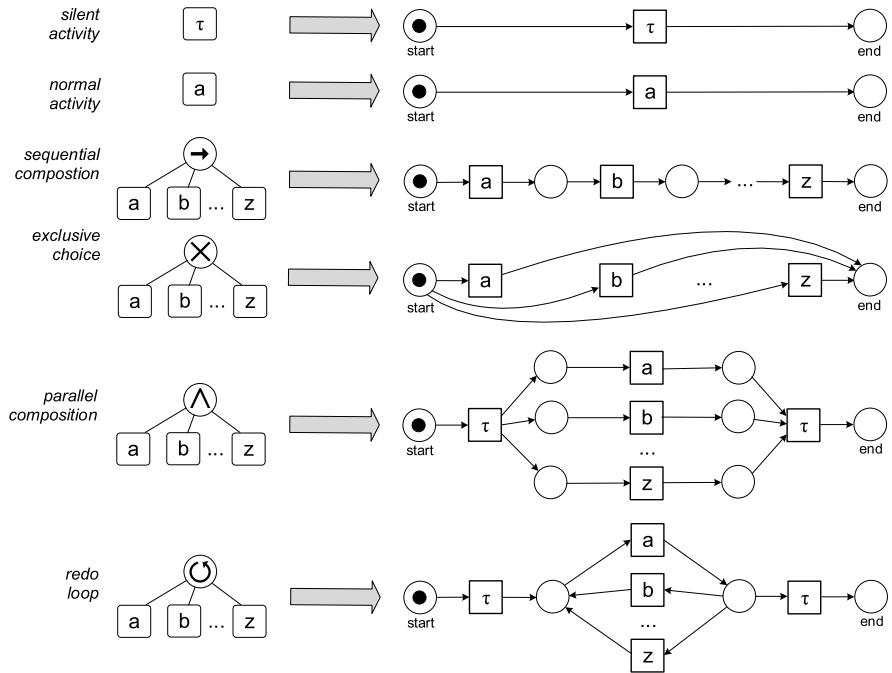


Fig. 3.18 Mapping process trees onto WF-nets

The redo loop operator \circlearrowleft is often used in conjunction with silent activity τ . For example, $\circlearrowleft(\tau, a, b, c, \dots, z)$ allows for any “word” involving activities a, b, c, \dots, z . Example traces are $\langle \rangle$, $\langle a, b, b, a \rangle$, and $\langle w, o, r, d \rangle$.

Process trees can be converted to WF-nets as shown in Fig. 3.18. A silent activity is mapped onto a transition having a τ label. The mappings for \rightarrow (sequential composition), \times (exclusive choice), and \wedge (parallel composition) are fairly straightforward. Silent transitions are used to model the start and end of the parallel composition. This is done to preserve the WF-net structure. The redo loop (\circlearrowleft) has one “do” part (activity a in Fig. 3.18) and one or more “redo” parts (activities b until z in Fig. 3.18). The direction of the arcs in the Petri net show the difference in semantics between the “do” and “redo” parts. Silent transitions are used to model the entry and exit of the redo loop. The mapping in Fig. 3.18 can be applied recursively and used to transform any process tree into a *sound* WF-net.

The mapping in Fig. 3.18 can easily be adapted for other representations such as BPMN, YAWL, EPCs, UML activity diagrams, statecharts, etc. The structured nature of process trees makes the conversion to other modeling notations straightforward. Conversions in the other direction (for example, from non-block-structured models to process trees) are more involved, but also less relevant since we only use process trees for process discovery. The mapping from process trees to WF-nets allows us to use existing conformance checking and performance analysis techniques.

The semantics of process trees can also be defined directly (without a mapping to WF-nets). To do this we first define two operators on sequences, concatenation (\cdot) and shuffle (\diamond).

Let $\sigma_1, \sigma_2 \in A^*$ be two sequences over A . $\sigma_1 \cdot \sigma_2 \in A^*$ *concatenates* two sequences, e.g., $\langle w, o \rangle \cdot \langle r, d \rangle = \langle w, o, r, d \rangle$. Concatenation can be generalized to sets of sequences. Let $S_1, S_2, \dots, S_n \subseteq A^*$ be sets of sequences over A . $S_1 \cdot S_2 = \{\sigma_1 \cdot \sigma_2 \mid \sigma_1 \in S_1 \wedge \sigma_2 \in S_2\}$. For example, $\{\langle w, o \rangle, \langle \rangle\} \cdot \{\langle r, d \rangle, \langle k \rangle\} = \{\langle w, o, r, d \rangle, \langle w, o, k \rangle, \langle r, d \rangle, \langle k \rangle\}$. $\bigodot_{1 \leq i \leq n} S_i = S_1 \cdot S_2 \cdots S_n$ concatenates an ordered collection of sets of sequences.

$\sigma_1 \diamond \sigma_2$ generates the set of all interleaved sequences (shuffle). For example, $\langle w, o \rangle \diamond \langle r, d \rangle = \{\langle w, o, r, d \rangle, \langle w, r, o, d \rangle, \langle r, w, o, d \rangle, \langle w, r, d, o \rangle, \langle r, w, d, o \rangle, \langle r, d, w, o \rangle\}$. Note that the ordering in the original sequences is preserved, e.g., d cannot appear before r . Another example is $\langle w, o, r \rangle \diamond \langle d \rangle = \{\langle w, o, r, d \rangle, \langle w, o, d, r \rangle, \langle w, d, o, r \rangle, \langle d, w, o, r \rangle\}$. The shuffle operator can also be generalized to sets of sequences. $S_1 \diamond S_2 = \{\sigma \in \sigma_1 \diamond \sigma_2 \mid \sigma_1 \in S_1 \wedge \sigma_2 \in S_2\}$. The shuffle operator is commutative and associative, i.e., $S_1 \diamond S_2 = S_2 \diamond S_1$ and $(S_1 \diamond S_2) \diamond S_3 = S_1 \diamond (S_2 \diamond S_3)$. We write $\diamond_{1 \leq i \leq n} S_i = S_1 \diamond S_2 \diamond \cdots \diamond S_n$ to interleave sets of sequences.

Definition 3.14 (Semantics) Let $Q \in \mathcal{Q}_A$ be a process tree over A . $\mathcal{L}(Q)$ is the *language* of Q , i.e., the set of traces that can be generated by it. $\mathcal{L}(Q)$ is defined recursively:

- $\mathcal{L}(Q) = \{\langle a \rangle\}$ if $Q = a \in A$,
- $\mathcal{L}(Q) = \{\langle \rangle\}$ if $Q = \tau$,
- $\mathcal{L}(Q) = \bigodot_{1 \leq i \leq n} \mathcal{L}(Q_i)$ if $Q = \rightarrow(Q_1, Q_2, \dots, Q_n)$,
- $\mathcal{L}(Q) = \bigcup_{1 \leq i \leq n} \mathcal{L}(Q_i)$ if $Q = \times(Q_1, Q_2, \dots, Q_n)$,
- $\mathcal{L}(Q) = \diamond_{1 \leq i \leq n} \mathcal{L}(Q_i)$ if $Q = \wedge(Q_1, Q_2, \dots, Q_n)$,
- $\mathcal{L}(Q) = \{\sigma_1 \cdot \sigma'_1 \cdot \sigma_2 \cdot \sigma'_2 \cdots \sigma_m \in A^* \mid m \geq 1 \wedge \forall_{1 \leq j \leq m} \sigma_j \in \mathcal{L}(Q_1) \wedge \forall_{1 \leq j < m} \sigma'_j \in \bigcup_{2 \leq i \leq n} \mathcal{L}(Q_i)\}$ if $Q = \circlearrowleft(Q_1, Q_2, \dots, Q_n)$.

The following examples further illustrate the process tree operators and their semantics:

- $\mathcal{L}(\tau) = \{\langle \rangle\}$,
- $\mathcal{L}(a) = \{\langle a \rangle\}$,
- $\mathcal{L}(\rightarrow(a, b, c)) = \{\langle a, b, c \rangle\}$,
- $\mathcal{L}(\times(a, b, c)) = \{\langle a \rangle, \langle b \rangle, \langle c \rangle\}$,
- $\mathcal{L}(\wedge(a, b, c)) = \{\langle a, b, c \rangle, \langle a, c, b \rangle, \langle b, a, c \rangle, \langle b, c, a \rangle, \langle c, a, b \rangle, \langle c, b, a \rangle\}$,
- $\mathcal{L}(\circlearrowleft(a, b, c)) = \{\langle a \rangle, \langle a, b, a \rangle, \langle a, c, a \rangle, \langle a, b, a, c, a \rangle, \langle a, c, a, b, a \rangle, \dots\}$,
- $\mathcal{L}(\rightarrow(a, \times(b, c), \wedge(a, a))) = \{\langle a, b, a, a \rangle, \langle a, c, a, a \rangle\}$,
- $\mathcal{L}(\times(\tau, a, \tau, \rightarrow(\tau, b), \wedge(c, \tau))) = \{\langle \rangle, \langle a \rangle, \langle b \rangle, \langle c \rangle\}$, and
- $\mathcal{L}(\circlearrowleft(a, \tau, c)) = \{\langle a \rangle, \langle a, a \rangle, \langle a, a, a \rangle, \langle a, c, a \rangle, \langle a, a, c, a \rangle, \langle a, c, a, c, a \rangle, \dots\}$.

Process trees are *sound by construction*. Process discovery algorithms may exploit this when searching for a process model describing the event data. There are some similarities with other notations. *Process calculi* such as CSP and CCS use similar operators to model processes. Process trees can be viewed as a carefully

chosen subset. *Regular expressions* can model regular languages, e.g., $a^*(b|c)d^*$ denotes the set of traces starting with zero or more a 's, followed by b or c , followed by zero or more d 's. Process trees are in-between process calculi and regular expressions, and are tailored towards process discovery. Process calculi can handle concurrency, but are difficult to discover from event data (unless a similar subset is chosen). Regular expressions do not provide operators for concurrency and redo loops. However, in terms of expressiveness, process trees are comparable to regular expressions. Process trees are also related to *soundness preserving reduction rules* for Petri nets [168]. Reductions rules are normally used to reduce the size of a Petri net while preserving essential properties (e.g., soundness, liveness, boundedness, etc.). Starting from a WF-net with one transition, they can also be applied in reverse direction to produce larger sound WF-nets.

Section 7.5 introduces inductive process discovery techniques. Then the rationale for the choice of operators will become clearer. For example, $\circlearrowleft(\tau, a, b, c, \dots, z)$ will be used as a last resort when all other operators are not applicable.

3.3 Model-Based Process Analysis

In Sect. 2.1, we discussed the different reasons for making models. Figure 2.4 illustrated the use of these models in the BPM life-cycle. Subsequent analysis showed that existing approaches using process models ignore event data. In later chapters we will show how to exploit event data when analyzing processes and their models. However, before doing so, we briefly summarize mainstream approaches for model-based analysis: *verification* and *performance analysis*. Verification is concerned with the correctness of a system or process. Performance analysis focuses on flow times, waiting times, utilization, and service levels.

3.3.1 Verification

In Sect. 3.2.3, we introduced the notion of soundness for WF-nets. This is a correctness criterion that can be checked using verification techniques. Consider, for example, the WF-net shown in Fig. 3.19. The model has been extended to model that *check ticket* should wait for the completion of *examine casually* but not for *examine thoroughly*. Therefore, place c_6 was added to model this dependency. However, a modeling error was made. One of the requirements listed in Definition 3.7, i.e., the “option to complete” requirement, is not satisfied. The marking $[c_2, c_3]$ is reached by executing the firing sequence $\langle a, b \rangle$ and from this marking the desired end marking $[end]$ is no longer reachable. Note that $[c_2, c_3]$ is a dead marking, e.g., d is not enabled because c_6 is empty.

Definition 3.12 defines a soundness notion for C-nets. The notion of soundness can easily be adapted for other languages such as YAWL, EPCs, and BPMN. When

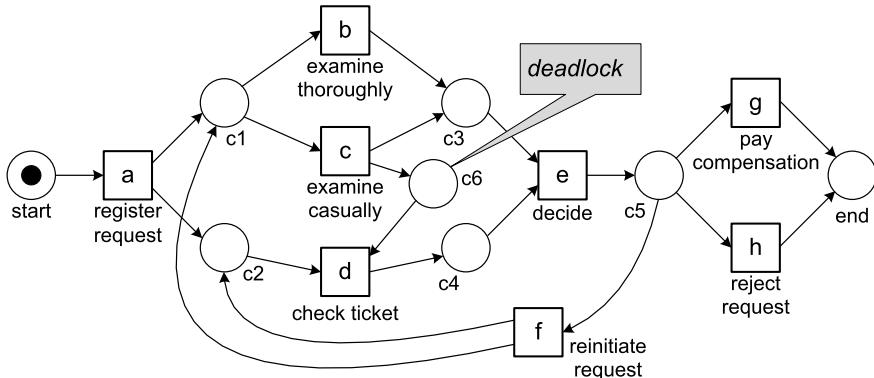


Fig. 3.19 A WF-net that is not sound

defining transition systems we already mentioned $S^{end} \subseteq S$ as the set of acceptable final states. Hence, we can define soundness as follows: a transition system is sound if and only if from any reachable state it is possible to reach a state in S^{end} . When introducing Petri nets we also defined generic properties such as liveness and boundedness. Some of these properties can be analyzed without constructing the state space. For example, for free-choice Petri nets, i.e., processes where choice and synchronization can be separated, liveness and boundedness can be checked by analyzing the rank of the corresponding incidence matrix [45]. Hence, soundness can be checked in polynomial time for free-choice WF-nets. Invariants can often be used to show boundedness or the unreachability of a particular marking. However, most of the more interesting verification questions require the exploration of (a part of) the state space.

Soundness is a generic property. Sometimes a more specific property needs to be investigated, e.g., “the ticket was checked for all rejected requests”. Such properties can be expressed in *temporal logic* [30, 93]. *Linear Temporal Logic* (LTL) is an example of a temporal logic that, in addition to classical logical operators, uses temporal operators such as: always (\Box), eventually (\Diamond), until (\sqcup), weak until (W), and next time (\bigcirc). The expression $\Diamond h \Rightarrow \Diamond d$ means that for all cases in which h (*reject request*) is executed also d (*check ticket*) is executed. Another example is $\Box(f \Rightarrow \Diamond e)$ that states that any occurrence of f will be followed by e . *Model checking* techniques can be used to check such properties [30].

Another verification task is the comparison of two models. For example, the implementation of a process needs is compared to the high-level specification of the process. As indicated before, there exist different equivalence notions (trace equivalence, branching bisimilarity, etc.) [176]. Moreover, there are also various simulation notions demanding that one model can “follow all moves” of the other but not vice versa (see also Sect. 6.3).

There are various tools to verify process models. A classical example is Woflan that is tailored towards checking soundness [179]. Also workflow systems such as YAWL provide verification capabilities. Consider, for example, the screenshot

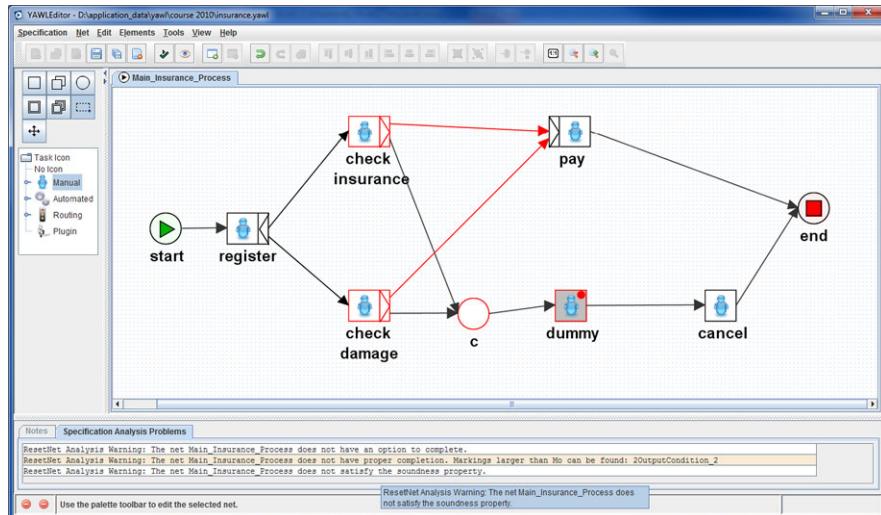


Fig. 3.20 An incorrect YAWL model: the cancelation region of *dummy* comprises of *check insurance*, *check damage*, condition *c* and the two implicit input conditions of *pay*. Hence, after cancellation, a token may be left on one of the output arcs of *register*

shown in Fig. 3.20. The figure shows the editor of YAWL while analyzing the model depicted. The process starts with task *register*. After this task, two checks can be done in parallel: *check insurance* and *check damage*. These tasks are XOR-splits; depending on the result of the check, one of the output arcs is selected. If both checks are OK, task *pay* is executed. If one of the checks indicates a problem, then the *dummy* task is executed. This task has a cancelation region consisting of *check insurance*, *check damage*, condition *c* and the two implicit input conditions of *pay*. The goal of this region is to remove all tokens, cancel the claim, and then end. However, the verifier of YAWL reports a problem. The YAWL model is not correct, because there may be a token pending in one of the implicit output conditions of *register*, i.e., there may be still a token on the arc connecting *register* and *check insurance* or on the arc connecting *register* and *check damage*. As a result the model may deadlock and “garbage” may be left behind. When these two implicit conditions are included in the cancelation region of the *dummy* task, then the verifier of YAWL will not find any problems and the model is indeed free of deadlocks and other anomalies.

3.3.2 Performance Analysis

The performance of a process or organization can be defined in different ways. Typically, three dimensions of performance are identified: *time*, *cost* and *quality*. For each of these performance dimensions different *Key Performance Indicators* (KPIs)

can be defined. When looking at the *time dimension* the following performance indicators can be identified:

- The *lead time* (also referred to as flow time) is the total time from the creation of the case to the completion of the case. In terms of a WF-net, this is the time it takes to go from source place i to sink place o . One can measure the average lead time over all cases. However, the degree of variance may also be important, i.e., it makes a difference whether all cases take more or less two weeks or if some take just a few hours whereas others take more than one month. The *service level* is the percentage of cases having a lead time lower than some threshold value, e.g., the percentage of cases handled within two weeks.
- The *service time* is the time actually worked on a case. One can measure the service time per activity, e.g., the average time needed to make a decision is 35 minutes, or for the entire case. Note that in case of concurrency the overall service time (i.e., summing up the times spent on the various activities) may be longer than the lead time. However, typically the service time is just a fraction of the lead time (minutes versus weeks).
- The *waiting time* is the time a case is waiting for a resource to become available. This time can be measured per activity or for the case as a whole. An example is the waiting time for a customer who wants to talk to a sales representative. Another example is the time a patient needs to wait before getting a knee operation. Again one may be interested in the average or variance of waiting times. It is also possible to focus on a service level, e.g., the percentage of patients that has a knee operation within three weeks after the initial diagnosis.
- The *synchronization time* is the time an activity is not yet fully enabled and waiting for an external trigger or another parallel branch. Unlike waiting time, the activity is not fully enabled yet, i.e., the case is waiting for synchronization rather than a resource. Consider, for example, a case at marking $[c2, c3]$ in the WF-net shown in Fig. 3.2. Activity e is waiting for *check ticket* to complete. The difference between the arrival time of the token in condition $c4$ and the arrival time of the token in condition $c3$ is the synchronization time.

Performance indicators can also be defined for the *cost dimension*. Different costing models can be used, e.g., Activity Based Costing (ABC), Time-Driven ABC, and Resource Consumption Accounting (RCA) [31]. The costs of executing an activity may be fixed or depend on the type of resource used, its utilization, or the duration of the activity. Resource costs may depend on the utilization of resources. A key performance indicator in most processes is the *average utilization* of resources over a given period, e.g., an operating room in a hospital has been used 85% of the time over the last two months. A detailed discussion of the various costing models is outside of the scope of this book.

The *quality dimension* typically focuses on the “product” or “service” delivered to the customer. Like costs, this can be measured in different ways. One example is customer satisfaction measured through questionnaires. Another example is the average number of complaints per case or the number of product defects.

Whereas verification focuses on the (logical) correctness of the modeled process, performance analysis aims at improving processes with respect to time, cost,

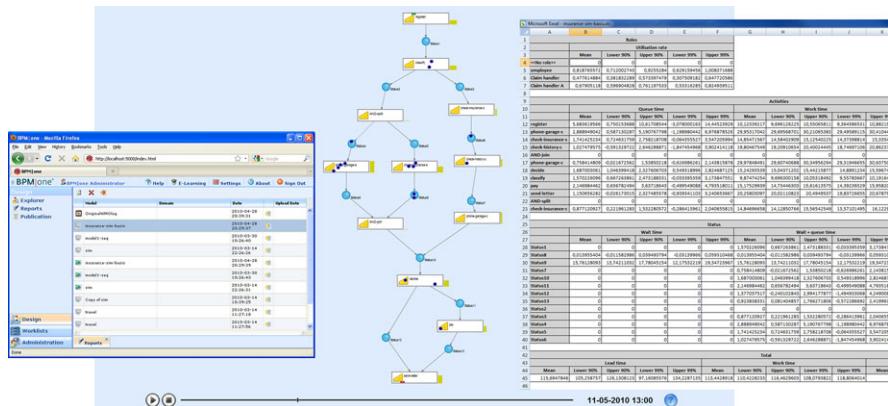


Fig. 3.21 Simulation using BPM|one of Pallas Athena: the modeled process can be animated and all kinds of KPIs of the simulated process are measured and stored in a spreadsheet

or quality. Within the context of operations management many analysis techniques have been developed. Some of these techniques “optimize” the model given a particular performance indicator. For example, integer programming or Markov decision problems can be used to find optimal policies. For the types of process models described in this chapter “what if” analyses using simulation, queueing models, or Markov models are most appropriate. Analytical models typically require many assumptions and can only be used to answer particular questions. Therefore, one needs to resort to *simulation*. Most BPM tools provide simulation capabilities. Figure 3.21 shows a screenshot of BPM|one while simulating a process for handling insurance claims. BPM|one can animate the simulation run and calculate all kinds of KPIs related to time and cost (e.g., lead time, service time, waiting time, utilization, and activity costs).

Although many organizations have tried to use simulation to analyze their business processes at some stage, *few are using simulation in a structured and effective manner*. This may be caused by a lack of training and limitations of existing tools. However, there are also several additional and more fundamental problems. First of all, simulation models tend to *oversimplify* things. In particular the behavior of resources is often modeled in a rather naïve manner. People do not work at constant speeds and need to distribute their attention over multiple processes. This can have dramatic effects on the performance of a process and, therefore, such aspects should not be “abstracted away” [139, 163]. Second, various *artifacts available are not used as input for simulation*. Modern organizations store events in logs and some may have accurate process models stored in their BPM/WFM systems. Also note that in many organizations, the state of the information system accurately reflects the state of the business processes supported by these systems. As discussed in Chap. 1, processes and information systems have become tightly coupled. Nevertheless, such information (i.e., event logs and status data) is rarely used for simulation or a lot of manual work is needed to feed this information into the model. Fortunately, as will

be shown later in this book, process mining can assist in extracting such information and use this to realize performance improvements (see Sect. 9.6). Third, the focus of simulation is mainly on “design” whereas managers would also like to use simulation for “*operational decision making*”, i.e., solving the concrete problem at hand rather than some abstract future problem. Fortunately, *short-term simulation* [139] can provide answers for questions related to “here and now”. The key idea is to start all simulation runs from the current state and focus on the analysis of the transient behavior. This way a “fast forward button” into the future is provided.

3.3.3 Limitations of Model-Based Analysis

Verification and performance analysis heavily rely on the availability of high quality models. When the models and reality have little in common, model-based analysis does not make much sense. For example, the process model can be internally consistent and satisfy all kinds of desirable properties. However, if the model describes an idealized version of reality, this is quite useless as in reality all kinds of deviations may take place. Similar comments hold for simulation models. It may be that the model predicts a significant improvement whereas in reality this is not the case because the model is based on flawed assumptions. All of these problems stem from *a lack of alignment between hand-made models and reality*. Process mining aims to address these problems by establishing a direct connection between the models and actual low-level event data about the process. Moreover, the *discovery techniques discussed in this book allow for viewing the same reality from different angles and at different levels of abstraction*.

Chapter 4

Data Mining

Process mining builds on two pillars: (a) process modeling and analysis (as described in Chap. 3) and (b) data mining. This chapter introduces some basic data mining approaches and structures the field. The motivation for doing so is twofold. On the one hand, some process mining techniques build on classical data mining techniques, e.g., discovery and enhancement approaches focusing on data and resources. On the other hand, ideas originating from the data mining field will be used for the evaluation of process mining results. For example, one can adopt various data mining approaches to measure the quality of the discovered or enhanced process models. Existing data mining techniques are of little use for control-flow discovery, conformance checking, and other process mining tasks. Nevertheless, a basic understanding of data mining is most helpful for fully understanding the process mining techniques presented in subsequent chapters.

4.1 Classification of Data Mining Techniques

In [69] data mining is defined as “the analysis of (often large) data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”. The input data is typically given as a table and the output may be rules, clusters, tree structures, graphs, equations, patterns, etc. The growth of the “digital universe” described in Chap. 2 is the main driver for the popularity of data mining. Initially, the term “data mining” had a negative connotation especially among statisticians. Terms like “data snooping”, “fishing”, and “data dredging” refer to ad-hoc techniques to extract conclusions from data without a sound statistical basis. However, over time the data mining discipline has become mature as characterized by solid scientific methods and many practical applications [9, 24, 69, 102, 190].

Table 4.1 Data set 1: Data about 860 recently deceased persons to study the effects of drinking, smoking, and body weight on the life expectancy

Drinker	Smoker	Weight	Age
yes	yes	120	44
no	no	70	96
yes	no	72	88
yes	yes	55	52
no	yes	94	56
no	no	62	93
...

4.1.1 Data Sets: Instances and Variables

Let us first look at three example data sets and possible questions. Table 4.1 shows part of a larger table containing information about 860 individuals that have recently deceased. For each person the age of death is recorded (column *age*). Column *drinker* indicates whether the person was drinking alcohol. Column *smoker* indicates whether the person was smoking. Column *weight* indicates the bodyweight of the deceased person. Each row in Table 4.1 corresponds to a person. Questions may be:

- What is the effect of smoking and drinking on a person's bodyweight?
 - Do people that smoke also drink?
 - What factors influence a person's life expectancy the most?
 - Can one identify groups of people having a similar lifestyle?

Table 4.2 shows another data set with information about 420 students that participated in a Bachelor program. Each row corresponds to a student. Students follow different courses. The table lists the highest mark for a particular course, e.g., the first student got a 9 for the course on linear algebra and an 8 for the course on logic. Table 4.2 uses the Dutch grading system, i.e., any mark is in-between 1 (lowest) and 10 (highest). Students who have a 5 or less, fail for the course. A “–” means

Table 4.2 Data set 2: Data about 420 students to investigate relationships among course grades and the student's overall performance in the Bachelor program

Table 4.3 Data set 3: Data on 240 customer orders in a coffee bar recorded by the cash register

Cappuccino	Latte	Espresso	Americano	Ristretto	Tea	Muffin	Bagel
1	0	0	0	0	0	1	0
0	2	0	0	0	0	1	1
0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	2	0
0	0	0	1	1	0	0	0
...

that the course was not taken. The table shows only a selection of courses. Besides mandatory courses there are dozens of elective courses. The last two columns refer to the overall performance. The *duration* column indicates how long the student was enrolled before getting a degree or dropping out. The *result* column shows the final result: cum laude, passed, or failed. The university may be interested in the following questions:

- Are the marks of certain courses highly correlated?
- Which electives do excellent students (cum laude) take?
- Which courses significantly delay the moment of graduation?
- Why do students drop out?
- Can one identify groups of students having a similar study behavior?

The third data set, partly shown in Table 4.3, contains data about 240 orders in a café. Each row corresponds to one customer order. The columns refer to products. For instance, the first customer ordered a cappuccino and a muffin. This example is quite generic and analyzing such a data set is generally referred to as *market basket analysis*. For example, one can think of analyzing the product combinations purchased in a supermarket or in an electronic bookstore. Cafés, supermarkets, bookstores, etc. may be interested in the following questions:

- Which products are frequently purchased together?
- When do people buy a particular product?
- Is it possible to characterize typical customer groups?
- How to promote the sales of products with a higher margin?

Tables 4.1, 4.2, and 4.3 show three typical *data sets* used as input for data mining algorithms. Such a data set is often referred to as *sample* or *table*. The rows in the three tables are called *instances*. Alternative terms are: *individuals*, *entities*, *cases*, *objects*, and *records*. Instances may correspond to deceased persons, students, customers, orders, orderlines, messages, etc. The columns in the three tables are called *variables*. Variables are often referred to as *attributes*, *features*, or *data elements*. The first data set (Table 4.1) has four variables: *drinker*, *smoker*, *weight*, and *age*.

We distinguish between *categorical* variables and *numerical* variables. Categorical variables have a limited set of possible values and can easily be enumerated, e.g.,

a Boolean variable that is either true or false. Numerical variables have an ordering and cannot be enumerated easily. Examples are temperature (e.g., 39.7 degrees centigrade), age (44 years), weight (56.3 kilograms), number of items (3 coffees), and altitude (11 meters below sea level). Categorical variables are typically subdivided into *ordinal* variables and *nominal* variables. Nominal variables have no logical ordering. For example Booleans (true and false), colors (Red, Yellow, Green), and EU countries (Germany, Italy, etc.) have no commonly agreed upon logical ordering. Ordinal variables have an ordering associated to it. For example, the *result* column in Table 4.2 refers to an ordinal variable that can have values “cum laude”, “passed”, and “failed”. For most applications it would make sense to consider the value “passed” in-between “cum laude” and “failed”.

Before applying any data mining technique the data is typically preprocessed, e.g., rows and columns may be removed for various reasons. For instance, columns with less relevant information should be removed beforehand to reduce the dimensionality of the problem. Instances that are clearly corrupted should also be removed. Moreover, the value of a variable for a particular instance may be missing or have the wrong type. This may be due to an error while recording the data, but it may also have a particular reason. For example, in Table 4.2 some course grades are missing (denoted by “–”). These missing values are not errors but contain valuable information. For some kinds of analysis, the missing course grade can be treated as “zero”, i.e., not taking the course is “lower” than the lowest grade. For other types of analysis it may be that the values in such a column are mapped onto “yes” (participated in the course) and “no” (the entries that now have a “–”).

When comparing Tables 4.1, 4.2, and 4.3 with the event log shown in Table 2.1 it becomes obvious that data mining techniques make less assumptions about the format of the input data than process mining techniques. For example, in Table 2.1 there are two notions, events and cases, rather than the single notion of an instance (i.e., row in table). Moreover, events are ordered in time whereas in Tables 4.1, 4.2, and 4.3 the ordering of the rows has no meaning. For particular questions it is possible to convert an event log into a simple data set for data mining. We will refer to this as *feature extraction*. Later, we will use feature extraction for various proposes, e.g., analyzing decisions in a discovered process models and clustering cases before process discovery so that each cluster has a dedicated process model.

After showing the basic input format for data mining and discussing typical questions, we classify data mining techniques into two main categories: *supervised learning* and *unsupervised learning*.

4.1.2 Supervised Learning: Classification and Regression

Supervised learning assumes *labeled data*, i.e., there is a *response variable* that labels each instance. For instance, in Table 4.2 the *result* column could be selected as the response variable. Hence, each student is labeled as “cum laude”, “passed”, or “failed”. The other variables are *predictor variables* and we are interested in

explaining the response variable in terms of the predictor variables. Sometimes the response variable is called the *dependent variable* and the predictor variables are called *independent variables*. The goal is to explain the dependent variable in terms of the independent variables. For example, we would like to predict the final result of a student in terms of the student's course grades.

Techniques for supervised learning can be further subdivided into *classification* and *regression* depending on the type of response variable (categorical or numerical).

Classification techniques assume a *categorical* response variable and the goal is to classify instances based on the predictor variables. Consider for example Table 4.1. We would like to classify people into the class of smokers and the class of non-smokers. Therefore, we select the categorical response variable *smoker*. Through classification we want to learn what the key differences between smokers and non-smokers are. For instance, we could find that most smokers drink and die young. By applying classification to the second data set (Table 4.2) while using column *result* as a response variable, we could find the obvious fact that cum laude students have high grades. In Sect. 4.2, we will show how to construct a so-called *decision tree* using classification.

Regression techniques assume a *numerical* response variable. The goal is to find a function that fits the data with the least error. For example, we could select *age* as response variable for the data set in Table 4.1 and (hypothetically) find the function $age = 124 - 0.8 \times weight$, e.g., a person of 50 kilogram is expected to live until the age of 84 whereas a person of 100 kilogram is expected to live until the age of 44. For the second data set we could find that the mark for the course on workflow systems heavily depends on the mark for linear algebra and logic, e.g., $workflow\ systems = 0.6 + 0.8 \times linear\ algebra + 0.2 \times logic$. For the third data set, we could (again hypothetically) find a function that predicts the number of bagels in terms of the numbers of different drinks.

The most frequently used regression technique is *linear regression*. Given a response variable y and predictor variables x_1, x_2, \dots, x_n a linear model $\hat{y} = f(x_1, x_2, \dots, x_n) = a_0 + \sum_{i=1}^n a_i x_i$ is learned over the data set. For every instance in the data set there is an error $|y - \hat{y}|$. A popular approach is to minimize the sum of squared errors, i.e., given m instances the goal is to find a function f such that $\sum_{j=1}^m (y_j - \hat{y}_j)^2$ is minimal. Other scoring functions are possible and more *general regression models* or even *neural networks* can be used. However, these techniques are out of the scope of this book and the interested reader is referred to [69].

Classification requires a categorical response variable. In some cases it makes sense to transform a numerical response variable into a categorical one. For example, for Table 4.1 one could decide to transform variable *age* into a categorical response variable by mapping values below 70 onto label "young" and values of 70 and above onto label "old". Now a decision tree can be constructed to classify instances into people that die(d) "young" and people that die(d) "old". Similarly, all values in Table 4.3 can be made categorical. For example, positive values are mapped onto "true" (the item was purchased) and value 0 is mapped onto "false" (the item was not purchased). After applying this mapping to Table 4.3, we can

apply classification to the coffee shop data while using e.g. column *muffin* as a response variable. We could, for instance, find that customers who drink lots of tea tend to eat muffins.

4.1.3 Unsupervised Learning: Clustering and Pattern Discovery

Unsupervised learning assumes *unlabeled data*, i.e., the variables are *not* split into response and predictor variables. In this chapter, we consider two types of unsupervised learning: *clustering* and *pattern discovery*.

Clustering algorithms examine the data to find groups of instances that are similar. Unlike classification the focus is not on some response variable but on the instance as a whole. For example, the goal could be to find homogeneous groups of students (Table 4.2) or customers (Table 4.3). Well-known techniques for clustering are *k-means clustering* and *agglomerative hierarchical clustering*. These will be briefly explained in Sect. 4.3.

There are many techniques to discover patterns in data. Often the goal is to find rules of the form *IF X THEN Y* where *X* and *Y* relate values of different variables. For example, *IF smoker = no AND age ≥ 70 THEN drinker = yes* for Table 4.1 or *IF logic ≤ 6 AND duration > 50 THEN result = failed* for Table 4.2. The most well-known technique is *association rule mining*. This technique will be explained in Sect. 4.4.

Note that decision trees can also be converted into rules. However, a decision tree is constructed for a particular response variable. Hence, rules extracted from a decision tree only say something about the response variable in terms of some of the predictor variables. Association rules are discovered using unsupervised learning, i.e., there is no need to select a response variable.

Data mining results may be both *descriptive* and *predictive*. Decision trees, association rules, regression functions say something about the data set used to learn the model. However, they can also be used to make predictions for new instances, e.g., predict the overall performance of students based on the course grades in the first semester.

In the remainder, we show some of the techniques mentioned in more detail. Moreover, at the end of this chapter we focus on measuring the quality of mining results.

4.2 Decision Tree Learning

Decision tree learning is a supervised learning technique aiming at the classification of instances based on predictor variables. There is one categorical response variable labeling the data and the result is arranged in the form of a tree. Figures 4.1, 4.2, and 4.3 show three decision trees computed for the data sets described earlier in this chapter. Leaf nodes correspond to possible values of the response variable. Non-leaf

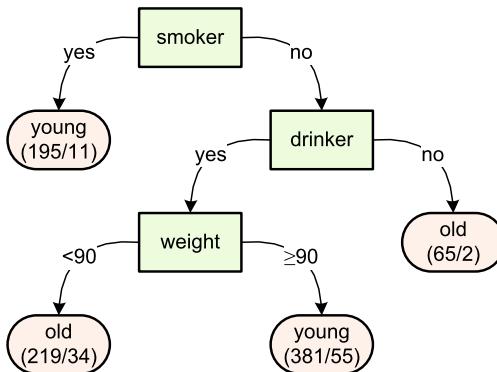


Fig. 4.1 A decision tree derived from Table 4.1. The 860 persons are classified into “young” (died before the age of 70) and “old” (died at 70 or later). People who smoke generally die young (195 persons of which 11 are misclassified). People who do not smoke and do not drink tend to live long (65 persons of which 2 are misclassified). People who only drink but are overweight (≥ 90) also die young (381 persons of which 55 are misclassified)

nodes correspond to predictor variables. In the context of decision tree learning, predictor variables are referred to as *attributes*. Every attribute node splits a set of instances into two or more subsets. The root node corresponds to all instances.

In Fig. 4.1, the root node represents all instances; in this case 860 persons. Based on the attribute *smoker* these instances are split into the ones that are smoking (195 persons) and the ones that not smoking ($860 - 195 = 665$ persons). The smokers are not further split. Based on this information instances are already labeled as “young”, i.e., smokers are expected to die before the age of 70. The non-smokers are split into drinkers and non-drinkers. The latter group of people is expected to live long and is thus labeled as “old”. All leaf nodes have two numbers. The first number indicates the number of instances classified as such. The second number indicates the number of instances corresponding to the leaf node but wrongly classified. Of the 195 smokers who were classified as “young” 11 people were misclassified, i.e., did not die before 70 while smoking.

The other two decision trees can be read in the same manner. Based on an attribute, a set of instances may also be split into three (or even more) subsets. An attribute may appear multiple times in a tree but not twice on the same path. For example, in Fig. 4.2 there are two nodes referring to the course on linear algebra. However, these are not on the same path and thus refer to disjoint sets of students. As mentioned before there are various ways to handle missing values depending on their assumed semantics. In Fig. 4.2, a missing course grade is treated as a kind of “zero” (see the left-most arc originating from the root node).

Decision trees such as the ones shown in Figs. 4.1, 4.2, and 4.3 can be obtained using a variety of techniques. Most of the techniques use a recursive top-down algorithm that works as follows:

1. Create the root node r and associate all instances to the root node. $X := \{r\}$ is the set of nodes to be traversed.

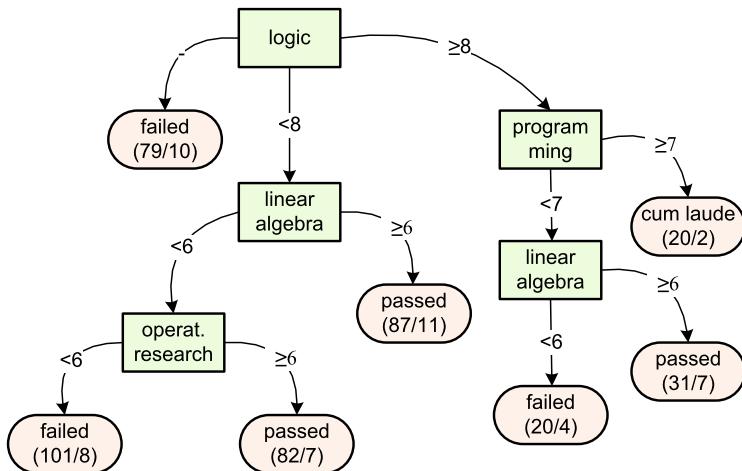
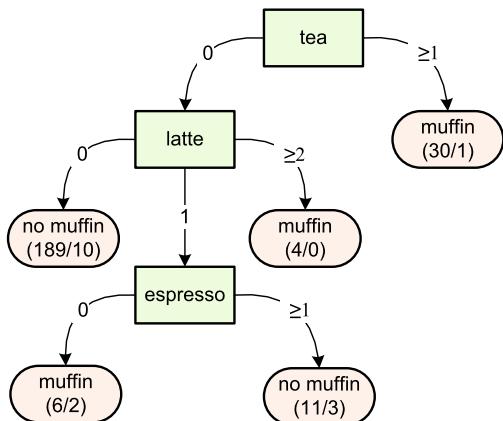


Fig. 4.2 A decision tree derived from Table 4.2. The 420 students are classified into “failed”, “passed”, and “cum laude” based on study results. Students that do not take the course on logic typically fail (79 students of which 10 are misclassified). Students that have a high mark for logic and programming, typically complete their degree cum laude (20 students of which 2 are misclassified)

Fig. 4.3 A decision tree derived from Table 4.3 after converting response variable *muffin* into a Boolean.
Customers who drink tea tend to eat muffins (30 customers of which 1 is misclassified). Customers who do not drink tea or latte typically do not eat muffins (189 customers of which 10 are misclassified)



2. If $X = \emptyset$, then return the tree with root r and end.
3. Select $x \in X$ and remove it from X , i.e., $X := X \setminus \{x\}$. Determine the “score” $s^{old}(x)$ of node x before splitting, e.g., based on entropy.
4. Determine if splitting is possible/needed. If not, go to step 2, otherwise continue with the next step.
5. For all possible attributes $a \in A$, evaluate the effects of splitting on the attribute. Select the attribute a providing the best improvement, i.e., maximize $s_a^{new}(x) - s^{old}(x)$. The same attribute should not appear multiple times on the same path

from the root. Also note that for numerical attributes, so-called “cut values” need to be determined (cf. < 8 and ≥ 8 in Fig. 4.2).

6. If the improvement is substantial enough, create a set of child nodes Y , add Y to X (i.e., $X := X \cup Y$), and connect x to all child nodes in Y .
7. Associate each node in Y to its corresponding set of instances and go to step 2.

Here, we only provide a rough sketch of the generic algorithm. Many design decisions are needed to make a concrete decision tree learner. For example, one needs to decide when to stop adding nodes. This can be based on the improvement of the scoring function or because the tree is restricted to a certain depth. There are also many ways to select attributes. This can be based on entropy (see below), the Gini index of diversity, etc. When selecting a numeric attribute to split on, cut values need to be determined because it is unreasonable/impossible to have a child node for every possible value. For example, a customer can purchase any number of latte's and it would be undesirable to enumerate all possibilities when using this attribute to split. As shown in Fig. 4.2, node *latte* has only three child nodes based on two cut values partitioning the domain of natural numbers in $\{0\}$, $\{1\}$, and $\{2, 3, \dots\}$.

These are just few of the many ingredients that determine a complete decision tree learning algorithm.

The crucial thing to see is that by *splitting the set of instances in subsets the variation within each subset becomes smaller*. This can be best illustrated using the notion of *entropy*.

Entropy: Encoding uncertainty

Entropy is an information-theoretic measure for the uncertainty in a multi-set of elements. If the multi-set contains many different elements and each element is unique, then variation is maximal and it takes many “bits” to encode the individual elements. Hence, the entropy is “high”. If all elements in the multi-set are the same, then actually no bits are needed to encode the individual elements. In this case the entropy is “low”. For example, the entropy of the multi-set $[a, b, c, d, e]$ is much higher than the entropy of the multi-set $[a^5]$ even though both multi-sets have the same number of elements (5).

Assume that there is a multi-set X with n elements and there are k possible values, say v_1, v_2, \dots, v_k , i.e., X is a multi-set over $V = \{v_1, v_2, \dots, v_k\}$ with $|X| = n$. Each value v_i appears c_i times in X , i.e., $X = [(v_1)^{c_1}, (v_2)^{c_2}, \dots, (v_k)^{c_k}]$. Without loss of generality, we can assume that $c_i \geq 1$ for all i , because values that do not appear in X can be removed from V upfront. The proportion of elements having value v_i is p_i , i.e., $p_i = c_i/n$. The entropy of X is measured in bits of information and is defined by the formula:

$$E = - \sum_{i=1}^k p_i \log_2 p_i$$

If all elements in X have the same value, i.e., $k = 1$ and $p_1 = 1$, then $E = -\log_2 1 = 0$. This means that no bits are needed to encode the value of an individual element; they are all the same anyway. If all elements in X are different, i.e., $k = n$ and $p_i = 1/k$, then $E = -\sum_{i=1}^k (1/k) \log_2(1/k) = \log_2 k$. For instance, if there are 4 possible values, then $E = \log_2 4 = 2$ bits are needed to encode each individual element. If there are 16 possible values, then $E = \log_2 16 = 4$ bits are needed to encode each individual element.

The proportion p_i can also be seen as a probability. Assume there is random stream of values such that there are four possible values $V = \{a, b, c, d\}$, e.g., a sequence like *bacaababadabaacada...* is generated. Value a has a probability of $p_1 = 0.5$, value b has a probability of $p_2 = 0.25$, value c has a probability of $p_3 = 0.125$, and value d has a probability of $p_4 = 0.125$. In this case $E = -((0.5 \log_2 0.5) + (0.25 \log_2 0.25) + (0.125 \log_2 0.125) + (0.125 \log_2 0.125)) = -((0.5 \times -1) + (0.25 \times -2) + (0.125 \times -3) + (0.125 \times -3)) = 0.5 + 0.5 + 0.375 + 0.375 = 1.75$ bits. This means that on average 1.75 bits are needed to encode one element. This is correct. Consider, for example, the following variable length binary encoding $a = 0$, $b = 11$, $c = 100$, and $d = 111$, i.e., a is encoded in one bit, b is encoded in two bits, and c and d are each encoded in three bits. Given the relative frequencies it is easy to see that this is (on average) the most compact encoding. Other encodings are either similar (e.g., $a = 1$, $b = 00$, $c = 011$, and $d = 000$) or require more bits on average. Suppose now that all four values have the same probability, i.e., $p_1 = p_2 = p_3 = p_4 = 0.25$. In this case $E = \log_2 4 = 2$. This is correct because there is no way to improve the encoding $a = 00$, $b = 01$, $c = 10$, and $d = 11$.

The example shows that by using information about the probability of each value, we can reduce the encoding from 2 bits to 1.75 bits on average. If the probabilities are more skewed, further reductions are possible. If value a has a probability of $p_1 = 0.9$, value b has a probability of $p_2 = 0.1$, value c has a probability of $p_3 = 0.05$, and value d has a probability of $p_4 = 0.05$, then $E = 0.901188$. This means that on average less than one bit is needed to encode each element.

Let us now apply the notion of entropy to decision tree learning. Fig. 4.4 shows three steps in the construction of a decision tree for the data set shown in Table 4.1. We label the instances into “old” and “young”. Moreover, for simplicity we abstract from the *weight* attribute. In the initial step, the tree consists only of a root. Since the majority of persons in our data set die before 70, we label this node as young. Since of the 860 persons in our data set only 546 actually die before 70, the remaining 314 persons are misclassified. Let us calculate the entropy for the root node: $E = -(((546/860) \log_2(546/860)) + ((314/860) \log_2(314/860))) = 0.946848$. This is a value close to the maximal value of one (in case both groups would have the same size).

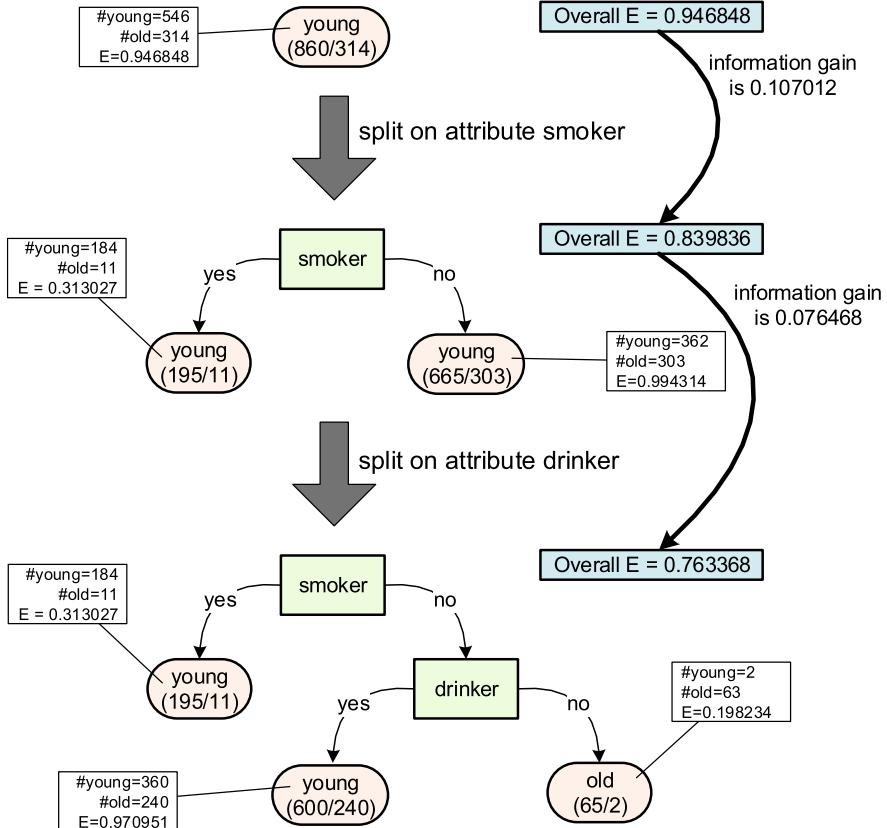


Fig. 4.4 Step-by-step construction of decision tree driven by information gain based on entropy

Next, Fig. 4.4 shows what happens if we split the data set based on attribute *smoker*. Now there are two leaf nodes both bearing the label *young*. Of the people that smoke (195), most die young (184). Hence, the entropy of this leaf node is very small: $E = -(((184/195)\log_2(184/195)) + ((11/195)\log_2(11/195))) = 0.313027$. This means that the variability is much smaller. The other leaf node is more heterogeneous: about half of the 665 non smokers (362 to be precise) die young. Indeed $E = -(((362/665)\log_2(362/665)) + ((303/665)\log_2(303/665))) = 0.994314$ is higher. However, the overall entropy is still lower. The overall entropy can be found by simply taking the weighted average, i.e., $E = (195/860) \times 0.313027 + (665/860) \times 0.994314 = 0.839836$.

As Fig. 4.4 shows the *information gain* is 0.107012. This is calculated by taking the old overall entropy (0.946848) minus the new overall entropy (0.839836). Note that still all persons are classified as *young*. However, we gained information by splitting on attribute *smoker*. The information gain, i.e., a reduction in entropy, was obtained because we were able to find a group of persons for which there is

less variability; most smokers die young. The goal is to *maximize the information gain* by selecting a particular attribute to split on. Maximizing the information gain corresponds to minimizing the entropy and heterogeneity in leaf nodes. We could also have chosen the attribute *drinker* first. However, this would have resulted in a smaller information gain.

The lower part of Fig. 4.4 shows what happens if we split the set of non-smokers based on attribute *drinker*. This results in two new leaf nodes. The node that corresponds to persons who do not smoke and do not drink has a low entropy value ($E = 0.198234$). This can be explained by the fact that indeed most of the people associated to this leaf node live long and there are only two exceptions to this rule. The entropy of the other new leaf node (people that drink but do not smoke) is again close to one. However, the overall entropy is clearly reduced. The information gain is 0.076468. Since we abstract from the *weight* attribute we cannot further split the leaf node corresponding to people that drink but do not smoke. Moreover, it makes no sense to split the leaf node with smokers because little can be gained as the entropy is already low.

Note that splitting nodes will always reduce the overall entropy. In the extreme case all the leaf nodes correspond to single individuals (or individuals having exactly the same attribute values). The overall entropy is then by definition zero. However, the resulting tree is not very useful and probably has little predictive value. It is vital to realize that the decision tree is learned based on *examples*. For instance, if in the data set no customer ever ordered six muffins, this does not imply that this is not possible. A decision tree is “overfitting” if it depends too much on the particularities of the data used to learn it (see also Sect. 4.6). An overfitting decision tree is overly complex and performs poorly on unseen instances. Therefore, it is important to select the right attributes and to stop splitting when little can be gained.

Entropy is just one of several measures that can be used to measure the diversity in a leaf node. Another measure is the *Gini index of diversity* that measures the “impurity” of a data set: $G = 1 - \sum_{i=1}^k (p_i)^2$. If all classifications are the same, then $G = 0$. G approaches 1 as there is more and more diversity. Hence, an approach can be to select the attribute that maximizes the reduction of the G value (rather than the E value).

See [9, 24, 69, 190] for more information (and pointers to the extensive literature) on the different strategies to build decision trees.

Decision tree learning is unrelated to process discovery, however it can be used in combination with process mining techniques. For example, process discovery techniques such as the α -algorithm help to locate all decision points in the process (e.g., the XOR/OR-splits discussed in Chap. 3). Subsequently, we can analyze each decision point using decision tree learning. The response variable is the path taken and the attributes are the data elements known at or before the decision point.

4.3 *k*-Means Clustering

Clustering is concerned with grouping instances into *clusters*. Instances in one cluster should be similar to each other and dissimilar to instances in other clusters. Clus-

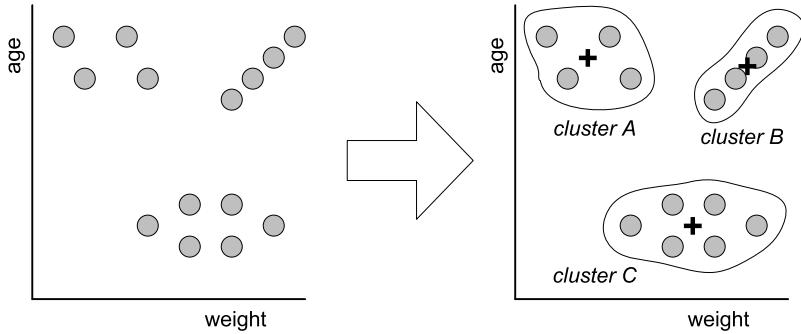


Fig. 4.5 Clustering instances in three clusters using k -means

tering uses unlabeled data and, hence, requires an unsupervised learning technique. Many clustering algorithms exist [9, 24, 69, 102, 190]. Here, we focus on *k-means clustering*.

Figure 4.5 illustrates the basic idea of clustering. Assume we have a data set with only two variables: *age* and *weight*. Such a data set could be obtained by projecting Table 4.1 onto the last two columns. The dots correspond to persons having a particular age and weight. Through a clustering technique like *k*-means, the three *clusters* shown on the right-hand-side of Fig. 4.5 can be discovered. Ideally, the instances in one cluster are close to one another while being further away from instances in other clusters. Each of the clusters has a *centroid* denoted by a +. The centroid denotes the “center” of the cluster and can be computed by taking the average of the coordinates of the instances in the cluster. Note that Fig. 4.5 shows only two dimensions. This is a bit misleading as typically there will be many dimensions (e.g., the number of courses or products). However, the two dimensional view helps to understand the basic idea.

Distance-based clustering algorithms like *k*-means and agglomerative hierarchical clustering assume a *distance notion*. The most common approach is to consider each instance to be an n -dimensional vector where n is the number of variables and then simply take the Euclidian distance. For this purpose ordinal values but also binary values need to be made numeric, e.g., *true* = 1, *false* = 0, *cum laude* = 2, *passed* = 1, *failed* = 0. Note that scaling is important when defining a distance metric. For example, if one variable represents the distance in meters ranging from 10 to 1,000,000 while another variable represents some utilization factor ranging from 0.2 to 0.8, then the distance variable will dominate the utilization variable. Hence, some normalization is needed.

Figure 4.6 shows the basic idea of *k*-means clustering. Here, we simplified things as much as possible, i.e., k = 2 and there are only 10 instances. The approach starts with a random initialization of two centroids denoted by the two + symbols. In Fig. 4.6(a) the centroids are randomly put onto the two dimensional space. Using the selected distance metric, all instances are assigned to the closest centroid. Here we use the standard Euclidian distance. All instances with an open dot are assigned to the centroid on the left whereas all the instances with a closed dot are assigned

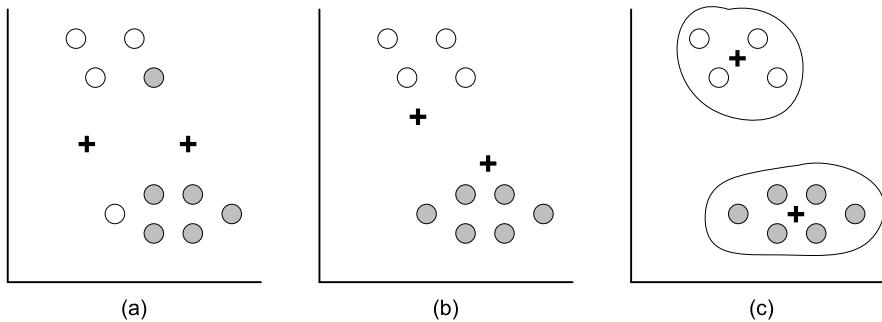


Fig. 4.6 Step-by-step evolution k -means

to the centroid on the right. Based on this assignment we get two initial clusters. Now we compute the real center of each cluster. These form the new positions of the two centroids. The centroids in Fig. 4.6(b) are based on the clusters shown in Fig. 4.6(a). In Fig. 4.6(b) we again assign all instances to the centroid that is closest. This results in the two new clusters shown in Fig. 4.6(b). All instances with an open dot are assigned to one centroid whereas all the instances with a closed dot are assigned to the other one. Now we compute the real centers of these two new clusters. This results in a relocation of the centroids as shown in Fig. 4.6(c). Again we assign the instances to the centroid that is closest. However, now nothing changes and the location of the centroids remains the same. After converging the k -means algorithm outputs the two clusters and related statistics.

The quality of a particular clustering can be defined as the average distance from an instance to its corresponding centroid. k -means clustering is only a heuristic and does not guarantee that it finds the k clusters that minimize the average distance from an instance to its corresponding centroid. In fact, the result depends on the initialization. Therefore, it is good to repeatedly execute the algorithm with different initializations and select the best one.

There are many variants of the algorithm just described. However, we refer to standard literature for details [9, 24, 69, 102, 190]. One of the problems when using the k -means algorithm is determining the number of clusters k . For k -means this is fixed from the beginning. Note that the average distance from an instance to its corresponding centroid decreases as k is increased. In the extreme case every instance has its own cluster and the average distance from an instance to its corresponding centroid is zero. This is not very useful. Therefore, a frequently used approach is to start with a small number of clusters and then gradually increase k as long as there are significant improvements.

Another popular clustering technique is *Agglomerative Hierarchical Clustering* (AHC). Here, a variable number of clusters is generated. Figure 4.7 illustrates the idea. The approach works as follows. Assign each instance to a dedicated singleton cluster. Now search for the two clusters that are closest to one another. Merge these two clusters into a new cluster. For example, the initial clusters consisting of just a and just b are merged into a new cluster ab . Now search again for the two clusters

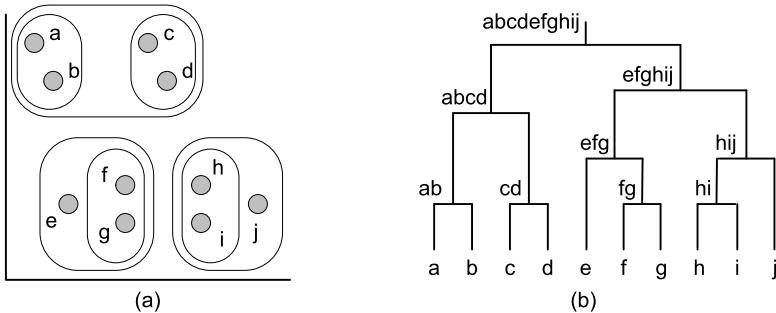


Fig. 4.7 Agglomerative hierarchical clustering: (a) clusters and (b) dendrogram

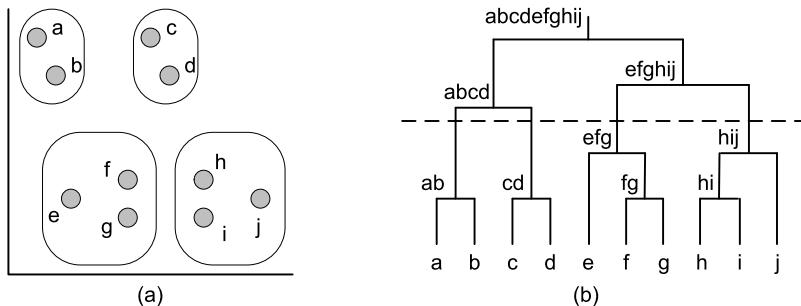


Fig. 4.8 Any horizontal line in dendrogram corresponds to a concrete clustering at a particular level of abstraction

that are closest to one another and merge them. This is repeated until all instances are in the same cluster. Figure 4.7(a) shows all intermediate clusters, i.e., all except the initial singleton clusters and the final overall cluster. Because of the hierarchical nature of the agglomerative hierarchical clustering we can visualize the clusters using a so-called *dendrogram* as shown in Fig. 4.7(b).

Any horizontal line cutting through the dendrogram corresponds to a concrete clustering. For example, Fig. 4.8(b) shows such a horizontal line. The clusters resulting from this are shown in Fig. 4.8(a). Moving the line to the bottom of the dendrogram results in many singleton clusters. Moving the line all the way up results in a single cluster containing all instances. By moving the horizontal line, the user can vary the abstraction level.

Clustering is only indirectly related to process discovery as described in Chap. 2. Nevertheless, clustering can be used as a preprocessing step for process mining [13, 62, 78]. By grouping similar cases together it may be possible to construct partial process models that are easier to understand. If the process model discovered for all cases is too complex to comprehend, then it may be useful to first identify clusters and then discover simpler models per cluster.

4.4 Association Rule Learning

Decision trees can be used to predict the value of some response variable that has been identified as being important. Driven by the response variable, rules like “people who drink and smoke die before 70” can be found. *Association rule learning* aims at finding similar rules but now *without* focusing on a particular response variable. The goal is to find rules of the form *IF X THEN Y* where *X* is often called the *antecedent* and *Y* the *consequent*. Such rules are also denoted as $X \Rightarrow Y$. *X* and *Y* can be any conjunction of “*variable = value*” terms. The only requirement is that *X* and *Y* are nonempty and any variable appears at most once in *X* and *Y*. Examples are *IF smoker = no AND age ≥ 70 THEN drinker = yes* for Table 4.1 or *IF logic ≤ 6 AND duration > 50 THEN result = failed* for Table 4.2. Typically, only categorical variables are considered. However, there are various techniques to transform numerical variables in categorical ones.

When learning association rules of the form $X \Rightarrow Y$, three metrics are frequently used: *support*, *confidence*, and *lift*. Let N_X be the number of instances for which *X* holds. N_Y is the number of instances for which *Y* holds. $N_{X \wedge Y}$ is the number of instances for which both *X* and *Y* hold. N is the total number of instances. The support of a rule $X \Rightarrow Y$ is defined as

$$\text{support}(X \Rightarrow Y) = N_{X \wedge Y} / N$$

The support indicates the applicability of the approach, i.e., the fraction of instances for which both antecedent and consequent hold. Typically a rule with high support is more useful than a rule with low support.

The confidence of a rule $X \Rightarrow Y$ is defined as

$$\text{confidence}(X \Rightarrow Y) = N_{X \wedge Y} / N_X$$

A rule with high confidence, i.e., a value close to 1, indicates that the rule is very reliable, i.e., if *X* holds, then *Y* will also hold. A rule with high confidence is definitely more useful than a rule with low confidence.

The lift of a rule $X \Rightarrow Y$ is defined as

$$\text{lift}(X \Rightarrow Y) = \frac{N_{X \wedge Y} / N}{(N_X / N) (N_Y / N)} = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

If *X* and *Y* are independent, then the lift will be close to 1. If $\text{lift}(X \Rightarrow Y) > 1$, then *X* and *Y* correlate positively. For example $\text{lift}(X \Rightarrow Y) = 5$ means that *X* and *Y* happen five times more together than what would be the case if they were independent. If $\text{lift}(X \Rightarrow Y) < 1$, then *X* and *Y* correlate negatively (i.e., the occurrence of *X* makes *Y* less likely and vice versa). Rules with a higher lift value are generally considered to be more interesting. However, typically lift values are only considered if certain thresholds with respect to support and confidence are met.

In the remainder of this section, we restrict ourselves to a special form of association rule learning known as *market basket analysis*. Here we only consider binary

variables that should be interpreted as present or not. For example, let us consider the first two columns in Table 4.1. This data set can be rewritten to so called *item-sets*: $\{ \{drinker, smoker\}, \{ \}, \{drinker\}, \{drinker, smoker\}, \{smoker\}, \{ \}, \dots \}$. If we ignore the number of items ordered in Table 4.3, then it is also straightforward to rewrite this data set in terms of item-sets: $\{ \{cappuccino, muffin\}, \{latte, muffin, bagel\}, \{espresso\}, \{cappuccino\}, \{tea, muffin\}, \{americano, ristretto\}, \dots \}$. The latter illustrates why the term “market basket” analysis is used for systematically analyzing such input. Based on item-sets, the goal is to generate rules of the form $X \Rightarrow Y$ where X and Y refer to disjoint non-empty sets of items. For example, $smoker \Rightarrow drinker$, $tea \wedge latte \Rightarrow muffin$, and $tea \Rightarrow muffin \wedge bagel$. Recall that there are $N = 240$ customer orders in Table 4.3. Assume that $N_{tea} = 50$ (i.e., 50 orders included at least one cup of tea), $N_{latte} = 40$, $N_{muffin} = 40$, $N_{tea \wedge latte} = 20$, and $N_{tea \wedge latte \wedge muffin} = 15$ (i.e., 15 orders included at least one tea, at least one latte, and at least one muffin). Let us consider the rule $tea \wedge latte \Rightarrow muffin$, i.e., $X = tea \wedge latte$ and $Y = muffin$. Given the numbers indicated we can easily compute the three metrics defined earlier:

$$support(X \Rightarrow Y) = N_{X \wedge Y} / N = N_{tea \wedge latte \wedge muffin} / N = 15 / 240 = 0.0625$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y} / N_X = N_{tea \wedge latte \wedge muffin} / N_{tea \wedge latte} = 15 / 20 = 0.75$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y} = \frac{N_{tea \wedge latte \wedge muffin} N}{N_{tea \wedge latte} N_{muffin}} = \frac{15 \times 240}{20 \times 40} = 4.5$$

Hence the $tea \wedge latte \Rightarrow muffin$ has a support of 0.0625, a confidence of 0.75, and a lift of 4.5.

If we also assume that $N_{tea \wedge muffin} = 25$, then we can deduce that the rule $tea \Rightarrow muffin$ has a support of 0.104167, a confidence of 0.5, and a lift of 3. Hence, this more compact rule has a better support but lower confidence and lift.

Let us also assume that $N_{latte \wedge muffin} = 35$. This implies that the rule $tea \Rightarrow latte \wedge muffin$ has a support of 0.0625, a confidence of 0.3, and a lift of 2.057. This rule has a rather poor performance compared to the original rule $tea \wedge latte \Rightarrow muffin$: the support is the same, but the confidence and lift are much lower.

To systematically generate association rules, one typically defines two parameters: *minsup* and *minconf*. The support of any rule $X \Rightarrow Y$ should be above the threshold *minsup*, i.e., $support(X \Rightarrow Y) \geq minsup$. Similarly the confidence of any rule $X \Rightarrow Y$ should be above the threshold *minconf*, i.e., $confidence(X \Rightarrow Y) \geq minconf$. Association rules can now be generated as follows:

1. Generate all frequent item-sets, i.e., all sets Z such that $N_Z / N \geq minsup$ and $|Z| \geq 2$.
2. For each frequent item-set Z consider all partitionings of Z into two non-empty subsets X and Y . If $confidence(X \Rightarrow Y) \geq minconf$, then keep the rule $X \Rightarrow Y$. If $confidence(X \Rightarrow Y) < minconf$, then discard the rule.
3. Output the rules found.

This simple algorithm has two problems. First of all, there is a computational problem related to the first step. If there are m variables, then there are $2^m - m - 1$

possible item-sets. Hence, for 100 products ($m = 100$) there are

1267650600228229401496703205275

candidate frequent item-sets. The second problem is that many uninteresting rules are generated. For example, after presenting the rule $tea \wedge latte \Rightarrow muffin$, there is no point in also showing $tea \Rightarrow latte \wedge muffin$ even when it meets the $minsup$ and $minconf$ thresholds. Many techniques have been developed to speed-up the generation of association rules and to select the most interesting rules. Here we only sketch the seminal *Apriori algorithm*.

Apriori: Efficiently generating frequent item-sets

The Apriori algorithm is one of the best known algorithms in computer science. The algorithm, initially developed by Agrawal and Srikant [7], is able to speed up the generation of association rules by exploiting the following two observations:

1. If an item-set is *frequent* (i.e., an item-set with a support above the threshold), then all of its non-empty subsets are also frequent. Formally, for any pair of non-empty item-sets X, Y : if $Y \subseteq X$ and $N_X/N \geq minsup$, then $N_Y/N \geq minsup$.
2. If, for any k , I_k is the set of all frequent item-sets with cardinality k and $I_l = \emptyset$ for some l , then $I_k = \emptyset$ for all $k \geq l$.

These two properties can be used to dramatically reduce the search-space when constructing the set of frequent item-sets. For example, if item-set $\{a, b\}$ is infrequent, then it does not make any sense to look at item-sets containing both a and b . The Apriori algorithm works as follows:

1. Create I_1 . This is the set of singleton frequent item-sets, i.e., item-sets with a support above the threshold $minsup$ containing just one element.
2. $k := 1$.
3. If $I_k = \emptyset$, then output $\bigcup_{i=1}^k I_i$ and end. If $I_k \neq \emptyset$, continue with the next step.
4. Create C_{k+1} from I_k . C_{k+1} is the candidate set containing item-sets of cardinality $k + 1$. Note that one only needs to consider elements that are the union of two item-sets A and B in I_k such that $|A \cap B| = k$ and $|A \cup B| = k + 1$.
5. For each candidate frequent item-set $c \in C_{k+1}$: examine all subsets of c with k elements; delete c from C_{k+1} if any of the subsets is not a member of I_k .
6. For each item-set c in the pruned candidate frequent item-set C_{k+1} , check whether c is indeed frequent. If so, add c to I_{k+1} . Otherwise, discard c .

7. $k := k + 1$ and return to Step 3.

The algorithm only considers candidates for I_{k+1} that are not ruled out by evidence in I_k . This way the number of traversals through the data set is reduced dramatically.

Association rules are related to process discovery. Recall that the α -algorithm also traverses the event log looking for patterns. However, association rules do not consider the ordering of activities and do not aim to build an overall process model.

4.5 Sequence and Episode Mining

The Apriori algorithm uses the monotonicity property that all subsets of a frequent item-set are also frequent. Many other pattern or rule discovery problems have similar monotonicity properties, thus enabling efficient implementations. A well-known example is the *mining of sequential patterns*. After introducing sequence mining, we also describe an approach to *discover frequent episodes* and mention some other data mining techniques relevant for process mining.

4.5.1 Sequence Mining

The Apriori algorithm does not consider the ordering of events. Sequence mining overcomes this problem by analyzing sequences of item-sets. One of the early approaches was developed by Srikant and Agrawal [131]. Here we sketch the essence of this approach. To explain the problem addressed by sequence mining, we consider the data set shown in Table 4.4. Each line corresponds to a customer ordering a set of items, e.g., at 9.02 on January 2nd 2011, Wil ordered a cappuccino, one day later he orders an espresso and a muffin. Per customer there is a sequence of orders. Orders have a sequence number, a timestamp, and an item-set. A more compact representation of the first customer sequence is $\langle \{cappuccino\}, \{espresso, muffin\}, \{americano, cappuccino\}, \{espresso, muffin\}, \{cappuccino\}, \{americano, cappuccino\} \rangle$. The goal is to find frequent sequences defined by a pattern like $\langle \{cappuccino\}, \{espresso, muffin\}, \{cappuccino\} \rangle$. A sequence is frequent if the pattern is contained in a predefined proportion of the customer sequences in the data set.

A sequence $\langle a_1, a_2, \dots, a_n \rangle$ is a *subsequence* of another sequence $\langle b_1, b_2, \dots, b_m \rangle$ if there exist integers $i_1 < i_2 < \dots < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, \dots , $a_n \subseteq b_{i_n}$. For example, the sequence $\langle \{x\}, \{x, y\}, \{y\} \rangle$ is a subsequence of $\langle \{z\}, \{x\}, \{z\}, \{x, y, z\}, \{y, z\}, \{z\} \rangle$ because $\{x\} \subseteq \{x\}$, $\{x, y\} \subseteq \{x, y, z\}$, and $\{y\} \subseteq \{y, z\}$. However, $\langle \{x\}, \{y\} \rangle$ is not a subsequence of $\langle \{x, y\} \rangle$ and vice versa. The

Table 4.4 A fragment of a data set used for sequence mining: each line corresponds to an order

Customer	Seq. number	Timestamp	Items
Wil	1	02-01-2011:09.02	{cappuccino}
	2	03-01-2011:10.06	{espresso, muffin}
	3	05-01-2011:15.12	{americano, cappuccino}
	4	06-01-2011:11.18	{espresso, muffin}
	5	07-01-2011:14.24	{cappuccino}
	6	07-01-2011:14.24	{americano, cappuccino}
Mary	1	30-12-2010:11.32	{tea}
	2	30-12-2010:12.12	{cappuccino}
	3	30-12-2010:14.16	{espresso, muffin}
	4	05-01-2011:11.22	{bagel, tea}
Bill	1	30-12-2010:14.32	{cappuccino}
	2	30-12-2010:15.06	{cappuccino}
	3	30-12-2010:16.34	{bagel, espresso, muffin}
	4	06-01-2011:09.18	{ristretto}
	5	06-01-2011:12.18	{cappuccino}
...

support of a sequence s is the fraction of sequences in the data set that has s as a subsequence. A sequence is *frequent* if its support meets some threshold minsup . Consider, for example, the data sets consisting of just the three visible customer sequences in Table 4.4. Pattern $\langle\{\text{tea}\}, \{\text{bagel, tea}\}\rangle$ has a support of $1/3$ as it is only a subsequence of Mary's sequence. Pattern $\langle\{\text{espresso}\}, \{\text{cappuccino}\}\rangle$ has a support of $2/3$ as it is a subsequence of both Wil's and Bill's subsequences, but not a subsequence of Mary's sequence. Pattern $\langle\{\text{cappuccino}\}, \{\text{espresso, muffin}\}\rangle$ has a support of $3/3 = 1$.

In principle, there is an infinite number of potential patterns. However, just like in the Apriori algorithm a monotonicity property can be exploited: if a sequence is frequent, then its subsequences are also frequent. Therefore, it is possible to efficiently generate patterns. Frequent sequences can also be used to create rules of the form $X \Rightarrow Y$ where X is a pattern and Y is an extension or continuation of the pattern. Consider for example $X = \langle\{\text{cappuccino}\}, \{\text{espresso}\}\rangle$ and $Y = \langle\{\text{cappuccino}\}, \{\text{espresso}\}, \{\text{latte, muffin}\}\rangle$. Suppose that X has a support of 0.05 and Y has a support of 0.04 . Then the confidence of $X \Rightarrow Y$ is $0.04/0.05 = 0.8$, i.e., 80% of the customer that ordered a cappuccino followed by an espresso later also order a muffin and latte.

In [131] several extensions of the above approach have been proposed. For example, it is possible to add taxonomies, sliding windows, and time constraints. For practical applications it is important to relax the strict subsequence requirement such that a one-to-one matching of item-sets is no longer needed.

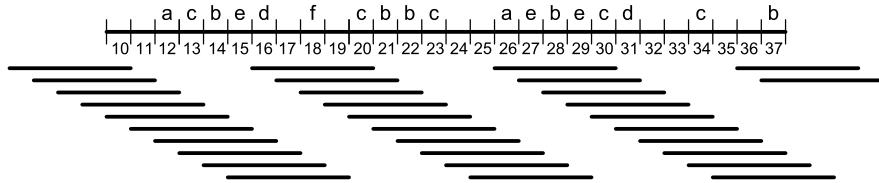


Fig. 4.9 A timed sequence of events and the corresponding time windows

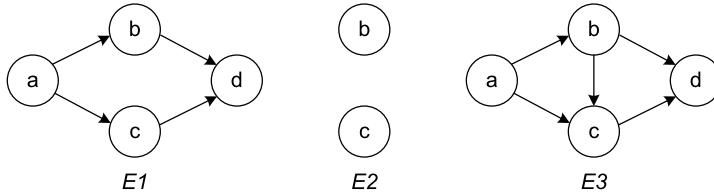


Fig. 4.10 Three episodes

4.5.2 Episode Mining

Another problem that can be solved using an Apriori-like approach is the *discovery of frequent episodes* [94]. Here a sliding window is used to analyze how frequent an *episode* is appearing. An episode defines a partial order. The goal is to discover frequent episodes.

Input for episode mining is a time sequence as shown in Fig. 4.9. The timed sequence starts at time 10 and ends at time 37. The sequence consists of discrete time points, and, as shown in Fig. 4.9, at some points in time an event occurs. An event has a type (e.g., the activity that happened) and a timestamp. For example, an event of type *a* occurs at time 12, an event of type *c* occurs at time 13, etc. Figure 4.9 also shows 32 *time windows* of length 5. These are all the windows (partially) overlapping with the timed sequence. The length 5 is a predefined parameter of the algorithm used to discover frequent patterns. An episode *occurs* in a time window if the partial order is “embedded” in it.

Figure 4.10 shows three episodes. An episode is described by a directed acyclic graph. The nodes refer to event types and the arcs define a partial order. For example, episode *E1* defines that *a* should be followed by *b* and *c*, *b* should be followed by *d*, and *c* should be followed by *d*. Episode *E2* merely states that *b* and *c* should both happen at least once. Episode *E3* states that *a* should be followed by *b* and *c*, *b* should be followed by *c*, *b* should be followed by *d*, and *c* should be followed by *d*. This episode contains two redundant arcs: the arc from *a* to *c* and the arc from *b* to *d* can be removed without changing the requirements. An episode *occurs* in a time window if it is possible to assign events to nodes in the episode such that the ordering relations are satisfied. Note that the episode only defines the minimal set of events, i.e., there may be all kinds of additional events. The key requirement is that the episode is embedded.

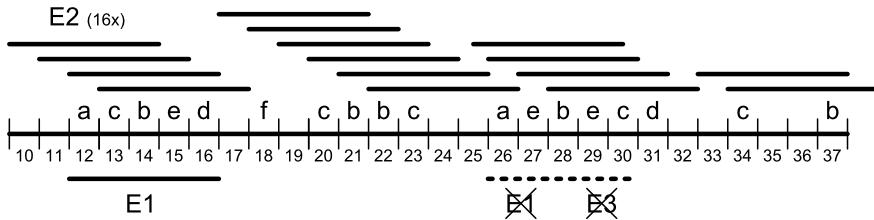


Fig. 4.11 Occurrences of episodes $E1$ and $E2$

To illustrate the notion of “occurring in a time window”, we use Fig. 4.11. Consider episode $E1$ and slide a window of length 5 from left to right. There are 32 possible positions. However, just one of the 32 windows embeds episode $E1$. This is the window starting at time 12 shown below the timed sequence in Fig. 4.11. Here we find the sequence $\langle a, c, b, e, d \rangle$. Clearly all the requirements are met in the sequence: a is followed by b and this b is followed by d , the same a is also followed by c and this c is followed by the same d .

Now consider episode $E2$ and again slide a window of length 5 from left to right. This pattern is much more frequent. Figure 4.11 shows all time windows in which the pattern occurs. In total there are 16 windows where $E2$ is embedded. Note that the only requirement is that both b and c occur: no ordering relation is defined.

Episode $E3$ does not occur in the time sequence if we use a window length of 5. There is no window of length 5 where the sequence $\langle a, b, c, d \rangle$ is embedded. If the window length is extended to 6, $E3$ occurs once. The corresponding window starts at time 26. Here we find the sequence $\langle a, e, b, e, c, d \rangle$.

The *support* of an episode is the fraction of windows in which the episode occurs. For a window size of 5 time units, the support of $E1$ is $1/32$, the support of $E2$ is $16/32 = 0.5$, and the support of $E3$ is $0/32 = 0$. Like for sequence mining and association rule learning, we define a threshold for the support. All episodes having a support of at least this threshold are *frequent*. For example, if the threshold is 0.2 then $E2$ is frequent whereas $E1$ and $E3$ are not.

The goal is to generate all frequent episodes. Note that there are typically many potential candidates (all partial orders over the set of event types). Fortunately, like in the Apriori algorithm, we can exploit a monotonicity property to quickly rule out bad candidates. To explain this property we need to define the notion of a *subepisode*. $E1$ is a subepisode of $E3$ because $E1$ is a subgraph of $E3$, i.e., the nodes and arcs of $E1$ are contained in $E3$. $E2$ is a subepisode of both $E1$ and $E3$. It is easy to see that, if an episode E is frequent, then also all of its subepisodes are frequent. This monotonicity property can be used to speed-up the search process.

Frequent episodes can also be used to create rules of the form $X \Rightarrow Y$ where X is a subepisode of Y . As before the confidence of such a rule can be computed. In our example, rule $E1 \Rightarrow E3$ has a confidence of $0/1 = 0$, i.e., a very poor rule. Rule $E2 \Rightarrow E1$ has a confidence of $1/16$.

Episode mining and sequence mining can be seen as variants of association rule learning. Because they take into account the ordering of events, they are related to

process discovery. However, there are many differences with process mining algorithms. First of all, *only local patterns* are considered, i.e., no overall process model is created. Second, the focus is on frequent behavior without trying to generate models that also *exclude* behavior. Consider, for example, episode $E1$ in Fig. 4.10. Also the time window $\langle a, b, d, c, d \rangle$ contains the episode despite the two occurrences of d . Therefore, episodes cannot be read as if they are process models. Moreover, episodes *cannot model choices, loops, etc.* Finally, episode mining and sequence mining *cannot handle concurrency* well. Sequence mining searches for sequential patterns only. Episode mining runs into problems if there are concurrent episodes, because it is unclear what time window to select to get meaningful episodes.

4.5.3 Other Approaches

In the data mining and machine learning communities several other techniques have been developed to analyze sequences of events. Applications are in text mining (sequences of letters and words), bio-informatics (analysis of DNA sequences), speech recognition, web analytics, etc. Examples of techniques that are used for this purpose are *neural networks* and *hidden Markov models* [9, 102].

Artificial neural networks try to mimic the human brain in order to learn complex tasks. An artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain. Different learning paradigms can be used to train the neural network: supervised learning, unsupervised learning, and reinforcement learning [9, 102]. Advantages are that neural networks can exploit parallel computing and that they can be used to solve ill-defined tasks, e.g., image and speech recognition. The main drawback is that the resulting model (e.g., a multi-layer perceptron), is typically not human readable. Hence there is no resulting process model in the sense of Chap. 3 (e.g., a WF-net or BPMN model).

Hidden Markov models are an extension of ordinary Markov processes. A hidden Markov model has a set of states and transition probabilities. Moreover, unlike standard Markov models, in each state an observation is possible, but the state itself remains hidden. Observations have probabilities per state as shown in Fig. 4.12. Three fundamental problems have been investigated for hidden Markov models [9]:

- Given an observation sequence, how to compute the probability of the sequence given a hidden Markov model?
- Given an observation sequence and a hidden Markov model, how to compute the most likely “hidden path” in the model?
- Given a set of observation sequences, how to derive the hidden Markov model that maximizes the probability of producing these sequences?

The last problem is most related to process mining but also the most difficult problem. The well-known Baum–Welch algorithm [9] is a so-called Expectation–Maximization (EM) algorithm that solves this problem iteratively for a fixed number of states. Although hidden Markov models are versatile and relevant for process

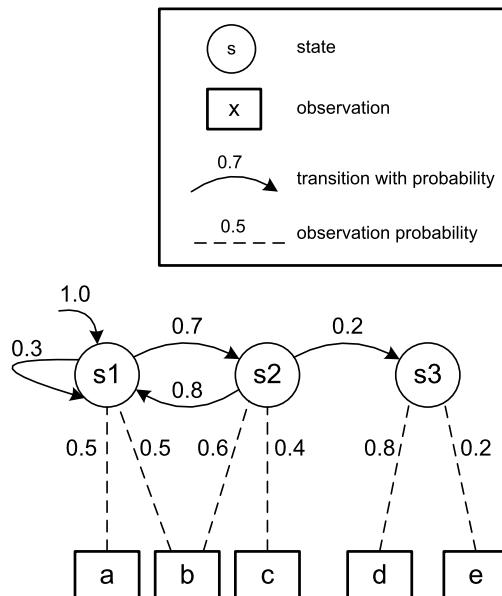


Fig. 4.12 A hidden Markov model with three states: s_1 , s_2 , and s_3 . The arcs have state transition probabilities as shown, e.g., in state s_2 the probability of moving to state s_3 is 0.2 and the probability of moving to s_1 is 0.8. Each visit to a state generates an observation. The observation probabilities are also given. When visiting s_2 the probability of observing b is 0.6 and the probability of observing c is 0.4. Possible observation sequences are $\langle a, b, c, d \rangle$, $\langle a, b, b, c \rangle$, and $\langle a, b, c, b, b, a, c, e \rangle$. For the observation sequence $\langle a, b, c, d \rangle$ it is clear what the hidden sequence is: $\langle s_1, s_2, s_2, s_3 \rangle$. For the other two observation sequences multiple hidden sequences are possible

mining, there are several complications. First of all, there are many computational challenges due to the time consuming iterative procedures. Second, one needs to guess an appropriate number of states as this is input to the algorithm. Third, the resulting hidden Markov model is typically not very accessible for the end user, i.e., accurate models are typically large and even for small examples the interpretation of the states is difficult. Clearly, hidden Markov models are at a lower abstraction level than the notations discussed in Chap. 3.

4.6 Quality of Resulting Models

This chapter provided an overview of the mainstream data mining techniques most relevant for process mining. Although some of these techniques can be exploited for process mining, they cannot be used for important process mining tasks such as process discovery, conformance checking, and process enhancement. However, there is an additional reason for showing a variety of data mining techniques. Like in data mining it is non-trivial to analyze the quality of process mining results. Here one can

Fig. 4.13 Confusion matrix for the decision tree shown in Fig. 4.2. Of the 200 students who failed, 178 are classified as failed and 22 are classified as passed. None of the failing students was classified as cum laude. Of the 198 students who passed, 175 are classified correctly, 21 were classified as failed, and 2 as cum laude

		predicted class		
		failed	passed	cum laude
actual class	failed	178	22	0
	passed	21	175	2
	cum laude	1	3	18

benefit from experiences in the data mining field. Therefore, we discuss some of the validation and evaluation techniques developed for the algorithms presented in this chapter. First, we focus on the quality of classification results, e.g., obtained through a decision tree. Second, we describe general techniques for cross-validation. Here, we concentrate on k -fold cross-validation. Finally, we conclude with a more general discussion on Occam's razor.

4.6.1 Measuring the Performance of a Classifier

In Sect. 4.2, we showed how to construct a decision tree. As discussed, there are many design decisions when developing a decision tree learner (e.g., selection of attributes to split on, when to stop splitting, and determining cut values). The question is how to evaluate the performance of a decision tree learner. This is relevant for judging the trustworthiness of the resulting decision tree and for comparing different approaches. A complication is that one can only judge the performance based on *seen* instances although the goal is also to predict good classifications for *unseen* instances. However, for simplicity, let us first assume that we first want to judge the result of a classifier (like a decision tree) on a given data set.

Given a data set consisting of N instances we know for each instance what the actual class is and what the predicted class is. For example, for a particular person that smokes, we may predict that the person will die young (predicted class is “young”), even though the person dies at age 104 (actual class is “old”). This can be visualized using a so-called *confusion matrix*. Figure 4.13 shows the confusion matrix for the data set shown in Table 4.2 and the decision tree shown in Fig. 4.2. The decision tree classifies each of the 420 students into an actual class and a predicted class. All elements on the diagonal are predicted correctly, i.e., $178 + 175 + 18 = 371$ of the 420 students are classified correctly (approximately 88%).

There are several performance measures based on the confusion matrix. To define these let us consider a data set with only two classes: “positive” (+) and “negative” (-). Figure 4.14(a) shows the corresponding 2×2 confusion matrix. The following entries are shown:

Figure 4.14(a) shows a confusion matrix for two classes. The columns represent the predicted class (+ or -) and the rows represent the actual class (+ or -). The matrix entries are labeled tp , fn , fp , and tn . The total number of instances is $N = tp + fn + fp + tn$. The counts for each row and column are also provided: $p = tp + fn$ (actual positive), $n = fn + tn$ (actual negative), $p' = tp + fp$ (predicted positive), and $n' = fn + tn$ (predicted negative).

		predicted class		
		+	-	
actual class	+	tp	fn	p
	-	fp	tn	n
		p'	n'	N

Figure 4.14(b) lists some performance measures with their formulas:

name	formula
error	$(fp+fn)/N$
accuracy	$(tp+tn)/N$
tp-rate	tp/p
fp-rate	fp/n
precision	tp/p'
recall	tp/p

(a)

(b)

Fig. 4.14 Confusion matrix for two classes and some performance measures for classifiers

- tp is the number of *true positives*, i.e., instances that are correctly classified as positive.
- fn is the number of *false negatives*, i.e., instances that are predicted to be negative but should have been classified as positive.
- fp is the number of *false positives*, i.e., instances that are predicted to be positive but should have been classified as negative.
- tn is the number of *true negatives*, i.e., instances that are correctly classified as negative.

Figure 4.14(a) also shows the sums of rows and columns, e.g., $p = tp + fn$ is the number of instances that are actually positive, $n' = fn + tn$ is the number of instances that are classified as negative by the classifier. $N = tp + fn + fp + tn$ is the total number of instances in the data set. Based on this it is easy to define the measures shown in Fig. 4.14(b). The *error* is defined as the proportion of instances misclassified: $(fp + fn)/N$. The *accuracy* measures the fraction of instances on the diagonal of the confusion matrix. The “true positive rate”, *tp-rate*, also known as “hit rate”, measures the proportion of positive instances indeed classified as positive. The “false positive rate”, *fp-rate*, also known as “false alarm rate”, measures the proportion of negative instances wrongly classified as positive. The terms *precision* and *recall* originate from information retrieval. Precision is defined as tp/p' . Here, one can think of p' as the number of documents that have been retrieved based on some search query and tp as the number of documents that have been retrieved and also should have been retrieved. Recall is defined as tp/p where p can be interpreted as the number of documents that should have been retrieved based on some search query. It is possible to have high precision and low recall; few of the documents searched for are returned by the query, but those that are returned are indeed relevant. It is also possible to have high recall and low precision; many documents are returned (including the ones relevant), but also many irrelevant documents are returned. Note that recall is the same as tp-rate. There is another frequently used metric not shown in Fig. 4.14(b): the so-called *F1 score*. The F1 score takes the harmonic mean of precision and recall: $(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$. If either the precision or recall is really poor (i.e., close to 0), then the F1 score is also

Fig. 4.15 Two confusion matrices for the decision trees in Fig. 4.4

		predicted class			predicted class	
		young	old		young	old
actual class	young	546	0		544	2
	old	314	0		251	63

(a)

(b)

close to 0. Only if both precision and recall are really good, the F1 score is close to 1.

To illustrate the different metrics let us consider the three decision trees depicted in Fig. 4.4. In the first two decision trees, all instances are classified as young. Note that even after splitting the root node based on the attribute *smoker*, still all instances are predicted to die before 70. Figure 4.15(a) shows the corresponding confusion matrix assuming “young = positive” and “old = negative”. $N = 860$, $tp = p = 546$, and $fp = n = 314$. Note that $n' = 0$ because all are classified as young. The error is $(314 + 0)/860 = 0.365$, the tp-rate is $546/546 = 1$, the fp-rate is $314/314 = 1$, precision is $546/860 = 0.635$, recall is $546/546 = 1$, and the F1 score is 0.777. Figure 4.15(b) shows the confusion matrix for the third decision tree in Fig. 4.4. The error is $(251 + 2)/860 = 0.292$, the tp-rate is $544/546 = 0.996$, the fp-rate is $251/314 = 0.799$, precision is $544/795 = 0.684$, recall is $544/546 = 0.996$, and the F1 score is 0.811. Hence, as expected, the classification improved: the error and fp rate decreased considerably and the tp-rate, precision and F1 score increased. Note that the recall went down slightly because of the two persons that are now predicted to live long but do not (despite not smoking nor drinking).

4.6.2 Cross-Validation

The various performance metrics computed using the confusion matrix in Fig. 4.15(b) are based on the same data set as the data set used to learn the third decision tree in Fig. 4.4. Therefore, the confusion matrix is only telling something about *seen* instances, i.e., instances used to learn the classifier. In general, it is trivial to provide classifiers that score perfectly (i.e., precision, recall and F1 score are all 1) on seen instances. (Here we assume that instances are unique or instances with identical attributes belong to the same class.) For example, if students have a unique registration number, then the decision tree could have a leaf node per student thus perfectly encoding the data set. However, this does not say anything about *unseen* instances, e.g., the registration number of a new student carries no information about expected performance of this student.

The most obvious criterion to estimate the performance of a classifier is its predictive accuracy on unseen instances. The number of unseen instances is potentially

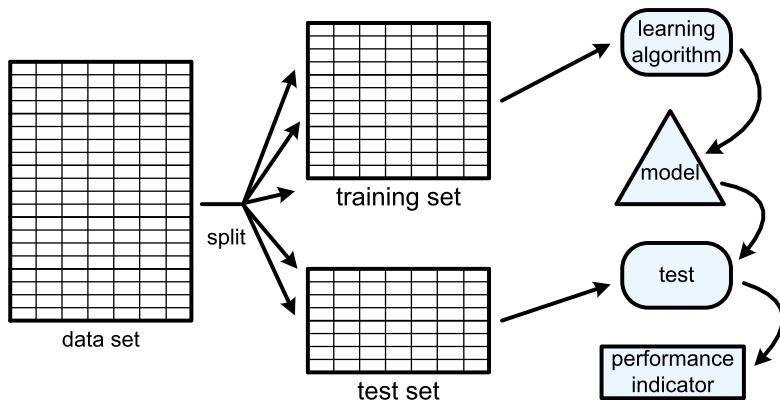


Fig. 4.16 Cross-validation using a test and training set

very large (if not infinite), therefore an estimate needs to be computed on a test set. This is commonly referred to as *cross-validation*. The data set is split into a *training set* and a *test set*. The training set is used to learn a model whereas the test set is used to evaluate this model based on unseen examples.

It is important to realize that cross-validation is not limited to classification but can be used for any data mining technique. The only requirement for cross-validation is that the performance of the result can be measured in some way. For classification we defined measures such as *precision*, *recall*, *F1 score*, and *error*.

For regression also various measures can be defined. In the context of linear regression the *mean square error* is a standard indicator of quality. If y_1, y_2, \dots, y_n are the actual values and $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ the predicted values according to the linear regression model, then $(\sum_{i=1}^n (y_i - \hat{y}_i)^2)/n$ is the mean square error.

Clustering is typically used in a more descriptive or explanatory manner, and rarely used to make direct predictions about unseen instances. Nevertheless, the clusters derived for a training set could also be tested on a test set. Assign all instances in the test set to the closest centroid. After doing this, the *average distance* of each instance to its centroid can be used as a performance measure.

In the context of association rule mining, we defined metrics such as *support*, *confidence*, and *lift*. One can learn association rules using a training set and then test the discovered rules using the test set. The confidence metric then indicates the proportion of instances for which the rule holds while being applicable. Later, we will also define such metrics for process mining. For example, given an event log that serves as a test set and a Petri net model, one can look at the proportion of instances that can be replayed by the model.

Figure 4.16 shows the basic setting for cross-validation. The data set is split into a test and training set. Based on the training set a model is generated (e.g., a decision tree or regression model). Then the performance is analyzed using the test set. If just one number is generated for the performance indicator, then this does not give an indication of the reliability of the result. For example, based on some test set the F1

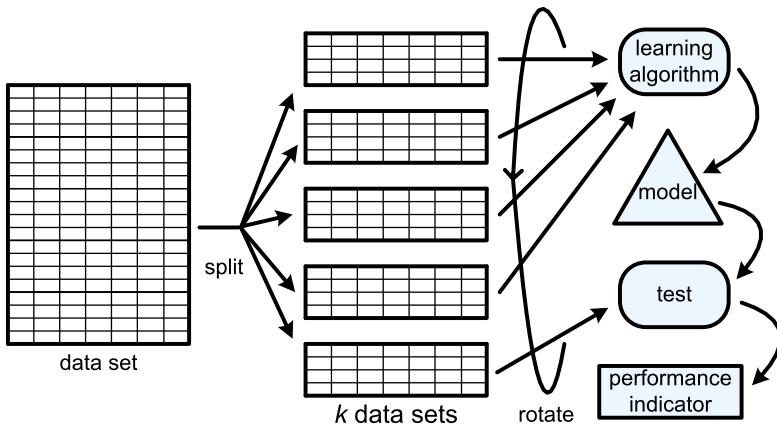


Fig. 4.17 *k*-fold cross-validation

score is 0.811. However, based on another test set the F1 score could be completely different even if the circumstances did not change. Therefore, one often wants to calculate a *confidence interval* for such a performance indicator. Confidence intervals can only be computed over multiple measurements. Here, we discuss two possibilities.

The first possibility is that one is measuring a performance indicator that is the average over of a large set of independent measurements. Consider for example classification. The test set consists of N instances that are mutually independent. Hence, each classification $1 \leq i \leq N$ can be seen as a separate test x_i where $x_i = 1$ means that the classification is wrong and $x_i = 0$ means that the classification is good. These tests can be seen as samples from a Bernoulli distribution with parameter p (p is the probability of a wrong classification). This distribution has an expected value p and variance $p(1 - p)$. If we assume that N is large, then the average error $(\sum_{i=1}^N x_i)/N$ is approximately normal distributed. This is due to the central limit theorem, also known as the “law of large numbers”. Using this assumption we find the 95% *confidence interval* which is $[p - \alpha_{0.95}\sqrt{p(1-p)/N}, p + \alpha_{0.95}\sqrt{p(1-p)/N}]$, i.e., with 95% certainty the real average error will lie within $p - \alpha_{0.95}\sqrt{p(1-p)/N}$ and $p + \alpha_{0.95}\sqrt{p(1-p)/N}$. $\alpha_{0.95} = 1.96$ is a standard value that can be found in any statistics textbook, p is the measured average error rate, and N is the number of tests. For calculating the 90% or 99% confidence interval one can use $\alpha_{0.90} = 1.64$ respectively $\alpha_{0.99} = 2.58$. Note that it is only possible to calculate such an interval if there are many independent measurements possible based on a single test run.

The second possibility is *k-fold cross-validation*. This approach is used when there are relatively few instances in the data set or when the performance indicator is defined on a set of instances rather than a single instance. For example, the F1 score cannot be defined for one instance in isolation. Figure 4.17 illustrates the idea behind *k*-fold cross-validation. The data set is split into *k* equal parts, e.g., $k = 10$. Then *k* tests are done. In each test, one of the subsets serves as a test set whereas

the other $k - 1$ subsets serve together as the training set. If subset $i \in \{1, 2, \dots, k\}$ is used as a test set, then the union of subsets $\{1, 2, \dots, i - 1, i + 1, \dots, k\}$ is used as the training set. One can inspect the individual tests or take the average of the k folds.

There are two advantages associated to k -fold cross-validation. First of all, all data is used both as training data and test data. Second, if desired, one gets k tests of the desired performance indicator rather than just one. Formally, the tests cannot be considered to be independent as the training sets used in the k folds overlap considerably. Nevertheless, the k folds make it possible to get more insight into the reliability.

An extreme variant of k -fold cross-validation is “leave-one-out” cross-validation, also known as jack-knifing. Here $k = N$, i.e., the test sets contain only one element at a time. See [9, 102] for more information on the various forms of cross-validation.

4.6.3 Occam’s Razor

Evaluating the quality of data mining results is far from trivial. In this subsection, we discuss some additional complications that are also relevant for process mining.

Learning is typically an “ill posed problem”, i.e., only examples are given. Some examples may rule out certain solutions, however, typically many possible models remain. Moreover, there is typically a *bias* in both the target representation and the learning algorithm. Consider, for example, the sequence 2, 3, 5, 7, 11, What is the next element in this sequence? Most readers will guess that it is 13, i.e., the next prime number, but there are infinitely many sequences that start with 2, 3, 5, 7, 11. Yet, there seems to be preference for hypothesizing about some solutions. The term *inductive bias* refers to a preference for one solution rather than another which cannot be determined by the data itself but which is driven by external factors.

A *representational bias* refers to choices that are implicitly made by selecting a particular representation. For example, in Sect. 4.2, we assumed that in a decision tree the same attribute may appear only once on a path. This representational bias rules out certain solutions, e.g., a decision tree where closer to the root a numerical attribute is used in a coarse-grained manner and in some subtrees it is used in a fine-grained manner. Linear regression also makes assumptions about the function used to best fit the data. The function is assumed to be linear although there may be other non-linear functions that fit the data much better. Note that a representational bias is not necessarily bad, e.g., linear regression has been successfully used in many application domains. However, it is important to realize that the search space is limited by the representation used. The limitations can guide the search process, but also exclude good solutions.

A *learning bias* refers to strategies used by the algorithm that give preference to particular solutions. For example, in Fig. 4.4, we used the criterion of information gain (reduction of entropy) to select attributes. However, we could also have used the Gini index of diversity G rather than entropy E to select attributes, thus resulting in different decision trees.

Both factors also play a role in process mining. Consider, for example, Fig. 2.6 in the first chapter. This process model was discovered using the α -algorithm [157] based on the set of traces $\{\langle a, b, d, e, h \rangle, \langle a, d, c, e, g \rangle, \langle a, c, d, e, f, b, d, e, g \rangle, \langle a, d, b, e, h \rangle, \langle a, c, d, e, f, d, c, e, f, b, d, e, h \rangle, \langle a, c, d, e, g \rangle\}$. Clearly, there is a representational bias. The assumption is that the process can be presented by a Petri net where every transition bears a unique and visible label. Many processes cannot be represented by such a Petri net. The α -algorithm also has a learning bias as it is focusing on “direct succession”. If a is directly followed by b in the event log, then this information is used. However, an observation such as “ a is eventually followed by b in the event log” is not exploited by the α -algorithm.

An inductive bias is not necessarily bad. In fact it is often needed to come to a solution. However, the analyst should be aware of this and reflect on the implicit choices made.

Curse of dimensionality

Some data sets have many variables. However, for most data mining problems the amount of data needed to maintain a specific level of accuracy is exponential in the number of parameters [69]. High-dimensional problems, i.e., analyzing a data set with many variables, may be computationally intractable or lead to incorrect conclusions. This is the “*curse of dimensionality*” that many real-life applications of data mining are confronted with. Consider, for example, a supermarket selling 1000 products. In this case, there are $2^{1000} - 1$ potential item-sets. Although the Apriori algorithm can quickly rule out many irrelevant candidates, the generation of association rules in such a setting is likely to encounter performance problems. Moreover, the interpretation of the results is typically difficult due to an excessive number of potential rules. In a supermarket having hundreds or thousands of products, there are many customers that purchase a unique combination of products. If there are 1000 different products, then there are $2^{1000} - 1 \approx 1.07 \times 10^{301}$ possible shopping lists (ignoring quantities). Although the probability that two customers purchase the same is small, the number of potential rules is very large. This problem is not restricted to association rule learning. Clustering or regression in a 1000 dimensional space will suffer from similar problems. Typical approaches to address this problem are *variable selection* and *transformation* [69]. The goal of variable selection is to simply remove irrelevant or redundant variables. For example, the student’s registration number and address are irrelevant when predicting study progress. Sometimes the data set can be transformed to reduce dimensionality, e.g., taking the average mark rather than individual marks per course.

Another problem is the delicate balance between *overfitting* and *underfitting*. A learned model is overfitting if it is too specific and too much driven by accidental information in the data set. For example, when constructing a decision tree

for a training set without conflicting input (i.e., instances with identical attributes belong to the same class), it is easy to construct a decision tree with a perfect F1 score. This tree can be obtained by continuing to split nodes until each leaf node corresponds to instances belonging to the same class. However, it is obvious that such a decision tree is too specific and has little predictive value.

A learned model is underfitting if it is too general and allows for things not “supported by evidence” in the data set. Whereas overfitting can be characterized by a lack of generalization, underfitting has the opposite problem: too much generalization. Consider, for example, the generation of association rules. Generating many detailed rules due to very low settings of *minsup* and *minconf*, corresponds to overfitting. Many rules are found, but these are probably rather specific for the training set. Generating very few rules due to very high settings of *minsup* and *minconf*, corresponds to underfitting. In the extreme case no association rules are found. Note that the model with no rules fits any data set and, hence, carries no information.

Underfitting is particularly problematic if the data set contains *no negative examples*. Consider, for example, the confusion matrix in Fig. 4.14(a). Suppose that we have a training set with only positive examples, i.e., $n = 0$ in the training set. How to construct a decision tree without negative examples? Most algorithms will simply classify everything as positive. This shows that classification assumes both positive and negative examples. This is not the case for association rule learning. Consider, for example, the data set shown in Table 4.3. Suppose that the item-set $\{latte, tea, bagel\}$ does not appear in the data set. This implies that no customer ordered these three items together in the training set. Can we conclude from this that it is not possible to order these three items together? Of course not! Therefore, association rule learning focuses on positive examples that are somehow frequent. Nevertheless, for some applications it would be useful to be able to discover “negative rules” such as the rule that customers are not allowed to order latte’s, teas, and bagels in a single order.

A good balance between overfitting and underfitting is of the utmost importance for process discovery. Consider again the Petri net model shown in Fig. 2.6. The model allows for the behavior seen in the event log. It also generalizes as it allows for more sequences than present in the training set. In the event log there is no trace $\langle a, h \rangle$, i.e., the scenario in which a request is registered and immediately rejected does not appear in the log. This does not necessarily imply that this is not possible. However, constructing a model that allows for $\langle a, h \rangle$ although it is not in the log would result in a model that is clearly underfitting. This dilemma is caused by the lack of negative examples in the event log. The traces in the event log show what has happened and not what could not happen. This problem will be addressed in later chapters.

We conclude this chapter with *Occam’s Razor*, a principle attributed to the 14th-century English logician William of Ockham. The principle states that “one should not increase, beyond what is necessary, the number of entities required to explain anything”, i.e., one should look for the “simplest model” that can explain what is observed in the data set. This principle is related to finding a natural balance between overfitting and underfitting. The *Minimal Description Length* (MDL) principle tries

to operationalize Occam’s Razor [63, 190]. According to the MDL paradigm, model quality is no longer only based on predicting performance (e.g., F1 score), but also on the simplicity of the model. Moreover, it does not aim at cross-validation in the sense of Sect. 4.6.2. In MDL performance is judged on the training data alone and not measured against new, unseen instances. The basic idea is that the “best” model is the one that *minimizes the encoding of both model and data set*. Here the insight is used that any regularity in the data can be used to compress the data, i.e., to describe it using fewer symbols than the number of symbols needed to describe the data literally. The more regularities there are, the more the data can be *compressed*. Equating “learning” with “finding regularity”, implies that the more we are able to compress the data, the more we have learned about the data [63]. Obviously, a data set can be encoded more compactly if valuable knowledge about the data set is captured in the model. However, encoding such knowledge also requires space. A complex and overfitting model helps to reduce the encoding of the data set. A simple and underfitting model can be stored compactly, but does not help in reducing the encoding of the data set. Note that this idea is related to the notion of entropy in decision tree learning. When building the decision tree, the goal is to find homogeneous leaf nodes that can be encoded compactly. However, when discussing algorithms for decision tree learning in Sect. 4.2 there was no penalty for the complexity of the decision tree itself. The goal of MDL is to minimize the entropy of (a) the data set encoded using the learned model and (b) the encoding of the model itself. To balance between overfitting and underfitting, variable weights may be associated to both encodings.

Applying Occam’s Razor is not easy. Extracting reliable and meaningful insights from complex data is far from trivial. In fact, it is much easier to transform complex data sets into “impressive looking garbage” by abusing the techniques presented in this chapter. However, when used wisely, data mining can add tremendous value. Moreover, process mining adds the “process dimension” to data and can be used to dissect event data from a more holistic perspective. As will be shown in the remainder, process mining creates a solid bridge between process modeling and analysis on the one hand and data mining on the other.

Part III

From Event Logs to Process Models

Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
**Process Mining:
The Missing Link**

Part II: Preliminaries

Chapter 3
**Process Modeling
and Analysis**

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
**Process Discovery:
An Introduction**

Chapter 7
**Advanced Process
Discovery Techniques**

Part IV: Beyond Process Discovery

Chapter 8
**Conformance
Checking**

Chapter 9
**Mining Additional
Perspectives**

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
**Process Mining
Software**

Chapter 12
**Process Mining in the
Large**

Chapter 13
**Analyzing “Lasagna
Processes”**

Chapter 14
**Analyzing “Spaghetti
Processes”**

Part VI: Reflection

Chapter 15
**Cartography and
Navigation**

Chapter 16
Epilogue

After providing preliminaries needed for a good understanding of the “roots” of process mining, we focus on the most challenging process mining task: discovering a process model from an event log. First, in Chap. 5 we describe the input required for process discovery. Then, Chap. 6 describes the α -algorithm in detail. This rather naïve algorithm helps to understand the basics and also sets the scene for discussing the challenges related to process mining. Finally, Chap. 7 gives an overview of state-of-the-art process discovery algorithms and shows how they address the challenges identified.

Chapter 5

Getting the Data

Process mining is impossible without proper event logs. This chapter describes the information that should be present in such event logs. Depending on the process mining technique used, these requirements may vary. The challenge is to extract such data from a variety of data sources, e.g., databases, flat files, message logs, transaction logs, ERP systems, and document management systems. When merging and extracting data, both syntax and semantics play an important role. Moreover, depending on the questions one seeks to answer, different views on the available data are needed. Process mining, like any other data-driven analysis approach, needs to deal with data quality problems. We discuss typical data quality challenges encountered in reality. The insights provided in this chapter help to get the event data assumed to be present in later chapters.

5.1 Data Sources

In Chap. 2, we introduced the concept of process mining. The idea is to analyze event data from a process-oriented perspective. The goal of process mining is to answer questions about operational processes. Examples are:

- What *really* happened in the past?
- Why did it happen?
- What is likely to happen in the future?
- When and why do organizations and people deviate?
- How to control a process better?
- How to redesign a process to improve its performance?

In subsequent chapters, we will discuss various techniques to answer the preceding questions. However, first we focus on the event data needed.

Figure 5.1 shows the overall “process mining workflow” emphasizing the role of event data. Starting point is the “raw” data hidden in all kinds of data sources. A data source may be a simple flat file, an Excel spreadsheet, a transaction log,

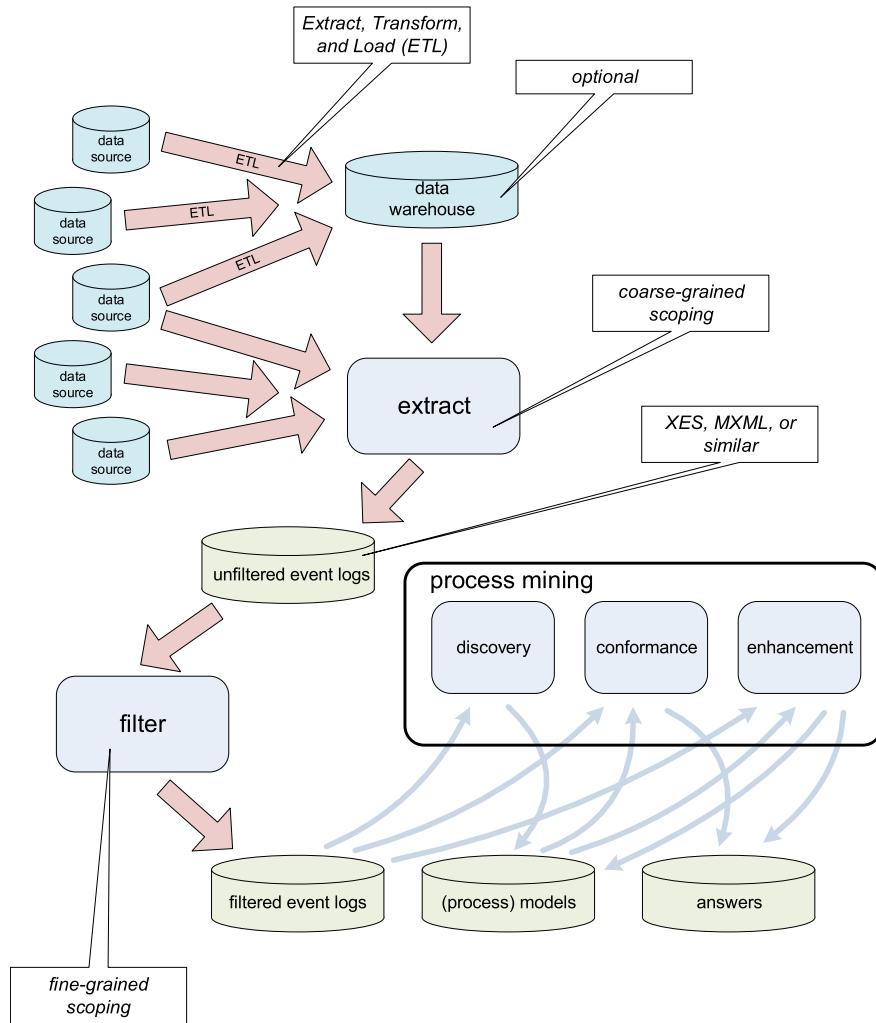


Fig. 5.1 Overview describing the workflow of getting from heterogeneous data sources to process mining results

or a database table. However, one should not expect all the data to be in a single well-structured data source. The reality is that event data is typically scattered over different data sources and often quite some efforts are needed to collect the relevant data. Consider, for example, a full SAP implementation that typically has more than 10,000 tables. Data may be scattered due to technical or organizational reasons. For example, there may be legacy systems holding crucial data or information systems used only at the departmental level. For cross-organizational process mining, e.g., to analyze supply chains, data may even be scattered over multiple organizations. Events can also be captured by tapping of message exchanges [161] (e.g., SOAP

messages) and recording read and write actions [47]. Data sources may be structured and well-described by meta data. Unfortunately, in many situations the data is unstructured or important meta data is missing. Data may originate from web pages, emails, PDF documents, scanned text, screen scraping, etc. Even if data is structured and described by meta data, the sheer complexity of enterprise information systems may be overwhelming. There is no point in trying to exhaustively extract event logs from thousands of tables and other data sources. Data extraction should be driven by questions rather than the availability of lots of data.

In the context of BI and data mining, the phrase “*Extract, Transform, and Load*” (ETL) is used to describe the process that involves: (a) *extracting* data from outside sources, (b) *transforming* it to fit operational needs (dealing with syntactical and semantical issues while ensuring predefined quality levels), and (c) *loading* it into the target system, e.g., a data warehouse or relational database. A *data warehouse* is a single logical repository of an organization’s transactional and operational data. The data warehouse does not produce data but simply taps off data from operational systems. The goal is to unify information such that it can be used for reporting, analysis, forecasting, etc. Figure 5.1 shows that ETL activities can be used to populate a data warehouse. It may require quite some efforts to create the common view required for a data warehouse. Different data sources may use different keys, formatting conventions, etc. For example, one data source may identify a patient by her last name and birth date while another data source uses her social security number. One data source may use the date format “31-12-2010” whereas another uses the format “2010/12/31”.

If a data warehouse already exists, it most likely holds valuable input for process mining. However, many organizations do not have a good data warehouse. The warehouse may contain only a subset of the information needed for end-to-end process mining, e.g., only data related to customers is stored. Moreover, if a data warehouse is present, it does not need to be process oriented. For example, the typical warehouse data used for *Online Analytical Processing* (OLAP) does not provide much process-related information. OLAP tools are excellent for viewing multidimensional data from different angles, drilling down, and for creating all kinds of reports (see Sect. 12.4). However, OLAP tools do not require the storage of business events and their ordering. The data sets used by the mainstream data mining approaches described in Chap. 4 also do not store such information. For example, a decision tree learner can be applied to any table consisting of rows (instances) and columns (variables). As will be shown in the next section, process mining requires information on relevant events and their order.

Whether there is a data warehouse or not, data needs to be extracted and converted into event logs. Here, *scoping* is of the utmost importance. Often the problem is not the syntactical conversion but the selection of suitable data. Questions like “Which of the more than 10,000 SAP tables to convert?” need to be answered first. Typical formats to store event logs are *XES* (eXtensible Event Stream) and *MXML* (Mining eXtensible Markup Language). These will be discussed in Sect. 5.3. For the moment, we assume that *one event log corresponds to one process*, i.e., when scoping the data in the extraction step, only events relevant for the process to be analyzed

should be included. In Sect. 5.5, we discuss the problem of converting “3-D data” into “2-D event logs”, i.e., events are projected onto the desired process model.

Depending on the questions and viewpoint chosen, different event logs may be extracted from the same data set. Consider for example the data in a hospital. One may be interested in the discovery of patient flows, i.e., typical diagnosis and treatment paths. However, one may also be interested in optimizing the workflow within the radiology department. Both questions require different event logs, although some events may be shared among the two required event logs. Once an event log is created, it is typically *filtered*. Filtering is an iterative process. *Coarse-grained scoping* was done when extracting the data into an event log. Filtering corresponds to *fine-grained scoping* based on initial analysis results. For example, for process discovery one can decide to focus on the 10 most frequent activities to keep the model manageable.

Based on the filtered log, the different types of process mining described in Sect. 2.2 can be applied: *discovery*, *conformance*, and *enhancement*.

Although Fig. 5.1 does not reflect the iterative nature of the whole process well, it should be noted that process mining results most likely trigger new questions and these questions may lead to the exploration of new data sources and more detailed data extractions. Typically, several iterations of the extraction, filtering, and mining phases are needed.

5.2 Event Logs

Table 5.1 shows a fragment of the event log already discussed in Chap. 2. This table illustrates the typical information present in an event log used for process mining. The table shows events related to the handling of requests for compensation. We assume that an event log contains data related to a *single process*, i.e., the first coarse-grained scoping step in Fig. 5.1 should make sure that all events can be related to this process. Moreover, each event in the log needs to refer to a *single process instance*, often referred to as *case*. In Table 5.1, each request corresponds to a case, e.g., case 1. We also assume that events can be related to some *activity*. In Table 5.1, events refer to activities like *register request*, *check ticket*, and *reject*. These assumptions are quite natural in the context of process mining. All mainstream process modeling notations, including the ones discussed in Chap. 3, specify a process as a collection of activities such that the life-cycle of a single instance is described. Hence, the “case id” and “activity” columns in Table 5.1 represent the bare minimum for process mining. Moreover, events within a case need to be ordered. For example, event 35654423 (the execution of activity *register request* for Case 1) occurs before event 35654424 (the execution of activity *examine thoroughly* for the same case). Without ordering information it is of course impossible to discover causal dependencies in process models.

Table 5.1 also shows additional information per event. For example, all events have a *timestamp* (i.e., date and time information such as “30-12-2010:11.02”). This

Table 5.1 A fragment of some event log: each line corresponds to an event

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	...	
...

information is useful when analyzing performance related properties, e.g., the waiting time between two activities. The events in Table 5.1 also refer to *resources*, i.e., the persons executing the activities. Also *costs* are associated to events. In the context of process mining, these properties are referred to as *attributes*. These attributes are similar to the notion of variables in Chap. 4.

Figure 5.2 shows the tree structure of an event log. Using this figure we can list our assumptions about event logs.

- A *process* consists of *cases*.
- A case consists of *events* such that each event relates to precisely one case.
- Events within a case are *ordered*.

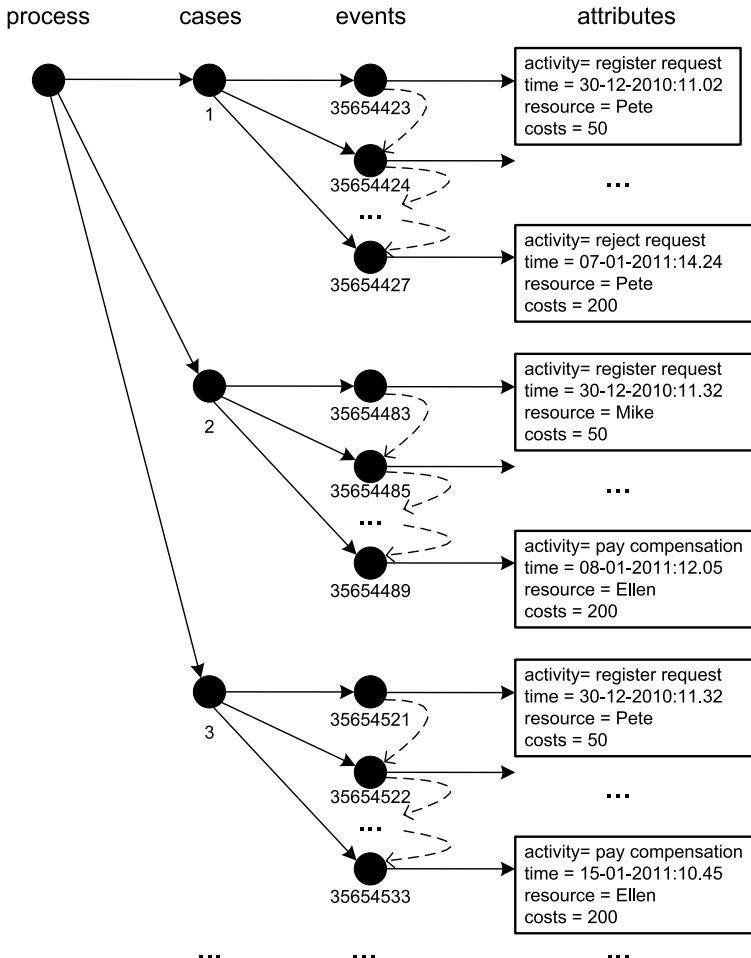


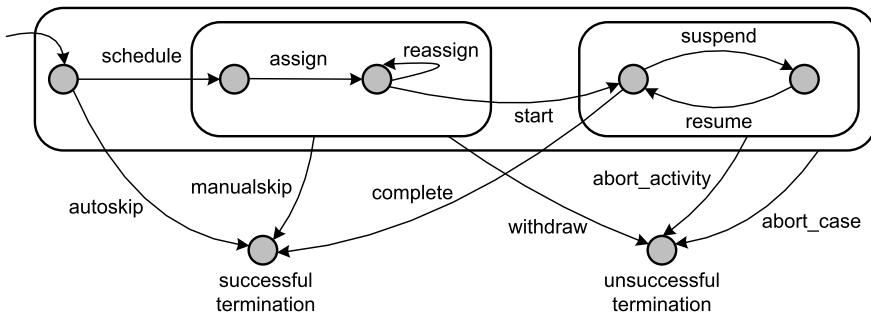
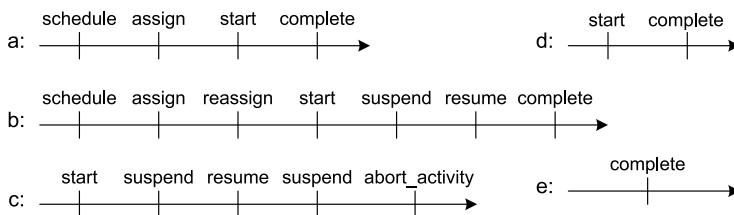
Fig. 5.2 Structure of event logs

- Events can have *attributes*. Examples of typical attribute names are activity, time, costs, and resource.

Not all events need to have the same set of attributes. However, typically, events referring to the same activity have the same set of attributes.

To be able to reason about logs and to precisely specify the requirements for event logs, we formalize the various notions.

Definition 5.1 (Event, attribute) Let \mathcal{E} be the *event universe*, i.e., the set of all possible event identifiers. Events may be characterized by various *attributes*, e.g., an event may have a timestamp, correspond to an activity, is executed by a particular person, has associated costs, etc. Let AN be a set of attribute names. For any event

**Fig. 5.3** Standard transactional life-cycle model**Fig. 5.4** Transactional events for five activity instances

$e \in \mathcal{E}$ and name $n \in AN$, $\#_n(e)$ is the value of attribute n for event e . If event e does not have an attribute named n , then $\#_n(e) = \perp$ (null value).

For convenience we assume the following standard attributes:

- $\#_{activity}(e)$ is the *activity* associated to event e .
- $\#_{time}(e)$ is the *timestamp* of event e .
- $\#_{resource}(e)$ is the *resource* associated to event e .
- $\#_{trans}(e)$ is the *transaction type* associated to event e , examples are *schedule*, *start*, *complete*, and *suspend*.

These are just examples. None of these attributes is mandatory. However, for these standard attributes we will assume some conventions. For example, timestamps should be non-descending in the event log. Moreover, we assume a time domain \mathcal{T} , i.e., $\#_{time}(e) \in \mathcal{T}$ for any $e \in \mathcal{E}$. The transaction type attribute $\#_{trans}(e)$ refers to the life-cycle of activities. In most situations, activities take time. Therefore, events may point out for example the start or completion of activities. In this book, we assume the *transactional life-cycle model* shown in Fig. 5.3.

Figure 5.4 shows some examples to explain the life-cycle model. The life-cycles of five activity instances are shown: a , b , c , d , and e . a is first scheduled for execution (i.e., an event e_1 with $\#_{trans}(e_1) = \text{schedule}$ and $\#_{activity}(e_1) = a$ occurs), then the activity is assigned to a resource (i.e., an event e_2 with $\#_{trans}(e_2) = \text{assign}$ and $\#_{activity}(e_2) = a$ occurs). Later the activity is started by this resource, and finally the

activity completes. Note that four events were recorded for this activity instance. Activity instance *b* has seven events associated to it. Compared to *a* the activity is reassigned (i.e., the resource that is supposed to execute the activity is changed), suspended (temporarily halted), and resumed. Of course it is possible to skip stages in the transactional life-cycle model, because events are not recorded or because certain steps are not necessary. Activity instance *d* in Fig. 5.4 has just two events; *e* just one, i.e., for *e* only the completion of the activity instance is recorded. Transaction type “autoskip” refers to an action by the system bypassing the activity. Transaction type “manualskip” refers to resource initiated skipping. Transaction types “abort_activity” and “abort_case” correspond to aborting the activity or the whole case. A “withdraw” event signals the situation in which the activity is canceled before it was started. Figure 5.3 shows all transaction types, their enabling, and their effect. For example, according to the transactional life-cycle model, “abort_activity” is only possible when the activity instance is running (i.e., started, suspended, or resumed).

Events can have many attributes. We often refer to the event by its activity name. Technically this is not correct. There may be many events that refer to the same activity name. Within a case these events may refer to the same activity instance (e.g., start and complete events) or different activity instances (e.g., in a loop). This distinction is particularly important when measuring service times, waiting times, etc. Consider, for example, the scenario in which the same activity is started twice for the same case, i.e., two activity instances are running in parallel, and then one of them completes. Did the activity that was started first complete or the second one? Fig. 5.5 illustrates the dilemma. Given the footprint of two starts followed by two completes of the same activity, there are two possible scenarios. In one scenario the durations of the two activity instances are 5 and 6. In the other scenario the durations of the activity instances are 9 and 2. Yet they leave the same footprint in the event log.

This problem can be addressed by adding information to the log or by using heuristics. This can be seen as a “secondary correlation problem”, i.e., relating two events within the same case. The primary correlation problem is to relate events to cases, i.e., process instances [53]. Figure 5.5 shows that even within one case there may be the need to correlate events because they belong to the same activity instance. When implementing systems, such information can easily be added to the logs; just provide an activity instance attribute to keep track of this. When dealing with existing systems this is not as simple as it seems. For example, when correlating messages between organizations there may be the need to scan the content of the message to find a suitable identifier (e.g., address or name). It is also possible to use heuristics to resolve most problems, e.g., in Fig. 5.5 one could just assume a first-in-first-out order and pick the first scenario. Moreover, one may introduce timeouts when the time between a start event and complete event is too long. For example, start events that are not followed by a corresponding complete event within 45 minutes are removed from the log.

Process mining techniques can be used to automatically discover process models. In these process models activities play a central role. These correspond to transitions

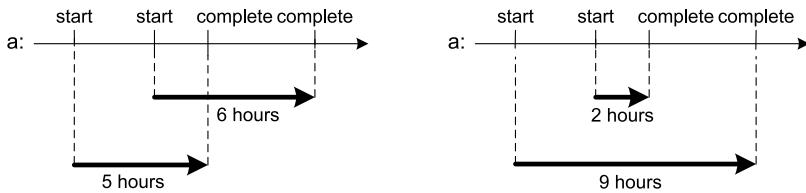


Fig. 5.5 Two scenarios involving two activity instances leaving the same footprint in the log

in Petri nets, tasks in YAWL, functions in EPCs, state transitions in transition systems, and tasks in BPMN. However, the transactional life-cycle model in Fig. 5.3 shows that there may be multiple events referring to the same activity. Some process mining techniques take into account the transactional model whereas others just consider atomic events. Moreover, sometimes we just want to focus on complete events whereas at other times the focus may be on withdrawals. This can be supported by filtering (e.g., removing events of a particular type) and by the concept of a *classifier*. A *classifier* is a function that maps the attributes of an event onto a label used in the resulting process model. This can be seen as the “name” of the event. In principle there can be many classifiers. However, only one is used at a time. Therefore, we can use the notation \underline{e} to refer to the name used in the process model.

Definition 5.2 (Classifier) For any event $e \in \mathcal{E}$, \underline{e} is the *name* of the event.

If events are simply identified by their activity name, then $\underline{e} = \#_{activity}(e)$. This means that activity instance a in Fig. 5.4 would be mapped onto $\langle a, a, a, a \rangle$. In this case the basic α -algorithm (not using transactional information) would create just one a transition. If events are identified by their activity name and transaction type, then $\underline{e} = (\#_{activity}(e), \#_{trans}(e))$. Now activity instance a would be mapped onto $\langle (a, schedule), (a, assign), (a, start), (a, complete) \rangle$ and the basic α -algorithm would create four transitions referring to a 's life-cycle. As shown in Sect. 6.2.4, transaction type attributes such as start, complete, etc. can be exploited to create a two-level process model that hides the transactional life-cycles of individual activities in subprocesses. It is also possible to use a completely different classifier, e.g., $\underline{e} = \#_{resource}(e)$. In this case events are named after the resources executing them. In this book we assume the classifier $\underline{e} = \#_{activity}(e)$ as the *default classifier*. This is why we considered the activity attribute to be mandatory in our initial examples. From now on, we only require a classifier.

Sequences

Sequences are the most natural way to present traces in an event log. When describing the operational semantics of Petri nets and transition systems, we also modeled behavior in terms of sequences. Given their importance, we introduce some useful operators on sequences.

For a given set A , A^* is the set of all finite sequences over A . A finite sequence over A of length n is a mapping $\sigma \in \{1, \dots, n\} \rightarrow A$. Such a sequence is represented by a string, i.e., $\sigma = \langle a_1, a_2, \dots, a_n \rangle$ where $a_i = \sigma(i)$ for $1 \leq i \leq n$. $|\sigma|$ denotes the length of the sequence, i.e., $|\sigma| = n$. $\sigma \oplus a' = \langle a_1, \dots, a_n, a' \rangle$ is the sequence with element a' appended at the end. Similarly, $\sigma_1 \oplus \sigma_2$ appends sequence σ_2 to σ_1 resulting a sequence of length $|\sigma_1| + |\sigma_2|$.

$hd^k(\sigma) = \langle a_1, a_2, \dots, a_{k \min n} \rangle$, i.e., the “head” of the sequence consisting of the first k elements (if possible). Note that $hd^0(\sigma)$ is the empty sequence and for $k \geq n$: $hd^k(\sigma) = \sigma$. $\text{pref}(\sigma) = \{hd^k(\sigma) \mid 0 \leq k \leq n\}$ is the set of prefixes of σ .

$tl^k(\sigma) = \langle a_{(n-k+1) \max 1}, a_{k+2}, \dots, a_n \rangle$, i.e., the “tail” of the sequence composed of the last k elements (if possible). Note that $tl^0(\sigma)$ is the empty sequence and for $k \geq n$: $tl^k(\sigma) = \sigma$.

$\sigma \uparrow X$ is the projection of σ onto some subset $X \subseteq A$, e.g., $\langle a, b, c, a, b, c, d \rangle \uparrow \{a, b\} = \langle a, b, a, b \rangle$ and $\langle d, a, a, a, a, a, a, d \rangle \uparrow \{d\} = \langle d, d \rangle$.

For any sequence $\sigma = \langle a_1, a_2, \dots, a_n \rangle$ over A , $\partial_{set}(\sigma) = \{a_1, a_2, \dots, a_n\}$ and $\partial_{multiset}(\sigma) = [a_1, a_2, \dots, a_n]$. ∂_{set} converts a sequence into a set, e.g., $\partial_{set}(\langle d, a, a, a, a, a, a, d \rangle) = \{a, d\}$. a is an element of σ , denoted as $a \in \sigma$, if and only if $a \in \partial_{set}(\sigma)$. $\partial_{multiset}$ converts a sequence into a multi-set, e.g., $\partial_{multiset}(\langle d, a, a, a, a, a, a, d \rangle) = [a^6, d^2]$. $\partial_{multiset}(\sigma)$ is also known as the *Parikh vector* of σ . These conversions allow us to treat sequences as sets or bags when needed.

An event log consists of cases and cases consist of events. The events for a case are represented in the form of a *trace*, i.e., a sequence of unique events. Moreover, cases, like events, can have attributes.

Definition 5.3 (Case, trace, event log) Let \mathcal{C} be the *case universe*, i.e., the set of all possible case identifiers. Cases, like events, have attributes. For any case $c \in \mathcal{C}$ and name $n \in AN$: $\#_n(c)$ is the value of attribute n for case c ($\#_n(c) = \perp$ if case c has no attribute named n). Each case has a special mandatory attribute *trace*, $\#_{\text{trace}}(c) \in \mathcal{E}^*$.¹ $\hat{c} = \#_{\text{trace}}(c)$ is a shorthand for referring to the trace of a case.

A *trace* is a finite sequence of events $\sigma \in \mathcal{E}^*$ such that each event appears only once, i.e., for $1 \leq i < j \leq |\sigma|$: $\sigma(i) \neq \sigma(j)$.

An *event log* is a set of cases $L \subseteq \mathcal{C}$ such that each event appears at most once in the entire log, i.e., for any $c_1, c_2 \in L$ such that $c_1 \neq c_2$: $\partial_{set}(\hat{c}_1) \cap \partial_{set}(\hat{c}_2) = \emptyset$.

If an event log contains timestamps, then the ordering in a trace should respect these timestamps, i.e., for any $c \in L$, i and j such that $1 \leq i < j \leq |\hat{c}|$: $\#_{\text{time}}(\hat{c}(i)) \leq \#_{\text{time}}(\hat{c}(j))$.

¹In the remainder, we assume $\#_{\text{trace}}(c) \neq \langle \rangle$, i.e., traces in a log contain at least one event.

Events and cases are represented using *unique* identifiers. An identifier $e \in \mathcal{E}$ refers to an event and an identifier $c \in \mathcal{C}$ refers to a case. This mechanism allows us to point to a specific event or a specific case. This is important as there may be many events having identical attributes, e.g., start events of some activity a may have been recorded for different cases and even within a case there may be multiple of such events. Similarly, there may be different cases that followed the same path in the process. These identifiers are just a technicality that helps us to point to particular events and cases. Therefore, they do not need to exist in the original data source and may be generated when extracting the data from different data sources.

Events and cases may have any number of attributes. Using the classifier mechanism, each event gets a name. Therefore, we often require events to have an activity attribute. Cases always have a trace attribute; $\hat{c} = \#_{trace}(c)$ is the sequence of events that have been recorded for c .

By formalizing event logs in this way, we *precisely* formulate the *requirements* we impose on event logs *without* discussing a concrete *syntax*. Moreover, we can use this formal representation to query the event log and use it as a starting point for analysis and reasoning. Some examples:

- $\{\#_{activity}(e) \mid c \in L \wedge e \in \hat{c}\}$ is the set of all activities appearing in log L .
- $\{\#_{resource}(e) \mid c \in L \wedge e \in \hat{c} \wedge \#_{trans}(e) = manualskip\}$ is the set of all resources that skipped an activity.
- $\{a \in \mathcal{A} \mid c \in L \wedge a = \#_{activity}(\hat{c}(1)) \wedge a = \#_{activity}(\hat{c}(|\hat{c}|))\}$ is the set of all activities that served as start and end activity for the same case.

Table 5.1 defines an event log in the sense of Definition 5.3. $L = \{1, 2, 3, 4, \dots\}$ is the set of cases shown in Table 5.1. $\hat{1} = \#_{trace}(1) = \langle 35654423, 35654424, 35654425, 35654426, 35654427 \rangle$ is the trace of case 1. $\#_{activity}(35654423) = register\ request$ is the activity associated to event 35654423. $\#_{time}(35654423) = 30-12-2010:11.02$ is the timestamp associated to this event. $\#_{resource}(35654423) = Pete$ is the resource doing the registration. $\#_{costs}(35654423) = 50$ are the costs associated to event 35654423. $\#_{activity}(35654424) = examine\ thoroughly$ is the activity associated to second event of case 1. Etc.

Depending on the attributes in the log, different types of analysis are possible. Figure 5.6 sketches possible results. The Petri net can be discovered by just using the activity attribute ($\#_{activity}(e)$). To measure durations of activities, one needs to have a transactional attribute ($\#_{trans}(e)$) to distinguish start from completion, and timestamps ($\#_{time}(e)$). To measure costs, the costs attribute is used ($\#_{costs}(e)$). Figure 5.6 also shows a *role* per activity and a *social network*. These have been discovered using the resource attribute ($\#_{resource}(e)$). For example, activities *decide* and *reinitiate request* require the role *manager* and Sara is the only one having this role. The social network in Fig. 5.6 shows how work is flowing through the organization, e.g., activities done by Sara are often followed by activities of Ellen. The thicker the connecting arc is, the more work is handed over from one person to another.

Table 5.1 happens to show unique id's for both events and cases, i.e., elements of the sets $\mathcal{E} = \{35654423, 35654424, 35654425, 35654426, 35654427, \dots\}$ (event universe) and $\mathcal{C} = \{1, 2, 3, 4, \dots\}$ (case universe) are shown *explicitly* in the table.

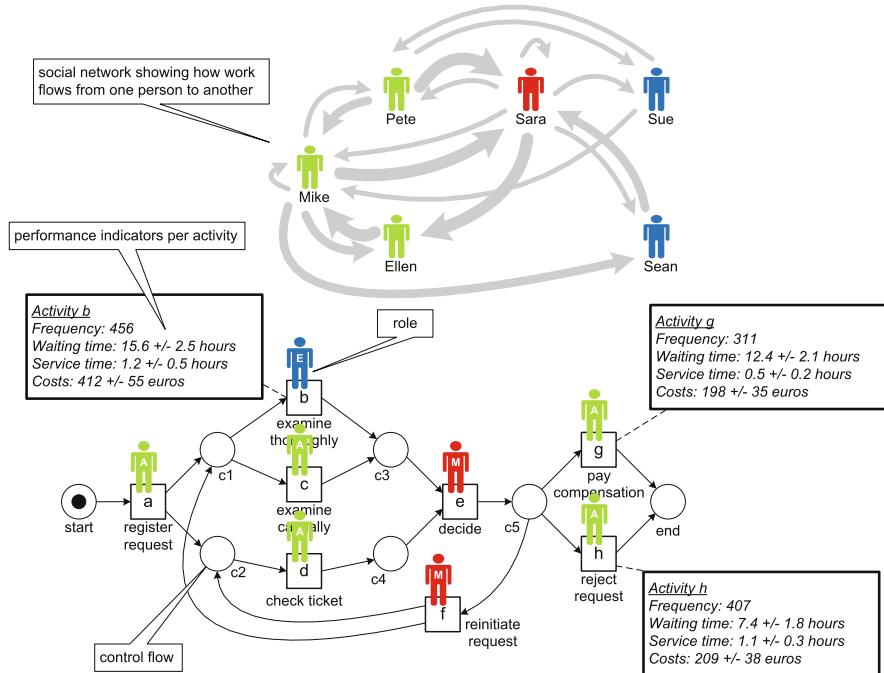


Fig. 5.6 Various types of process mining results based on the attributes in the event log

This is not mandatory; these identities are just used for mathematical convenience and have no further meaning. One can think of them as a symbolic key in a table or a position in an XML document. The reason for adding them is that this way it becomes easy to refer to a particular case or event. In fact, for simple algorithms like the α -algorithm, Definition 5.3 is a bit of overkill. See, for example, Table 2.2 in Sect. 2.3 showing the essential information used to construct a Petri net. If one is just interested in activity names (or some other classifier), the definition can be simplified drastically as is shown next.

Definition 5.4 (Simple event log) Let \mathcal{A} be a set of activity names. A *simple* trace σ is a sequence of activities, i.e., $\sigma \in \mathcal{A}^*$. A *simple* event log L is a multi-set of traces over \mathcal{A} , i.e., $L \in \mathbb{B}(\mathcal{A}^*)$.²

A simple event log is just a multi-set of traces over some set \mathcal{A} . For example $[\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$ defines a log containing 6 cases. In total there are $3 \times 4 + 2 \times 4 + 1 \times 3 = 23$ events. All cases start with a and end with d . In a simple log there are no attributes, e.g., timestamps and resource information are abstracted from. Moreover, cases and events are no longer uniquely identifiable. For

²Note that we still assume that each trace contains at least one element, i.e., $\sigma \in L$ implies $\sigma \neq \langle \rangle$.

example, the three cases following the sequence $\langle a, b, c, d \rangle$ in the simple event log $[\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$ cannot be distinguished.

Definition 5.5 (Transforming an event log into a simple event log) Let $L \subseteq \mathcal{C}$ be an event log as defined in Definition 5.3. Assume that a classifier has been defined: \underline{e} is the name of event $e \in \mathcal{E}$. This classifier can also be applied to sequences, i.e., $\langle e_1, e_2, \dots, e_n \rangle = \langle \underline{e}_1, \underline{e}_2, \dots, \underline{e}_n \rangle$. $\underline{L} = [\langle \hat{c} \rangle \mid c \in L]$ is the simple event log corresponding to L .

All cases in L are converted into sequences of (activity) names using the classifier. A case $c \in L$ is an identifier from the case universe \mathcal{C} . $\hat{c} = \#_{trace}(c) = \langle e_1, e_2, \dots, e_n \rangle \in \mathcal{E}^*$ is the sequence of events executed for c . $\langle \hat{c} \rangle = \langle \underline{e}_1, \underline{e}_2, \dots, \underline{e}_n \rangle$ maps these events onto (activity) names using the classifier.

If we apply this transformation to the event log shown in Table 5.1 while assuming the default classifier ($\underline{e} = \#_{activity}(e)$), then we obtain the event log

$$\begin{aligned}\underline{L} = & [\langle \text{register request}, \text{examine thoroughly, check ticket, decide, reject request} \rangle, \\ & \langle \text{register request}, \text{check ticket, examine casually, decide, pay compensation} \rangle, \\ & \langle \text{register request}, \text{examine casually, check ticket, decide, reinitiate request,} \\ & \quad \text{examine thoroughly, check ticket, decide, pay compensation} \rangle, \\ & \langle \text{register request}, \text{check ticket, examine thoroughly, decide, reject request} \rangle, \\ & \dots]\end{aligned}$$

Another classifier could have been used to create a simple log. For example, when using the classifier $\underline{e} = \#_{resource}(e)$, the following log is obtained:

$$\begin{aligned}\underline{L} = & [\langle \text{Pete, Sue, Mike, Sara, Pete} \rangle, \\ & \langle \text{Mike, Mike, Pete, Sara, Ellen} \rangle, \\ & \langle \text{Pete, Mike, Ellen, Sara, Sara, Sean, Pete, Sara, Ellen} \rangle, \\ & \langle \text{Pete, Mike, Sean, Sara, Ellen} \rangle, \\ & \dots]\end{aligned}$$

In this event log, the activity names have been replaced by the names of the people executing the activities. Such projections are used when constructing a social network.

In the remainder, we will use whatever notation is most suitable. Definition 5.3 specifies a precise but very generic description of an event log that can be used for various purposes. Definition 5.4 describes a very simple format without any attributes. This format is useful for explaining simple process discovery algorithms that are not using the information stored in additional attributes. For simple event logs we focus on a single attribute (typically the activity name). As shown, any event log L can be easily converted into a simple event log \underline{L} .

5.3 XES

Until 2010 the de facto standard for storing and exchanging event logs was *MXML* (Mining eXtensible Markup Language). MXML emerged in 2003 and was later adopted by the process mining tool ProM. Using MXML it is possible to store event logs such as the one shown in Table 5.1 using an XML-based syntax. *ProMimport* is a tool supporting the conversion of different data sources to MXML, e.g., MS Access, Aris PPM, CSV, Apache, Adept, PeopleSoft, Subversion, SAP R/3, Protos, CPN Tools, Cognos, and Staffware. MXML has a standard notation for storing timestamps, resources, and transaction types. Moreover, one can add arbitrary data elements to events and cases. The latter resulted in ad-hoc extensions of MXML where certain data attributes were interpreted in a specific manner. For example, *SA-MXML* (Semantically Annotated Mining eXtensible Markup Language) is a semantic annotated version of the MXML format used by the ProM framework. SA-MXML incorporates references between elements in logs and concepts in ontologies. For example, a resource can have a reference to a concept in an ontology describing a hierarchy of roles, organizational entities, and positions. To realize these semantic annotations, existing XML elements were interpreted in a new manner. Other extensions were realized in a similar manner. Although this approach worked quite well in practice, the various ad-hoc extensions also revealed shortcomings of the MXML format. This triggered the development of *XES* (eXtensible Event Stream) [64].

XES is the successor of MXML. Based on many practical experiences with MXML, the XES format has been made less restrictive and truly extendible. In September 2010, the format was adopted by the *IEEE Task Force on Process Mining* and became the de facto exchange format for process mining. The IEEE Standards Organization is currently evaluating XES with the aim to turn XES into an official IEEE standard. The format is supported by tools such as ProM (as of version 6), see www.xes-standard.org for detailed information about the standard.

Figure 5.7 shows the XES meta model expressed in terms of a UML class diagram. A XES document (i.e., XML file) contains one log consisting of any number of traces. Each trace describes a sequential list of events corresponding to a particular case. The log, its traces, and its events may have any number of attributes. Attributes may be nested. There are five core types: *String*, *Date*, *Int*, *Float*, and *Boolean*. These correspond to the standard XML types: *xs:string*, *xs:dateTime*, *xs:long*, *xs:double*, and *xs:boolean*. For example, *2011-12-17T21:00:00.000+02:00* is a value of type *xs:dateTime* representing nine o'clock in the evening of December 17th, 2011 in timezone GMT+2.

XES does not prescribe a fixed set of mandatory attributes for each element (log, trace, and event); an event can have any number of attributes. However, to provide semantics for such attributes, the log refers to so-called *extensions*. An extension gives semantics to particular attributes. For example the *Time extension* defines a timestamp attribute of type *xs:dateTime*. This corresponds to the $\#_{time}(e)$ attribute used in Sect. 5.2. The *Organizational extension* defines a resource attribute of type *xs:string*. This corresponds to the $\#_{resource}(e)$ attribute used in Sect. 5.2. Users can

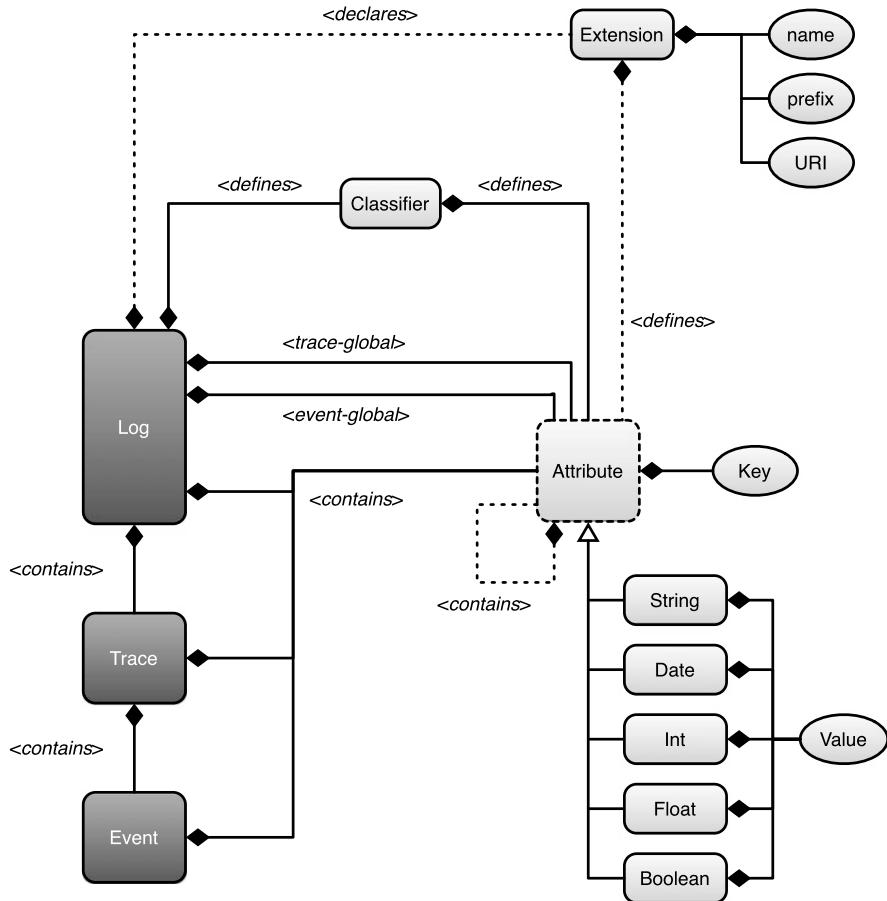


Fig. 5.7 Meta model of XES [64]. A log contains traces and each trace contains events. Logs, traces, and events have attributes. Extensions may define new attributes and a log should declare the extensions used in it. Global attributes are attributes that are declared to be mandatory. Such attributes reside at the trace or event level. Attributes may be nested. Event classifiers are defined for the log and assign a “label” (e.g., activity name) to each event. There may be multiple classifiers

define their own extensions. For example, it is possible to develop domain-specific or even organization-specific extensions. Figure 5.7 shows that a log declares the set of extensions to be used. Each extension may define attributes that are considered to be standard when the extension is used.

In Sect. 5.2, we used \mathcal{C} and \mathcal{E} to denote the case respectively event universe. This was used to be able to refer to a case and event. In XES such unique identifiers are not necessary. In fact, one can think of the position in the log as the identifier of an event or case.

XES may declare particular attributes to be *mandatory*. For example, it may be stated that any trace should have a name or that any event should have a timestamp.

For this purpose a log holds two lists of *global attributes*: one for the traces and one for the events.

XES supports the *classifier* concept described earlier (Definition 5.2). A XES log defines an arbitrary number of classifiers. Each classifier is specified by a list of attributes. Any two events that have the identical values with respect to these attributes are considered to be equal for that classifier. These attributes should be mandatory event attributes. For example, if a classifier is specified by both a name attribute and a resource attribute, then two events are mapped onto the same class if their name and resource attributes coincide.

The XES meta model shown in Fig. 5.7 does not prescribe a concrete syntax. In principle many serializations are possible. However, to exchange XES documents, a standard XML serialization is used. Figure 5.8 shows a fragment of the XES XML serialization of the event log of Table 5.1. In the example XES log three extensions are declared: *Concept*, *Time*, and *Organizational*. For each of these extensions a shorter prefix is given. These prefixes are used in the attribute names. For example, the *Time* extension defines an attribute *timestamp*. As shown in Fig. 5.8, this extension uses prefix *time*, therefore the timestamp of an event is stored using the key *time:timestamp*.

The example log in Fig. 5.8 specifies two lists of global attributes. Traces have one global attribute: attribute *concept:name* is mandatory for all traces. Events have three global attributes: attributes *time:timestamp*, *concept:name* and *org:resource* are mandatory for all events.

Three classifiers are defined in the XES log shown in Fig. 5.8. Classifier *Activity* classifies events based on the *concept:name* attribute. Classifier *Resource* classifies events based on the *org:resource* attribute. Classifier *Both* classifies events based on two attributes: *concept:name* and *org:resource*. Recall that Definition 5.2 already introduced the concept of a classifier: an event $e \in \mathcal{E}$ is classified as \underline{e} . For example, $\underline{e} = \#_{\text{resource}}(e)$ classifies events based on the resource executing the event.

For more information about the concrete syntax of XES we refer to www.xes-standard.org. However, the fragment shown in Fig. 5.8 already demonstrates that XES indeed operationalizes the concept of an event log as described in Definition 5.3. Moreover, the extension mechanism makes the format extendible while at the same time providing semantics for commonly used attributes. In the context of XES, five *standard extensions* have been defined. These extensions are described in the so-called *XESEXT XML* format [64]. Here we only mention a subset of standard attributes defined by these extensions.

- The *concept extension* defines the *name* attribute for traces and events. Note that the example XES file indeed uses *concept:name* attributes for traces and events. For traces, the attribute typically represents some identifier for the case. For events, the attribute typically represents the activity name. The concept extension also defines the *instance* attribute for events. This is used to distinguish different activity instances in the same trace. This extension can be used to resolve the dilemma shown in Fig. 5.5.

```

<?xml version="1.0" encoding="UTF-8" ?>
<extension name="Concept" prefix="concept" uri="http://.../concept.xesext"/>
<extension name="Time" prefix="time" uri="http://.../time.xesext"/>
<extension name="Organizational" prefix="org" uri="http://.../org.xesext"/>
<global scope="trace">
    <string key="concept:name" value="name"/>
</global>
<global scope="event">
    <date key="time:timestamp" value="2010-12-17T20:01:02.229+02:00"/>
    <string key="concept:name" value="name"/>
    <string key="org:resource" value="resource"/>
</global>
<classifier name="Activity" keys="concept:name"/>
<classifier name="Resource" keys="org:resource"/>
<classifier name="Both" keys="concept:name org:resource"/>
<trace>
    <string key="concept:name" value="1"/>
    <event>
        <string key="concept:name" value="register request"/>
        <string key="org:resource" value="Pete"/>
        <date key="time:timestamp" value="2010-12-30T11:02:00.000+01:00"/>
        <string key="Event_ID" value="35654423"/>
        <string key="Costs" value="50"/>
    </event>
    <event>
        <string key="concept:name" value="examine thoroughly"/>
        <string key="org:resource" value="Sue"/>
        <date key="time:timestamp" value="2010-12-31T10:06:00.000+01:00"/>
        <string key="Event_ID" value="35654424"/>
        <string key="Costs" value="400"/>
    </event>
    <event>
        <string key="concept:name" value="check ticket"/>
        <string key="org:resource" value="Mike"/>
        <date key="time:timestamp" value="2011-01-05T15:12:00.000+01:00"/>
        <string key="Event_ID" value="35654425"/>
        <string key="Costs" value="100"/>
    </event>
    <event>
        <string key="concept:name" value="decide"/>
        <string key="org:resource" value="Sara"/>
        <date key="time:timestamp" value="2011-01-06T11:18:00.000+01:00"/>
        <string key="Event_ID" value="35654426"/>
        <string key="Costs" value="200"/>
    </event>
    <event>
        <string key="concept:name" value="reject request"/>
        <string key="org:resource" value="Pete"/>
        <date key="time:timestamp" value="2011-01-07T14:24:00.000+01:00"/>
        <string key="Event_ID" value="35654427"/>
        <string key="Costs" value="200"/>
    </event>
</trace>
<trace>
    <string key="concept:name" value="2"/>
    <event>
        <string key="concept:name" value="register request"/>
        <string key="org:resource" value="Mike"/>
        <date key="time:timestamp" value="2010-12-30T11:32:00.000+01:00"/>
        <string key="Event_ID" value="35654483"/>
        <string key="Costs" value="50"/>
    </event>
    ...
</trace>
...
</log>
```

Fig. 5.8 Fragment of a XES file

- The *life-cycle extension* defines the *transition* attribute for events. When using the standard *transactional life-cycle model* shown in Fig. 5.3, possible values of this attribute are “schedule”, “start”, “complete”, “autoskip”, etc.
- The *organizational extension* defines three standard attributes for events: *resource*, *role*, and *group*. The resource attribute refers to the resource that triggered or executed the event. The role and group attributes characterize the (required) capabilities of the resource and the resource’s position in the organization. For example, an event executed by a sales manager may have role “manager” and group “sales department” associated to it.
- The *time extension* defines the *timestamp* attribute for events. Since such a timestamp is of type *xs:dateTime*, both a date and time are recorded.
- The *semantic extension* defines the *modelReference* attribute for all elements in the log. This extension is inspired by *SA-MXML*. The references in the log point to concepts in an ontology. For example, there may be an ontology describing different kinds of customers, e.g., Silver, Gold, and Platinum customers. Using the *modelReference* attribute a trace can point to this ontology thus classifying the customer.

Users and organizations can add new extensions and share these with others. For example, general extensions referring to costs, risks, context, etc. can be added. However, extensions may also be domain-specific (e.g., healthcare, customs, or retail) or organization-specific.

Currently, XES is supported by tools such as ProM, Nitro, XESame, Disco, Celonis, Minit, SNP Business Process Analysis, Rialto Process, and the reference implementation OpenXES. ProM is probably the most widely used process mining tool (see Sect. 11.3) providing a wide variety of process mining techniques. ProM 6 can load both MXML and XES files. Tools like Disco (www.fluxicon.com) can be used to quickly convert event logs into the XES (or MXML) format. XESame (www.processmining.org) generates XES files from collections of database tables. Here, the idea is that given a set of tables there may be different views possible (see also Sect. 5.5). Therefore, XES files serve as a view on the event data. OpenXES (www.openxes.org) is the XES reference implementation, an open source java library for reading, storing, and writing XES logs. OpenXES can easily be embedded in other tools and is able to efficiently (de)serialize large event logs into/from XML files. This frees software developers from developing tedious code to import and export event data.

Challenges when extracting event logs

Definition 5.3 provides a succinct formal definition of the requirements an event log needs to satisfy. XES operationalizes these requirements and provides a concrete syntax. Hence, the target format is well-defined. Nonetheless, extracting event logs may be very challenging. Here, we list the five most important challenges.

- **Challenge 1: Correlation**

Events in an event log are grouped per case. This simple requirement can be quite challenging as it requires *event correlation*, i.e., events need to be related to each other. Consider, for example, event data scattered over multiple tables or even multiple systems. How to identify events and their corresponding cases? Also consider messages exchanged with other organizations. How to relate responses to the original requests? When designing logging functionality from scratch, it is quite easy to address this problem. However, when dealing with legacy and a variety of interconnected systems, additional efforts are needed to correlate events; see [53] for an example of an approach to correlate events without any a-priori information.

- **Challenge 2: Timestamps**

Events need to be ordered per case. In principle, such ordering does not require timestamps. However, when merging data from different sources, one typically needs to depend on timestamps to sort events (in order of occurrence). This may be problematic because of multiple clocks and delayed recording. For example, in an X-ray machine the different components have local clocks and events are often queueing before being recorded. Therefore, there may be significant differences between the actual time an event takes place and its timestamp in the log. As a result the ordering of events is unreliable, e.g., cause and effect may be reversed. In other applications, timestamps may be too coarse. In fact, many information systems only record a date and not a timestamp. For example, most events in a hospital are recorded in the hospital information system based on a patient id and a date, without storing the actual time of the test or visit. As a result it is impossible to reconstruct the order of events on a given day. One way to address this problem is to assume only a partial ordering of events (i.e., not a total order) and subsequently use dedicated process mining algorithms for this. Another way to (partially) address the problem is to “guess” the order based on domain knowledge or frequent patterns across days.

- **Challenge 3: Snapshots**

Cases may have a lifetime extending beyond the recorded period, e.g., a case was started before the beginning of the event log or was still running when the recording stopped. Therefore, it is important to realize that event logs typically just provide a *snapshot* of a longer running process. When the average duration of a case is short compared to the length of the recording, it is best to solve this problem by removing incomplete cases. In many cases, the initial and final activities are known, thus making it easy to filter the event log: simply remove all cases with a missing “head” or “tail”. However, when the average duration of a case is of the same order of magnitude as the length of the recording, it becomes difficult to discover end-to-end processes.

- **Challenge 4: Scoping**

The fourth problem is the scoping of the event log. Enterprise information systems may have thousands of tables with business-relevant data (cf. a typical SAP installation). How to decide which tables to incorporate? Domain knowledge is needed to locate the required data and to scope it. Obviously, the desired scope depends on both the available data and the questions that need to be answered.

- **Challenge 5: Granularity**

In many applications, the events in the event log are at a different level of granularity than the activities relevant for end users. Some systems produce low-level events that are too detailed to be presented to stakeholders interested in managing or improving the process. Fortunately, there are several approaches to preprocess low-level event logs. For example, in [77] it is shown that frequently appearing low-level patterns can be abstracted into events representing activities.

The availability of high-quality event logs is essential for process mining. Moreover, good event logs can serve many other purposes. Sometimes the term *business process provenance* is used to refer to the systematic collection of the information needed to reconstruct what has actually happened in the business process. From an auditing point of view the systematic, reliable, and trustworthy recording of events is essential. The term “provenance” originates from scientific computing [39]. Here, provenance information is recorded to ensure that scientific experiments are reproducible. High-quality event logs that cannot be tampered with make sure that “history cannot be rewritten or obscured” and serve as a solid basis for process improvement and auditing. Therefore, XES should be seen in a provenance context that extends beyond process discovery and includes topics such as conformance checking. We will elaborate on this topic in Part IV.

5.4 Data Quality

From a practical point of view *data quality* is of the utmost importance for the success of process mining. If event data is missing or cannot be trusted, then the results of process mining are less valuable. To be able to discuss the quality of event data, we first conceptualize event data using the definitions in Sect. 5.2 and the XES meta model presented in Sect. 5.3. The conceptualization is used to systematically classify data quality problems. After the problems have been identified, 12 guidelines for logging are given [147].

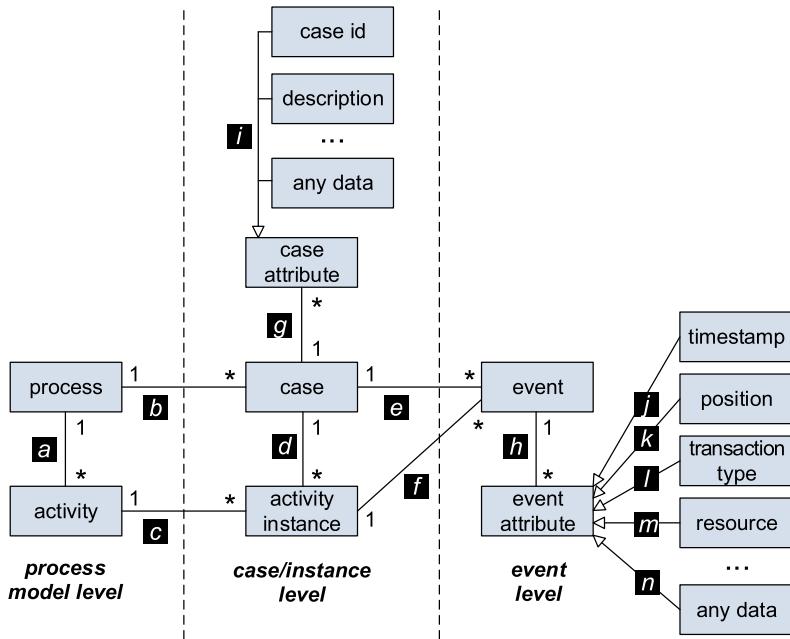


Fig. 5.9 Basic logging concepts conceptualized using a class diagram

5.4.1 Conceptualizing Event Logs

Section 5.2 introduced various event-log-related notions: events, event attribute names and values, activities, timestamps, transaction types, resources, cases, traces, and case attribute names and values. Also subtle notions such as classifiers and activity instances were discussed. More or less the same constructs were discussed in the context of XES in Sect. 5.3. XES is extendible and allows for the definition of domain-specific logs; however, to discuss data quality issues, we consolidate things in a simple class diagram.

Figure 5.9 aims to list the key ingredients of an event log. Anything shown in this class model can be captured using XES or the formal definitions provided before. However, using Fig. 5.9 we can discuss the key concepts without being distracted by the formalities of Sect. 5.2 and the technicalities of XES described in Sect. 5.3.

In Fig. 5.9, three levels are identified: *process model level*, *case/instance level*, and *event level*. The *case/instance level* shown in Fig. 5.9 consists of *cases* and *activity instances* that connect *processes* and *activities* in the model to *events* in the event log. When modeling, cases and activity instances only exist in abstract form. Only when a process model is instantiated, we can point to concrete cases, activity instances, and events. When observing a real process we are confronted with concrete instances of the process (i.e., cases). The same holds for activities and activity instances. Within the same case (i.e., process instance) there may be

multiple instances of the same activity. For instance, some check activity may be performed multiple times for the same customer request.

The class diagram shown in Fig. 5.9 shows the following associations and cardinalities:

- a Each process may have an arbitrary number of activities, but each activity belongs to precisely one process.
- b Each case belongs to precisely one process.
- c Each activity instance refers to precisely one activity.
- d Each activity instance belongs to precisely one case; there may be several activity instances for each activity/case combination.
- e Each event refers to precisely one case.
- f Each event corresponds to one activity instance; for the same activity instance there may be multiple events.
- g Each case attribute refers to one case; each attribute has a name and a value, e.g., “(birthdate, 29-01-1966)”.
- h Each event attribute refers to one event and is characterized by a name and a corresponding value, e.g., “(costs, \$199.99)”.
- i There are different subclasses of case attributes, e.g., the description of a case, case identifier, start time of case, etc. Case attributes are invariant, i.e., they do not change while the corresponding events of the case occur.
- j–n There are different subclasses of event attributes, e.g., the time of occurrence of the event (j), the position in trace (k), the transaction type (l), the resource causing the event (m), or any other type of attribute data (costs, risk, age, etc.).

The attributes attached to events provide valuable information that can be aggregated and mapped onto the process model level. For instance, timestamps can be used to compute the mean waiting time for an activity. Resource attributes attached to events can be used to learn working patterns and allocation rules. Cost information can be projected onto process models to see inefficiencies.

In most cases, event data are provided as a table, a CSV (Comma Separated Values) file, or a spreadsheet (e.g., Excel file) where each row corresponds to an event. This is illustrated by Fig. 5.10. The “dashed boxes” refer to attributes that are derivable from the relations between events, cases, activity instances, activities and processes. For example, the *process* attribute of an event can be derived by following relations *e* and *b*. The *activity* attribute of an event can be derived via relations *f* and *c*. Note that we abstract from case attributes in the latter classification of quality problems. We will focus on the key entities *case*, *activity instance*, and *event* as highlighted in Fig. 5.10.

Per event attribute, it is indicated whether the attribute is mandatory. The *process* attribute of an event is optional. If the attribute is missing, we assume there is just one process.

The *activity instance* attribute is also optional. In many data sets this information will be missing. For example, only complete events are recorded, making activity

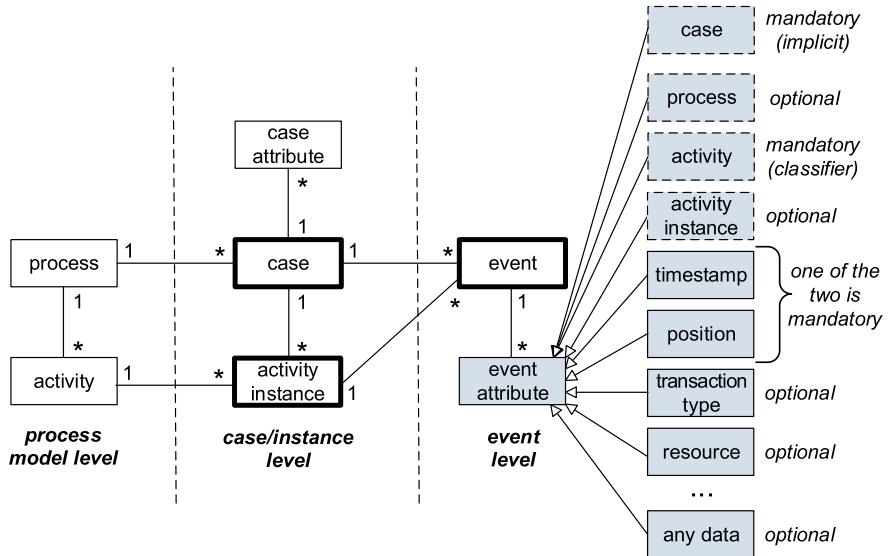


Fig. 5.10 Class diagram zooming in on event attributes and later used to classify data quality problems

instances singletons. If there are start and complete events but not explicit activity instances, then one may use heuristics to derive activity instances. Under the condition that start and complete events alternate, it is possible to deterministically derive activity instances. However, Fig. 5.5 shows an example where alternative explanations exist. In case of overlapping activity instances, transactional information is not sufficient. If start or complete events are missing, similar problems emerge. Heuristics may be used to solve these problems (see Sect. 5.2). Moreover, it is always possible to consider each event as a singleton activity instance. Obviously, such solutions may introduce data quality problems. When start events cannot be related to complete events, it is impossible to measure service times and resource utilization accurately.

In most cases, events have *timestamps* determining the *position* in a trace. Most control-flow discovery algorithms only use the ordering of events within a case as input. Moreover, multiple events may have the same timestamp. In some cases, timestamps are rather coarse (minutes or even days). If timestamps are not fine-grained enough, then the precision of the results is impacted (e.g., mean waiting time). Some process mining algorithms may be based on partial ordered traces rather than totally ordered traces. The lack of a total order may be caused by a lack of precision or by explicit information on causalities. As indicated in Fig. 5.10, we assume that at least the timestamp or position in the trace is known.

The other two standard event attributes—*transaction type* and *resource*—are also optional (see Sect. 5.2). Moreover, there may be additional data attributes related to costs, volume, risks, context, etc. (named *any data* in Fig. 5.10).

5.4.2 Classification of Data Quality Issues

Figure 5.10 summarizes the key concepts related to event data. These are used to discuss *data quality* issues. We consider three main entities (*case*, *activity instance*, and *event*) and nine event attributes (*case*, *process*, *activity*, *activity instance*, *timestamp*, *position*, *transaction type*, *resource*, and *any data*). This allows us to create a *classification* of data quality problems. This classification is related to the challenges mentioned in the context of XES (Sect. 5.3) and is inspired by [80, 98].

First, we consider the main entities (case, activity instance, and event) and not the attributes. At the entity level there are three potential problems:

- (*Missing in log*) The entity exists (or existed) in reality, but was not recorded. For example, an event (e.g., taking a blood sample) occurred but it was not captured by the information system.
- (*Missing in reality*) The entity does not exist and never existed in reality, but was recorded. For example, a scheduled doctor’s appointment never took place due to an emergency, but it was recorded by the information system anyway.
- (*Concealed in log*) The entity was recorded and exists (or existed) in reality, but it is hidden in a larger less structured data set. For example, the same entity may appear multiple times in the event log. The scope of the data set may also be much larger than needed for analysis. The event log may be a “mashup” of different data sources creating such challenges. It may be far from trivial to select, identify, and deduplicate entities.

Table 5.2 provides an overview of data quality problems at the entity level. For example, the cell *EV-MIL* refers to missing events.

Table 5.3 classifies problems related to *event attributes*. There are three potential problems related to such attributes:

- (*Missing attribute*) The attribute has not been recorded for a particular event. For example, the timestamp of an event is missing.
- (*Incorrect attribute*) The recorded value of the event attribute is wrong. For example, an event is related to another case.
- (*Imprecise attribute*) The value of the event attribute is too imprecise. For example, the value of a timestamp is too coarse-grained or the address is incomplete.

Table 5.3 combines the above problem types with the different types of event attributes identified in Fig. 5.10. For example, the cell *CASE-MIS* refers to events that cannot be related to a case.

Cell *TS-INC* refers to incorrect timestamps. For example, a patient in a hospital gets his medication at noon, but the doctor enters this in the hospital information system later in the afternoon. As a result, the timestamp of recording is different from the actual time of the event thus potentially creating a data quality problem.

Cell *TS-IMP* refers to timestamps that are too coarse-grained. In a hospital some events may only have a date (“24-3-2016”). Hence, the ordering of events on the same day is lost. The desired precision depends on the process to be analyzed. For example, when analyzing software systems millisecond precision may still be too

Table 5.2 Type of problem (*MIL*, *MIR*, or *CIL*) versus the entity (*CASE*, *AI*, or *EV*) affected

Entity	Type of problem		
	Missing in log (<i>MIL</i>)	Missing in reality (<i>MIR</i>)	Concealed in log (<i>CIL</i>)
Case (<i>CASE</i>)	A case is missing in the event log, e.g., a customer order got flushed.	A case that never existed was added to the log, e.g., by inadvertently entering an improper identifier a fictive case is created.	A case is “hidden” in larger data set, e.g., customer orders, order lines, and deliveries with overlapping identifiers are intermingled in a single log.
Activity instance (<i>AI</i>)	An activity instance is missing in the event log, e.g., the start and complete of a production step are not related through an activity instance.	An activity instance that never existed was added to the log.	An activity instance is “hidden” in larger data set.
Event (<i>EV</i>)	An event is missing in the event log, e.g., a medical test was not recorded in the event log.	An event that never occurred was inadvertently added to the log, e.g., a check that was never conducted was recorded to feign compliance.	An event is “hidden” in larger data set, e.g., identifying a security breach from transactional data having a much broader scope.

coarse-grained. Nanosecond precision may be required to analyze delays in automated processes.

Cell *RES-IMP* refers to events that do not refer to a specific resource, but a group of resources having the same role or working in the same department. This makes it impossible to analyze queueing and workload at the level of individuals.

Table 5.3 focuses on event attributes. When attributes are related to cases, the same problems may appear (missing, incorrect, or imprecise), but such data quality problems are not listed here.

Table 5.4 shows another quality dimension orthogonal to the classification given thus far. Data quality problems may persist and occur *continuously* throughout the event log that is analyzed. Problems may seem irregular and occur *intermittent*. Unknown intermittent data quality problems may be more problematic than known problems that occur continuously. For example, if a check is recorded in 80% of cases, one may derive incorrect conclusions assuming that all checks were logged. Data quality problems are classified as *changing* if clear patterns can be identified. For example, approvals are never recorded on Sunday or the granularity of timestamps improved after the installation of the new system.

The three recurrence categories in Table 5.4 (*CONT*, *INT*, and *CHNG*) can be combined with Tables 5.2 and 5.3. Some examples:

Table 5.3 Type of problem (*MIS*, *INC*, *IMP*) versus the event attribute affected, e.g., cell *TS-MIS* refers to missing timestamps

Attribute	Type of problem		
	Missing attribute (<i>MIS</i>)	Incorrect attribute (<i>INC</i>)	Imprecise attribute (<i>IMP</i>)
Case (<i>CASE</i>)	The event does not refer to a case.	The event refers to the wrong case.	The event may be related to multiple cases due to ambiguity.
Process (<i>PROC</i>)	The event cannot be related to a specific process.	The event refers to an unrelated process.	The event may be associated to multiple processes in an ambiguous manner.
Activity (<i>ACT</i>)	The event does not refer to an activity.	The event refers to another activity.	The event may be related to multiple activities due to imprecise labels.
Activity instance (<i>AI</i>)	The event does not refer to an activity instance.	The event refers to the wrong activity instance.	The event may be inadvertently related to multiple activity instances.
Timestamp (<i>TS</i>)	The event has no timestamp.	The event has an incorrect timestamp, e.g., a wrong date was entered or the event was recorded at a later point in time.	The event has a timestamp that is too coarse-grained, e.g., only a date was recorded.
Position (<i>POS</i>)	Ordering information for the event is missing (may be reconstructed using timestamps).	The event appears at the wrong position in the event log (e.g., the ordering of events is not consistent with the timestamps).	Ordering information for the event is partially lost.
Transaction type (<i>TT</i>)	The event has no transaction type (start, complete, etc.).	The transaction information is wrong.	The event may be inadvertently related to multiple transaction types.
Resource (<i>RES</i>)	The event is not related to any resource.	The event is related to the wrong resource.	The event may be related to multiple resources, e.g., only the role or department of the resources is recorded.
Any data (<i>ANY</i>)	A data attribute (e.g., costs) is missing for the event.	A data attribute has the wrong value (e.g., wrong amount).	The event has an attribute value that is too coarse-grained (e.g., street name is given but number is missing).

Table 5.4 Recurrence of data quality problems: A particular problem (e.g., a missing attribute) may persist, repeat in an unpredictable manner, or return periodically

Recurrence	Examples
Continuous (CONT)	A precise timestamp is missing for all medical examination events. All cases handled by the Eindhoven branch are missing in the log. The name of the responsible doctor is never recorded.
Intermittent (INT)	Some events have precise timestamps whereas for other events only the date is known. Blood pressure measurements that have been performed are not always recorded (e.g., depending on workload). Some nurses repeatedly forget to enter the name of the responsible doctor.
Changing (CHNG)	In the second week of January, timestamps were missing due to software problems. In weekends, the resource attribute is not recorded due to understaffing. In the second semester, intermediate tests were not recorded.

- *RES-MIS-CONT* refers to the problem that the resource attribute is never recorded.
- *EV-MIL-INT* refers to the problem that events are sometimes missing from the log.
- *TS-IMP-CHNG* refers to the problem that events have imprecise timestamps in certain periods, e.g., before the new software system was installed only dates were recorded.

In total $((3 \times 3) + (9 \times 3)) \times 3 = 108$ data quality problems can be identified using the three tables. These include the 17 problems identified in [80] and the 27 problems identified in [98]. The empirical investigation reported in [98] shows that *EV-MIL-**, *TS-IMP-**, and *RES-IMP-** are among the most frequent data quality problems in hospitals.

5.4.3 Guidelines for Logging

The data quality problems just described illustrate that the input side of data analysis is often neglected. Event data are often seen as a by-product. For example, the data is there for financial reasons or simply because a programmer decided to put a write statement in the code. Since the “input side of process mining” is vital, we now discuss the *12 guidelines for logging* introduced in [147]. These guidelines make no assumptions on the underlying technology used to record event data.

In this section, we use a rather loose definition of event data: events simply refer to “things that happen” and are described by *references* and *attributes*. *References* have a *reference name* and an *identifier* that refers to some object (person, case, ticket, machine, room, etc.) in the universe of discourse. *Attributes* have a *name* and a *value*, e.g., *age* = 49 or *time* = “19-12-2015 03:14:00”. In Fig. 5.10, we did not make a distinction between references and attributes. However, it is easy to add

this dimension. Based on these concepts, we define our 12 *Guidelines for Logging (GL1–GL12)* [147].

To create an event log from such “raw events”, (1) we need to select the events relevant for the process at hand, (2) events need to be correlated to form process instances (cases), (3) events need to be ordered using timestamp information (or have an explicit order), and (4) event attributes need to be selected or computed based on the raw data (resource, cost, etc.). The guidelines for logging refer both to the availability of raw event data and the transformation process. The aim is to improve data quality and avoid the issues listed in Sect. 5.4.2. Moreover, the guidelines also emphasize the correct interpretation of event data.

- GL1** *Reference and attribute names should have clear semantics, i.e., they should have the same meaning for all people involved in creating and analyzing event data.* Different stakeholders should interpret event data in the same way.
- GL2** *There should be a structured and managed collection of reference and attribute names.* Ideally, names are grouped hierarchically (like a taxonomy or ontology). A new reference or attribute name can only be added after there is consensus on its value and meaning. Also consider adding domain or organization-specific extensions (see, for example, the extension mechanism of XES described in Sect. 5.3).
- GL3** *References should be stable (e.g., identifiers should not be reused or rely on the context).* For example, references should not be time, region, or language dependent. Some systems create different logs depending on the language settings. This is unnecessarily complicating analysis.
- GL4** *Attribute values should be as precise as possible. If the value does not have the desired precision, then this should be indicated explicitly (e.g., through a qualifier).* For example, if for some events only the date is known but not the exact timestamp, then this should be stated explicitly.
- GL5** *Uncertainty with respect to the occurrence of the event or its references or attributes should be captured through appropriate qualifiers.* For example, due to communication errors, some values may be less reliable than usual. Note that uncertainty is different from imprecision.
- GL6** *Events should be at least partially ordered. The ordering of events may be stored explicitly (e.g., using a list) or implicitly through an attribute denoting the event’s timestamp.* If the recording of timestamps is unreliable or imprecise, there may still be ways to order events based on observed causalities (e.g., usage of data).
- GL7** *If possible, also store transactional information about the event (start, complete, abort, schedule, assign, suspend, resume, withdraw, etc.).* Having start and complete events allows for the computation of activity durations. It is recommended to explicitly link events to activity instances to be able to relate events belonging to the same activity occurrence. Without references to activity instances it may not always be clear which events belong together, e.g., which start event corresponds to which complete event.

- GL8** *Perform regularly automated consistency and correctness checks to ensure the syntactical correctness of the event log.* Check for missing references or attributes, and reference/attribute names not agreed upon. Event quality assurance is a continuous process (to avoid degradation of log quality over time).
- GL9** *Ensure comparability of event logs over time and different groups of cases or process variants.* The logging itself should not change over time (without being reported). For comparative process mining, it is vital that the same logging principles are used. If for some groups of cases, some events are not recorded even though they occur, then this may suggest differences that do not actually exist.
- GL10** *Do not aggregate events in the event log used as input for the analysis process.* Aggregation should be done during analysis and not before (since it cannot be undone). Event data should be as “raw” as possible.
- GL11** *Do not remove events and ensure provenance. Reproducibility is key for process mining.* For example, do not remove a student from the database after he dropped out since this may lead to misleading analysis results. Mark objects as not relevant (a so-called “soft delete”) rather than deleting them: concerts are not deleted—they are canceled; employees are not deleted—they are fired, etc.
- GL12** *Ensure privacy without losing meaningful correlations.* Sensitive or private data should be removed as early as possible (i.e., before analysis). However, if possible, one should avoid removing correlations. For example, it is often not useful to know the name of a student, but it may be important to still be able to use his high school marks and know what other courses he failed. Hashing can be a powerful tool in the trade-off between privacy and analysis.

The guidelines and classification aim to make the reader aware of data quality problems directly influencing the results of process mining.

5.5 Flattening Reality into Event Logs

In order to do process mining, events need to be related to cases. As indicated before, this is natural as a process model describes the life-cycle of a case of a particular type. All activities in a conventional process model (independent of the notation used) correspond to status changes of such a case. We will refer to such process models as *flat models*. In this book, we adopt this (often hidden) assumption associated to all mainstream process modeling notations. However, it is important to realize that *real-life processes are not flat*. We use a simple example to illustrate this.

Consider the class diagram shown in Fig. 5.11 describing a database consisting of four tables. Table *Order* contains information about orders. For example, each record in the *Order* table has a unique order number, refers to a customer, and has an associated amount. Multiple products can be ordered in one order. Therefore,

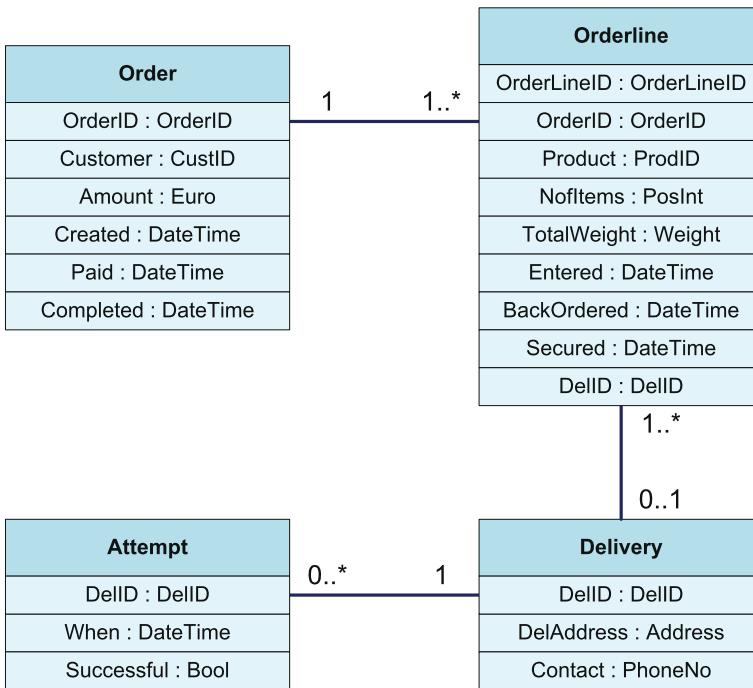


Fig. 5.11 Class diagram showing the relations between orders, order lines, deliveries, and delivery attempts

Table *Orderline* holds information about individual order lines. Records in the *Orderline* table refer to orders in the *Order* table. Figure 5.11 shows that each order line corresponds to one order and each order corresponds to one or more order lines. One order line describes which type of product is ordered, the quantity, and weight. Table *Delivery* holds information about deliveries. Each delivery corresponds to a collection of order lines. These order lines are delivered to a particular address. To deliver the corresponding collection of products, multiple attempts may be needed. Table *Attempts* stores information about these attempted deliveries. An attempt may be successful or not. If not, another attempt is made at a later point in time. Figure 5.11 shows that each delivery corresponds to zero or more attempts and one or more order lines. Each order line corresponds to at most one delivery.

Table 5.5 shows a small fragment of a larger *Order* table. For each order, up to three timestamps are recorded. The timestamp in the *Created* column denotes the time at which the order was created. The *Paid* column denotes the time at which the order was paid and the *Completed* column denotes the time at which the order was completed. Table 5.5 shows several *null* timestamps. This indicates that the corresponding events did not take place yet.

Table 5.6 shows some example records of the *Orderline* table. Each line refers to a particular product. For example, order line 112346 corresponds to two iPod nanos that are part of order 91245. Each order line refers to an order and to a delivery

Table 5.5 Some records of the *Order* table

Order					
OrderID	Customer	Amount	Created	Paid	Completed
91245	John	100	28-11-2011:08.12	02-12-2011:13.45	05-12-2011:11.33
91561	Mike	530	28-11-2011:12.22	03-12-2011:14.34	05-12-2011:09.32
91812	Mary	234	29-11-2011:09.45	02-12-2011:09.44	04-12-2011:13.33
92233	Sue	110	29-11-2011:10.12	null	null
92345	Kirsten	195	29-11-2011:14.45	02-12-2011:13.45	null
92355	Pete	320	29-11-2011:16.32	null	null
...

(if already created). For each order line, up to three timestamps are recorded: the time of entering the order line (column *Entered*), the time of back ordering (column *BackOrdered*), and the time of securing the item (column *Secured*). A *null* value indicates that the corresponding event did not take place (yet). Typically, only few order lines will be back-ordered (i.e., most rows will have a null value in the *BackOrdered* column). A backorder is an order line that cannot be delivered because of a lack of sufficient inventory. Therefore, the inventory needs to be replenished before the backorder can be delivered. Since only few order lines become backorders, column *BackOrdered* has many null values. Once the products are available and reserved for a particular order line, the corresponding timestamp is added in column *Secured*.

Information about deliveries is stored in the *Delivery* table shown in Table 5.7. For each delivery, an address and a phone number are recorded. Each delivery refers to a collection of order lines and may require multiple attempts.

Attempts to deliver products are recorded in the *Attempt* table. Table 5.8 shows some example attempts. An attempt has a timestamp and refers to a delivery (column *DellID*). Delivery 882345 required three attempts before the corresponding set of order lines could be delivered successfully. Delivery 882346 required only one attempt.

The four tables show only a snapshot of the available data. Orders that have not yet been fully handled may have many null values.

The database consisting of tables *Order*, *Orderline*, *Delivery*, and *Attempts* is a bit artificial and its design could be improved. For example, the tables with multiple timestamps could have been split into multiple tables. Moreover, in an ERP system like SAP much more detailed information is stored. Hence, the four tables are an oversimplification of reality and only serve as a means to explain the problem of *flattening reality for process mining*.

Clearly the timestamps in the four tables correspond to events related to the “overall ordering and delivery” process. However, when creating an event log, each event needs to be associated to a particular case. Therefore, we need to flatten the four tables into one table with a “case id” column. However, one can choose from four types of cases: orders, order lines, deliveries, and attempts. Any record in one of the four tables potentially corresponds to a case. Which one to choose?

Table 5.6 Part of the *Orderline* table: each record corresponds to an order line. An order line refers to an order (column *OrderID*) and a delivery (column *DelID*)

Table 5.7 Some records of the *Delivery* table

Delivery		
DellID	DelAddress	Contact
882345	5513VJ-22a	0497-2553660
882346	5513XG-45	040-2298761
...

Table 5.8 Part of the *Attempt* table

Attempt		
DellID	When	Successful
882345	05-12-2011:08.55	false
882345	06-12-2011:09.12	false
882345	07-12-2011:08.56	true
882346	05-12-2011:08.43	true
...

Let us assume that we are mainly interested in orders. Therefore, we let each case correspond to a record in table *Order*. Table *Order* has up to three timestamps per record. Hence, only three events per case can be found if only the *Order* table is considered and information about order lines and deliveries related to an order remains unused. When using only records from the *Order* table, control-flow discovery will most likely return a sequential process consisting of three steps: *create*, *pay*, and *complete*. To obtain a process model containing more activities, we need to consider the other tables. By using the references in the tables, orders can be related to order lines. In turn, order lines can be related to deliveries and the corresponding attempts. For example, order lines 112345, 112346 and 112347 refer to order 91245. Figure 5.12 shows all events that can be found by searching for all records that can be related to order 91245. The rectangles refer to concrete records in the four tables. The rounded rectangles refer to possible events and their attributes. All events in the figure refer to case 91245. As Fig. 5.12 shows, order 91245 is related to order lines 112345, 112346 and 112347, and deliveries 882345 and 882346. For delivery 882345 there are three corresponding attempt records and for delivery 882346 only one.

The top three events in Fig. 5.12 (the events directly connected to the root node) would have been the only events if only the *Order* table would have been considered. There are also various intermediate selections possible, resulting in subsets of the events shown in Fig. 5.12. For example, only the top ten events remain if only the *Order* and *Orderline* tables are considered, thus abstracting from deliveries.

Table 5.9 shows an event log using the selection illustrated by Fig. 5.12. As before, each line corresponds to an event. The *case id* column shows how events are correlated, i.e., each event refers to an order. The *activity* column names events as already shown Fig. 5.12. The *timestamp* column shows the date and time associated to the event. The *other attributes* column shows additional attributes. Depending on the type of activity, different attributes are recorded. Table 5.9 shows that there is quite some redundancy in the event log. This is partly unavoidable due to the

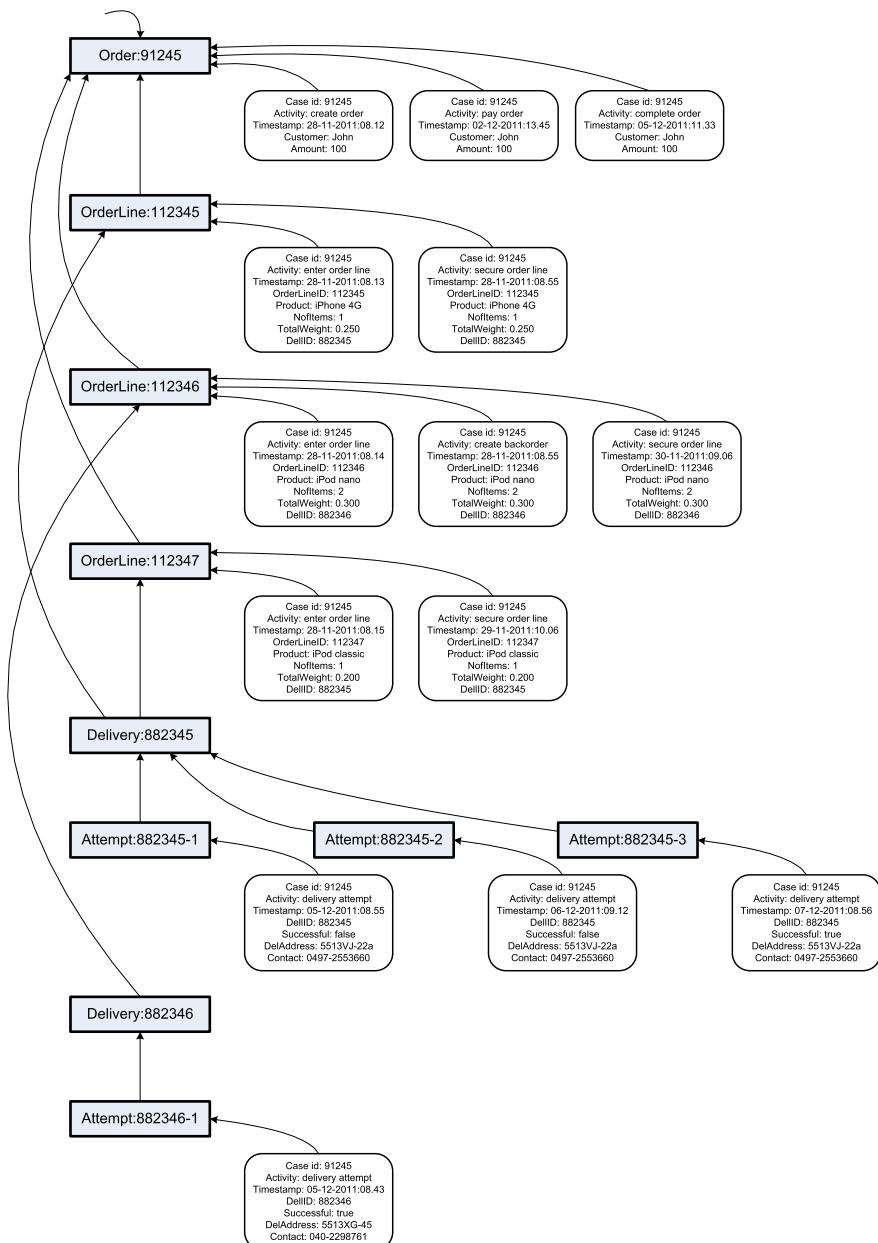


Fig. 5.12 All events that can be related to order 91245. The 14 rounded rectangles correspond to events associated to case 91245. The squared rectangles represent records in one of the four tables

Table 5.9 Events extracted from all four tables using order records from the *Order* table as a starting point

Attempt			
Case id	Activity	Timestamp	Other attributes
91245	create order	28-11-2011:08.12	Customer: John, Amount: 100
91245	enter order line	28-11-2011:08.13	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
91245	enter order line	28-11-2011:08.14	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	enter order line	28-11-2011:08.15	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DellID: 882345
91245	secure order line	28-11-2011:08.55	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
91245	create backorder	28-11-2011:08.55	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	secure order line	29-11-2011:10.06	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DellID: 882345
91245	secure order line	30-11-2011:09.06	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	pay order	02-12-2011:13.45	Customer: John, Amount: 100
91245	delivery attempt	05-12-2011:08.43	DellID: 882346, Successful: true, DelAddress: 5513XG-45, Contact: 040-2298761
91245	delivery attempt	05-12-2011:08.55	DellID: 882345, Successful: false, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91245	complete order	05-12-2011:11.33	Customer: John, Amount: 100
91245	delivery attempt	06-12-2011:09.12	DellID: 882345, Successful: false, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91245	delivery attempt	07-12-2011:08.56	DellID: 882345, Successful: true, DelAddress: 5513VJ-22a, Contact: 0497-2553660
91561	create order	28-11-2011:12.22	Customer: Mike, Amount: 530
91561	enter order line	28-11-2011:12.23	OrderLineID: 112448, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
...
...

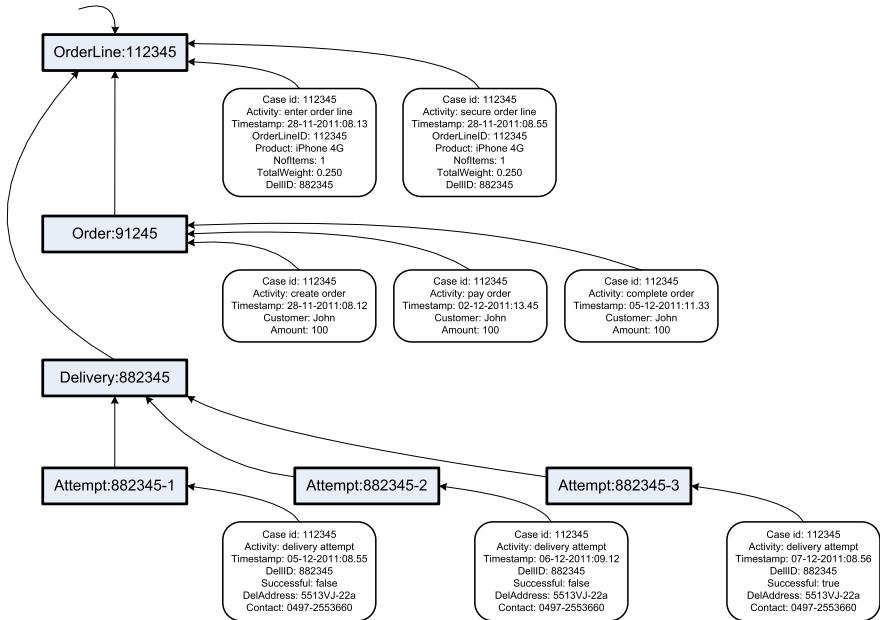


Fig. 5.13 All events that can be related to order line 112345

structure that logs should have. However, the case attributes *Customer* and *John* do not need to be repeated for each event; one could have used case attributes rather than event attributes.

Table 5.9 flattens the original database consisting of four tables. The flattened event log is like a view on the complete data set. Alternative views are possible. For example, Fig. 5.13 shows another way to flatten the original database. Now cases correspond to order lines rather than orders. Hence, the root node is order line 112345. This is an order line of order 91245 and three attempts were needed to deliver the iPhone 4G. The timestamps in the *Order* table have been used to create events associated to order line cases rather than orders. Based on the view sketched in Fig. 5.13, one can generate another event log.

In Fig. 5.12, the root node is an order. In Fig. 5.13, the root node is an order line. Similarly, it is possible to take a delivery or delivery attempts as root node. Moreover, various selections of events can be used, e.g., the three order-related events or the three delivery-related events in Fig. 5.13 could have been left out from the selection. This shows that many views on the original data set are possible, i.e., there are many ways to flatten reality as recorded into a single event log.

Flattening a data set into an event log can be compared to aggregating multidimensional data in Online Analytical Processing (OLAP) (see Sect. 12.4). For example, using a typical OLAP tool, sales data can be viewed by product categories, by region, and/or by quarter. Depending on the type of question, a different view on the data can be chosen. One important difference is that in process mining we analyze

processes rather than a simple OLAP cube. Therefore, we need to correlate events and order them, thus making the extraction process more complex.

Proplets: Seeing in 3-D

Process mining shows that the assumptions made by classical process modeling languages such as BPMN, UML ADs, Statecharts, BPEL, YAWL, WF-nets, and EPCs are somewhat artificial. They only provide *one monolithic view* on the real process of interest. The process is flattened to allow for a diagram that describes the life-cycle of one case in isolation. The application of process mining to real-life processes shows that squeezing a process into such a single monolithic flat model is problematic. Like in physics, where experiments help to (in)validate models, process discovery also helps to reveal the limitations of oversimplified models. The empirical nature of process mining helps managers, consultants, and process analysts to better understand the “fabric of real business processes” and, thus, also see the limitations of conventional process modeling languages.

Proplets [153] are one of the few business process modeling languages allowing for *3-D process models*. Rather than describing the whole process in terms of one monolithic 2-D process model, the process is modeled as a collection of interacting Proplets. For example, when modeling orders and deliveries, the class diagram of Fig. 5.11 can be used as a starting point. Based on this class diagram four classes of Proplets are identified: orders, order lines, deliveries, and delivery attempts. Each Proplet class is modeled separately. The Proplets interact and are related by following the real anatomy of the process.

See [153] for more examples illustrating that classical notations force the modeler to *straightjacket* processes into one monolithic model. Unfortunately, hierarchy concepts in conventional languages do not support one-to-many or many-to-many relationships. In Fig. 5.11, orders and deliveries are in a many-to-many relationship: one order may result in multiple deliveries and one delivery many involve order lines of different orders. This cannot be handled by the refinement of activities; order and delivery Proplets need to coexist independent of one another.

Object-oriented modeling and artifact-centric modeling use ideas related to Proplets. However, mainstream process modeling notations and BPM systems still use conventional 2-D notations. The ACSI project [1] aims to promote the use of Proplets and develop new process mining techniques for non-monolithic processes.

Although it is important to view business processes in 3-D, we often need to resort to 2-D models for a variety of reasons. Here we mention three of them. First of all, the data sources provided may only allow for a 2-D view, e.g., only one table is provided as input. Second, users expect process models in terms of classical

2-D process modeling languages such as BPMN, UML ADs, Statecharts, BPEL, YAWL, WF-nets, and EPCs. Last but not least, most process mining techniques require flattening the data. Therefore, we advocate the following approach.

- Create a *process-oriented data warehouse* containing information about relevant events. The data warehouse should avoid storing aggregated data, and gather the raw business events instead. In traditional data warehouses, events are aggregated into quantitative data, thus hampering process analysis.
- Depending on the questions, define an appropriate *view*. Based on the chosen view, *flatten* the required data to produce an event log (e.g., in XES format). This corresponds to taking a *2-D slice* from the 3-D data.
- Use the 2-D slice to apply a variety of *process mining* techniques. If needed, *filter* the event log further (e.g., removing infrequent activities). Continue extracting, filtering, and mining until the questions are answered.

Depending on the questions, it may be the case that multiple 2-D slices need to be taken to create a 3-D view on the overall process. This view is consistent with Fig. 5.1; by extracting the event data the scope of the process is determined.

Chapter 6

Process Discovery: An Introduction

Process discovery is one of the most challenging process mining tasks. Based on an event log a process model is constructed thus capturing the behavior seen in the log. This chapter introduces the topic using the rather naïve α -algorithm. This algorithm nicely illustrates some of the general ideas used by many process mining algorithms and helps to understand the notion of process discovery. Moreover, the α -algorithm serves as a stepping stone for discussing challenges related to process discovery.

6.1 Problem Statement

As discussed in Chap. 2, there are three types of process mining: discovery, conformance, and enhancement. Moreover, we identified various perspectives, e.g., the control-flow perspective, the organizational or resource perspective, the data perspective, and the time perspective. In this chapter, we focus on the *discovery* task and the *control-flow* perspective. This combination is often referred to as *process discovery*. The general process discovery problem can be formulated as follows.

Definition 6.1 (General process discovery problem) Let L be an event log as defined in Definition 5.3 or as specified by the XES standard (cf. Sect. 5.3). A *process discovery algorithm* is a function that maps L onto a process model such that the model is “representative” for the behavior seen in the event log. The challenge is to find such an algorithm.

This definition does not specify what kind of process model should be generated, e.g., a BPMN, EPC, YAWL, or Petri net model. Moreover, event logs with potentially many attributes may be used as input. Recall that the XES format allows for storing information related to all perspectives whereas here the focus is on the control-flow perspective. The only requirement is that the behavior is “representative”, but it is unclear what this means.

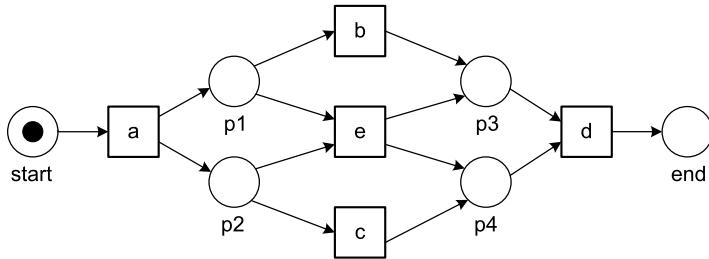


Fig. 6.1 WF-net N_1 discovered for $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$

Definition 6.1 is rather broad and vague. The target format is not specified and a potentially “rich” event log is used as input without specifying tangible requirements. To make things more concrete, we define the target to be a Petri net model. Moreover, we use a *simple event log* as input (cf. Definition 5.4). A simple event log L is a multi-set of traces over some set of activities \mathcal{A} , i.e., $L \in \mathbb{B}(\mathcal{A}^*)$. For example,

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

L_1 is a simple log describing the history of six cases. The goal is now to discover a Petri net that can “replay” event log L_1 . Ideally, the Petri net is a sound WF-net as defined in Sect. 3.2.3. Based on these choices we reformulate the process discovery problem and make it more concrete.

Definition 6.2 (Specific process discovery problem) A *process discovery algorithm* is a function γ that maps a log $L \in \mathbb{B}(\mathcal{A}^*)$ onto a marked Petri net $\gamma(L) = (N, M)$. Ideally, N is a *sound WF-net* and all traces in L correspond to possible firing sequences of (N, M) .

Function γ defines a so-called “Play-In” technique as described in Chap. 2. Based on L_1 , a process discovery algorithm γ could discover the WF-net shown in Fig. 6.1, i.e., $\gamma(L_1) = (N_1, [start])$. Each trace in L_1 corresponds to a possible firing sequence of WF-net N_1 shown in Fig. 6.1. Therefore, it is easy to see that the WF-net can indeed replay all traces in the event log. In fact, each of the three possible firing sequences of WF-net N_1 appears in L_1 .

Let us now consider another event log,

$$\begin{aligned} L_2 = & [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \\ & \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, e, f, c, b, d \rangle] \end{aligned}$$

L_2 is a simple event log consisting of 13 cases represented by 6 different traces. Based on event log L_2 , some γ could discover WF-net N_2 shown in Fig. 6.2. This WF-net can indeed replay all traces in the log. However, not all firing sequences of N_2 correspond to traces in L_2 . For example, the firing sequence $\langle a, c, b, e, f, c, b, d \rangle$

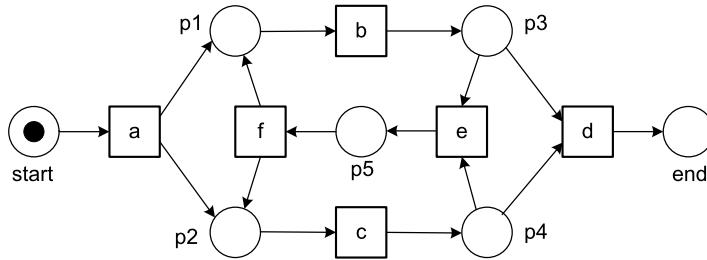


Fig. 6.2 WF-net N_2 discovered for $L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, e, f, c, b, d \rangle]$

does not appear in L_2 . In fact, there are infinitely many firing sequences because of the loop construct in N_2 . Clearly, these cannot all appear in the event log. Therefore, Definition 6.2 does not require all firing sequences of (N, M) to be traces in L .

In this chapter, we focus on the discovery of Petri nets. The reason is that Petri nets are simple and graphical while still allowing for the modeling of concurrency, choices, and iteration. This is illustrated by Figs. 6.1 and 6.2. In both models activities b and c are concurrent. In N_1 , there is choice following a . In N_2 , there is choice between d and e each time both b and c complete. Both N_1 and N_2 are sound WF-nets. As explained in Chap. 3, WF-nets are a natural subclass of Petri nets tailored toward the modeling and analysis of operational processes. A process model describes the life-cycle of one case. Therefore, WF-nets explicitly model the creation and the completion of the cases. The creation is modeled by putting a token in the unique source place i (place $start$ in Figs. 6.1 and 6.2). The completion is modeled by reaching the state marking the unique sink place o (place end in Figs. 6.1 and 6.2). Given a unique source place i and a unique sink place o , the soundness requirement described in Definition 3.7 follows naturally. Recall that a WF-net N is *sound* if and only if

- $(N, [i])$ is *safe*, i.e., places cannot hold multiple tokens at the same time;
- For any marking $M \in [N, [i]]$, $o \in M$ implies $M = [o]$, i.e., if the sink place is marked, all other places should be empty (*proper completion*);
- For any marking $M \in [N, [i]]$, $[o] \in [N, M]$, i.e., it is always possible to mark the sink place (*option to complete*); and
- $(N, [i])$ contains *no dead transitions*, i.e., all parts of the model are potentially reachable.

Most process modeling notations use or assume correctness criteria similar to soundness. For instance, deadlocks and livelocks are symptoms of a process that cannot complete (properly). These phenomena are undesired, independent of the notation used.

Although we use WF-nets in this chapter, this does not imply that discovered process models cannot be presented using other notations. As discussed in Chap. 3, there exist many translations from Petri nets into other notations and vice versa.

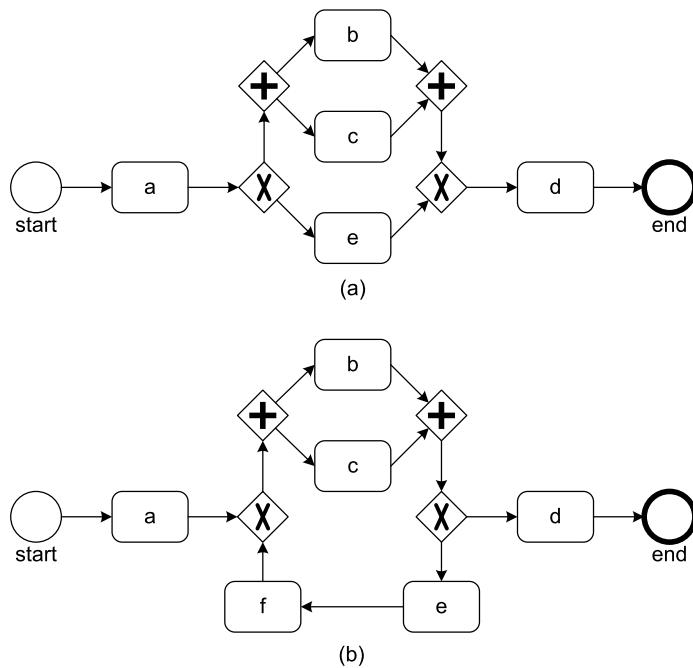


Fig. 6.3 Two BPMN models: (a) the model corresponding to WF-net N_1 discovered for L_1 , and (b) the model corresponding to WF-net N_2 discovered for L_2

Compact formalisms with formal semantics like Petri nets are most suitable to develop and explain process mining algorithms. The representation used to show results to end users is less relevant for the actual process discovery task. For example, the WF-nets depicted in Figs. 6.1 and 6.2 can also be presented in terms of the two trace equivalent BPMN models shown in Fig. 6.3. Similarly, the discovered models could have been translated into equivalent EPCs, UML activity diagrams, statecharts, YAWL models, BPEL specifications, etc.

In the general problem formulation (Definition 6.1) we stated that the discovered model should be “representative” for the behavior seen in the event log. In Definition 6.2, this was operationalized by requiring that the model is able to replay all behavior in this log, i.e., any trace in the event log is a possible firing sequence of the WF-net. This is the so-called “fitness” requirement. In general, there is a trade-off between the following four quality criteria:

- (*Fitness*) The discovered model should allow for the behavior seen in the event log.
- (*Precision*) The discovered model should not allow for behavior completely unrelated to what was seen in the event log.
- (*Generalization*) The discovered model should generalize the example behavior seen in the event log.
- (*Simplicity*) The discovered model should be as simple as possible.

A model having a good fitness is able to replay most of the traces in the log. Precision is related to the notion of *underfitting* presented in the context of data mining (see Sect. 4.6.3). A model having a poor precision is underfitting, i.e., it allows for behavior that is very different from what was seen in the event log. Generalization is related to the notion of *overfitting*. An overfitting model does not generalize enough, i.e., it is too specific and too much driven by examples in the event log. The fourth quality criterion is related to Occam’s Razor which states that “one should not increase, beyond what is necessary, the number of entities required to explain anything” (see Sect. 4.6.3). Following this principle, we look for the “simplest process model” that can explain what is observed in the event log.

It turns out to be challenging to balance the four quality criteria. For instance, an oversimplified model is likely to have a low fitness or lack of precision. Moreover, there is an obvious trade-off between underfitting and overfitting. We discuss these four quality criteria later in this chapter. However, we first introduce a concrete process discovery algorithm.

6.2 A Simple Algorithm for Process Discovery

This section introduces the α -algorithm [157]. This algorithm is an example of a γ function as mentioned in Definition 6.2, i.e., given a simple event log it produces a Petri net that (hopefully) can replay the log. The α -algorithm was one of the first process discovery algorithms that could adequately deal with concurrency (see Sect. 7.6). However, the α -algorithm should not be seen as a very practical mining technique as it has problems with noise, infrequent/incomplete behavior, and complex routing constructs. Nevertheless, it provides a good introduction into the topic. The α -algorithm is simple and many of its ideas have been embedded in more complex and robust techniques. We will use the algorithm as a baseline for discussing the challenges related to process discovery and for introducing more practical algorithms.

6.2.1 Basic Idea

Input for the α -algorithm is a simple event log L over \mathcal{A} , i.e., $L \in \mathbb{B}(\mathcal{A}^*)$. In the remainder, we will simply refer to L as the event log. We refer to the elements of \mathcal{A} as *activities*, see Sect. 3.2. These activities will correspond to transitions in the discovered Petri net. In this chapter, we will use the convention that capital letters refer to sets of activities (e.g., $A, B \subseteq \mathcal{A}$), whereas for individual activities no capitalization is used (e.g., $a, b, c, \dots \in \mathcal{A}$). The output of the α -algorithm is a marked Petri net, i.e., $\alpha(L) = (N, M)$. We aim at the discovery of WF-nets. Therefore, we can omit the initial marking and write $\alpha(L) = N$ (the initial marking is implied; $M = [i]$).

Table 6.1 Footprint of L_1 :
 $a\#_{L_1}a$, $a \rightarrow_{L_1} b$, $a \rightarrow_{L_1} c$,
etc.

	a	b	c	d	e
a	$\#_{L_1}$	\rightarrow_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}
b	\leftarrow_{L_1}	$\#_{L_1}$	\parallel_{L_1}	\rightarrow_{L_1}	$\#_{L_1}$
c	\leftarrow_{L_1}	\parallel_{L_1}	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$
d	$\#_{L_1}$	\leftarrow_{L_1}	\leftarrow_{L_1}	$\#_{L_1}$	\leftarrow_{L_1}
e	\leftarrow_{L_1}	$\#_{L_1}$	$\#_{L_1}$	\rightarrow_{L_1}	$\#_{L_1}$

The α -algorithm scans the event log for particular patterns. For example, if activity a is followed by b but b is never followed by a , then it is assumed that there is a causal dependency between a and b . To reflect this dependency, the corresponding Petri net should have a place connecting a to b . We distinguish four *log-based ordering relations* that aim to capture relevant patterns in the log.

Definition 6.3 (Log-based ordering relations) Let L be an event log over \mathcal{A} , i.e., $L \in \mathbb{B}(\mathcal{A}^*)$. Let $a, b \in \mathcal{A}$.

- $a >_L b$ if and only if there is a trace $\sigma = \langle t_1, t_2, t_3, \dots, t_n \rangle$ and $i \in \{1, \dots, n - 1\}$ such that $\sigma \in L$ and $t_i = a$ and $t_{i+1} = b$;
- $a \rightarrow_L b$ if and only if $a >_L b$ and $b \not>_L a$;
- $a\#_L b$ if and only if $a \not>_L b$ and $b \not>_L a$; and
- $a \parallel_L b$ if and only if $a >_L b$ and $b >_L a$.

Consider, for instance, $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$ again. For this event log, the following log-based ordering relations can be found:

$$>_{L_1} = \{(a, b), (a, c), (a, e), (b, c), (c, b), (b, d), (c, d), (e, d)\}$$

$$\rightarrow_{L_1} = \{(a, b), (a, c), (a, e), (b, d), (c, d), (e, d)\}$$

$$\#_{L_1} = \{(a, a), (a, d), (b, b), (b, e), (c, c), (c, e), (d, a), (d, d), (e, b), (e, c), (e, e)\}$$

$$\parallel_{L_1} = \{(b, c), (c, b)\}$$

Relation $>_{L_1}$ contains all pairs of activities in a “directly follows” relation. $c >_{L_1} d$ because d directly follows c in trace $\langle a, b, c, d \rangle$. However, $d \not>_{L_1} c$ because c never directly follows d in any trace in the log. \rightarrow_{L_1} contains all pairs of activities in a “causality” relation, e.g., $c \rightarrow_{L_1} d$ because sometimes d directly follows c and never the other way around ($c >_{L_1} d$ and $d \not>_{L_1} c$). $b \parallel_{L_1} c$ because $b >_{L_1} c$ and $c >_{L_1} b$, i.e., sometimes c follows b and sometimes the other way around. $b\#_{L_1} e$ because $b \not>_{L_1} e$ and $e \not>_{L_1} b$.

For any log L over \mathcal{A} and $x, y \in \mathcal{A}$, $x \rightarrow_L y$, $y \rightarrow_L x$, $x\#_L y$, or $x \parallel_L y$, i.e., precisely one of these relations holds for any pair of activities. Therefore, the *footprint* of a log can be captured in a matrix as shown in Table 6.1.

The footprint of event log L_2 is shown in Table 6.2. The subscripts have been removed to not clutter the table. When comparing the footprints of L_1 and L_2 one can see that only the e and f columns and rows differ.

Table 6.2 Footprint of
$$L_2 = [\langle a, b, c, d \rangle^3,$$

$$\langle a, c, b, d \rangle^4,$$

$$\langle a, b, c, e, f, b, c, d \rangle^2,$$

$$\langle a, b, c, e, f, c, b, d \rangle,$$

$$\langle a, c, b, e, f, b, c, d \rangle^2,$$

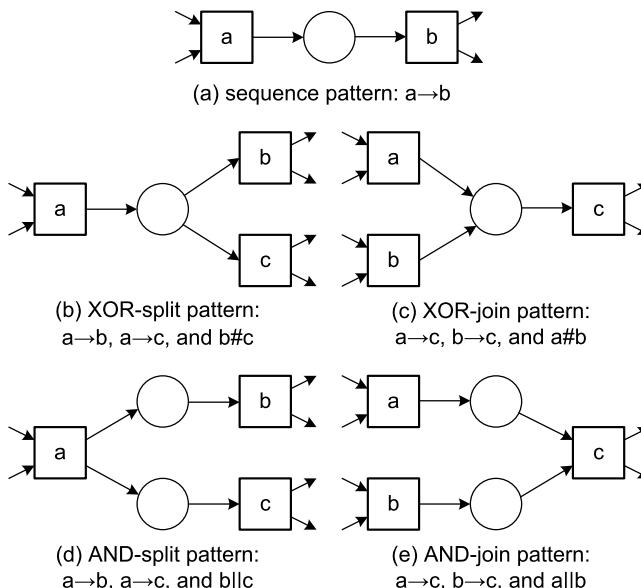
$$\langle a, c, b, e, f, b, c, e, f, c,$$

$$b, d \rangle]$$

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
<i>a</i>	#	→	→	#	#	#
<i>b</i>	←	#		→	→	←
<i>c</i>	←		#	→	→	←
<i>d</i>	#	←	←	#	#	#
<i>e</i>	#	←	←	#	#	→
<i>f</i>	#	→	→	#	←	#

The log-based ordering relations can be used to *discover patterns* in the corresponding process model as is illustrated in Fig. 6.4. If *a* and *b* are in sequence, the log will show $a \rightarrow_L b$. If after *a* there is a choice between *b* and *c*, the log will show $a \rightarrow_L b$, $a \rightarrow_L c$, and $b \#_L c$ because *a* can be followed by *b* and *c*, but *b* will not be followed by *c* and vice versa. The logical counterpart of this so-called XOR-split pattern is the XOR-join pattern as shown in Fig. 6.4(b)–(c). If $a \rightarrow_L c$, $b \rightarrow_L c$, and $a \#_L b$, then this suggests that after the occurrence of either *a* or *b*, *c* should happen. Figure 6.4(d)–(e) shows the so-called AND-split and AND-join patterns. If $a \rightarrow_L b$, $a \rightarrow_L c$, and $b \parallel_L c$, then it appears that after *a* both *b* and *c* can be executed in parallel (AND-split pattern). If $a \rightarrow_L c$, $b \rightarrow_L c$, and $a \parallel_L b$, then the log suggests that *c* needs to synchronize *a* and *b* (AND-join pattern).

Figure 6.4 only shows simple patterns and does not present the additional conditions needed to extract the patterns. However, the figure nicely illustrates the basic idea.

**Fig. 6.4** Typical process patterns and the footprints they leave in the event log

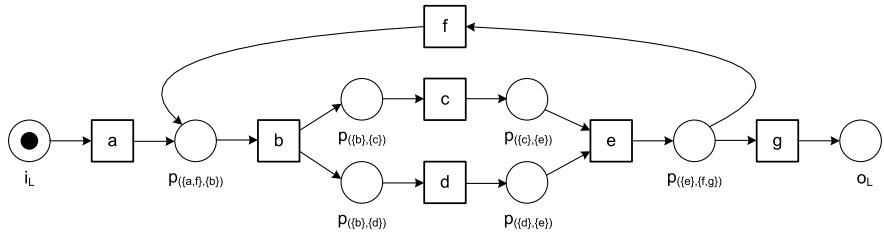


Fig. 6.5 WF-net \$N_3\$ derived from \$L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle, \langle a, b, d, c, e, g \rangle^2, \langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]\$

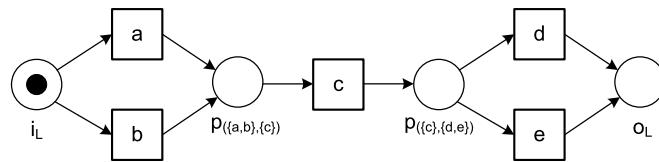


Fig. 6.6 WF-net \$N_4\$ derived from \$L_4 = [\langle a, c, d \rangle^{45}, \langle b, c, d \rangle^{42}, \langle a, c, e \rangle^{38}, \langle b, c, e \rangle^{22}]\$

Table 6.3 Footprint of \$L_3\$

	\$a\$	\$b\$	\$c\$	\$d\$	\$e\$	\$f\$	\$g\$
\$a\$	#	\$\rightarrow\$	#	#	#	#	#
\$b\$	\$\leftarrow\$	#	\$\rightarrow\$	\$\rightarrow\$	#	\$\leftarrow\$	#
\$c\$	#	\$\leftarrow\$	#	\$\parallel\$	\$\rightarrow\$	#	#
\$d\$	#	\$\leftarrow\$	\$\parallel\$	#	\$\rightarrow\$	#	#
\$e\$	#	#	\$\leftarrow\$	\$\leftarrow\$	#	\$\rightarrow\$	\$\rightarrow\$
\$f\$	#	\$\rightarrow\$	#	#	\$\leftarrow\$	#	#
\$g\$	#	#	#	#	\$\leftarrow\$	#	#

Consider, for example, WF-net \$N_3\$ depicted in Fig. 6.5 and the event log \$L_3\$ describing four cases,

$$\begin{aligned}
 L_3 = & [\langle a, b, c, d, e, f, b, d, c, e, g \rangle, \\
 & \langle a, b, d, c, e, g \rangle^2, \\
 & \langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]
 \end{aligned}$$

The \$\alpha\$-algorithm constructs WF-net \$N_3\$ based on \$L_3\$ (see Fig. 6.5).

Table 6.3 shows the footprint of \$L_3\$. Note that the patterns in the model indeed match the log-based ordering relations extracted from the event log. Consider, for example, the process fragment involving \$b\$, \$c\$, \$d\$, and \$e\$. Obviously, this fragment can be constructed based on \$b \rightarrow_{L_3} c\$, \$b \rightarrow_{L_3} d\$, \$c \parallel_{L_3} d\$, \$c \rightarrow_{L_3} e\$, and \$d \rightarrow_{L_3} e\$. The choice following \$e\$ is revealed by \$e \rightarrow_{L_3} f\$, \$e \rightarrow_{L_3} g\$, and \$f \#_{L_3} g\$; etc.

Another example is shown in Fig. 6.6. WF-net N_4 can be derived from L_4 ,

$$L_4 = [\langle a, c, d \rangle^{45}, \langle b, c, d \rangle^{42}, \langle a, c, e \rangle^{38}, \langle b, c, e \rangle^{22}]$$

L_4 contains information about 147 cases that follow one of the four possible traces. There are two start and two end activities. These can be detected easily by looking for the first and last activities in traces.

6.2.2 Algorithm

After showing the basic idea and some examples, we describe the α -algorithm [157].

Definition 6.4 (α -algorithm) Let L be an event log over $T \subseteq \mathcal{A}$. $\alpha(L)$ is defined as follows:

1. $T_L = \{t \in T \mid \exists_{\sigma \in L} t \in \sigma\}$,
2. $T_I = \{t \in T \mid \exists_{\sigma \in L} t = \text{first}(\sigma)\}$,
3. $T_O = \{t \in T \mid \exists_{\sigma \in L} t = \text{last}(\sigma)\}$,
4. $X_L = \{(A, B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \wedge \forall_{a \in A} \forall_{b \in B} a \rightarrow_L b \wedge \forall_{a_1, a_2 \in A} a_1 \#_L a_2 \wedge \forall_{b_1, b_2 \in B} b_1 \#_L b_2\}$,
5. $Y_L = \{(A, B) \in X_L \mid \forall_{(A', B') \in X_L} A \subseteq A' \wedge B \subseteq B' \implies (A, B) = (A', B')\}$,
6. $P_L = \{p_{(A, B)} \mid (A, B) \in Y_L\} \cup \{i_L, o_L\}$,
7. $F_L = \{(a, p_{(A, B)}) \mid (A, B) \in Y_L \wedge a \in A\} \cup \{(p_{(A, B)}, b) \mid (A, B) \in Y_L \wedge b \in B\} \cup \{(i_L, t) \mid t \in T_I\} \cup \{(t, o_L) \mid t \in T_O\}$, and
8. $\alpha(L) = (P_L, T_L, F_L)$.

L is an event log over some set T of activities. In Step 1, it is checked which activities do appear in the log (T_L). These will correspond to the transitions of the generated WF-net. T_I is the set of start activities, i.e., all activities that appear first in some trace (Step 2). T_O is the set of end activities, i.e., all activities that appear last in some trace (Step 3). Steps 4 and 5 form the core of the α -algorithm. The challenge is to determine the places of the WF-net and their connections. We aim at constructing places named $p_{(A, B)}$ such that A is the set of input transitions ($\bullet p_{(A, B)} = A$) and B is the set of output transitions ($p_{(A, B)} \bullet = B$) of $p_{(A, B)}$.

The basic motivation for finding $p_{(A, B)}$ is illustrated by Fig. 6.7. All elements of A should have causal dependencies with all elements of B , i.e., for all $(a, b) \in A \times B: a \rightarrow_L b$. Moreover, the elements of A should never follow one another, i.e., for all $a_1, a_2 \in A: a_1 \#_L a_2$. A similar requirement holds for B .

Table 6.4 shows the structure in terms of the footprint matrix introduced earlier. If we *only* consider the columns and rows related to $A \cup B$ and group the rows and columns belonging to A respectively B , we get the pattern shown in Table 6.4. There are four quadrants. Two quadrants only contain the symbol $\#$. This shows that the elements of A should never follow another (upper-left quadrant) and that the elements of B should never follow another (lower-right quadrant). The upper-right

Fig. 6.7 Place $p_{(A,B)}$ connects the transitions in set A to the transitions in set B

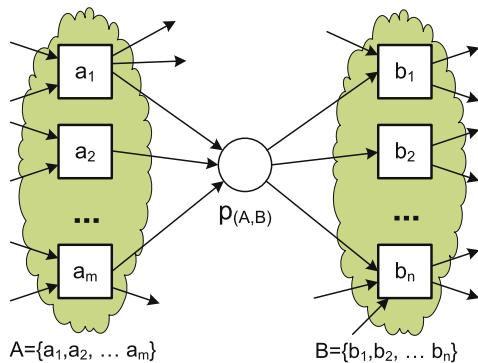


Table 6.4 How to identify $(A, B) \in X_L$? Rearrange the rows and columns corresponding to $A = \{a_1, a_2, \dots, a_m\}$ and $B = \{b_1, b_2, \dots, b_n\}$ and remove the other rows and columns from the footprint

	a_1	a_2	\dots	a_m	b_1	b_2	\dots	b_n
a_1	#	#	\dots	#	\rightarrow	\rightarrow	\dots	\rightarrow
a_2	#	#	\dots	#	\rightarrow	\rightarrow	\dots	\rightarrow
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
a_m	#	#	\dots	#	\rightarrow	\rightarrow	\dots	\rightarrow
b_1	\leftarrow	\leftarrow	\dots	\leftarrow	#	#	\dots	#
b_2	\leftarrow	\leftarrow	\dots	\leftarrow	#	#	\dots	#
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
b_n	\leftarrow	\leftarrow	\dots	\leftarrow	#	#	\dots	#

quadrant only contains the symbol \rightarrow , any of the elements in A can be followed by any of the elements in B but never the other way around. By symmetry, the lower-left quadrant only contains the symbol \leftarrow .

Let us consider L_1 again. Clearly, $A = \{a\}$ and $B = \{b, e\}$ meet the requirements stated in Step 4. Also $A' = \{a\}$ and $B' = \{b\}$ meet the same requirements. X_L is the set of all such pairs that meet the requirements just mentioned. In this case,

$$X_{L_1} = \{(\{a\}, \{b\}), (\{a\}, \{c\}), (\{a\}, \{e\}), (\{a\}, \{b, e\}), (\{a\}, \{c, e\}), \\ (\{b\}, \{d\}), (\{c\}, \{d\}), (\{e\}, \{d\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$$

If one would insert a place for any element in X_{L_1} , there would be too many places. Therefore, only the “maximal pairs” (A, B) should be included. Note that for any pair $(A, B) \in X_L$, non-empty set $A' \subseteq A$, and non-empty set $B' \subseteq B$, it is implied that $(A', B') \in X_L$. In Step 5, all non-maximal pairs are removed, thus yielding

$$Y_{L_1} = \{(\{a\}, \{b, e\}), (\{a\}, \{c, e\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$$

Step 5 can also be understood in terms the footprint matrix. Consider Table 6.4 and let A' and B' be such that $\emptyset \subset A' \subseteq A$ and $\emptyset \subset B' \subseteq B$. Removing rows and columns $A \cup B \setminus (A' \cup B')$ results in a matrix still having the pattern shown in Table 6.4. Therefore, we only consider maximal matrices for constructing Y_L .

Table 6.5 Footprint of L_5

	a	b	c	d	e	f
a	#	\rightarrow	#	#	\rightarrow	#
b	\leftarrow	#	\rightarrow	\leftarrow	\parallel	\rightarrow
c	#	\leftarrow	#	\rightarrow	\parallel	#
d	#	\rightarrow	\leftarrow	#	\parallel	#
e	\leftarrow	\parallel	\parallel	\parallel	#	\rightarrow
f	#	\leftarrow	#	#	\leftarrow	#

Every element of $(A, B) \in Y_L$ corresponds to a place $p_{(A, B)}$ connecting transitions A to transitions B . In addition P_L also contains a unique source place i_L and a unique sink place o_L (cf. Step 6). Remember that the goal is to create a WF-net.¹

In Step 7, the arcs of the WF-net are generated. All start transitions in T_I have i_L as an input place and all end transitions T_O have o_L as output place. All places $p_{(A, B)}$ have A as input nodes and B as output nodes. The result is a Petri net $\alpha(L) = (P_L, T_L, F_L)$ that describes the behavior seen in event log L .

Thus far we presented four logs and four WF-nets. Application of the α -algorithm shows that indeed $\alpha(L_3) = N_3$ and $\alpha(L_4) = N_4$. In Figs. 6.5 and 6.6, the places are named based on the sets Y_{L_3} and Y_{L_4} . Moreover, $\alpha(L_1) = N_1$ and $\alpha(L_2) = N_2$ modulo renaming of places (because different place names are used in Figs. 6.1 and 6.2). These examples show that the α -algorithm is indeed able to discover WF-nets based on event logs.

Let us now consider event log L_5 ,

$$L_5 = [\langle a, b, e, f \rangle^2, \langle a, b, e, c, d, b, f \rangle^3, \langle a, b, c, e, d, b, f \rangle^2, \\ \langle a, b, c, d, e, b, f \rangle^4, \langle a, e, b, c, d, b, f \rangle^3]$$

Table 6.5 shows the footprint of the log.

Let us now apply the 8 steps of the algorithm for $L = L_5$:

$$T_L = \{a, b, c, d, e, f\}$$

$$T_I = \{a\}$$

$$T_O = \{f\}$$

$$X_L = \{(\{a\}, \{b\}), (\{a\}, \{e\}), (\{b\}, \{c\}), (\{b\}, \{f\}), (\{c\}, \{d\}), \\ (\{d\}, \{b\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$$

$$Y_L = \{(\{a\}, \{e\}), (\{c\}, \{d\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$$

$$P_L = \{p_{(\{a\}, \{e\})}, p_{(\{c\}, \{d\})}, p_{(\{e\}, \{f\})}, p_{(\{a, d\}, \{b\})}, p_{(\{b\}, \{c, f\})}, i_L, o_L\}$$

$$F_L = \{(a, p_{(\{a\}, \{e\})}), (p_{(\{a\}, \{e\})}, e), (c, p_{(\{c\}, \{d\})}), (p_{(\{c\}, \{d\})}, d),$$

¹ Nevertheless, the α -algorithm may construct a Petri net that is not a WF-net (see, for instance, Fig. 6.12). Later, we will discuss such problems in detail.

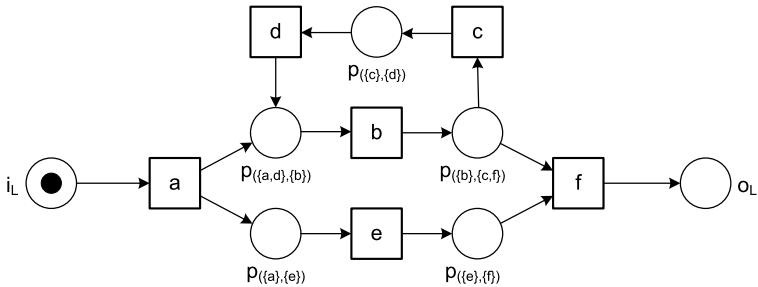


Fig. 6.8 WF-net N_5 derived from $L_5 = [\langle a, b, e, f \rangle^2, \langle a, b, e, c, d, b, f \rangle^3, \langle a, b, c, e, d, b, f \rangle^2, \langle a, b, c, d, e, b, f \rangle^4, \langle a, e, b, c, d, b, f \rangle^3]$

$$\begin{aligned} & (e, p_{\{e\}, \{f\}}), (p_{\{e\}, \{f\}}), f, (a, p_{\{a, d\}, \{b\}}), (d, p_{\{a, d\}, \{b\}}), \\ & (p_{\{a, d\}, \{b\}}), b, (b, p_{\{b\}, \{c, f\}}), (p_{\{b\}, \{c, f\}}), c, (p_{\{b\}, \{c, f\}}), f, \\ & (i_L, a), (f, o_L) \} \end{aligned}$$

$$\alpha(L) = (P_L, T_L, F_L)$$

Figure 6.8 shows $N_5 = \alpha(L_5)$, i.e., the model just computed. N_5 can indeed replay the traces in L_5 . Place names are not shown in Fig. 6.8, and we will also not show them in later WF-nets, because they can be derived from the surrounding transition names and just clutter the diagram.

6.2.3 Limitations of the α -Algorithm

In [157], it was shown that the α -algorithm can discover a large class of WF-nets if one assumes that the log is *complete* with respect to the log-based ordering relation $>_L$. This assumption implies that, for any complete event log L , $a >_L b$ if a can be directly followed by b . Consequently, a footprint like the one shown in Table 6.5 is assumed to be valid. We revisit the notion of completeness later in this chapter.

Even if we assume that the log is complete, the α -algorithm has some problems. There are many different WF-nets that have the same possible behavior, i.e., two models can be structurally different but trace equivalent. Consider, for instance, the following event log:

$$L_6 = [\langle a, c, e, g \rangle^2, \langle a, e, c, g \rangle^3, \langle b, d, f, g \rangle^2, \langle b, f, d, g \rangle^4]$$

$\alpha(L_6)$ is shown in Fig. 6.9. Although the model is able to generate the observed behavior, the resulting WF-net is needlessly complex. Two of the input places of g are redundant, i.e., they can be removed without changing the behavior. The places denoted as p_1 and p_2 are so-called *implicit places* and can be removed without

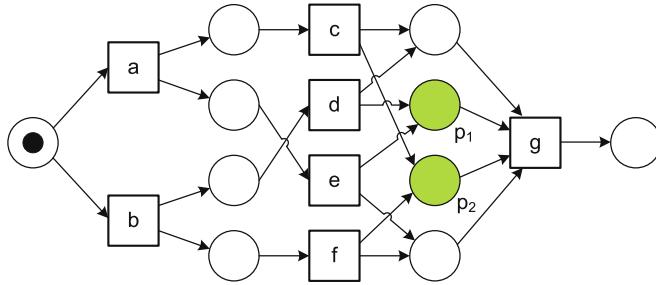


Fig. 6.9 WF-net N_6 derived from $L_6 = [\langle a, c, e, g \rangle^2, \langle a, e, c, g \rangle^3, \langle b, d, f, g \rangle^2, \langle b, f, d, g \rangle^4]$. The two highlighted places are redundant, i.e., removing them will simplify the model without changing its behavior

Fig. 6.10 Incorrect WF-net

N_7 derived from

$$L_7 = [\langle a, c \rangle^2, \langle a, b, c \rangle^3, \\ \langle a, b, b, c \rangle^2, \\ \langle a, b, b, b, b, c \rangle^1]$$

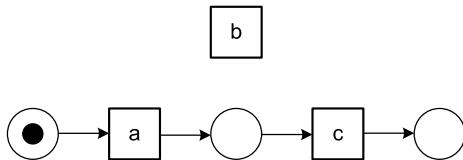
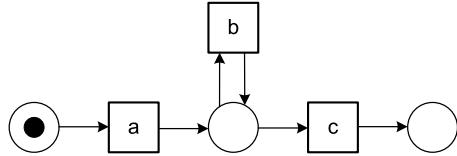


Fig. 6.11 WF-net N'_7 having a so-called “short-loop” of length one



affecting the set of possible firing sequences. In fact, Fig. 6.9 shows only one of many possible trace equivalent WF-nets.

The original α -algorithm (as presented in Sect. 6.2.2) has problems dealing with short loops, i.e., loops of length one or two. For a loop of length one, this is illustrated by WF-net N_7 in Fig. 6.10, which shows the result of applying the basic algorithm to L_7 ,

$$L_7 = [\langle a, c \rangle^2, \langle a, b, c \rangle^3, \langle a, b, b, c \rangle^2, \langle a, b, b, b, b, c \rangle^1]$$

The resulting model is not a WF-net as transition b is disconnected from the rest of the model. The models allows for the execution of b before a and after c . This is not consistent with the event log. This problem can be addressed easily as shown in [11]. Using an improved version of the α -algorithm one can discover the WF-net shown in Fig. 6.11.

The problem with loops of length two is illustrated by Petri net N_8 in Fig. 6.12 which shows the result of applying the basic algorithm to L_8 ,

$$L_8 = [\langle a, b, d \rangle^3, \langle a, b, c, b, d \rangle^2, \langle a, b, c, b, c, b, d \rangle]$$

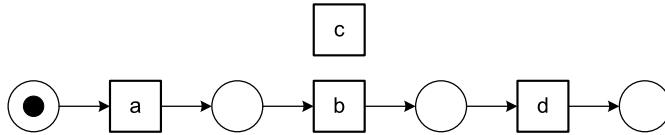


Fig. 6.12 Incorrect WF-net N_8 derived from $L_8 = [\langle a, b, d \rangle^3, \langle a, b, c, b, d \rangle^2, \langle a, b, c, b, c, b, d \rangle]$

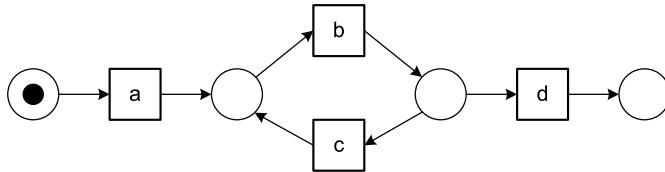


Fig. 6.13 Corrected WF-net N'_8 having a so-called “short-loop” of length two

The following log-based ordering relations are derived from this event log: $a \rightarrow_{L_8} b$, $b \rightarrow_{L_8} d$, and $b \parallel_{L_8} c$. Hence the basic algorithm incorrectly assumes that b and c are in parallel because they follow one another. The model shown in Fig. 6.12 is not even a WF-net because c is not on a path from source to sink. Using the extension described in [11], the improved α -algorithm correctly discovers the WF-net shown in Fig. 6.13.

There are various ways to improve the basic α -algorithm to be able to deal with loops. The α^+ -algorithm described in [11] is one of several alternatives to address problems related to the original algorithm presented in Sect. 6.2.2. The α^+ -algorithm uses a pre- and post-processing phase. The pre-processing phase deals with loops of length two whereas the pre-processing phase inserts loops of length one.

The basic algorithm has no problems mining loops of length three or more. For a loop of involving at least three activities (say a , b , and c), concurrency can be distinguished from loops using relation $>_L$. For a loop we find only $a >_L b$, $b >_L c$, and $c >_L a$. If the three activities are concurrent, we find $a >_L b$, $a >_L c$, $b >_L a$, $b >_L c$, $c >_L a$, and $c >_L b$. Hence, it is easy to detect the difference. Note that for a loop of length two this is not the case. For a loop involving a and b , we find $a >_L b$ and $b >_L a$. If a and b are concurrent, we find the same relations. Hence, both constructs leave the same footprint in the event log.

A more difficult problem is the discovery of so-called *non-local dependencies* resulting from non-free choice process constructs. An example is shown in Fig. 6.14. This net would be a good candidate after observing the following event log:

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

However, the α -algorithm will derive the WF-net without the places labeled p_1 and p_2 . Hence, $\alpha(L_9) = N_4$, as shown in Fig. 6.6, although the traces $\langle a, c, e \rangle$ and $\langle b, c, d \rangle$ do not appear in L_9 . Such problems can be (partially) resolved using refined versions of the α -algorithm such as the one presented in [185].

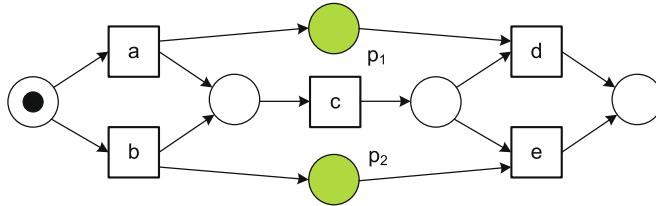
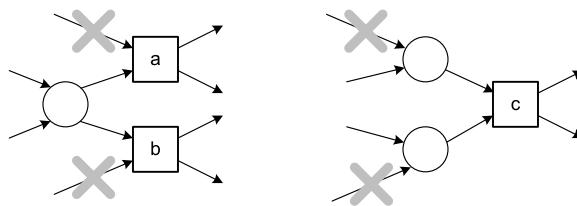


Fig. 6.14 WF-net N_9 having a non-local dependency

Fig. 6.15 Two constructs that may jeopardize the correctness of the discovered WF-net



Another limitation of the α -algorithm is that *frequencies are not taken into account*. Therefore, the algorithm is very sensitive to noise and incompleteness (see Sect. 6.4.2).

The α -algorithm is able to discover a large class of models. The basic 8-line algorithm has some limitations when it comes to particular process patterns (e.g., short-loops and non-local dependencies). Some of these problems can be solved using various refinements. As shown in [11, 157], the α -algorithm guarantees to produce a correct process model provided that the underlying process can be described by a WF-net that does not contain duplicate activities (two transitions with the same activity label) and silent transitions (activities that are not recorded in the event log), and does not use the two constructs shown in Fig. 6.15. See [11, 157] for the precise requirements.

Even if the underlying process is using constructs as shown in Fig. 6.15, the α -algorithm may still produce a useful process model. For instance, the α -algorithm is unable to discover the highlighted places (p_1 and p_2) in Fig. 6.14, but still produces a sound process model that is able to replay the log.

6.2.4 Taking the Transactional Life-Cycle into Account

When describing the typical information in event logs in Chap. 5, we discussed the *transactional life-cycle model* of an *activity instance*. Figure 5.3 shows examples of transaction types, e.g., schedule, start, complete, and suspend. Events often have such a transaction type attribute, e.g., $\#_{trans}(e) = \text{complete}$. The standard life-cycle extension of XES also provides such an attribute. The α -algorithm can be easily adapted to take this information into account. First of all, the log could be projected

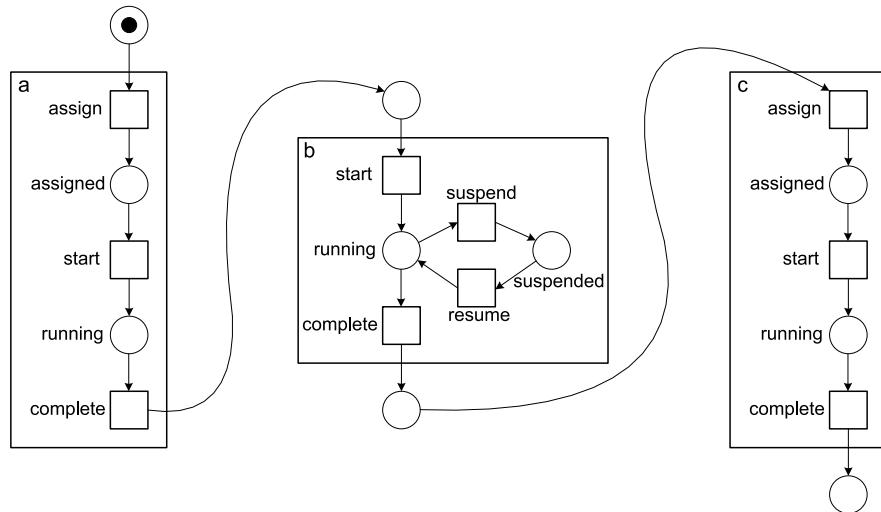


Fig. 6.16 Mining event logs with transactional information; the life-cycle of each activity is represented as a subprocess

onto smaller event logs in which each of the smaller logs contains all events related to a specific activity. This information can be used to discover the transactional life-cycle for each activity. Second, when mining the overall process, information about the general transactional life-cycle (e.g., Fig. 5.3) or information about an activity-specific transactional life-cycle can be exploited. Figure 6.16 illustrates the latter. All events related to an activity are mapped onto transitions embedded in a subprocess. The relations between the transitions for each subprocess are either discovered separately or modeled using domain knowledge. Figure 6.16 shows a sequence of three activities. Activities *a* and *c* share a common transactional life-cycle involving the event types *assign*, *start*, and *complete*. Activity *b* has a transactional life-cycle involving the event types *start*, *suspend*, *resume*, and *complete*.

6.3 Rediscovering Process Models

In Chap. 8, we will describe *conformance checking* techniques for measuring the quality of a process model *with respect to an event log*. However, when discussing the results of the α -algorithm, we already concluded that some WF-nets “could not be discovered” based on an event log. This assumes that we aim to discover a particular, known, model. In reality, we often do not know the “real” model. In fact, in practice, there is no such thing as *the* model describing a process. There may be many models (i.e., views on the same reality) and the process being studied may change while being discovered. However, as sketched in Fig. 6.17, we can create the experimental setting for testing process discovery algorithms in which we assume the original model to be known.

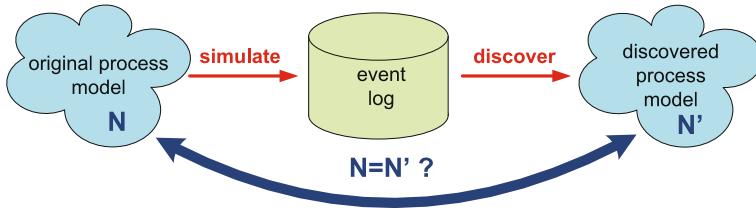


Fig. 6.17 The rediscovery problem: Is the discovered model N' equivalent to the original model N ?

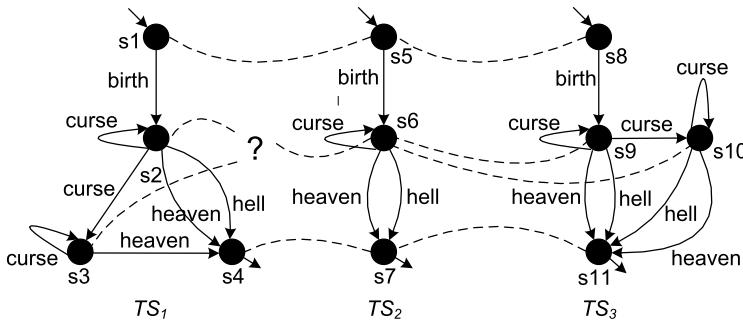


Fig. 6.18 Three trace equivalent transition systems: TS_1 and TS_2 are not bisimilar, but TS_2 and TS_3 are bisimilar

Starting point in Fig. 6.17 is a process model, e.g., a WF-net N . Based on this model we can run many simulation experiments and record the simulated events in an event log. Let us assume that the event log is complete with respect to some criterion, e.g., if x can be followed by y in N it happened at least once according to log. Using the complete event log as input for a process discovery algorithm (e.g., the α -algorithm), we can construct a new model. Now the question is: “What do the discovered model N' and the original model N have in common? Are they equivalent?” Equivalence can be viewed at different levels. For example, it is unreasonable to expect that a discovery algorithm is able to reconstruct the original layout as this information is not in the log; layout information is irrelevant for the behavior of a process. For the same reason, it is unreasonable to expect that the original place names of the WF-net can be reconstructed. The α -algorithm generates places named $p_{(A,B)}$. These are of course not intended to match original place names. Therefore, we need to focus on *behavior* (and not on layout and syntax) when comparing the discovered model N' and the original model N .

Three notions of behavioral equivalence

As shown in [176], many equivalence notions can be defined. Here, we informally describe three well-known notions: *trace equivalence*, *bisimilarity*,

and *branching bisimilarity*. These notions are defined for a pair of transition systems TS_1 and TS_2 (Sect. 3.2.1) and not for higher-level languages such as WF-nets, BPMN, EPCs, and YAWL. However, any model with executable semantics can be transformed into a transition system. Therefore, we can assume that the original process model N and the discovered process model N' mentioned in Fig. 6.17 define two transition systems that can be used as a basis for comparison.

Trace equivalence considers two transition systems to be equivalent if their sets of execution sequences are identical. Let TS_2 be the transition system corresponding to WF-net $N_6 = \alpha(L_6)$ shown in Fig. 6.9 and let TS_1 be the transition system corresponding to the same WF-net but now without places p_1 and p_2 . Although both WF-nets are syntactically different, the sets of execution sequences of TS_1 and TS_2 are the same. However, two transition systems that allow for the same set of execution sequences may also be quite different as illustrated by Fig. 6.18.

The three transition systems in Fig. 6.18 are trace equivalent: any trace in one transition system is also possible in any of the other transition systems. For instance, the trace $\langle birth, curse, curse, curse, heaven \rangle$ is possible in all three transition systems. However, there is a relevant difference between TS_1 and TS_2 . In TS_1 one can end up in state $s3$ where one will always go to heaven despite the cursing. Such a state does not exist in TS_2 ; while cursing in state $s6$ one can still go to hell. When moving from state $s2$ to state $s3$ in TS_1 a choice was made which cannot be seen in the set of traces but that is highly relevant for understanding the process.

Bisimulation equivalence, or bisimilarity for short, is a more refined notion taking into account the moment of choice. Two transition systems are bisimilar if the first system can “mimic any move” of the second, and vice versa (using the same relation). Consider, for example, TS_2 and TS_3 in Fig. 6.18. TS_2 can simulate TS_3 and vice versa. The states of both transition systems are related by dashed lines; $s5$ is related to $s8$, $s6$ is related to both $s9$ and $s10$, and $s7$ is related to $s11$. In two related states the same set of actions needs to be possible and taking any of these actions on one side should lead to a related state when taking the same action on the other side. Because TS_2 can move from $s5$ to $s6$ via action *birth*, TS_3 should also be able to take a *birth* action in $s8$ resulting in a related state ($s9$). TS_2 and TS_3 are bisimilar because any action by one can be mimicked by the other. Now consider TS_1 and TS_2 . Here, it is impossible to relate $s3$ in TS_1 to a corresponding state in TS_2 . If $s3$ is related to $s6$, then in $s3$ it should be possible to do a *hell* action, but this is not the case. Hence, TS_2 can simulate TS_1 , i.e., any action in TS_1 can be mimicked by TS_2 , but TS_1 cannot simulate TS_2 . Therefore, TS_1 and TS_2 are not bisimilar. Bisimulation equivalence is a stronger equivalence relation than trace equivalence, i.e., if two transition systems are bisimilar, then they are also trace equivalent.

Branching bisimulation equivalence, or branching bisimilarity for short, takes *silent actions* into account. In Chap. 3 we introduced already the label τ for this purpose. A τ action is “invisible”, i.e., cannot be observed. In terms of process mining this means that the corresponding activity is not recorded in the event log. As before, two transition systems are branching bisimilar if the first system can “follow any move” of the second and vice versa, but now taking τ actions into account. (Here, we do not address subtle differences between weak bisimulation, also known as observational equivalence, and branching bisimulation equivalence [176].) If one system takes a τ action, then the second system may also take a τ action or do nothing (as long as the states between both systems remain related). If one system takes a non- τ action, then the second system should also be able to take the same non- τ action possibly preceded by sequence of τ actions. The states before and after the non- τ action, need to be related. Figure 6.19 shows two YAWL models and their corresponding transition systems TS_1 and TS_2 . The two transition systems are *not* branching bisimilar. The reason is that in the YAWL model on the left, a choice is made after task *check*, whereas in the other model the choice is postponed until either *reject* or *accept* happens. Therefore, the YAWL model on the left cannot simulate the model on the right. Technically, states s_3 and s_4 in TS_1 do not have a corresponding state in TS_2 . It is impossible to relate s_3 and s_4 to s_7 since s_7 allows for both actions whereas s_3 and s_4 allow for only one action. The YAWL model on the right models the so-called *deferred choice* workflow pattern whereas the YAWL model on the left models the more common *exclusive choice* pattern [155].

Branching bisimulation equivalence is highly relevant for process mining since typically not all actions are recorded in the event log. For example, if the choice made in task *check* is not recorded in the event log, then one discovers the YAWL model on the right, i.e., the right moment of choice cannot be captured.

Although both models in Fig. 6.19 are not branching bisimilar they are trace equivalent. In both models there are only two possible (visible) traces: $\langle \text{check}, \text{reject} \rangle$ and $\langle \text{check}, \text{accept} \rangle$.

We refer to [176] for formal definitions of the preceding concepts. Here we discuss these concepts because they are quite important when judging process mining results.

The different notions of equivalence show that the comparison of the original model and the discovered model in Fig. 6.17 is not a simple syntactical check. Instead a choice must be made with respect to the type of behavioral equivalence that is appropriate.

As mentioned before, the experimental setting shown in Fig. 6.17 can only be used in the situation in which the model is known beforehand. In most applica-

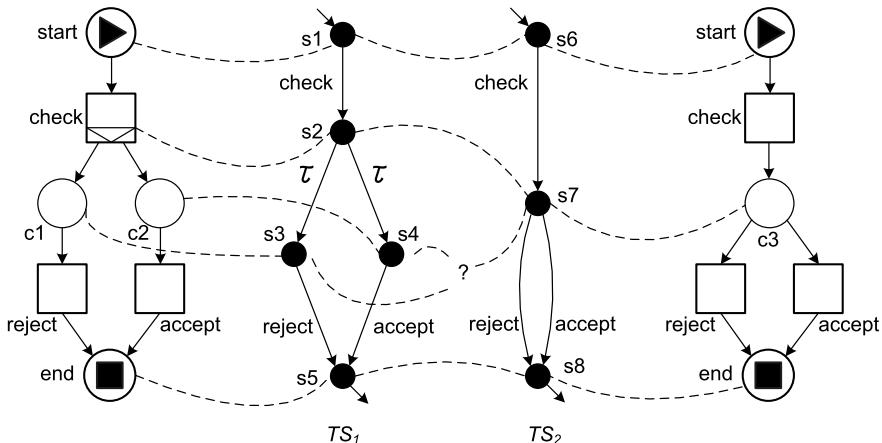


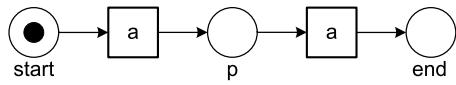
Fig. 6.19 Two YAWL models and the corresponding transition systems

tions such a model is not known. Moreover, classical notions such as trace equivalence, bisimilarity, and branching bisimilarity provide only true/false answers. As discussed in [14], a binary equivalence is not very useful in the context of process mining. If two processes are very similar (identical except for some exceptional paths), classical equivalence checks will simply conclude that the processes are not equivalent rather than stating that the processes are, e.g., 95% similar. Therefore, this book will focus on the comparison of a model and an event log rather than comparing two models. For instance, in Chap. 8 we will show techniques that can conclude that 95% of the event log “fits” the model.

6.4 Challenges

The α -algorithm was one of the first process discovery algorithms to adequately capture concurrency (see also Sect. 7.6). Today there are much better algorithms that overcome the weaknesses of the α -algorithm. These are either variants of the α -algorithm or algorithms that use a completely different approach, e.g., genetic mining or synthesis based on regions. In Chap. 7, we review some of these alternative approaches. However, before presenting new process discovery techniques, we first elaborate on the main challenges. For this purpose we show the effect that a representational bias can have (Sect. 6.4.1). Then we discuss problems related to the input event log that may be noisy or incomplete (Sect. 6.4.2). In Sect. 6.4.3, we discuss the four quality criteria mentioned earlier: fitness, precision, generalization, and simplicity. Finally, Sect. 6.4.4 again emphasizes that discovered models are just a view on reality. Hence, the usefulness of the model strongly depends on the questions one seeks to answer.

Fig. 6.20 A WF-net having two transitions with the same label describing event log $L_{10} = [\langle a, a \rangle^{55}]$



6.4.1 Representational Bias

At the beginning of the chapter we decided to focus on a mining algorithm that produces a WF-net, i.e., we assumed that the underlying process can be adequately described by a WF-net. Any discovery technique requires such a *representational bias*. For example, algorithms for learning decision trees (see Sect. 4.2) make similar assumptions about the structure of the resulting tree. For instance, most decision tree learners can only split once on an attribute on every path in the tree.

When discussing the α -algorithm we assumed that the process to be discovered is a sound WF-net. More specifically, we assumed that the underlying process can be described by a WF-net where each transition bears a unique and visible label. In such a WF-net it is not possible to have two transitions with the same label (i.e., $l(t_1) = l(t_2)$ implies $t_1 = t_2$) or transitions whose occurrences remain invisible (i.e., it is not possible to have a so-called silent transition, so for all transitions t , $l(t) \neq \tau$). (See Sect. 3.2.2 and the earlier discussion on branching bisimulation equivalence.) These assumptions may seem harmless, but have a noticeable effect on the class of process models that can be discovered. We show two examples illustrating the impact of such a representational bias.

For an event log like $L_{10} = [\langle a, a \rangle^{55}]$, i.e., for all cases precisely two a 's are executed, ideally one would like to discover the WF-net shown in Fig. 6.20. Unfortunately, this process model will not be discovered due to the representational bias of the α -algorithm. There is no WF-net without duplicate and τ labels that has the desired behavior and the α -algorithm can only discover such WF-nets (i.e., each transition needs to have unique visible label).

Let us now consider event log $L_{11} = [\langle a, b, c \rangle^{20}, \langle a, c \rangle^{30}]$. Figure 6.21(a) describes the underlying process well: activity b can be skipped by executing the τ transition. Figure 6.21(b) shows an alternative WF-net using two a transitions and no τ transition. These two models are trace equivalent. (They are not branching bisimilar because the moment of choice is different.) However, it is not possible to construct a WF-net without duplicate and τ labels that is trace equivalent to these two models. Figure 6.21(c) shows the model produced by the α -algorithm; because of the representational bias, the algorithm is destined to fail for this log. The WF-net in Fig. 6.21(c) can only reproduce trace $\langle a, b, c \rangle$ and not $\langle a, c \rangle$.

Event logs L_{10} and L_{11} illustrate the effect a representational bias can have. However, from the viewpoint of the α -algorithm, the choice to not consider duplicate labels and τ transitions is sensible. τ transitions are not recorded in the log and hence any algorithm will have problems reconstructing their behavior. Multiple transitions with the same label are undistinguishable in the event log. Therefore, any algorithm will have problems associating the corresponding events to one of these transitions.

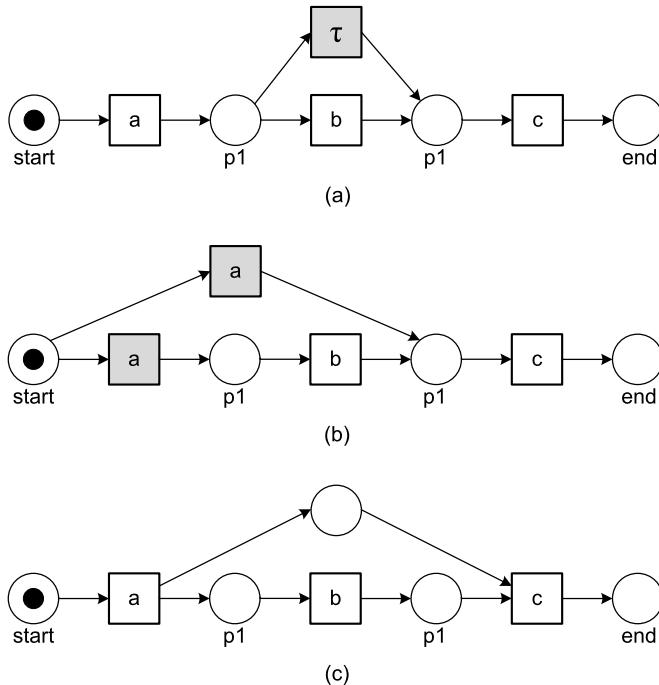


Fig. 6.21 Three WF-nets for the event log $L_{11} = [(\langle a, b, c \rangle^{20}, (\langle a, c \rangle^{30})]$

The problems sketched previously apply to many process discovery algorithms. For example, the choice between the concurrent execution of b and c or the execution of just e shown in Fig. 6.1 cannot be handled by many algorithms. Most algorithms do *not* allow for so-called “non-free-choice constructs” where concurrency and choice meet. The concept of *free-choice nets* is well-defined in the Petri net domain [45]. A Petri net is free choice if any two transitions sharing an input place have identical input sets, i.e., $\bullet t_1 \cap \bullet t_2 \neq \emptyset$ implies $\bullet t_1 = \bullet t_2$ for any $t_1, t_2 \in T$. Most analysis questions (e.g., soundness) can be answered in polynomial time for free-choice nets [136, 168]. Moreover, many process modeling languages are inherently free-choice, thus making this an interesting subclass. Unfortunately, in reality processes tend to be non-free-choice. The example of Fig. 6.1 shows that sometimes the α -algorithm is able to deal with non-free-choice constructs. However, there are many non-free-choice processes that cannot be discovered by the α -algorithm (see for example N_9 in Fig. 6.14). The non-free-choice construct is just one of many constructs that existing process mining algorithms have problems with. Other examples are arbitrary nested loops, cancelation, unbalanced splits and joins, and partial synchronization. In this context it is important to observe *process discovery is, by definition, restricted by the expressive power of the target language*, i.e., the representational bias.

For the reader interested in the topic, we refer to the *workflow patterns* [155, 191] mentioned earlier. These patterns help to discuss and identify the representational bias of a language.

The representational bias helps limiting the search space of possible candidate models. This can make discovery algorithms more efficient. However, it can also be used to give preference to particular types of models. It seems that existing approaches can benefit from selecting a more suitable representational bias. For instance, the α -algorithm may yield models that have deadlocks or livelocks. Here it would be nice to have a representational bias to limit the search space to only sound models (i.e., free of deadlocks and other anomalies). Unfortunately, currently, this can typically only be achieved by severely limiting the expressiveness of the modeling language or by using more time-consuming analysis techniques. Consider, for example, the so-called *block-structured* process models. A model is block-structured if it satisfies a number of syntactical requirements such that soundness is guaranteed by these requirements. Different definitions exist [49, 132, 187]. Most of these definitions require a one-to-one correspondence between splits and joins, e.g., concurrent paths created by an AND-split need to be synchronized by the corresponding AND-join. Since many real-life processes are not block structured (see for example Figs. 14.1 and 14.10), one should be careful to not limit the expressiveness too much. Note that techniques that turn unstructured models into block-structured process models tend to introduce many duplicate or silent activities. Therefore, such transformations do not alleviate the core problems.

6.4.2 Noise and Incompleteness

To discover a suitable process model it is assumed that the event log contains a *representative sample of behavior*. Besides the issues mentioned in Chap. 5 (e.g., correlating events and scoping the log), there are two related phenomena that may make an event log less representative for the process being studied:

- (*Noise*) The event log contains rare and infrequent behavior not representative for the typical behavior of the process.²
- (*Incompleteness*) The event log contains too few events to be able to discover some of the underlying control-flow structures.

6.4.2.1 Noise

Noise, as defined in this book, does not refer to incorrect logging. When extracting event logs from various data sources one needs to try to locate data problems

²Note that the definition of noise may be a bit counter-intuitive. Sometimes the term “noise” is used to refer to incorrectly logged events, i.e., errors that occurred while recording the events. Such a definition is not very meaningful as no event log will explicitly reveal such errors. Hence, we consider “outliers” as noise. Moreover, we assume that such outliers correspond to exceptional behavior rather than logging errors.

as early as possible. However, at some stage one needs to assume that the event log contains information on what really happened. It is impossible for a discovery algorithm to distinguish incorrect logging from exceptional events. This requires human judgment and pre- and postprocessing of the log. Therefore, we use the term “noise” to refer to rare and infrequent behavior (“outliers”) rather than errors related to event logging. For process mining it is important to filter out noise and several process discovery approaches specialize in doing so, e.g., heuristic mining, genetic mining, and fuzzy mining.

Recall the *support* and *confidence* metrics defined in the context of learning association rules. The support of a rule $X \Rightarrow Y$ indicates the applicability of the rule, i.e., the fraction of instances for which both antecedent and consequent hold. The confidence of a rule $X \Rightarrow Y$ indicates the reliability of the rule. If rule $tea \wedge latte \Rightarrow muffin$ has a support of 0.2 and a confidence of 0.9, then 20% of the customers actually order tea, latte and muffins at the same time and 90% of the customers that order tea and latte also order a muffin. For learning association rules we defined a threshold for both confidence and support, i.e., rules with low confidence or support are considered to be noise.

Let us informally apply the idea of confidence and support to the basic α -algorithm. Starting point for the α -algorithm is the $>_L$ relation. Recall that $a >_L b$ if and only if there is a trace in L in which a is directly followed by b . Now we can define the support of $a >_L b$ based on number of times the pattern $\langle \dots, a, b, \dots \rangle$ appears in the log, e.g., the fraction of cases in which the pattern occurs. Subsequently, we can use a threshold for cleaning the $>_L$ relation. The confidence of $a >_L b$ can be defined by comparing the number of times the pattern $\langle \dots, a, b, \dots \rangle$ appears in the log divided by the frequency of a and b . For example, suppose that $a >_L b$ has a reasonable support, e.g., the pattern $\langle \dots, a, b, \dots \rangle$ occurs 1000 times in the log. Moreover, a occurs 1500 times and b occurs 1200 times. Clearly, $a >_L b$ has a good confidence. However, if the pattern $\langle \dots, a, b, \dots \rangle$ occurs 1000 times and a and b are very frequent and occur each more than 100,000 times, then the confidence in $a >_L b$ is much lower. The $>_L$ relation is the basis for the footprint matrices as shown in Tables 6.1, 6.2, 6.3, and 6.5. Hence, by removing “noisy $a >_L b$ rules”, we obtain a more representative footprint, and a better starting point for the α -algorithm. (There are several complications when doing this, however, the basic idea should be clear.) This simplified discussion shows how “noise” can be quantified and addressed when discovering process models. When presenting heuristic mining in Sect. 7.2 we return to this topic.

In the context of noise, we also talk about the *80/20 model*. Often we are interested in the process model that can describe 80% of the behavior seen in the log. This model is typically relatively simple because the remaining 20% of the log account for 80% of the variability in the process.

6.4.2.2 Incompleteness

When it comes to process mining the notion of *completeness* is also very important. It is related to noise. However, whereas noise refers to the problem of having “too

much data” (describing rare behavior), completeness refers to the problem of having “too little data”.

Like in any data mining or machine learning context one cannot assume to have seen all possibilities in the “training material” (i.e., the event log at hand). For WF-net N_1 in Fig. 6.1 and event log $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$, the set of possible traces found in the log is exactly the same as the set of possible traces in the model. In general, this is not the case. For instance, the trace $\langle a, b, e, c, d \rangle$ may be possible but did not (yet) occur in the log. Process models typically allow for an exponential or even infinite number of different traces (in case of loops). Moreover, some traces may have a much lower probability than others. Therefore, it is unrealistic to assume that every possible trace is present in the event log.

The α -algorithm assumes a relatively weak notion of completeness to avoid this problem. Although N_3 has infinitely many possible firing sequences, a small log like $L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle, \langle a, b, d, c, e, g \rangle^2, \langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]$ can be used to construct N_3 . The α -algorithm uses a local completeness notion based on $>_L$, i.e., if there are two activities a and b , and a can be directly followed by b , then this should be observed at least once in the log.

To illustrate the relevance of completeness, consider a process consisting of 10 activities that can be executed in parallel and a corresponding log that contains information about 10,000 cases. The total number of possible interleavings in the model with 10 concurrent activities is $10! = 3,628,800$. Hence, it is impossible that each interleaving is present in the log as there are fewer cases (10,000) than potential traces (3,628,800). Even if there are 3,628,800 cases in the log, it is extremely unlikely that all possible variations are present. To motivate this consider the following analogy. In a group of 365 people it is very unlikely that everyone has a different birthdate. The probability is $365!/365^{365} \approx 1.454955 \times 10^{-157} \approx 0$, i.e., incredibly small. The number of atoms in the universe is often estimated to be approximately 10^{79} [189]. Hence, the probability of picking a particular atom from the entire universe is much higher than covering all 365 days. Similarly, it is unlikely that all possible traces will occur for any process of some complexity because most processes have much more than 365 possible execution paths. In fact, because typically some sequences are less probable than others, the probability of finding all traces is even smaller. Therefore, weaker completeness notions are needed. For the process in which 10 activities can be executed in parallel, local completeness can reduce the required number of observations dramatically. For example, for the α -algorithm only $10 \times (10 - 1) = 90$ rather than 3,628,800 different observations are needed to construct the model.

6.4.2.3 Cross-Validation

The preceding discussion on completeness and noise shows the need for *cross-validation* as discussed in Sect. 4.6.2. The event log can be split into a *training log* and a *test log*. The training log is used to learn a process model whereas the test log is used to evaluate this model based on unseen cases. Chapter 8 will present con-

crete techniques for evaluating the quality of a model with respect to an event log. For example, if many traces of the test log do not correspond to possible firing sequences of the WF-net discovered based on the training log, then one can conclude that the quality of the model is low.

Also *k-fold cross-validation* can be used, i.e., the event log is split into k equal parts, e.g., $k = 10$. Then k tests are done. In each test, one of the subsets serves as a test log whereas the other $k - 1$ subsets serve together as the training log.

One of the problems for cross validation is the lack of negative examples, i.e., the log only provides examples of possible behavior and does not provide explicit examples describing scenarios that are impossible (see discussion in Sect. 4.6.3). This is complicating cross-validation. One possibility is to insert artificially generated negative events [59, 60, 122]. The basic idea is to compare the quality of the discovered model with respect to the test log containing *actual behavior* with the quality of the discovered model with respect to a test log containing *random behavior*. Ideally, the model scores much better on the log containing actual behavior than on the log containing random behavior.

Cross-validation can also be applied at the level of the footprint matrix. Simply split the event log in k parts and construct the footprint matrix for each of the k parts. If the k footprint matrices are very different (even for smaller values of k), then one can be sure that the event log does not meet the completeness requirement imposed by the α -algorithm. Such a validation can be done before constructing the process model. If there are strong indications that $>_L$ is far from complete, more advanced process mining techniques need to be applied and the results need to be interpreted with care (see also Chap. 7).

6.4.3 Four Competing Quality Criteria

Completeness and noise refer to qualities of the event log and do not say much about the quality of the discovered model. Determining the quality of a process mining result is difficult and is characterized by many dimensions. In this book, we refer to four main quality dimensions: *fitness*, *simplicity*, *precision*, and *generalization*. In this section, we review these four dimensions without providing concrete metrics. Some of the dimensions will be discussed in later chapters in more detail. However, after reading this section it should already be clear that they can indeed be quantified.

Figure 6.22 gives a high-level characterization of the four quality dimensions. A model with good *fitness* allows for the behavior seen in the event log. A model has a perfect fitness if all traces in the log can be replayed by the model from beginning to end. There are various ways of defining fitness. It can be defined at the case level, e.g., the fraction of traces in the log that can be fully replayed. It can also be defined at the event level, e.g., the fraction of events in the log that are indeed possible according to the model. When defining fitness many design decisions need to be made. For example: What is the penalty if a step needs to be skipped and what is the penalty if tokens remain in the WF-net after replay? Later, we will give concrete definitions for fitness.

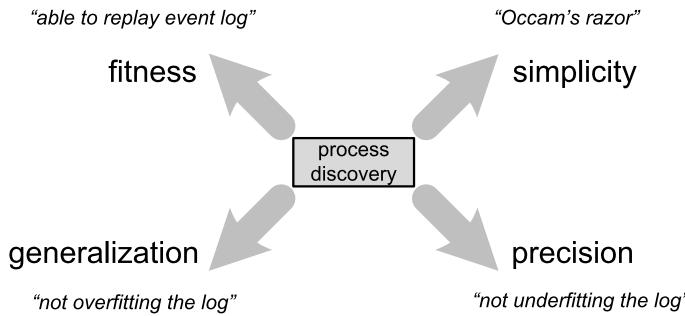
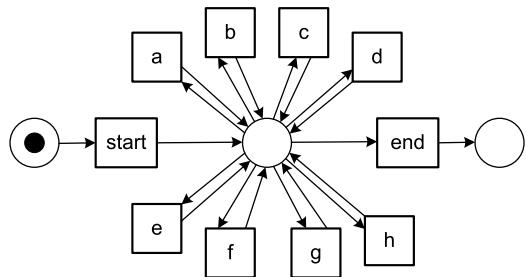


Fig. 6.22 Balancing the four quality dimensions: *fitness*, *simplicity*, *precision*, and *generalization*

Fig. 6.23 The so-called “flower Petri net” allowing for any log containing activities $\{a, b, \dots, h\}$



In Sect. 4.6.1, we defined performance measures like error, accuracy, tp-rate, fp-rate, precision, recall, and F1 score. Recall, also known as the tp-rate, measures the proportion of positive instances indeed classified as positive (tp/p). The traces in the log are positive instances. When such an instance can be replayed by the model, then the instance is indeed classified as positive. Hence, the various notions of fitness can be seen as variants of the recall measure. Most of the notions defined in Sect. 4.6.1 cannot be used because there are *no negative examples*, i.e., fp and tn are unknown (see Fig. 4.14). Since the event log does not contain information about events that could *not* happen at a particular point in time, other notations are needed.

The *simplicity* dimension refers to *Occam’s Razor*. This principle was already discussed in Sect. 4.6.3. In the context of process discovery this means that the simplest model that can explain the behavior seen in the log, is the best model. The complexity of the model could be defined by the number of nodes and arcs in the underlying graph. Also more sophisticated metrics can be used, e.g., metrics that take the “structuredness” or “entropy” of the model into account. See [101] for an empirical evaluation of the *model complexity metrics* defined in literature. In Sect. 4.6.3, we also mentioned that this principle can be operationalized using the *Minimal Description Length* (MDL) principle [63, 190].

Fitness and simplicity alone are not adequate. This is illustrated by the so-called “flower model” shown in Fig. 6.23. The “flower Petri net” allows for any sequence starting with *start* and ending with *end* and containing any ordering of activities in between. Clearly, this model allows for all event logs used to introduce the

α -algorithm. The added *start* and *end* activities in Fig. 6.23 are just a technicality to turn the “flower model” into a WF-net. Surprisingly, all event logs shown thus far (L_1, L_2, \dots, L_{11}) can be replayed by this single model. This shows that the model is not very useful. In fact, the “flower model” does not contain any knowledge other than the activities in the event log. The “flower model” can be constructed based on the occurrences of activities only. The resulting model is simple and has a perfect fitness. Based on the first two quality dimensions this model is acceptable. This shows that the fitness and simplicity criteria are necessary, but not sufficient.

If the “flower model” is on one end of the spectrum, then the “enumerating model” is on the other end of the spectrum. The enumerating model of a log simply lists all the sequences possible, i.e., there is a separate sequential process fragment for each trace in the model. At the start there is one big XOR split selecting one of the sequences and at the end these sequences are joined using one big XOR join. If such a model is represented by a Petri net and all traces are unique, then the number of transitions is equal to the number of events in the log. The “enumerating model” is simply an encoding of the log. Such a model is complex but, like the “flower model”, has a perfect fitness.

Extreme models such as the “flower model” (anything is possible) and the “enumerating model” (only the log is possible) show the need for two additional dimensions. A model is *precise* if it does not allow for “too much” behavior. Clearly, the “flower model” lacks precision. A model that is not precise is “underfitting”. Underfitting is the problem that the model over-generalizes the example behavior in the log, i.e., the model allows for behaviors very different from what was seen in the log.

A model should *generalize* and not restrict behavior to the examples seen in the log (like the “enumerating model”). A model that does not generalize is “overfitting”. Overfitting is the problem that a very specific model is generated whereas it is obvious that the log only holds example behavior, i.e., the model explains the particular sample log, but a next sample log of the same process may produce a completely different process model.

Process mining algorithms need to strike a balance between “overfitting” and “underfitting”. A model is overfitting if it does not generalize and only allows for the exact behavior recorded in the log. This means that the corresponding mining technique assumes a very strong notion of completeness: “If the sequence is not in the event log, it is not possible!”. An underfitting model over-generalizes the things seen in the log, i.e., it allows for more behavior even when there are no indications in the log that suggest this additional behavior (like in Fig. 6.23).

Let us now consider some examples showing that it is difficult to balance between being too general and too specific. Consider, for example, WF-net N_4 shown in Fig. 6.6 and N_9 shown in Fig. 6.14. Both nets can produce the log $L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$, but only N_4 can produce $L_4 = [\langle a, c, d \rangle^{45}, \langle b, c, d \rangle^{42}, \langle a, c, e \rangle^{38}, \langle b, c, e \rangle^{22}]$. Clearly, N_4 is the logical choice for L_4 . Moreover, although both nets can produce L_9 , it is obvious that N_9 is a better model for L_9 as none of the 87 cases follows one of the two additional paths ($\langle b, c, d \rangle$ and $\langle a, c, e \rangle$). However, now consider $L_{12} = [\langle a, c, d \rangle^{99}, \langle b, c, d \rangle^1, \langle a, c, e \rangle^2, \langle b, c, e \rangle^{98}]$. One can argue that

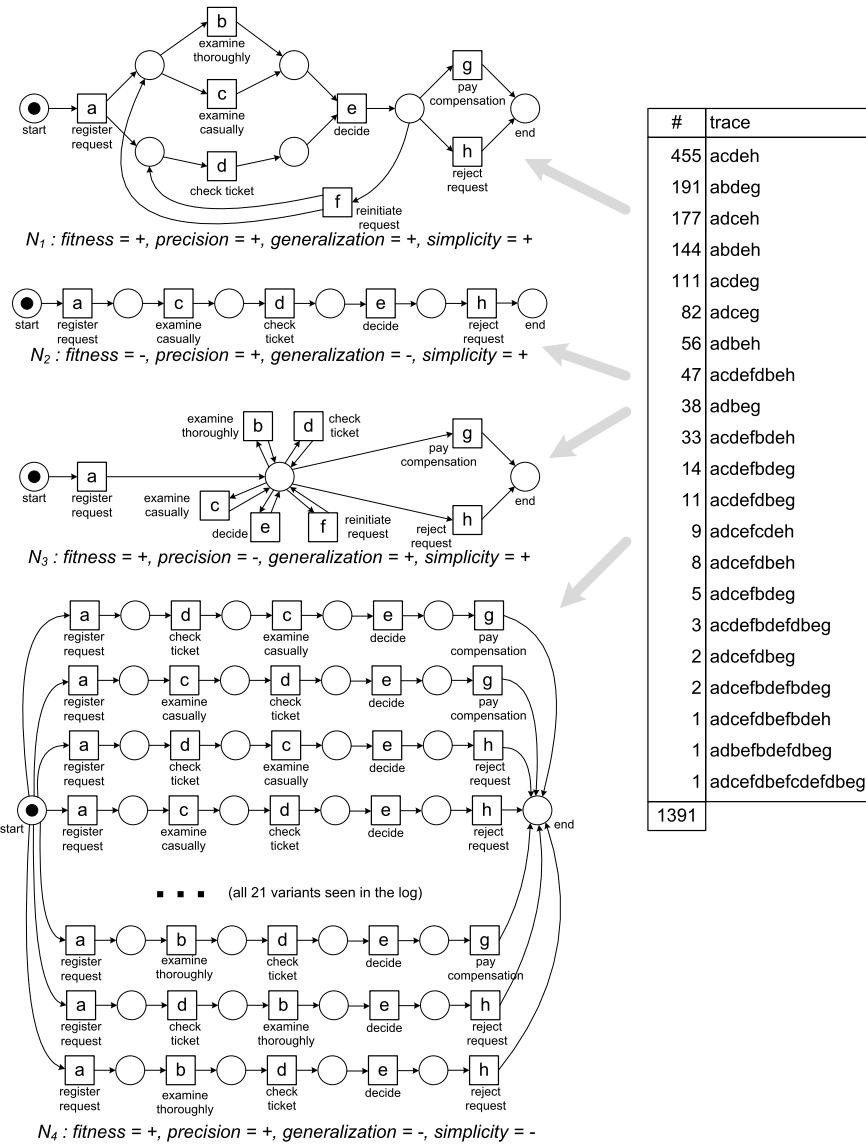


Fig. 6.24 Four alternative models for the same log

N_4 is a better model for L_{12} as all traces can be reproduced. However, 197 out of 200 traces can be explained by the more precise model N_9 . If the three traces are seen as noise, the main behavior is captured by N_9 and not N_4 . Such considerations show that there is a delicate balance between “overfitting” and “underfitting”. Hence, it is difficult, if not impossible, to select “the best” model.

Figure 6.24 illustrates the preceding discussion using the example from Chap. 2. Assume that the four models that are shown are discovered based on the event log also depicted in the figure. There are 1391 cases. Of these 1391 cases, 455 followed the trace $\langle a, c, d, e, h \rangle$. The second most frequent trace is $\langle a, b, d, e, g \rangle$ which was followed by 191 cases.

If we apply the α -algorithm to this event log, we obtain model N_1 shown in Fig. 6.24. A comparison of the WF-net N_1 and the log shows that this model is quite good; it is simple and has a good fitness. Moreover, it balances between overfitting and underfitting.

The other three models in Fig. 6.24 have problems with respect to one or more quality dimensions. WF-net N_2 models only the most frequent trace, i.e., it only allows for the sequence $\langle a, c, d, e, h \rangle$. Hence, none of the other $1391 - 455 = 936$ traces fits. Moreover, the model does not generalize, i.e., N_2 is also overfitting.

WF-net N_3 is a variant of the “flower model”. Only the start and end transitions are captured well. The fitness is good, the model is simple, and not overfitting. However, N_3 lacks precision, i.e., is underfitting, as for example the trace $\langle a, b, b, b, b, b, f, f, f, f, f, g \rangle$ is possible. This behavior seems to be very different from any of the traces in the log.

Figure 6.24 shows only a part of WF-net N_4 . This model simply enumerates the 21 different traces seen in the event log. This model is precise and has a good fitness. However, WF-net N_4 is overly complex and is overfitting.

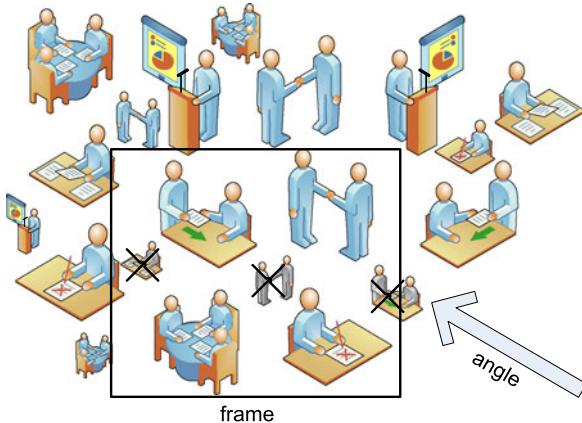
The four models in Fig. 6.24 illustrate the four quality dimensions. Each of these dimensions can be quantified as shown in [121]. In [121], a replay technique is described to quantify fitness resulting in a value between 0 (very poor fitness) to 1 (perfect fitness). A notion called “structural appropriateness” considers the simplicity dimension; the model is analyzed to see whether it is “minimal in structure”. Another notion called “behavioral appropriateness” analyzes the balance between overfitting and underfitting. There are different ways to operationalize the four quality dimensions shown in Fig. 6.22. Depending on the representational bias and goals of the analyst, different metrics can be quantified.

6.4.4 Taking the Right 2-D Slice of a 3-D Reality

The simple examples shown in this chapter already illustrate that process discovery is a non-trivial problem that requires sophisticated analysis techniques. Why is process mining such a difficult problem? There are obvious reasons that also apply to many other data mining and machine learning problems, e.g., dealing with noise and a complex and large search space. However, there are also some specific problems:

- There are *no negative examples* (i.e., a log shows what has happened but does not show what could not happen);
- Due to concurrency, loops, and choices the *search space has a complex structure* and the log typically contains only a *fraction* of all possible behaviors; and

Fig. 6.25 Creating a 2-D slice of a 3-D reality: the process is viewed from a specific angle, the process is scoped using a *frame*, and the *resolution* determines the granularity of the resulting model



- There is *no clear relation between the size of a model and its behavior* (i.e., a smaller model may generate more or less behavior although classical analysis and evaluation methods typically assume some monotonicity property).

The next chapter will show several process discovery techniques that adequately address these problems.

As we will see in Part IV, the discovered process model is *just the starting point* for analysis. By relating events in the log to the discovered model, all kinds of analysis are possible, e.g., checking conformance, finding bottlenecks, optimizing resource allocation, reducing undesired variability, time prediction, and generating recommendations.

One should not seek to discover *the* process model. Process models are just a *view on reality*. Whether a process model is suitable or not, ultimately depends on the questions one would like to answer. Real-life processes are complex and may have many dimensions; models only provide a view on this reality. As discussed in Sect. 5.5, this means that the “3-D reality needs to be flattened into a 2-D process model” in order to apply process mining techniques. For instance, there are many “2-D slices” that one could take of a data set involving customer orders, order lines, deliveries, payments, replenishment orders, etc. Obviously, the different slices result in the discovery of different process models. Using the metaphor of a “process view”, a discovered process model views reality from a particular “angle”, is “framed”, and is shown using a particular “resolution”:

- A discovered model views reality from a particular *angle*. For example, the same process may be analyzed from the viewpoint of a complete order, a delivery, a customer, or an order line.
- A discovered model *frames* reality. The frame determines the boundaries of the process and selects the perspectives of interest (control-flow, information, resources, etc.).
- A discovered model provides a view at a specific *resolution*. The same process can be viewed using a coarser or finer granularity showing less or more details.

Figure 6.25 illustrates the “process view” metaphor. Given a data set it is possible to *zoom in*, i.e., selecting a smaller frame and increasing resolution, resulting in a more fine-grained model of a selected part of the process. It is also possible to *zoom out*, i.e., selecting a larger frame and decreasing resolution, resulting in a more coarse-grained model covering a larger part of the end-to-end process. Both the data set used as input and the questions that need to be answered determine which 2-D slices are most useful.

Chapter 7

Advanced Process Discovery Techniques

The α -algorithm nicely illustrates some of the main ideas behind process discovery. However, this simple algorithm is unable to manage the trade-offs involving the four quality dimensions described in Chap. 6 (fitness, simplicity, precision, and generalization). To successfully apply process mining in practice, one needs to deal with noise and incompleteness. This chapter focuses on more advanced process discovery techniques. The goal is not to present one particular technique in detail, but to provide an overview of the most relevant approaches. This will assist the reader in selecting the appropriate process discovery technique. Moreover, insights into the strengths and weaknesses of the various approaches support the correct interpretation and effective use of the discovered models.

7.1 Overview

Figure 7.1 summarizes the problems mentioned in the context of the α -algorithm. Each back dot represents a trace (i.e., a sequence of activities) corresponding to one or more cases in the event log. (Recall that multiple cases may have the same corresponding trace.) An event log typically contains only a fraction of the possible behavior, i.e., the dots should only be seen as *samples* of a much larger set of possible behaviors. Moreover, one is typically primarily interested in frequent behavior and not in all possible behavior, i.e., one wants to abstract from noise and therefore not all dots need to be relevant for the process model to be constructed.

Recall that we defined noise as infrequent or exceptional behavior. It is interesting to analyze such noisy behaviors, however, when constructing the overall process model, the inclusion of infrequent or exceptional behavior leads to complex diagrams. Moreover, it is typically impossible to make reliable statements about noisy behavior given the small set of observations. Figure 7.1 distinguishes between frequent behavior (solid rectangle with rounded corners) and all behavior (dashed rectangle), i.e., normal and noisy behavior. The difference between normal and noisy behavior is a matter of definition, e.g., normal behavior could be defined as the 80%

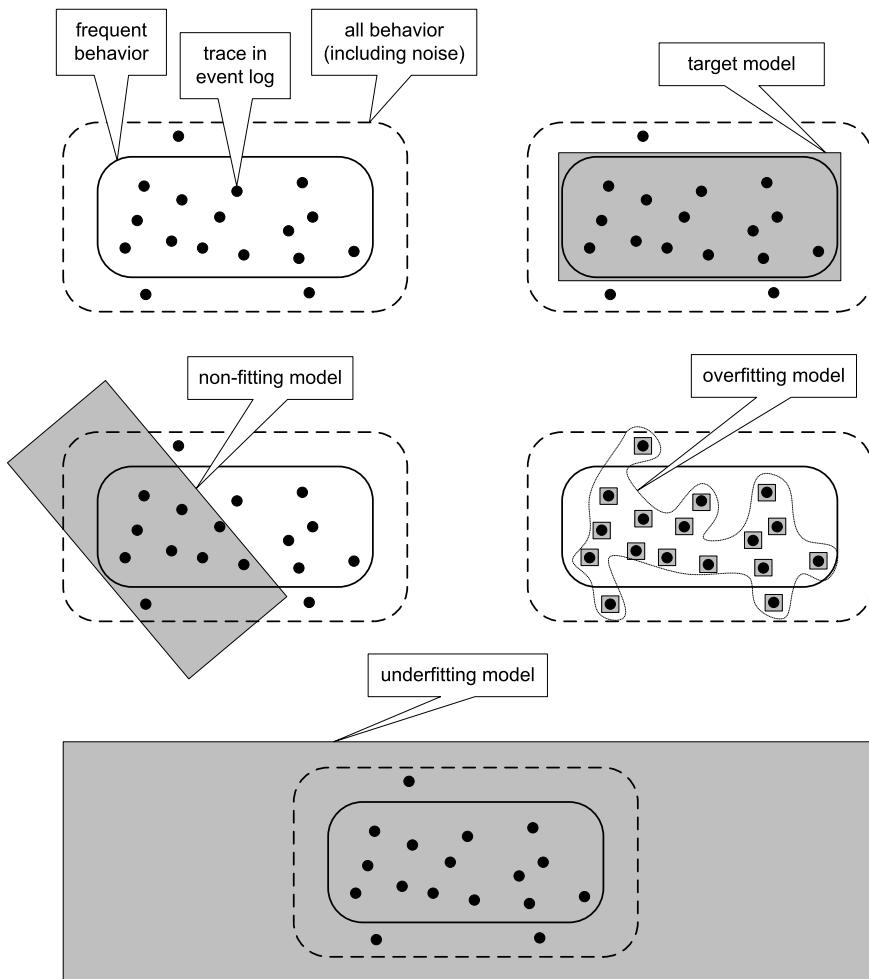


Fig. 7.1 Overview of the challenges that process discovery techniques need to address

most frequently occurring traces. Earlier we mentioned the *80/20 model*, i.e., the process model that is able to describe 80% of the behavior seen in the log. This model is typically relatively simple because the remaining 20% of the log may easily account for 80% of the variability in the process.

Let us assume that the two rectangles with rounded corners can be determined by observing the process infinitely long and that the process does not change (i.e., no concept drift). Based on these assumptions, Fig. 7.1 sketches four discovered models depicted by shaded rectangles. These discovered models are based on the example traces in the log, i.e., the black dots. The “ideal process model” allows for the behavior coinciding with the frequent behavior seen when the process would be observed ad infinitum while being in steady state. The “non-fitting model” in

Fig. 7.1 is unable to characterize the process well as it is not even able to capture the examples in the event log used to learn the model. The “overfitting model” does not generalize and only says something about the examples in the current event log. New examples will most likely not fit into this model. The “underfitting model” lacks precision and allows for behavior that would never be seen if the process would be observed ad infinitum.

Figure 7.1 illustrates the challenges process discovery techniques need to address: How to extract a simple target model that is not underfitting, overfitting, or non-fitting? Clearly, the α -algorithm is unable to do so. Therefore, we present more advanced approaches. However, before doing so, we describe typical characteristics of process discovery algorithms.

7.1.1 Characteristic 1: Representational Bias

The first, and probably most important, characteristic of a process discovery algorithm is its *representational bias*, i.e., the class of process models that can be discovered. For instance, the α -algorithm is only able to discover Petri nets in which each transition has a unique and visible label. Instead of Petri nets, some other representation can be used, e.g., a subclass of BPMN, EPCs, YAWL, hidden Markov models, transition systems, and causal nets. The representational bias determines the search space and potentially limits the expressiveness of the discovered model. Consider, for example, the three process models for event log $L_{11} = [\langle a, b, c \rangle^{20}, \langle a, c \rangle^{30}]$ in Fig. 6.21. If the representational bias allows for duplicate labels (two transitions with the same label) or silent (τ) transitions, a suitable WF-net can be discovered. However, if the representational bias does not allow for this, the discovery algorithm is destined to fail and will not find a suitable WF-net. The *workflow patterns* [155, 191] are a tool to discuss and identify the representational bias of a language. Here, we do not discuss the more than 40 control-flow patterns. Instead, we mention some typical representational limitations imposed by process discovery algorithms:

- *Inability to represent concurrency.* Low-level models, such as Markov models, flow charts, and transition systems, do not allow for the modeling of concurrency other than enumerating all possible interleavings. Recall that such a low-level model will need to show $2^{10} = 1024$ states and $10 \times 2^{10-1} = 5120$ transitions to model a process with 10 parallel activities. Higher level models (like Petri nets and BPMN) only need to depict 10 activities and $2 \times 10 = 20$ “local” states (states before and after each activity).
- *Inability to deal with (arbitrary) loops.* Many process discovery algorithms impose some limitations on loops, e.g., the α -algorithm needs a pre- and post-processing step to deal with shorts loops (see Figs. 6.11 and 6.13). The “Arbitrary Cycles” pattern [155, 191] is typically not supported by algorithms that assume the underlying model to be block-structured.

- *Inability to represent silent actions.* In some notations, it is impossible to model silent actions like the skipping of an activity. Although such events are not explicitly recorded in the event log, they need to be reflected in the model. This limits the expressive power as illustrated by Fig. 6.21.
- *Inability to represent duplicate actions.* In many notations there cannot be two activities having the same label. If the same activity appears in different parts of the process, but these different instances of the same activity cannot be distinguished in the event log, then most algorithms will assume a single activity thus creating causal dependencies (e.g., non-existing loops) that do not exist in the actual process.
- *Inability to model OR-splits/joins.* As shown in Chap. 3, YAWL, BPMN, EPCs, causal nets, etc. allow for the modeling of OR-splits and OR-joins; see for example the models depicted in Figs. 3.6, 3.10 and 3.13 using such constructs. If the representational bias of a discovery algorithm does not allow for OR-splits and OR-joins, then the discovered model may be more complex or the algorithm is unable to find a suitable model.
- *Inability to represent non-free-choice behavior.* Most algorithms do not allow for non-free-choice constructs, i.e., constructs where concurrency and choice meet. Figure 6.1 uses a non-free-choice construct, because places p_1 and p_2 serve both as an XOR-split (to choose between doing just e or both b and c) and as an AND-split (to start the concurrent activities b and c). This WF-net can be discovered by the α -algorithm. However, non-free-choice constructs can also represent non-local dependencies as is illustrated by the WF-net in Fig. 6.14. Such WF-nets cannot be discovered by the basic α -algorithm. Whereas WF-nets can express non-free-choice behavior, many discovery algorithms use a representation that cannot do so.
- *Inability to represent hierarchy.* Most process discovery algorithms work on “flat” models. A notable exception is the Fuzzy Miner [66] that extracts hierarchical models. Activities that have a lower frequency but that are closely related to other low frequent activities are grouped into subprocesses. The representational bias determines whether, in principle, hierarchical models can be discovered or not.

7.1.2 Characteristic 2: Ability to Deal With Noise

Noisy behavior, i.e., exceptional/infrequent behavior, should not be included in the discovered model (see Sect. 6.4.2). First of all, users typically want to see the mainstream behavior. Second, it is impossible to infer meaningful information on activities or patterns that are extremely rare. Therefore, the more mature algorithms address this issue by abstracting from exceptional/infrequent behavior. Noise can be removed by preprocessing the log, or the discovery algorithm can abstract from noise while constructing the model. The ability or inability to deal with noise is an important characteristic of a process discovery algorithm.

7.1.3 Characteristic 3: Completeness Notion Assumed

Related to noise is the issue of *completeness*. Most process discovery algorithms make an implicit or explicit completeness assumption. For example, the α -algorithm assumes that the relation $>_L$ is complete, i.e., if one activity can be directly followed by another activity, then this should be seen at least once in the log. Other algorithms make other completeness assumptions. Some algorithms assume that the event log contains all possible traces, i.e., a very strong completeness assumption. This is very unrealistic and results in overfitting models. Algorithms that are characterized by a strong completeness assumption tend to overfit the log. A completeness assumption that is too weak tends to result in underfitting models.

7.1.4 Characteristic 4: Approach Used

There are many different approaches to do the actual discovery. It is impossible to give a complete overview. Moreover, several approaches are partially overlapping in terms of the techniques used. Nevertheless, we briefly discuss five characteristic families of approaches.

7.1.4.1 Direct Algorithmic Approaches

The first family of process discovery approaches extracts some *footprint* from the event log and uses this footprint to *directly* construct a process model. The α -algorithm [157] is an example of such an approach: relation $>_L$ is extracted from the log and based on this relation a Petri net is constructed. There are several variants of the α -algorithm [11, 185, 186] using a similar approach. Approaches using so-called “language-based regions” [19, 28, 170] infer places by converting the event log into a system of inequations. In this case, the system of inequations can be seen as the footprint used to construct the Petri net. See [174] for a survey of process mining approaches producing a Petri net. The approaches described in [66, 183, 184] also extract footprints from event logs. However, these approaches take frequencies into account to address issues related to noise and incompleteness.

7.1.4.2 Two-Phase Approaches

The second family of process discovery approaches uses a two-step approach in which first a “low-level model” (e.g., a transition system or Markov model) is constructed. In the second step the low-level model is converted into a “high-level model” that can express concurrency and other (more advanced) control-flow patterns. An example of such an approach is described in [165]. Here a transition system is extracted from the log using a customizable abstraction mechanism. Subsequently, the transition system is converted into a Petri net using so-called “state-based regions” [34]. The resulting model can be visualized as a Petri net, but can also

be converted into other notations (e.g. BPMN and EPCs). Similar approaches can be envisioned using hidden Markov models [9]. Using an Expectation-Maximization (EM) algorithm such as the Baum–Welch algorithm, the “most likely” Markov model can be derived from a log. Subsequently this model is converted into high-level model. A drawback of such approaches is that the representational bias cannot be exploited during discovery. Moreover, some of the mappings are “lossy”, i.e., the process model needs to be slightly modified to fit the target language. These algorithms also tend to be rather slow compared to more direct algorithmic approaches.

7.1.4.3 Divide-and-Conquer Approaches

Rather than using a single pass through the event log, it is also possible to try and break the problem into smaller problems. The inductive miner [88] aims to split the event log recursively into sublogs. For example, if one group of activities is preceded by another group of activities, but never the other way around, then we may deduce that these groups are in a sequence relation. Subsequently, the event log is decomposed based on the two groups of activities. Next to the sequence relation, the inductive miner also detects choices, concurrency and loops. The sublogs are decomposed until they refer to single activity. The way that the log is decomposed provides a structured process model. Various inductive process discovery techniques have been developed for process trees (Sect. 3.2.8) [88–91].

7.1.4.4 Computational Intelligence Approaches

Techniques originating from the field of *computational intelligence* form the basis for the third family of process discovery approaches. Examples of techniques are ant colony optimization, genetic programming, genetic algorithms, simulated annealing, reinforcement learning, machine learning, neural networks, fuzzy sets, rough sets, and swarm intelligence. These techniques have in common that they use an evolutionary approach, i.e., the log is not directly converted into a model but uses an iterative procedure to mimic the process of natural evaluation. It is impossible to provide an overview of computational intelligence techniques here. Instead we refer to [9, 102] and use the *genetic process mining* approach described in [12] as an example. This approach starts with an initial population of individuals. Each individual corresponds to a randomly generated process model. For each individual a fitness value is computed describing how well the model fits with the log. Populations evolve by selecting the fittest individuals and generating new individuals using genetic operators such as crossover (combining parts of two individuals) and mutation (random modification of an individual). The fitness gradually increases from generation to generation. The process stops once an individual (i.e., model) of acceptable quality is found.

7.1.4.5 Partial Approaches

The approaches described thus far produce a complete end-to-end process model. It is also possible to focus on rules or frequent patterns. In Sect. 4.5.1, an approach for *mining of sequential patterns* was described [131]. This approach is similar to the discovery of association rules, however, now the order of events is taken into account. Another technique using an Apriori-like approach is the *discovery of frequent episodes* [94] described in Sect. 4.5.2. Here a sliding window is used to analyze how frequent an “episode” (i.e., a partial order) is appearing. Similar approaches exist to learn declarative (LTL-based) languages like *Declare* [162].

In the remainder, we discuss four approaches in more detail: heuristic mining (Sect. 7.2), genetic process mining (Sect. 7.3), region-based mining (Sect. 7.4), and inductive mining (Sect. 7.5). The chapter concludes with a historical perspective on process discovery going back to the classical work of Marc Gold, Anil Nerode, Alan Biermann, and others.

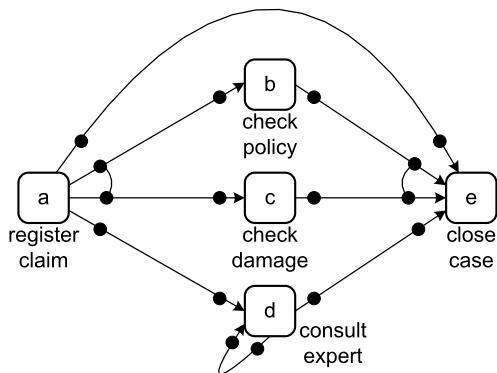
7.2 Heuristic Mining

Heuristic mining algorithms as described in [183, 184] use a representation similar to causal nets (see Sect. 3.2.7). Moreover, these algorithms take frequencies of events and sequences into account when constructing a process model. The basic idea is that infrequent paths should not be incorporated into the model. Both the representational bias provided by causal nets and the usage of frequencies makes the approach much more robust than most other approaches.

7.2.1 Causal Nets Revisited

In Sect. 3.2.7, we introduced the notion of causal nets, also referred to as C-nets. Figure 7.2 shows another example of a C-net. There is one start activity a representing the registration of an insurance claim. There is one end activity e that closes the case. Activity a has three output bindings: $\{b, c\}$, $\{d\}$ and $\{e\}$, indicating that after completing a , activities b and c are activated, d is activated, or e is activated. Recall that only valid sequences are considered (see Definition 3.11) when reasoning about the behavior of a C-net. A binding sequence is valid if the sequence (a) starts with start activity $a_i = a$, (b) ends with end activity $a_o = e$, (c) only removes obligations that are pending, and (d) ends without any pending obligations. Suppose that a occurs with output binding $\{b, c\}$. After executing $\langle(a, \emptyset, \{b, c\})\rangle$, there are two pending obligations: (a, b) and (a, c) . This indicates that in the future b should occur with a in its input binding. Similarly, c should occur with a in its input binding. Executing b removes the obligation (a, b) , but creates a new obligation (b, e) , etc. An example of a valid sequence is $\langle(a, \emptyset, \{b, c\}), (b, \{a\}, \{e\}), (c, \{a\}, \{e\}), (e, \{b, c\}, \emptyset)\rangle$. At the end, there are no

Fig. 7.2 Causal net modeling the handling of insurance claims



pending obligations. $\langle(a, \emptyset, \{d\}), (d, \{a\}, \{d\}), (d, \{d\}, \{e\}), (e, \{d\}, \emptyset)\rangle$ is another valid sequence. Because of the loop involving d there are infinitely many valid sequences.

The process modeled by Fig. 7.2 cannot be expressed as a WF-net (assuming that each transition has a unique visible label). This illustrates that C-nets are a more suitable representation for process discovery.

There are subtle differences between the notation used in [183, 184] and the C-nets used in this book. Whereas C-nets are very similar to the notation used in [183], there are relevant differences with [184]. In the original heuristic mining algorithm input and output bindings are a conjunction of mutually exclusive disjunctions, e.g., $O(t) = \{\{a, b\}, \{b, c\}, \{b, d\}\}$ means that t will activate a or b , and b or c , and b or d . These are exclusive or's. Hence, using the C-net semantics provided in Sect. 3.2.7 this corresponds to $O(t) = \{\{a, c, d\}, \{b\}\}$, i.e., either just b is activated or a, c and d are activated. C-nets are more intuitive and also more expressive (in a practical sense) than the original heuristic nets. Therefore, we use C-nets in the remainder.

7.2.2 Learning the Dependency Graph

To illustrate the basic concepts used by heuristic mining algorithms, we use the following event log:

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

If we assume the three traces with frequency one to be noise, then the remaining 37 traces in the log correspond to valid sequences of the C-net in Fig. 7.2. Before explaining how to derive such a C-net, we first apply the α -algorithm to event log L . The result is shown in Fig. 7.3.

As expected, the α -algorithm does not infer a suitable model. The model does not allow for frequent traces, such as $\langle a, e \rangle$ and $\langle a, d, e \rangle$. By accident the model also

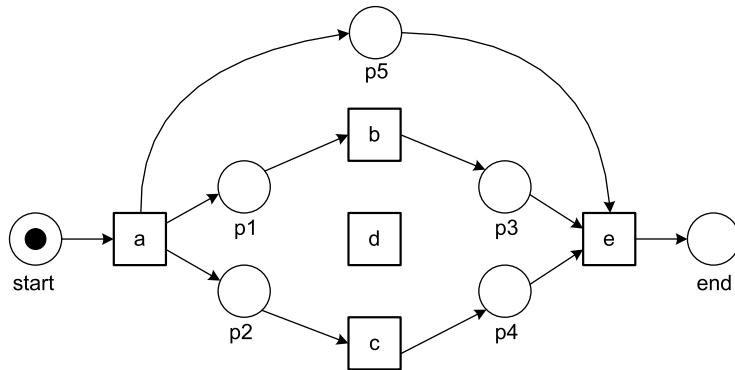


Fig. 7.3 WF-net constructed by the α -algorithm. The resulting model does not allow for $\langle a, e \rangle$, $\langle a, b, e \rangle$, $\langle a, c, e \rangle$, $\langle a, d, e \rangle$, $\langle a, d, d, e \rangle$, and $\langle a, d, d, d, e \rangle$

Table 7.1 Frequency of the “directly follows” relation in event log L : $|x >_L y|$ is the number of times x is directly followed by y in L

$ >_L $	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0

does not allow for infrequent traces such as $\langle a, b, e \rangle$, $\langle a, c, e \rangle$, and $\langle a, d, d, d, e \rangle$. There are two main problems. One problem is that the α -algorithm has a representational bias that does not allow for skipping activities (e.g., jumping from a to e) and cannot handle the requirement that d should be executed at least once when selected. The other problem is that the α -algorithm does not consider frequencies. Therefore, we use C-nets and take frequencies into account for heuristic mining.

Table 7.1 shows the number of times one activity is directly followed by another activity. For instance, $|d >_L d| = 4$, i.e., in the entire log d is followed four times by another d (two times in $\langle a, d, d, e \rangle^2$ and two times in $\langle a, d, d, d, e \rangle^1$). Using Table 7.1 we can calculate the value of the *dependency relation* between any pair of activities.

Definition 7.1 (Dependency measure) Let L be an event log¹ over \mathcal{A} and $a, b \in \mathcal{A}$. $|a >_L b|$ is the number of times a is directly followed by b in L , i.e.,

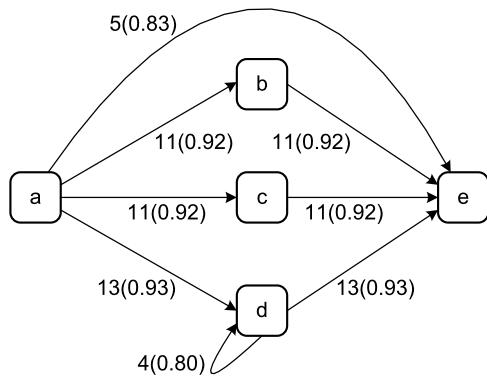
$$|a >_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

¹Note that in this chapter we again assume that the event log is simple (like in Chap. 6) because at this stage we still abstract from the other perspectives.

Table 7.2 Dependency measures between the five activities based on event log L

$ \Rightarrow_L $	a	b	c	d	e
a	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
b	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
c	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
d	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
e	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

Fig. 7.4 Dependency graph using a threshold of 2 for $| >_L |$ and 0.7 for $| \Rightarrow_L |$: each arc shows the $| >_L |$ value and the $| \Rightarrow_L |$ value between brackets. For example, $|a >_L d| = 13$ and $|a \Rightarrow_L d| = 0.93$



$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b :

$$|a \Rightarrow_L b| = \begin{cases} \frac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \frac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

$|a \Rightarrow_L b|$ produces a value between -1 and 1 . If $|a \Rightarrow_L b|$ is close to 1 , then there is a strong positive dependency between a and b , i.e., a is often the cause of b . A value close to 1 can only be reached if a is often directly followed by b but b is hardly ever directly followed by a . If $|a \Rightarrow_L b|$ is close to -1 , then there is a strong negative dependency between a and b , i.e., b is often the cause of a . There is a special case for $|a \Rightarrow_L a|$. If a is often followed by a this suggests a loop and a strong reflexive dependency. However, $\frac{|a >_L a| - |a >_L a|}{|a >_L a| + |a >_L a| + 1} = 0$ by definition. Therefore, the following formula is used: $|a \Rightarrow_L a| = \frac{|a >_L a|}{|a >_L a| + 1}$. Table 7.2 shows the dependency measures for event log L .

Using the information in Tables 7.1 and 7.2 we can derive the so-called *dependency graph*. The dependency graph corresponds to the dependency relation $D \subseteq A \times A$ in Definition 3.8. In a dependency graph only arcs are shown that meet certain *thresholds*. The dependency graph shown in Fig. 7.4 uses a threshold of 2 for $| >_L |$ and 0.7 for $| \Rightarrow_L |$, i.e., an arc between x and y is only included if $|x >_L y| \geq 2$ and $|x \Rightarrow_L y| \geq 0.7$.

Fig. 7.5 Dependency graph using a threshold of 5 for $|>_L|$ and 0.9 for $|⇒_L|$. The self loop involving d disappeared because $|d >_L d| = 4 < 5$ and $|d \Rightarrow_L d| = 0.80 < 0.9$. The connection between a and e disappeared because $|a \Rightarrow_L e| = 0.83 < 0.9$

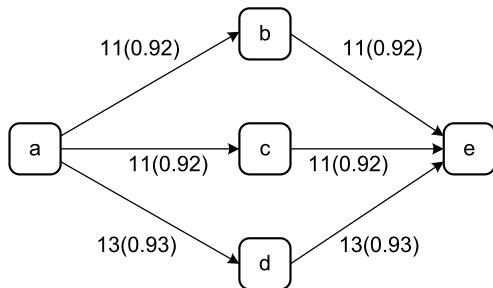


Figure 7.5 shows another dependency graph based on Tables 7.1 and 7.2 using higher thresholds. As a result two arcs disappear. Obviously, the dependency graph does not show the routing logic, e.g., one cannot see that after a , both b and c can be executed concurrently. Nevertheless, the dependency graph reveals the “backbone” of the process model.

The two dependency graphs show that, for a given event log, different models can be generated by adjusting the thresholds. This way the user can decide to focus on the mainstream behavior or to also include low frequent (i.e., noisy) behavior. In Figs. 7.4 and 7.5 the set of activities is the same. The two thresholds cannot be used to remove low frequent activities. This should be done by preprocessing the event log. For example, one could decide to concentrate on the most frequent activities and simply remove all other activities from the event log before calculating the dependency measures. Other techniques such as the one used by the Fuzzy Miner [66] remove such activities while realizing the dependency graph.

As shown in [183, 184], various refinements can be used to improve the dependency graph. For instance, it is possible to better deal with loops of length two and long distance dependencies. (See discussion in context of the processes shown in Figs. 6.13 and 6.14.)

7.2.3 Learning Splits and Joins

The goal of heuristic mining is to extract a C-net $C = (A, a_i, a_o, D, I, O)$ from the event log. The nodes of the dependency graph correspond to the set of activities A . The arcs of the dependency graph correspond to the dependency relation D . In a C-net, there is a unique start activity a_i and a unique end activity a_o . This is just a technicality. One can preprocess the log and insert artificial start and end events to each trace. Hence the assumption that there is a unique start activity a_i and a unique end activity a_o imposes no practical limitations. In fact, it is convenient to have a clear start and end. We also assume that in the dependency graph all activities are on a path from a_i to a_o . Activities that are not on such a path should be removed or the thresholds need to be adjusted locally such that a minimal set of connections is established. It makes no sense to include activities that are not on a path from a_i

to a_o : such an activity would be dead or could be active before the case starts, and does not contribute to the completion of the case. Therefore, we can assume that, by constructing the dependency graph, we already have the core structure of the C-net: (A, a_i, a_o, D) . Hence, only the functions $I \in A \rightarrow AS$ and $O \in A \rightarrow AS$ need to be derived to complete the C-net.

Given a dependency graph (A, a_i, a_o, D) , we define $oa = \{a' \in A \mid (a', a) \in D\}$ and $ao = \{a' \in A \mid (a, a') \in D\}$ for any $a \in A$. Clearly, $I(a_i) = O(a_o) = \{\emptyset\}$. There are $2^{|oa|} - 1$ potential elements for $I(a)$ for any $a \neq a_i$ and $2^{|ao|} - 1$ potential elements for $O(a)$ for any $a \neq a_o$. Consider, for example, the dependency graph shown in Fig. 7.4. $ao = \{b, c, d, e\}$. Hence, $O(a)$ has $2^4 - 1 = 15$ potential output bindings: $\{b\}, \{c\}, \{d\}, \{e\}, \{b, c\}, \{b, d\}, \dots, \{b, c, d, e\}$. $O(b)$ has only $2^1 - 1 = 1$ possible element, $\{e\}$. $I(b)$ also has just one possible element, $\{a\}$. $O(d)$ has $2^2 - 1 = 3$ potential output bindings: $\{d\}, \{e\}$, and $\{d, e\}$. $I(d)$ also has $2^2 - 1 = 3$ potential input bindings, $\{a\}, \{d\}$, and $\{a, d\}$.

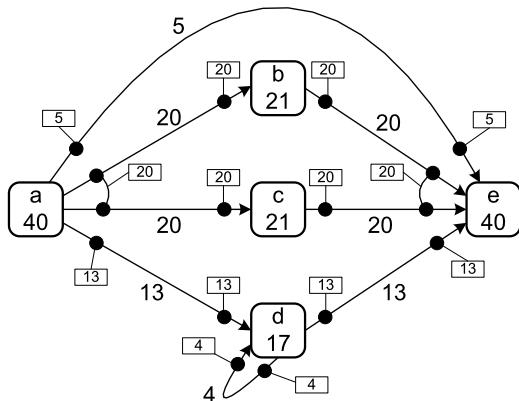
If there is just one potential binding element, then this element should be taken. Hence, $I(b) = \{\{a\}\}$, $I(c) = \{\{a\}\}$, $O(b) = \{\{e\}\}$, and $O(c) = \{\{e\}\}$. For the other input and output bindings, subsets need to be selected based on the event log. To do this, the event log is replayed on the dependency graph to see how frequent output sets are triggered.

Consider, for example, $O(d)$. In event log L , activity d is four times followed by just d and 13 times by just e ; d is never followed by both d and e . Therefore, $\{e\}$ is definitely included in $O(d)$ because it is the most frequent output binding. $\{d\}$ may be included depending on the threshold for including bindings. If we assume that both possible bindings are included, then $O(d) = \{\{d\}, \{e\}\}$. Similarly, we find $I(d) = \{\{a\}, \{d\}\}$. Let us now consider $O(a)$. As indicated earlier there are $2^4 - 1 = 15$ possible output bindings. Replayng the event log on the dependency graph shows that a is 5 times followed by e (in traces $\langle a, e \rangle^5$), a is 20 times followed by both b and c (in traces $\langle a, b, c, e \rangle^{10}$ and $\langle a, c, b, e \rangle^{10}$), and a is 13 times followed by d (in traces $\langle a, d, e \rangle^{10}$, $\langle a, d, d, e \rangle^2$, and $\langle a, d, d, d, e \rangle^1$). Activity a is once followed by just b (in trace $\langle a, b, e \rangle$) and is once followed by just c (in trace $\langle a, c, e \rangle$). Let us assume that the latter two output bindings are below a preset threshold. Then $O(a) = \{\{b, c\}, \{d\}, \{e\}\}$, i.e., of the 15 possible output bindings only three are frequent enough to be included.

Many replay strategies are possible to determine the frequency of a binding. In [119, 183, 184] heuristics are used to select the bindings to be included. In [4], a variant of the A^* algorithm is used to find an “optimal” replay of traces on the dependency graph. The semantics of a C-net are global, i.e., the validity of a binding sequence cannot be determined locally (like in a Petri net). We refer to [4, 119, 183, 184] for example replay strategies.

By replaying the event log on the dependency graph, we can estimate the frequencies of input and output bindings. Using thresholds, it is possible to exclude bindings based on their frequencies. This results in functions I and O , thus completing the C-net. Figure 7.6 shows the C-net based on the dependency graph in Fig. 7.4. As shown $O(a) = \{\{b, c\}, \{d\}, \{e\}\}$ and $I(e) = \{\{a\}, \{b, c\}, \{d\}\}$. Bindings $\{b\}$ and $\{c\}$ are not included in $O(a)$ and $I(e)$ because they occur only once (below

Fig. 7.6 C-net derived from the event log L . Each node shows the frequency of the corresponding activity. Every arc has a frequency showing how often both activities agreed on a common binding. The frequencies of input and output bindings are also depicted, e.g., 20 of the 40 occurrences of a were followed by the concurrent execution of b and c



threshold). Figure 7.6 also shows the frequencies of activities, dependencies, and bindings. For example, activity a occurred 40 times. The output binding $\{b, c\}$ of a occurred 20 times. Activity d occurred 17 times: 13 times triggered by a and 4 times by d itself. Activity b occurred 21 times. The frequency of the only input binding $\{a\}$ is only 20. This difference is caused by the exclusion of the infrequent output binding $\{b\}$ of a (this binding occurs only in trace $\langle a, b, e \rangle$). A similar difference can be found for activity c .

Figure 7.7 provides a more intuitive visualization of the C-net of Fig. 7.6. Now the thickness of the arcs corresponds to the frequencies of the corresponding paths. Such visualizations are important to get insight into the main process flows. In Chap. 15, we will adopt the metaphor of a roadmap to visualize process models. A roadmap highlights highways using thick lines and bright colors. At the same time insignificant roads are not shown. Figure 7.7 illustrates that the same can be done using heuristic mining.

The approach presented in this section is quite generic and can be applied to other representations. A notable example is the *fuzzy mining* approach described in [65, 66]. This approach provides an extensible set of parameters to determine which activities and arcs need to be included. Moreover, the approach can construct hierarchical models, i.e., less frequent activities may be moved to subprocesses. Also the metaphor of a roadmap is exploited to create process models that can be understood easily while providing information on the frequency and importance of activities and paths (cf. Sect. 15.1.3).

7.3 Genetic Process Mining

The α -algorithm and techniques for heuristic and fuzzy mining provide process models in a direct and deterministic manner. *Evolutionary approaches* use an iterative procedure to mimic the process of natural evolution. Such approaches are not deterministic and depend on randomization to find new alternatives. This sec-

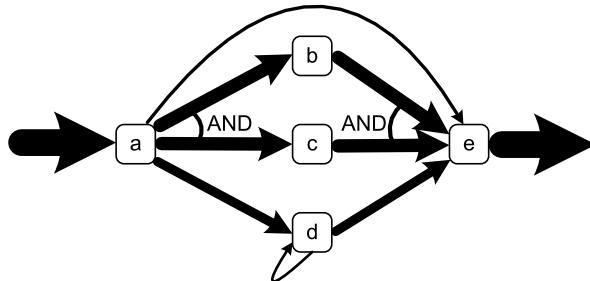


Fig. 7.7 Alternative visualization of the C-net clearly showing the “highways” in the process model

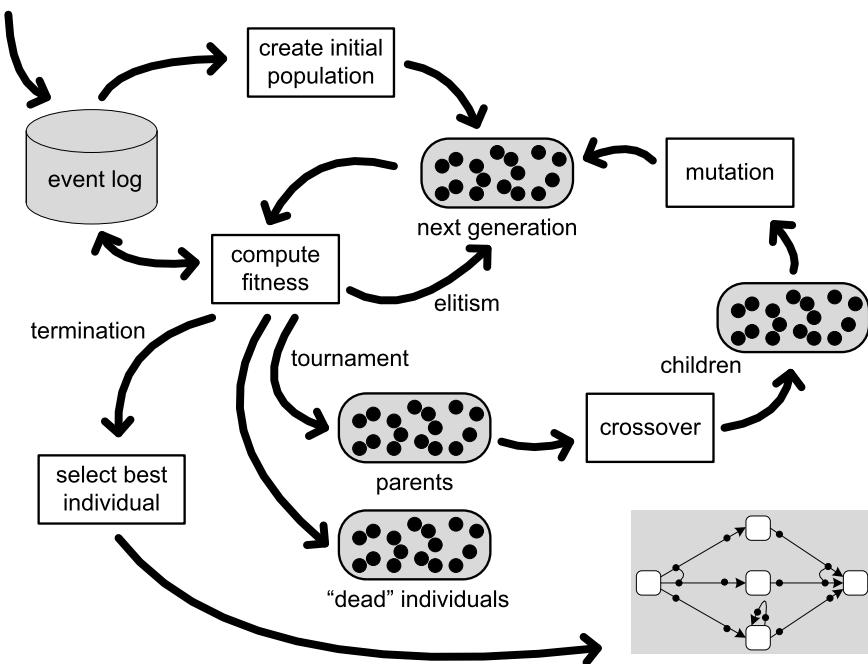


Fig. 7.8 Overview of the approach used for genetic process mining

tion describes *genetic process mining* [12] as an example of a process discovery approach using a technique from the field of computational intelligence.

Figure 7.8 shows an overview of the approach used in [12]. Like in any genetic algorithm there are four main steps: (a) initialization, (b) selection, (c) reproduction, and (d) termination.

In the *initialization* step the initial population is created. This is the first generation of individuals to be used. Here an individual is a process model. Using the activity names appearing in the log, process models are created *randomly*. There may

be hundreds or even thousands individuals in each generation. The process models (i.e., individuals) in the initial population may have little to do with the event log; the activity names are the same but the behaviors of the initial models are likely to be very different from the behavior seen in the event log. However, “by accident” the generated models may have parts that fit parts of the event log due random effects and the large number of individuals.

In the *selection* step, the fitness of each individual is computed. A fitness function determines the quality of the individual in relation to the log.² In Sect. 6.4.3, we discussed different ways of measuring the quality of a model. A simple criterion is the proportion of traces in the log that can be replayed by the model. This is not a good fitness function, because it is very likely that none of the models in the initial population can replay any of the traces in the event log. Moreover, using this criterion an over-general model like the “flower model” would have a high fitness. Therefore, a more refined fitness function needs to be used that also rewards the partial correctness of the model and takes into account all four competing quality criteria described in Sect. 6.4.3. The best individuals, i.e., the process models having the highest fitness value are moved to the next generation. This is called *elitism*. For instance, the best 1% of the current generation is passed on to the next generation without any modifications. Through *tournaments* “parents” are selected for creating new individuals. Tournaments among individuals and elitism should make sure that the “genetic material of the best process models” has the highest probability of being used for the next generation: *survival of the fittest*. As a result, individuals with a poor fitness are unlikely to survive. Figure 7.8 refers to such models as “dead” individuals.

In the *reproduction* phase the selected parent individuals are used to create new offspring. Here two genetic operators are used: *crossover* and *mutation*. For crossover two individuals are taken and used to create two new models; these end up in the pool with “child models” shown in Fig. 7.8. These child models share parts of the genetic material of their parents. The resulting children are then modified using mutation, e.g., randomly adding or deleting a causal dependency. Mutation is used to insert new generic material in the next generation. Without mutation, evolution beyond the genetic material in the initial population is impossible.

Through reproduction (i.e., crossover and mutation) and elitism a new generation is created. For the models in this generation the fitness is computed. Again the best individuals move on to the next round (elitism) or are used to produce new offspring. This is repeated and the expectation is that the “quality” of each generation gets better and better. The evolution process *terminates* when a satisfactory solution is found, i.e., a model having at least the desired fitness. Depending on the event log it may take a very long time to converge. In fact, due to the representational bias and noise in the event log there may not be a model that has the desired level of

²Note that we overload the term “fitness” in this book. On the one hand, we use it to refer to the ability to replay the event log (see Sects. 6.4.3 and 8.2). On the other hand, we use it for the selection of individuals in genetic process mining. Note that the latter interpretation includes the former, but also adds other elements of the four criteria mentioned in Sect. 6.4.3.

fitness. Therefore, other termination criteria may be added (e.g., a maximum number of generations or stopping when 10 successive generations do not produce better individuals). When terminating, a model with the best fitness is returned.

The approach described in Fig. 7.8 is very general. When actually implementing a genetic process mining algorithm the following design choices need to made:

- *Representation of individuals.* Each individual corresponds to a process model described in a particular language, e.g., Petri nets, C-nets, BPMN, or EPCs. This choice is important as it determines the class of processes that can be discovered (representational bias). Moreover, it should be possible to define suitable genetic operators for the representation chosen. In [12], a variant of C-nets is used.
- *Initialization.* For the initial population, models need to be generated randomly. In [12], two approaches are proposed: (a) an approach where with a certain probability a causal dependency between two activities is inserted to create C-nets and (b) an approach in which a randomized variant of heuristic mining is used to create an initial population with a higher average fitness than purely randomly generated C-nets.
- *Fitness function.* Here, the challenge is to define a function that balances the four competing quality criteria described in Sect. 6.4.3. Many fitness functions can be defined. The fitness function drives the evolution process and can be used to favor particular models. In [12], the proportion of events in the log that can be parsed by the model is computed. This is combined with penalties for having many enabled activities (cf. the flower model in Fig. 6.23).
- *Selection strategy (tournament and elitism).* The genetic algorithm needs to determine the fraction of individuals that go to the next round without any changes. Through elitism it is ensured that good models do not get lost due to crossover or mutation. There are different approaches to select parents for crossover. In [12], parents are selected by randomly taking five individuals and then selecting the best one, i.e., a tournament among five randomly selected models is used.
- *Crossover.* The goal of crossover is to recombine existing genetic material. The basic idea is to create a new process model that uses parts of its two parent models. In [10, 12], both parents are C-nets having the same set of activities. One of these common activities is selected randomly, say a . Let $I_1(a)$ and $O_1(a)$ be the possible bindings of one parent, and let $I_2(a)$ and $O_2(a)$ be the potential bindings of the other parent. Now parts of $I_1(a)$ are swapped with parts of $I_2(a)$ and parts of $O_1(a)$ are swapped with parts of $O_2(a)$. Subsequently, both C-nets are repaired as bindings need to be consistent among activities. The crossover of two parent models results in two new child models. These child models may be mutated before being added to the next generation.
- *Mutation.* The goal of mutation is to randomly insert new genetic material. In [10, 12], each activity in each child resulting from crossover has a small probability of being selected for mutation. If this is the case, say a is selected for mutation, then $I(a)$ or $O(a)$ is randomly modified by adding or removing potential bindings.

The above list shows that many design decisions need to be taken when developing a genetic process mining algorithm. We refer to [10, 12] for concrete examples.

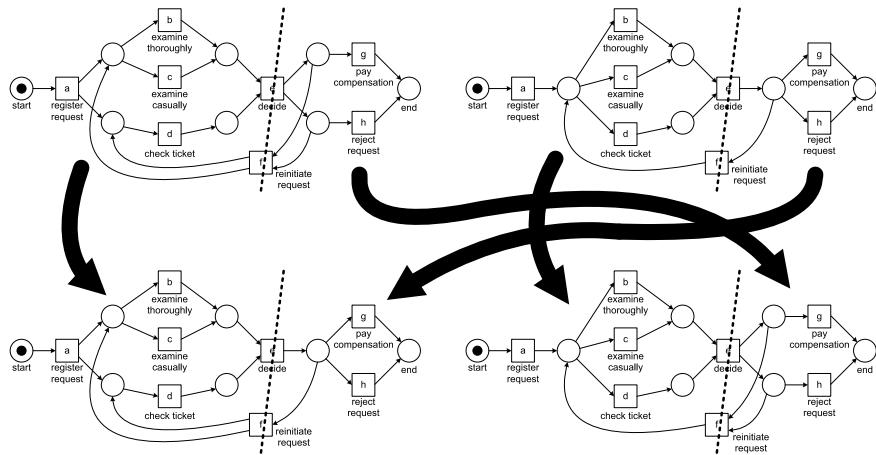


Fig. 7.9 Two parent models (top) and two child models resulting from a crossover. The crossover points are indicated by the dashed lines

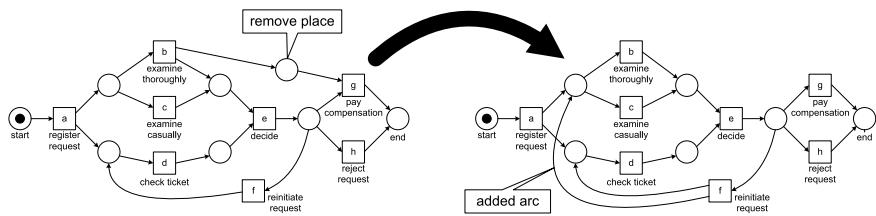


Fig. 7.10 Mutation: a place is removed and an arc is added

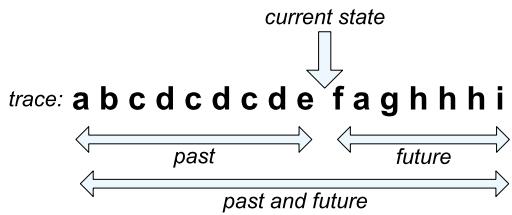
An essential choice is the representation of individuals. The approach described in [10, 12] uses a variant of C-nets similar to the notation used for the initial heuristic mining algorithm [184]. However, many other representations are possible.

To illustrate the genetic operators, we show a crossover example and a mutation example. For clarity we use Petri nets to describe the individuals before and after modification. Figure 7.9 shows two “parent” models and two “child” models resulting from crossover. In this example, the crossover point is the line through activities *e* and *f*. Figure 7.10 shows an example of mutation: one place is removed and one arc is added.

Figures 7.9 and 7.10 nicely illustrate the idea behind the two genetic operators: crossover and mutation. However, the realization of such operators is not as simple as these examples suggest. Typically repair actions are needed after crossover and mutation. For instance, the resulting model may no longer be a WF-net or C-net. Again we refer to [10, 12] for concrete examples.

Genetic process mining is *flexible* and *robust*. Like heuristic mining techniques, it can deal with noise and incompleteness. The approach can also be adapted and extended easily. By changing the fitness function it is possible to give preference to

Fig. 7.11 Every position in a trace corresponds to a state, e.g., the state after executing the first nine events of a trace consisting of 16 events. To characterize the state, the past and/or future can be used as “ingredients”



particular constructs. Unfortunately, like most evolutionary approaches, genetic process mining is *not very efficient* for larger models and logs. It may take a very long time to discover a model having an acceptable fitness. In theory, it can be shown that suitably chosen genetic operators guarantee that eventually a model with optimal fitness will be produced. However, in practice this argument is not useful given the potentially excessive computation times. It is also useful to combine heuristics with genetic process mining. In this case, genetic process mining is used to improve a process model obtained using heuristic mining. This saves computation time and may result in models that could never have been obtained through conventional algorithms searching only for local dependencies.

7.4 Region-Based Mining

In the context of Petri nets, researchers have been looking at the so-called *synthesis problem*, i.e., constructing a system model from a description of its behavior. State-based regions can be used to construct a Petri net from a transition system. Language-based regions can be used to construct a Petri net from a prefix-closed language. Synthesis approaches using language-based regions can be applied directly to event logs. To apply state-based regions, one first needs to create a transition system.

7.4.1 Learning Transition Systems

To construct a Petri net using state-based regions, we first need to discover a transition system based on the traces in the event log. Recall that a transition system can be described by a triplet $TS = (S, A, T)$ where S is the set of *states*, $A \subseteq \mathcal{A}$ is the set of *activities*, and $T \subseteq S \times A \times S$ is the set of *transitions*. $S^{start} \subseteq S$ is the set of *initial states*. $S^{end} \subseteq S$ is the set of *final states*. (See Sect. 3.2.1 for an introduction to transition systems.)

How to construct $TS = (S, A, T)$ based on some simple event log L over \mathcal{A} , i.e., $L \in \mathbb{B}(\mathcal{A}^)$?* An obvious choice is to take A to be the set of activities in the simple event log. In order to determine the set of states, each “position” in each trace in the log needs to be mapped onto a corresponding state. This is illustrated by Fig. 7.11.

Let $\sigma' = \langle a, b, c, d, c, d, c, d, e, f, a, g, h, h, h, i \rangle \in L$ be a trace in the event log. Every position in this trace, i.e., before the first event, in-between two events, or after the last event should correspond to a state in the transition system. Consider, for example, the state shown in Fig. 7.11. The partial trace $\sigma'_{past} = \langle a, b, c, d, c, d, c, d, e \rangle$ describes the past of the corresponding case. $\sigma'_{future} = \langle f, a, g, h, h, h, i \rangle$ describes the future of this case. A *state representation* function $l_1^{state}()$ is a function that, given some sequence σ and a number k indicating the number of events of σ that have occurred, produces some state, e.g., the set of activities that have occurred in the first k events.

Let $\sigma = \langle a_1, a_2, \dots, a_n \rangle \in L$ be a trace of length n . $l_1^{state}(\sigma, k) = hd^k(\sigma) = \langle a_1, a_2, \dots, a_k \rangle$ is an example of a state representation function. Recall that $hd^k(\sigma)$ was defined in Sect. 5.2; the function returns the ‘‘head’’ of the sequence σ consisting of the first k elements. $l_1^{state}(\sigma, k)$ describes the current state by the *full history* of the case after k events. For instance, $l_1^{state}(\sigma', 9) = \langle a, b, c, d, c, d, c, d, e \rangle$.

$l_2^{state}(\sigma, k) = tl^{n-k}(\sigma) = \langle a_{k+1}, a_{k+2}, \dots, a_n \rangle$ is another example of a state representation function. $l_2^{state}(\sigma, k)$ describes the current state by the *full future* of the case after k events. $l_2^{state}(\sigma', 9) = \langle f, a, g, h, h, h, i \rangle$.

$l_3^{state}(\sigma, k) = \partial_{multiset}(hd^k(\sigma)) = [a_1, a_2, \dots, a_k]$ is a state representation function converting the full history into a multi-set. This function assumes that for the current state the order of events is not important, only the frequency of activities matters. $l_3^{state}(\sigma', 9) = [a^1, b^1, c^3, d^3, e^1]$, i.e., in the state shown in Fig. 7.11 *a*, *b*, and *e* have been executed once and both *c* and *d* have been executed three times.

$l_4^{state}(\sigma, k) = \partial_{set}(hd^k(\sigma)) = \{a_1, a_2, \dots, a_k\}$ is a state representation function taking a set representation of the full history. For this state representation function the order and frequency of activities do not matter. For the current state it only matters which activities have been executed at least once. $l_4^{state}(\sigma', 9) = \{a, b, c, d, e\}$.

Functions $l_1^{state}()$, $l_3^{state}()$, and $l_4^{state}()$ all consider the full history of the case after k events: $l_1^{state}()$ does not abstract from the order and frequency of past activities, $l_3^{state}()$ abstracts from the order, and $l_4^{state}()$ abstracts from both order and frequency. Hence, $l_4^{state}()$ provides a coarser abstraction than $l_1^{state}()$. By definition $l_4^{state}(\sigma_1, k) = l_4^{state}(\sigma_2, k)$ if $l_1^{state}(\sigma_1, k) = l_1^{state}(\sigma_2, k)$ (but not the other way around). Function $l_2^{state}()$ is based on the future rather than the past.

Using some state representation function $l^{state}()$ we can automatically construct a transition system based on some event log L .

Definition 7.2 (Transition system based on event log) Let $L \in \mathbb{B}(\mathcal{A}^*)$ be an event log and $l^{state}()$ a state representation function. $TS_{L, l^{state}()} = (S, A, T)$ is a transition system based on L and $l^{state}()$ with:

- $S = \{l^{state}(\sigma, k) \mid \sigma \in L \wedge 0 \leq k \leq |\sigma|\}$ is the state space;
- $A = \{\sigma(k) \mid \sigma \in L \wedge 1 \leq k \leq |\sigma|\}$ is the set of activities;
- $T = \{(l^{state}(\sigma, k), \sigma(k+1), l^{state}(\sigma, k+1)) \mid \sigma \in L \wedge 0 \leq k < |\sigma|\}$ is the set of transitions;
- $S^{start} = \{l^{state}(\sigma, 0) \mid \sigma \in L\}$ is the set of initial states; and
- $S^{end} = \{l^{state}(\sigma, |\sigma|) \mid \sigma \in L\}$ is the set of final states.

Fig. 7.12 Transition system $TS_{L_1, l_1^{state}()}$ derived from $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$ using $l_1^{state}(\sigma, k) = hd^k(\sigma)$

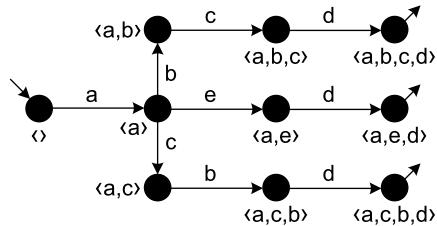


Fig. 7.13 Transition system $TS_{L_1, l_2^{state}()}$ derived from L_1 using $l_2^{state}(\sigma, k) = tl^{|\sigma|-k}(\sigma)$

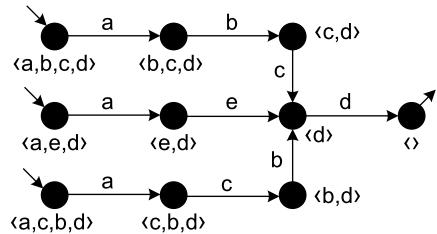
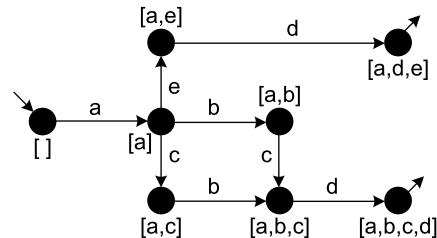


Fig. 7.14 Transition system $TS_{L_1, l_3^{state}()}$ derived from L_1 using $l_3^{state}(\sigma, k) = \partial_{multiset}(hd^k(\sigma))$

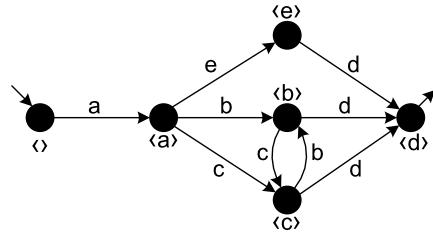


Let us consider event log $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$. Figure 7.12 shows transition system $TS_{L_1, l_1^{state}()}$. Consider, for example, a case with trace $\sigma = \langle a, b, c, d \rangle$. Initially, the case is in state $l_1^{state}(\sigma, 0) = \langle \rangle$. After executing a the case is in state $l_1^{state}(\sigma, 1) = \langle a \rangle$. After executing b state $l_1^{state}(\sigma, 2) = \langle a, b \rangle$ is reached. Executing c results in state $l_1^{state}(\sigma, 3) = \langle a, b, c \rangle$. Executing the last event d results in state $l_1^{state}(\sigma, 4) = \langle a, b, c, d \rangle$. The five states visited by this case are added to the transition system. The corresponding transitions are also added. The same is done for $\langle a, c, b, d \rangle$ and $\langle a, e, d \rangle$, thus resulting in the transition system of Fig. 7.12.

Using state representation function $l_2^{state}()$ we obtain transition system $TS_{L_1, l_2^{state}()}$ shown in Fig. 7.13. In this transition system there are three initial states and only one final state, because this abstraction uses the future rather than the past. Consider, for example, a case with trace $\sigma = \langle a, e, d \rangle$. Initially, the case is in state $l_2^{state}(\sigma, 0) = \langle a, e, d \rangle$, i.e., all three activities still need to occur. After executing a the case is in state $l_2^{state}(\sigma, 1) = \langle e, d \rangle$. After executing e state $l_2^{state}(\sigma, 2) = \langle d \rangle$ is reached. Executing the last event d results in state $l_2^{state}(\sigma, 3) = \langle \rangle$.

Transition system $TS_{L_1, l_3^{state}()}$ is shown in Fig. 7.14. Here, the states are represented by the multi-sets of activities that have been executed before. For instance, $l_3^{state}(\langle a, b, c, d \rangle, 3) = [a, b, c]$. Because there are no repeated activities $TS_{L_1, l_4^{state}()}$

Fig. 7.15 Transition system $TS_{L_1, l_5^{state}()}^{}()$ derived from L_1 using $l_5^{state}(\sigma, k) = tl^1(hd^k(\sigma))$



is identical to $TS_{L_1, l_3^{state}()}^{}()$ apart from the naming of states, e.g., $l_4^{state}(\langle a, b, c, d \rangle, 3) = \{a, b, c\}$ rather than $[a, b, c]$.

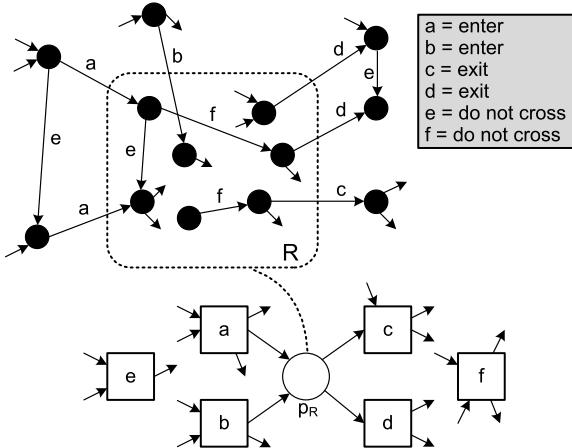
The sets of traces allowed by the three transition systems shown in Figs. 7.12, 7.13, and 7.14 are the same: $\langle a, b, c, d \rangle$, $\langle a, c, b, d \rangle$, $\langle a, e, d \rangle$. This is not always the case. Add, for example, the trace $\langle a, c, b, f, f \rangle$ to L_1 . In this case, $TS_{L_1, l_4^{state}()}^{}()$ allows for traces $\langle a, b, c, f, f \rangle$ and $\langle a, c, b, f, f, f, f, f \rangle$, i.e., b and c may be swapped and any number of f events is allowed at the end. $TS_{L_1, l_3^{state}()}^{}()$ allows for traces $\langle a, b, c, f, f \rangle$ and $\langle a, c, b, f, f \rangle$, but not $\langle a, c, b, f, f, f, f, f \rangle$. $TS_{L_1, l_1^{state}()}^{}()$ allows for trace $\langle a, c, b, f, f \rangle$, but not $\langle a, b, c, f, f \rangle$. Since $l_4^{state}()$ provides a coarser abstraction than $l_1^{state}()$, it generalizes more.

The state representation functions mentioned thus far are just examples. Depending on the desired abstraction, another state representation function can be defined. The essential question is whether partially executed cases are considered to be in the same state or not. For instance, if we assume that only the last activity matters, we can use state representation function $l_5^{state}(\sigma, k) = tl^1(hd^k(\sigma))$. This results in transition system $TS_{L_1, l_5^{state}()}^{}()$ shown in Fig. 7.15. Now, states in the transition system are labeled with the last activity executed. For the initial state this results in the empty sequence. $TS_{L_1, l_5^{state}()}^{}()$ allows for the traces in the event log, but also traces such as $\langle a, b, c, b, c, d \rangle$. Another example is $l_6^{state}(\sigma, k) = hd^3(tl^{|\sigma|-k}(\sigma))$, i.e., the state is determined by the next three events.

Thus far we only considered a simple event log as input. Real-life event logs contain much more information as was shown in Chap. 5 (cf. Definition 5.3 and the XES format). Information about resources and data can also be taken into account when constructing a transition system. This information can be used to identify states and to label transitions. For example, states may encode whether the customer being served is a gold or silver customer. Transitions can be labeled with resource names rather than activity names. See [165] for a systematic treatment of the topic.

A transition system defines a “low-level” process model. Unfortunately, such models cannot express higher level constructs and suffer from the “state explosion” problem. As indicated before, a simple process with 10 parallel activities already results in a transition system with $2^{10} = 1024$ states and $10 \times 2^{10-1} = 5120$ transitions. Fortunately, state-based regions can be used to synthesize a more compact model from such transition systems.

Fig. 7.16 Region R corresponding to place p_R . All activities can be classified into *entering* the region (a and b), *leaving* the region (c and d), and *non-crossing* (e and f)



7.4.2 Process Discovery Using State-Based Regions

After transforming an event log into a low-level transition system we can synthesize a Petri net from it. In turn, this Petri net can be used to construct a process model in some other high-level notation (e.g., BPMN, UML activity diagrams, YAWL, and EPCs). The challenge is to fold a large transition system into a smaller Petri net by detecting concurrency. The core idea is to discover *regions* that correspond to *places*. A region is a set of states such that all activities in the transition system “agree” on the region.

Definition 7.3 (State-based region) Let $TS = (S, A, T)$ be a transition system and $R \subseteq S$ be a subset of states. R is a *region* if for each activity $a \in A$ one of the following conditions hold:

1. All transitions $(s_1, a, s_2) \in T$ enter R , i.e., $s_1 \notin R$ and $s_2 \in R$;
2. All transitions $(s_1, a, s_2) \in T$ exit R , i.e., $s_1 \in R$ and $s_2 \notin R$; or
3. All transitions $(s_1, a, s_2) \in T$ do not cross R , i.e., $s_1, s_2 \in R$ or $s_1, s_2 \notin R$.

Let R be a region. In this case all activities can be classified into *entering* the region, *leaving* the region, and *non-crossing*. An activity cannot be entering the region in one part of the transition system and exiting the region in some other part. Figure 7.16 illustrates the concept. The dashed rectangle describes a region R , i.e., a set of states in the transition system. All activities need to take a position with respect to this region. All a -labeled transitions enter region R . If there would be a transition with an a label not connecting a state outside the region to a state inside the region, then R would not be a region. All b -labeled transitions enter the region, all c and d labeled transitions exit the region. All e and f labeled transitions do not cross R , i.e., they always connect two states outside the region or two states inside the region.

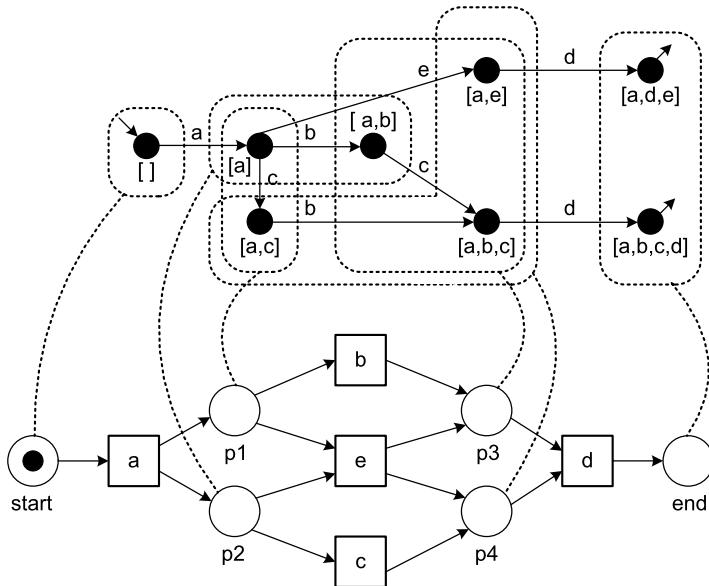


Fig. 7.17 Transition system $TS_{L, lstate_0}$ derived from $L_1 = [(a, b, c, d)^3, (a, c, b, d)^2, (a, e, d)]$ is converted into a Petri net using state-based regions

By definition, the union of two regions is again a region. Therefore, we are only interested in *minimal* regions. The basic idea is that each minimal region R corresponds to a place p_R in a Petri net as shown in Fig. 7.16. The activities entering the region become Petri-net transitions having p_R as output place, activities leaving the region become output transitions of p_R , and activities that do not cross the region correspond to Petri-net transitions not connected to p_R . Hence, the minimal regions fully encode a Petri net.

Figure 7.17 illustrates the concept of state-based regions using a concrete example. By applying Definition 7.3, we find six minimal regions. Consider for example $R_1 = \{[a], [a, c]\}$. All a labeled transitions in the transition system enter R_1 (there is only one), all b labeled transitions exit R_1 (there are two such transitions), all e labeled transitions exit R_1 (there is only one), and all other transitions in the transition system do not cross R_1 . Hence, R_1 is a region corresponding to place p_1 with input transition a and output transitions b and e . $R_2 = \{[a], [a, b]\}$ is another region: a enters R_2 , c and e exit R_2 , and all other transitions in the transition system do not cross R_2 . R_2 is the region corresponding to place p_2 in Fig. 7.17. In the Petri net constructed based on the six minimal regions, b and c are concurrent.

Figure 7.17 shows a small process with very little concurrency. Therefore, the transition system and Petri net have similar sizes. However, for larger processes with lots of concurrency the reduction can be spectacular. The transition system modeling 10 parallel activities having $2^{10} = 1024$ states and $10 \times 2^{10-1} = 5120$ transitions, can be reduced into a Petri net with only 20 places and 10 transitions.

The transition system in Fig. 7.17 was obtained from log L_1 using state representation function $l_3^{state}()$. In fact, in this example, the discovered process model using this two-step approach is identical to the model discovered by the α -algorithm. This demonstrates that a two-step approach can be used to convert an event log into a Petri net. Therefore, process discovery using transition system construction and state-based regions is an alternative to the approaches presented thus far.

Figure 7.17 only conveys the basic idea behind regions [51]. The synthesis of Petri nets using state-based regions is actually more involved and can be customized to favor particular process patterns. As shown in [34], any finite transition system can be converted into a bisimilar Petri net, i.e., the behaviors of the transition system and Petri net are equivalent even if we consider the moment of choice (see Sect. 6.3). However, for some Petri nets it may be necessary to perform “label splitting”. As a result the Petri net may have multiple transitions referring to the same activity. This way the WF-net shown in Fig. 6.20 can be discovered. Moreover, it is also possible to enforce the resulting Petri net to have particular properties, e.g., free-choice [45]. See [165] for more information about the two-step approach.

Classical state-based regions aim at producing a Petri net that is bisimilar to the transition system. This means that while constructing the Petri net the behavior is not generalized. Therefore, it is important to select a coarser state representation function when constructing the transition system. For larger processes, a state representation function like $l_1^{state}()$ definitely results in an overfitting model that can only replay the log without any form of generalization. Many abstractions (i.e., state representation functions) are possible to balance between overfitting and underfitting. In [165], these are described systematically.

7.4.3 Process Discovery Using Language-Based Regions

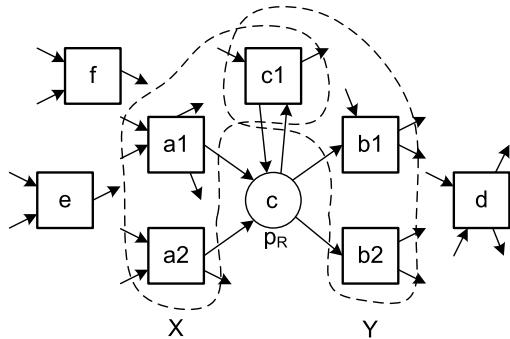
As illustrated by Fig. 7.16, the goal of state-based regions is to determine the places in a Petri net. Language-based regions also aim at finding such places but do not use a transition system as input; instead some “language” is used as input. Several techniques and variants of the problem have been defined. In this section we only present the basic idea and refer to literature for details [16, 19, 28, 170].

Suppose, we have an event log in which the events refer to a set of activities A . For this log one could construct a Petri net N_\emptyset with the set of transitions being A and no places. Since a transition without any input places is continuously enabled, this Petri net is able to reproduce the log. In fact, the Petri net N_\emptyset can reproduce any log over A . In Sect. 6.4.3, we referred to such a model as the “flower model”. There we added places and transitions to model this behavior in terms of a WF-net. However, the idea is the same. Adding places to the Petri net N_\emptyset can only limit the behavior.

Consider, for example, place p_R in Fig. 7.18. Removing place p_R will not remove any behavior. However, adding p_R may remove behavior possible in the Petri net without this place. The behavior gets restricted when a place is empty while

Fig. 7.18 Region

$R = (X, Y, c)$ corresponding to place p_R :
 $X = \{a1, a2, c1\} = \bullet p_R$,
 $Y = \{b1, b2, c1\} = p_R \bullet$, and
 c is the initial marking of p_R



one of its output transitions wants to consume a token from it. For example, $b1$ is blocked if p_R is unmarked while all other input places of $b1$ are marked. Suppose now that we have a set of traces L . If these traces are possible in the net with place p_R , then they are also possible in the net without p_R . The reverse does not always hold. This triggers the question whether p_R can be added without disabling any of the traces in L . This is what regions are all about.

Definition 7.4 (Language-based region) Let $L \in \mathbb{B}(\mathcal{A}^*)$ be a simple event log. $R = (X, Y, c)$ is a *region* of L if and only if:

- $X \subseteq \mathcal{A}$ is the set of input transitions of R ;
- $Y \subseteq \mathcal{A}$ is the set of output transitions of R ;
- $c \in \{0, 1\}$ is the initial marking of R ; and
- For any $\sigma \in L$, $k \in \{1, \dots, |\sigma|\}$, $\sigma_1 = hd^{k-1}(\sigma)$, $a = \sigma(k)$, $\sigma_2 = hd^k(\sigma) = \sigma_1 \oplus a$:

$$c + \sum_{t \in X} \partial_{multiset}(\sigma_1)(t) - \sum_{t \in Y} \partial_{multiset}(\sigma_2)(t) \geq 0$$

$R = (X, Y, c)$ is a region of L if and only if inserting a place p_R with $\bullet p_R = A$, $p_R \bullet = B$, and initially c tokens does not disable the execution of any of the traces in L . To check this, Definition 7.4 inspects all events in the event log. Let $\sigma \in L$ be a trace in the log. $a = \sigma(k)$ is the k -th event in this trace. This event should not be disabled by place p_R . Therefore, we calculate the number of tokens $M(p_R)$ that are in this place just before the occurrence of the k -th event.

$$M(p_R) = c + \sum_{t \in X} \partial_{multiset}(\sigma_1)(t) - \sum_{t \in Y} \partial_{multiset}(\sigma_1)(t)$$

$\sigma_1 = hd^{k-1}(\sigma)$ is the partial trace of events that occurred before the occurrence of the k -th event. $\partial_{multiset}(\sigma_1)$ converts this partial trace into a multi-set. $\partial_{multiset}(\sigma_1)$ is also known as the *Parikh vector* of σ_1 . $\sum_{t \in X} \partial_{multiset}(\sigma_1)(t)$ counts the number of tokens produced for place p_R , $\sum_{t \in Y} \partial_{multiset}(\sigma_1)(t)$ counts the number of tokens consumed from this place, and c is the initial number of tokens in p_R . Therefore,

$M(p_R)$ is indeed the number of tokens in p_R just before the occurrence of the k -th event. This number should be positive. In fact, there should be at least one token in p_R if $a \in Y$. In other words, $M(p_R)$ minus the number of tokens consumed from p_R by the k -th event should be non-negative. Hence

$$M(p_R) - \sum_{t \in Y} \partial_{multiset}(\{a\})(t) = c + \sum_{t \in X} \partial_{multiset}(\sigma_1)(t) - \sum_{t \in Y} \partial_{multiset}(\sigma_2)(t) \geq 0$$

This shows that a region R , according to Definition 7.4, indeed corresponds to a so-called *feasible place* p_R , i.e., a place that can be added without disabling any of the traces in the event log.

The requirement stated in Definition 7.4 can also be formulated in terms of an inequation system. To illustrate this we use an example log from Chap. 6,

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

There are five activities. For each activity t we introduce two variables, x_t and y_t . $x_t = 1$ if transition t produces a token for p_R and $x_t = 0$ if not. $y_t = 1$ if transition t consumes a token from p_R and $y_t = 0$ if not. A potential region $R = (X, Y, c)$ corresponds to an assignment for all of these variables: $x_t = 1$ if $t \in X$, $x_t = 0$ if $t \notin X$, $y_t = 1$ if $t \in Y$, $y_t = 0$ if $t \notin Y$. The requirement stated in Definition 7.4 can now be reformulated in terms of the variables $x_a, x_b, x_c, x_d, x_e, y_a, y_b, y_c, y_d, y_e$, and c :

$$\begin{aligned} c - y_a &\geq 0 \\ c + x_a - (y_a + y_c) &\geq 0 \\ c + x_a + x_c - (y_a + y_c + y_d) &\geq 0 \\ c - y_b &\geq 0 \\ c + x_b - (y_b + y_c) &\geq 0 \\ c + x_b + x_c - (y_b + y_c + y_e) &\geq 0 \\ c, x_a, \dots, x_e, y_a, \dots, y_e &\in \{0, 1\} \end{aligned}$$

Note that these inequations are based on all non-empty prefixes of $\langle a, c, d \rangle$ and $\langle b, c, e \rangle$. Any solution of this linear inequation system corresponds to a region. Some example solutions are:

$$R_1 = (\emptyset, \{a, b\}, 1)$$

$$c = y_a = y_b = 1, \quad x_a = x_b = x_c = x_d = x_e = y_c = y_d = y_e = 0$$

$$R_2 = (\{a, b\}, \{c\}, 0)$$

$$x_a = x_b = y_c = 1, \quad c = x_c = x_d = x_e = y_a = y_b = y_d = y_e = 0$$

$$R_3 = (\{c\}, \{d, e\}, 0)$$

$$x_c = y_d = y_e = 1, \quad c = x_a = x_b = x_d = x_e = y_a = y_b = y_c = 0$$

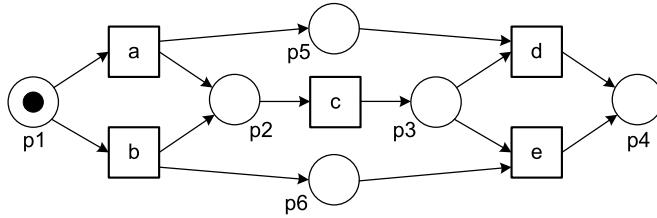


Fig. 7.19 WF-net constructed using regions R_1, \dots, R_6 : p_1 corresponds to $R_1 = (\emptyset, \{a, b\}, 1)$, p_2 corresponds to $R_2 = (\{a, b\}, \{c\}, 0)$, etc.

$$R_4 = (\{d, e\}, \emptyset, 0)$$

$$x_d = x_e = 1, \quad c = x_a = x_b = x_c = y_a = y_b = y_c = y_d = y_e = 0$$

$$R_5 = (\{a\}, \{d\}, 0)$$

$$x_a = y_d = 1, \quad c = x_b = x_c = x_d = x_e = y_a = y_b = y_c = y_e = 0$$

$$R_6 = (\{b\}, \{e\}, 0)$$

$$x_b = y_e = 1, \quad c = x_a = x_c = x_d = x_e = y_a = y_b = y_c = y_d = 0$$

Consider, for example, $R_6 = (\{b\}, \{e\}, 0)$. This corresponds to the solution $x_b = y_e = 1$ and $c = x_a = x_c = x_d = x_e = y_a = y_b = y_c = y_d = 0$. If we fill out the values in the inequation system, we can see that this is indeed a solution. If we construct a Petri net based on these six regions, we obtain the WF-net shown in Fig. 7.19.

Suppose that the trace $\langle a, c, e \rangle$ is added to event log L_9 . This results in three additional inequations:

$$c - y_a \geq 0$$

$$c + x_a - (y_a + y_c) \geq 0$$

$$c + x_a + x_c - (y_a + y_c + y_e) \geq 0$$

Only the last inequation is new. Because of this inequation, $x_b = y_e = 1$ and $c = x_a = x_c = x_d = x_e = y_a = y_b = y_c = y_d = 0$ is no longer a solution. Hence, $R_6 = (\{b\}, \{e\}, 0)$ is not a region anymore and place p_6 needs to be removed from the WF-net shown in Fig. 7.19. After removing this place, the resulting WF-net indeed allows for $\langle a, c, e \rangle$.

One of the problems of directly applying language-based regions is that the linear inequation system has many solutions. Few of these solutions correspond to sensible places. For example, $x_a = x_b = y_d = y_e = 1$ and $c = x_c = x_d = x_e = y_a = y_b = y_c = 0$ also defines a region: $R_7 = (\{a, b\}, \{d, e\}, 0)$. However, adding this place to Fig. 7.19 would only clutter the diagram. Another example is $c = x_a = x_b = y_c = 1$ and $x_c = x_d = x_e = y_a = y_b = y_d = y_e = 0$, i.e., region $R_8 = (\{a, b\}, \{c\}, 1)$. This region is a weaker variant R_2 as the place is initially marked.

Another problem is that classical techniques for language-based regions aim at a Petri net that does not allow for any behavior not seen in the log [19]. This means that the log is considered to be complete. As shown before, this is very unrealistic and results in models that are complex and overfitting. To address these problems dedicated techniques have been proposed. For instance, in [170] it is shown how to avoid overfitting and how to ensure that the resulting model has desirable properties (WF-net, free-choice, etc.). Nevertheless, pure region-based techniques tend to have problems handling noise and incompleteness. Therefore, combinations of heuristic mining and region-based techniques seem more suitable for practical applications.

7.5 Inductive Mining

A range of *inductive process discovery* techniques exist for the *process trees* introduced in Sect. 3.2.8 [88–91]. Whereas Petri nets, WF-nets, BPMN models, EPCs, YAWL models, and UML activity diagrams may suffer from deadlocks, livelocks, and other anomalies, process trees are *sound by construction*. This section introduces the basic inductive mining approach. The inductive mining framework is highly extendible and allows for many variants of the basic approach. The “family” of inductive mining techniques includes members that can handle infrequent behavior and deal with huge models and logs while ensuring formal correctness criteria such as the ability to rediscover the original model (in the limit). The results returned by these techniques can easily be converted to other notations ranging from Petri nets and BPMN models to process calculi and statecharts. Inductive mining is currently one of the leading process discovery approaches due to its flexibility, formal guarantees and scalability.

7.5.1 Inductive Miner Based on Event Log Splitting

Given a simple event log $L \in \mathbb{B}(A^*)$ (i.e., a multi-set of traces over some set of activities A) we would like to discover a process tree $Q \in \mathcal{Q}_A$. Consider, for example, event log $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$ consisting of 6 cases and 23 events. The α -algorithm creates the WF-net $N_1 = \alpha(L_1)$ shown in Fig. 7.20 (left-hand side). The basic *Inductive Miner* (IM) will produce the equivalent process tree $Q_1 = \text{IM}(L_1) = \rightarrow(a, \times(\wedge(b, c), e), d)$ also shown in Fig. 7.20 (right-hand side). The process tree can be automatically converted into the WF-net produced by the α -algorithm using the conversion shown in Fig. 3.18 followed by a reduction removing superfluous silent transitions. Any process tree can be converted to an equivalent WF-net, BPMN model, etc. Moreover, the basic Inductive Miner (IM) can discover a much wider class of processes and learn “correct” process models in situations where the α -algorithm and many other algorithms fail.

We use several simple examples to explain the approach. For clarity we first assume that there are no duplicate or silent activities, i.e., in the process trees used

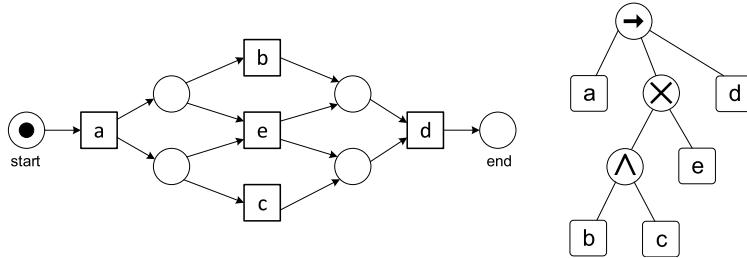


Fig. 7.20 WF-net N_1 (left) and process tree Q_1 (right) discovered for $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$

to generate the example logs there are no two leaves with the same activity label and no leaves with a τ label. Later we relax this assumption.

The basic IM algorithm uses the *directly-follows graph* that corresponds to the “directly follows” relation ($>_L$) used by the α -algorithm. Other elements such as “eventually follows” or the dependency measures used for the dependency graph can also be used in the Inductive Miner (IM) framework. Frequencies provide important information for process discovery. However, we start by looking at ingredients similar to the ones used by the α -algorithm: the directly follows relation, start activities, and end activities.

Definition 7.5 (Directly-follows graph) Let L be an event log, i.e., $L \in \mathbb{B}(\mathcal{A}^*)$. The *directly-follows graph* of L is $G(L) = (A_L, \mapsto_L, A_L^{start}, A_L^{end})$ with:

- $A_L = \{a \in \sigma \mid \sigma \in L\}$ is the set of activities in L ,
- $\mapsto_L = \{(a, b) \in A \times A \mid a >_L b\}$ is the directly follows relation,³
- $A_L^{start} = \{a \in A \mid \exists_{\sigma \in L} a = \text{first}(\sigma)\}$ is the set of start activities, and
- $A_L^{end} = \{a \in A \mid \exists_{\sigma \in L} a = \text{last}(\sigma)\}$ is the set of end activities.

The IM algorithm iteratively splits the initial event log into smaller *sublogs*. For any sublog L we can create a directly-follows graph $G(L)$. $a \mapsto_L b$ if a was directly followed by b somewhere in L . $a \not\mapsto_L b$ if a was never directly followed by b . \mapsto_L^+ is the transitive closure of \mapsto_L . $a \mapsto_L^+ b$ if there is a non-empty path from a to b in $G(L)$, i.e., there exists a sequence of activities a_1, a_2, \dots, a_k such that $k \geq 2$, $a_1 = a$ and $a_k = b$ and $a_i \mapsto_L a_{i+1}$ for $i \in \{1, \dots, k-1\}$. $a \not\mapsto_L^+ b$ if there is no path from a to b in the directly-follows graph.

To understand how the IM algorithm learns $Q_1 = \text{IM}(L_1) = \rightarrow(a, \times(\wedge(b, c), e), d)$ shown in Fig. 7.20 from event log $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$, consider Fig. 7.21 and Fig. 7.22. $G_1 = G(L_1)$ in Fig. 7.21 is the directly-follows graph of L_1 . Note that $a \mapsto_{L_1} b$ because of the arc between a and b in G_1 . $a \not\mapsto_{L_1} d$ and $a \mapsto_{L_1}^+ d$ because there is a path from a to d , but no arc between a and d .

³ $a >_L b$ if and only if there is a trace $\sigma = \langle t_1, t_2, t_3, \dots, t_n \rangle$ and $i \in \{1, \dots, n-1\}$ such that $\sigma \in L$ and $t_i = a$ and $t_{i+1} = b$ (see Definition 6.3).

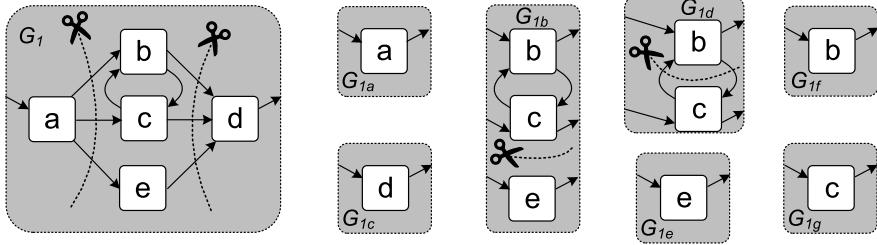
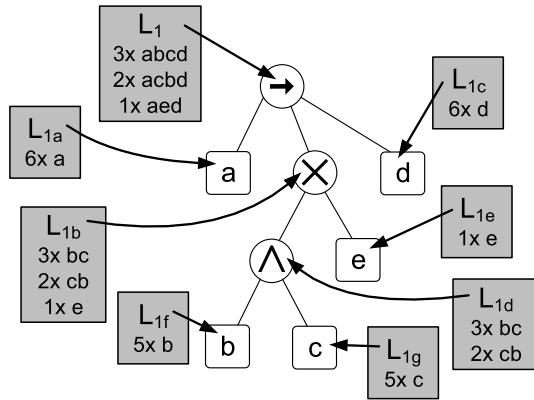


Fig. 7.21 G_1 is the directly-follows graph for $L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$. The event log is recursively cut into smaller sublogs using the directly-follows graphs of these sublogs

Fig. 7.22 The different sublogs created when learning process tree
 $Q_1 = \rightarrow(a, \times(\wedge(b, c), e), d)$
for $L_1 = [\langle a, b, c, d \rangle^3,$
 $\langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$



$A_{L_1}^{start} = \{a\}$ as denoted by the incoming arc. $A_{L_1}^{end} = \{d\}$ as denoted by the outgoing arc.

We would like to split L_1 recursively until we find sublogs of the form $[\langle x \rangle^k]$, i.e., sublogs corresponding to the execution of activity x . To find out how to split the event log in each step, we try to find so-called *cuts* in the directly-follows graph of the (sub)log we would like to split. We consider *exclusive-choice cuts*, *sequence cuts*, *parallel cuts*, and *redo-loop cuts* corresponding to the four process tree operators (\times , \rightarrow , \wedge , and \odot).

Directly-follows graph $G_1 = G(L_1)$ in Fig. 7.21 is cut into three smaller directly-follows graphs (G_{1a} , G_{1b} , and G_{1c}) using *sequence cut* ($\rightarrow, \{a\}, \{b, c, e\}, \{d\}$). The cut splits the set of activities in three disjoint subsets such that arcs only go from left to right. The formal definition of a sequence cut is given later. Based on the sequence cut, event log L_1 is split into $L_{1a} = [\langle a \rangle^6]$, $L_{1b} = [\langle b, c \rangle^3, \langle c, b \rangle^2, \langle e \rangle]$, and $L_{1c} = [\langle d \rangle^6]$. These sublogs are created by projecting the original event log on the three disjoint subsets of activities in cut ($\rightarrow, \{a\}, \{b, c, e\}, \{d\}$). Note that each event in L_1 ends up in precisely one of the sublogs. Using the three sublogs, we create three new directly-follows graphs, $G_{1a} = G(L_{1a})$, $G_{1b} = G(L_{1b})$, and $G_{1c} = G(L_{1c})$. These graphs are shown in Fig. 7.21. $G_{1a} = G(L_{1a})$ and $G_{1c} = G(L_{1c})$ represent

base cases, i.e., subprocesses consisting of a single activity executed once per case. G_{1b} is not a base case and needs to be split further.

$\text{IM}(L_1) = \rightarrow(\text{IM}(L_{1a}), \text{IM}(L_{1b}), \text{IM}(L_{1c}))$ because of the sequence cut. Since $\text{IM}(L_{1a}) = a$ and $\text{IM}(L_{1c}) = d$, this can be rewritten as $\text{IM}(L_1) = \rightarrow(a, \text{IM}(L_{1b}), d)$. Next we compute $\text{IM}(L_{1b})$ using $G_{1b} = G(L_{1b})$ in Fig. 7.21. Directly-follows graph $G_{1b} = G(L_{1b})$ is cut into two smaller directly-follows graphs using *exclusive-choice cut* ($\times, \{b, c\}, \{e\}$). The exclusive-choice cut splits the set of activities in two disjoint subsets such that there are no arcs going from one set to the other (and vice versa). Based on the sequence cut, event log L_{1b} is split into $L_{1d} = [\langle b, c \rangle^3, \langle c, b \rangle^2]$ and $L_{1e} = [\langle e \rangle]$. Again each event ends up in precisely one of the sublogs. However, because of the nature of the exclusive-choice cut, we partitioned the traces based on the activities they contain rather than projecting trace on disjoint activity sets.

$\text{IM}(L_{1b}) = \times(\text{IM}(L_{1d}), \text{IM}(L_{1e}))$ because of the exclusive-choice cut. $\text{IM}(L_{1e}) = e$ corresponds again to the base case: In each trace in the sublog, activity d is executed once (compare G_{1e} with G_{1a} and G_{1c}). It remains to compute $\text{IM}(L_{1d})$. Figure 7.21 shows directly-follows graph $G_{1d} = G(L_{1d})$ which is cut into two smaller directly-follows graphs using *parallel cut* ($\wedge, \{b\}, \{c\}$). The parallel cut splits the set of activities in two disjoint subsets such that every activity in one set is connected to all activities in the other set (and vice versa). Based on the parallel cut, event log L_{1d} is split into $L_{1f} = [\langle b \rangle^5]$ and $L_{1g} = [\langle c \rangle^5]$. $G_{1f} = G(L_{1f})$ and $G_{1g} = G(L_{1g})$ are shown in Fig. 7.21 and correspond to the base case. Hence, $\text{IM}(L_{1f}) = b$ and $\text{IM}(L_{1g}) = c$. Therefore, $\text{IM}(L_{1d}) = \wedge(b, c)$, $\text{IM}(L_{1b}) = \times(\wedge(b, c), e)$, $\text{IM}(L_1) = \rightarrow(a, \times(\wedge(b, c), e), d)$.

Process tree $Q_1 = \text{IM}(L_1) = \rightarrow(a, \times(\wedge(b, c), e), d)$ was computed by recursively applying a sequence cut, an exclusive-choice cut, and a parallel cut. Figure 7.22 shows the process tree and the sublogs created during discovery. The leaves correspond to base cases. The inner nodes correspond to operators used to cut the event log in sublogs.

Definition 7.6 (Cut) Let L be an event log with corresponding directly-follows graph $G(L) = (A_L, \mapsto_L, A_L^{start}, A_L^{end})$. Let $n \geq 1$. An n -ary cut of $G(L)$ is a partition of A_L into pairwise disjoint sets A_1, A_2, \dots, A_n : $A_L = \bigcup_{i \in \{1, \dots, n\}} A_i$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. Notation is $(\oplus, A_1, A_2, \dots, A_n)$ with $\oplus \in \{\rightarrow, \times, \wedge, \circlearrowright\}$. For each type of operator ($\rightarrow, \times, \wedge$, and \circlearrowright) specific conditions apply:

- An *exclusive-choice cut* of $G(L)$ is a cut $(\times, A_1, A_2, \dots, A_n)$ such that
 - $\forall_{i,j \in \{1, \dots, n\}} \forall_{a \in A_i} \forall_{b \in A_j} i \neq j \Rightarrow a \not\mapsto_L b$.
- A *sequence cut* of $G(L)$ is a cut $(\rightarrow, A_1, A_2, \dots, A_n)$ such that
 - $\forall_{i,j \in \{1, \dots, n\}} \forall_{a \in A_i} \forall_{b \in A_j} i < j \Rightarrow (a \mapsto_L^+ b \wedge b \not\mapsto_L^+ a)$.
- A *parallel cut* of $G(L)$ is a cut $(\wedge, A_1, A_2, \dots, A_n)$ such that
 - $\forall_{i \in \{1, \dots, n\}} A_i \cap A_L^{start} \neq \emptyset \wedge A_i \cap A_L^{end} \neq \emptyset$ and
 - $\forall_{i,j \in \{1, \dots, n\}} \forall_{a \in A_i} \forall_{b \in A_j} i \neq j \Rightarrow a \mapsto_L b$.

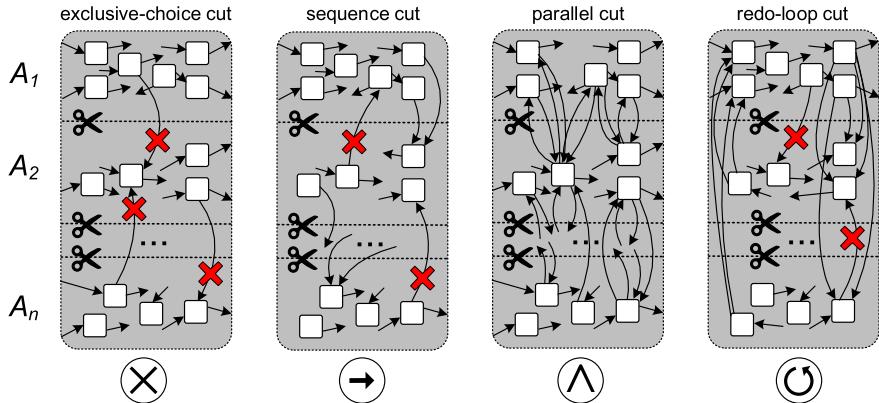


Fig. 7.23 Four types of cuts, $(\oplus, A_1, A_2, \dots, A_n)$ with $\oplus \in \{\times, \rightarrow, \wedge, \circlearrowright\}$

- A *redo-loop cut* of $G(L)$ is a cut $(\circlearrowright, A_1, A_2, \dots, A_n)$ such that

- $n \geq 2$,
- $A_L^{start} \cup A_L^{end} \subseteq A_1$,
- $\{a \in A_1 \mid \exists_{i \in \{2, \dots, n\}} \exists_{b \in A_i} a \mapsto_L b\} \subseteq A_L^{end}$,
- $\{a \in A_1 \mid \exists_{i \in \{2, \dots, n\}} \exists_{b \in A_i} b \mapsto_L a\} \subseteq A_L^{start}$,
- $\forall_{i, j \in \{2, \dots, n\}} \forall_{a \in A_i} \forall_{b \in A_j} i \neq j \Rightarrow a \not\mapsto_L b$,
- $\forall_{i \in \{2, \dots, n\}} \forall_{b \in A_i} \exists_{a \in A_L^{end}} a \mapsto_L b \Rightarrow \forall_{a' \in A_L^{end}} a' \mapsto_L b$, and
- $\forall_{i \in \{2, \dots, n\}} \forall_{b \in A_i} \exists_{a \in A_L^{start}} b \mapsto_L a \Rightarrow \forall_{a' \in A_L^{start}} b \mapsto_L a'$.

A cut $(\oplus, A_1, A_2, \dots, A_n)$ with $\oplus \in \{\rightarrow, \times, \wedge, \circlearrowright\}$ of directly-follows graph $G(L)$ is *maximal* if there is no cut $(\oplus, A'_1, A'_2, \dots, A'_m)$ with $m > n$. Cut $(\oplus, A_1, A_2, \dots, A_n)$ is called *trivial* if $n = 1$.

The four types of cuts are illustrated in Fig. 7.23. These are based on the characteristics of the four process tree operators assuming that there are no duplicate or silent activities. Definition 3.14 describes the semantics of operator $\oplus \in \{\times, \rightarrow, \wedge, \circlearrowright\}$. It is easy to verify that these operators indeed leave the “fingerprints” shown in Fig. 7.23.

Consider four simple process trees, $Q_{ab} = \times(a, b)$, $Q_{cd} = \rightarrow(c, d)$, $Q_{ef} = \wedge(e, f)$, and $Q_{gh} = \circlearrowright(g, h)$. $\mathcal{L}(Q_{ab}) = \{\langle a \rangle, \langle b \rangle\}$, $\mathcal{L}(Q_{cd}) = \{\langle c, d \rangle\}$, $\mathcal{L}(Q_{ef}) = \{\langle e, f \rangle, \langle f, e \rangle\}$, $\mathcal{L}(Q_{gh}) = \{\langle g \rangle, \langle g, h, g \rangle, \langle g, h, g, h, g \rangle, \dots\}$.

Consider now the directly-follows graph of an event log L generated by $\times(Q_{ab}, Q_{cd}, Q_{ef}, Q_{gh})$ that is *directly-follows complete*. An event log L generated from a process tree is directly-follows complete if directly-follows graph $G(L)$ is maximal, i.e., all activities, all start activities, all end activities, and all possible direct successions have been observed. Clearly, the exclusive-choice cut ($\times, \{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}$) meets the requirement stated in Definition 7.6 for any directly-follows complete log. For example, $a \not\mapsto_L c$, $d \not\mapsto_L h$, etc. The activities in the pairwise disjoint activity sets never follow one another directly.

Next, we consider the directly-follows graph of an event log L generated by $\rightarrow(Q_{ab}, Q_{cd}, Q_{ef}, Q_{gh})$ that is directly-follows complete. The sequence cut ($\rightarrow, \{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}$) meets the requirements stated in Definition 7.6. For example, $a \mapsto_L^+ c, c \not\mapsto_L^+ a, a \mapsto_L^+ e, e \not\mapsto_L^+ a$, etc. Activities in different subsets need to be strictly ordered to apply this cut.

The directly-follows graph of a directly-follows complete event log L generated by $\wedge(Q_{ab}, Q_{cd}, Q_{ef}, Q_{gh})$ allows for parallel cut ($\wedge, \{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}$). The first requirement stated in Definition 7.6 ensures that each of the activity subsets has at least one start and one end activity. The second requirement states that any two activities in different subsets can directly follow one another (e.g., $a \mapsto_L c, c \mapsto_L a, a \mapsto_L e, e \mapsto_L a$, etc.). Both requirements are satisfied by the nature of parallel composition.

The directly-follows graph of a directly-follows complete event log L generated by $\circlearrowleft(Q_{ab}, Q_{cd}, Q_{ef}, Q_{gh})$ allows for redo-loop cut ($\circlearrowleft, \{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}$). Each of the seven requirements in Definition 7.6 is satisfied. All start and end activities are in the “do part” of the redo-loop, i.e., $A_L^{start} \cup A_L^{end} = \{a, b\} \cup \{a, b\} \subseteq \{a, b\}$. All connections run via a and b , e.g., $a \mapsto_L c, d \mapsto_L a, b \mapsto_L c, d \mapsto_L b$, etc. In a redo loop, the directly-follows graph must contain a clear set of start and end activities. All connections between the different child nodes must go through these activities.

The IM algorithm works as follows. Given an event log, the directly-follows graph is constructed. If there is a non-trivial exclusive-choice cut, then a maximal exclusive-choice cut is applied splitting the event log into smaller event logs. If there is no non-trivial exclusive-choice cut, but there is a non-trivial sequence cut, then a maximal sequence cut is applied splitting the event log into smaller event logs. If there are no non-trivial exclusive-choice and sequence cuts, but there is a non-trivial parallel cut, then a maximal parallel cut is applied splitting the event log into smaller event logs. If there are no non-trivial exclusive-choice, sequence and parallel cuts, but there is a redo-loop cut, then a maximal redo-loop cut is applied splitting the event log into smaller event logs. After splitting the event log into sublogs the procedure is repeated until a base case (sublog with only one activity) is reached.

How the event log is split into sublogs depends on the operator. In case of an exclusive-choice cut, the traces are split as a whole. In case of a sequence cut and parallel cut, the traces are projected on the respective sets of activities, i.e., each sublog has a projected trace for each trace in the log that needs to be split. In case of a redo-loop cut, the loops are unfolded and a trace is created for every iteration. Empty traces are handled in a dedicated manner (based on the operator) and result in the insertion of τ activities. For example, exclusive choice cut ($\times, A_1, A_2, \dots, A_n$) may result in $\times(Q_1, Q_2, \dots, Q_n, \tau)$ if there are empty traces in the log to be split.

If there are no non-trivial cuts meeting the requirements in Definition 7.6, a *fall-through* is selected. The part that cannot be split is presented by a so-called *flower model* (“anything can happen”), similar to the one introduced in Sect. 6.4.3. For example, the flower model in Fig. 6.23 can be represented as process tree $\circlearrowleft(\tau, a, b, c, d, e, f, g, h)$ allowing for any trace involving activities $a-h$.

Suppose the sublog L' for which no cut is applicable contains activities $\{a_1, a_2, \dots, a_m\}$. Fall-through IM(L') = $\circlearrowleft(\tau, a_1, a_2, \dots, a_m)$ is selected, i.e., the subpro-

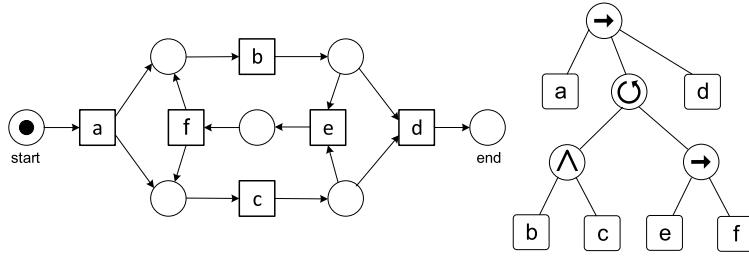


Fig. 7.24 WF-net N_2 (left) and process tree Q_2 (right) discovered for $L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \langle a, c, b, e, f, b, c, d \rangle, \langle e, f, c, b, d \rangle]$

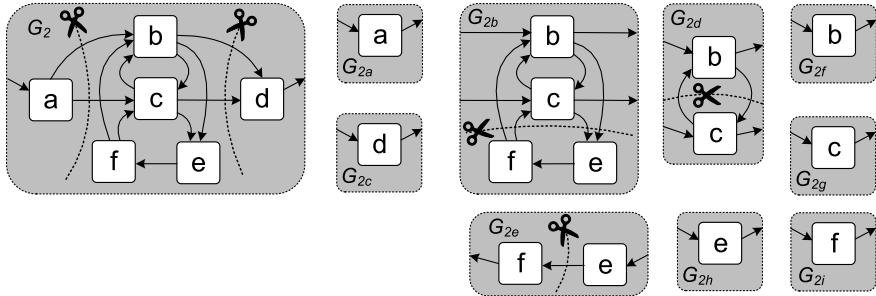


Fig. 7.25 G_2 is the directly-follows graph for L_2 . The other directly-follows graphs correspond to the various sublogs

cess is represented by a subtree that allows for any behavior involving activities $\{a_1, a_2, \dots, a_m\}$. The fall-through serves as a last resort ensuring fitness, but possibly resulting in lower precision.

In the base case, the sublog L' contains only events corresponding to a particular activity, say a . If the sublog is of the form $L' = [\langle a \rangle^k]$ with $k \geq 1$ (i.e., a occurs once in each trace), then $IM(L') = a$. If the sublog is of the form $L' = [\langle \cdot \rangle^k, \langle a \rangle^l]$ with $k, l \geq 1$, then $IM(L') = \times(a, \tau)$ because a is sometimes skipped. If a is executed at least once in each trace in the sublog and sometimes multiple times (e.g., $L' = [\langle a \rangle^9, \langle a, a \rangle^2, \langle a, a, a \rangle]$), then $IM(L') = \circlearrowleft(a, \tau)$. In all other cases (e.g., $L' = [\langle \cdot \rangle^3, \langle a \rangle^4, \langle a, a, a \rangle]$), $IM(L') = \circlearrowright(\tau, a)$ because a is executed zero or more times in the traces of sublog L .

To illustrate the IM algorithm better, we consider a slightly larger event log,

$$L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \langle a, b, c, e, f, c, b, d \rangle]$$

This log could have been generated by the WF-net N_2 or process tree $Q_2 = \rightarrow(a, \circlearrowleft(\wedge(b, c), \rightarrow(e, f))), d$ both shown in Fig 7.24.

The IM algorithm starts by creating the directly-follows graph $G_2 = G(L_2)$ for event log L_2 (see Fig. 7.25). Since there is no non-trivial exclusive-choice cut, we

try a sequence cut. The maximal sequence cut ($\rightarrow, \{a\}, \{b, c, e, f\}, \{d\}$) is shown in Fig. 7.25. Based on this cut, three sublogs are created:

$$L_{2a} = [\langle a \rangle^{13}]$$

$$L_{2b} = [\langle b, c \rangle^3, \langle c, b \rangle^4, \langle b, c, e, f, b, c \rangle^2, \langle c, b, e, f, b, c \rangle^2, \langle b, c, e, f, c, b \rangle, \langle c, b, e, f, b, c, e, f, c, b \rangle]$$

$$L_{2c} = [\langle d \rangle^{13}]$$

Event log L_{2a} (L_{2c}) has directly-follows graph $G_{2a} = G(L_{2a})$ ($G_{2c} = G(L_{2c})$). Both correspond to base cases and are represented by subtrees $\text{IM}(L_{2a}) = a$ and $\text{IM}(L_{2c}) = d$. $G_{2b} = G(L_{2b})$ in Fig. 7.25 is the directly-follows graph for sublog L_{2b} . There are no non-trivial exclusive-choice, sequence or parallel cuts. Therefore, we apply the maximal redo-loop cut ($\circlearrowleft, \{b, c\}, \{e, f\}$) shown in Fig. 7.25. Note that all seven requirements stated in Definition 7.6 are satisfied. Using the redo-loop cut, sublog L_{2b} is split into two smaller sublogs, $L_{2d} = [\langle b, c \rangle^{11}, \langle c, b \rangle^9]$ and $L_{2e} = [\langle e, f \rangle^7]$. Note that some traces in L_{2b} correspond to multiple traces in L_{2d} and L_{2e} . Consider, for example, $\langle c, b, e, f, b, c, e, f, c, b \rangle \in L_{2b}$ which is split into five smaller traces ($\langle c, b \rangle, \langle e, f \rangle, \langle b, c \rangle, \langle e, f \rangle$, and $\langle c, b \rangle$) distributed over L_{2d} and L_{2e} . Subsequently, the IM algorithm selects the parallel cut ($\wedge, \{b\}, \{c\}$) in G_{2d} , the directly-follows graph created for L_{2d} . The resulting sublogs $L_{2f} = [\langle b \rangle^{20}]$ and $L_{2g} = [\langle c \rangle^{20}]$ correspond to base cases. Hence, $\text{IM}(L_{2f}) = b$, $\text{IM}(L_{2g}) = c$, and $\text{IM}(L_{2d}) = \wedge(b, c)$. G_{2e} is the directly-follows graph created for L_{2e} . The IM algorithm selects the sequence cut ($\rightarrow, \{e\}, \{f\}$). The resulting sublogs $L_{2h} = [\langle e \rangle^7]$ and $L_{2i} = [\langle f \rangle^7]$ correspond to base cases. Hence, $\text{IM}(L_{2h}) = e$, $\text{IM}(L_{2i}) = f$, and $\text{IM}(L_{2e}) = \rightarrow(e, f)$. $\text{IM}(L_{2b}) = \circlearrowleft(\wedge(b, c), \rightarrow(e, f))$. By combining the results for the subtrees, we find $Q_2 = \text{IM}(L_2) = \rightarrow(a, \circlearrowleft(\wedge(b, c), \rightarrow(e, f)), d)$.

Finally, we revisit the example from Chap. 2 (cf. Table 2.2). The IM algorithm is able to discover the same model as the α -algorithm. For any directly-follows complete event log generated by the WF-net in Fig. 7.27, the IM algorithm discovers a process tree equivalent to $\rightarrow(a, \circlearrowleft(\rightarrow(\wedge(\times(b, c), d), e), f), \times(g, h))$.

7.5.2 Characteristics of the Inductive Miner

The IM algorithm returns a specific process tree. However, there may be several process trees having a same behavior. For example, $\times(a, b, c)$ and $\times(c, \times(b, a))$ are indistinguishable, i.e., $\mathcal{L}(\times(a, b, c)) = \mathcal{L}(\times(c, \times(b, a))) = \{\langle a \rangle, \langle b \rangle, \langle c \rangle\}$. $\wedge(a, b)$ and $\wedge(b, a, \tau)$ are also indistinguishable, i.e., $\mathcal{L}(\wedge(a, b)) = \mathcal{L}(\wedge(b, a, \tau)) = \{\langle a, b \rangle, \langle b, a \rangle\}$. We consider two process trees Q_1 and Q_2 to be *equivalent* if $\mathcal{L}(Q_1) = \mathcal{L}(Q_2)$ (i.e., trace equivalence).

The ordering of the children of a parallel or exclusive choice operator does not matter. We assume the IM algorithm to be deterministic and pick a particular order. Therefore, we can write $\text{IM}(L)$.

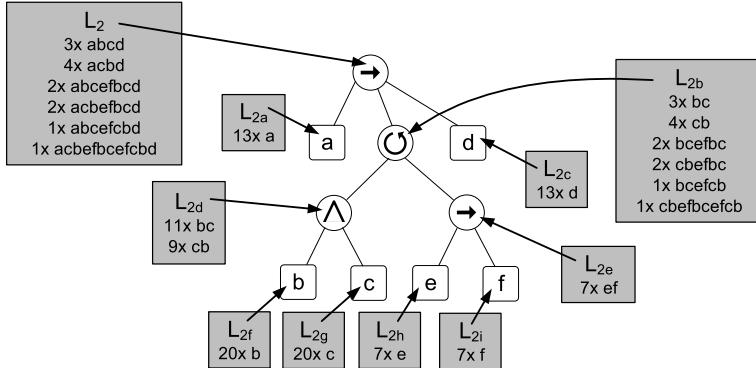


Fig. 7.26 The different sublogs created when learning process tree $Q_2 = \rightarrow(a, \circlearrowleft(\wedge(b, c), \rightarrow(e, f)), d)$ for $L_2 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^4, \langle a, b, c, e, f, b, c, d \rangle^2, \langle a, c, b, e, f, b, c, d \rangle^2, \langle a, b, c, e, f, c, b, d \rangle, \langle a, c, b, e, f, b, c, e, f, c, b, d \rangle]$

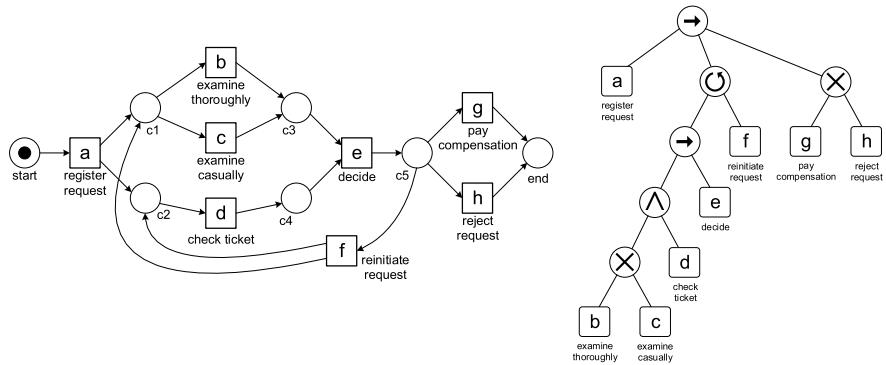


Fig. 7.27 The process model discovered by the α -algorithm for event log $[(a, b, d, e, h), (a, d, c, e, g), (a, c, d, e, f, b, d, e, g), (a, d, b, e, h), (a, c, d, e, f, d, c, e, f, c, d, e, h), (a, c, d, e, g)]$ and the corresponding process tree $\rightarrow(a, \circlearrowleft(\rightarrow(\wedge(\times(b, c), d), e), f), \times(g, h))$

A process tree Q is called *language-rediscoverable* by the IM algorithm if for any directly-follows complete event log L generated from Q , $\mathcal{L}(\text{IM}(L)) = \mathcal{L}(Q)$. Recall that an event log L is *directly-follows complete* for process tree Q if the directly-follows graph $G(L)$ is *maximal*, i.e., all activities in Q appear in the log, for every start (end) activity in Q there is a trace starting (ending) with it in the log, and $a \mapsto_L b$ if b can directly follow a in Q .

In [88], it is shown that almost all process trees *without duplicate and silent activities* are *language-rediscoverable* using the basic algorithm described in Sect. 7.5.1. The only exception is the situation where both a redo-loop cut and parallel cut are possible. An example is shown in Fig. 7.28. $Q_{rd} = \circlearrowleft(\wedge(a, b), \wedge(c, d))$ and $Q_{par} = \wedge(\circlearrowleft(a, c), \circlearrowleft(b, d))$ are not trace equivalent, but have the same directly-follows graph for any directly-follows complete event log. Since the IM algorithm

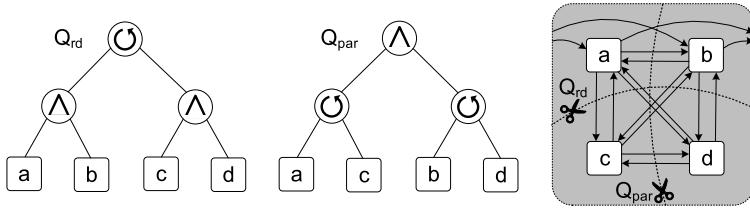


Fig. 7.28 Two process trees $Q_{rd} = \bigcirc(\wedge(a, b), \wedge(c, d))$ and $Q_{par} = \wedge(\bigcirc(a, c), \bigcirc(b, d))$ having different behaviors but identical directly-follows graphs: redo-loop cut ($\bigcirc, \{a, b\}, \{c, d\}$) and parallel cut ($\wedge, \{a, c\}, \{b, d\}$) are both possible

only uses the directly-follows graph, it cannot distinguish both processes. Fortunately, this is a rather peculiar situation and still does not jeopardize fitness. In almost all cases, the directly-follows graph is informative enough. For example, if the start and end activities in the “do part” of the loop are disjoint, then language-rediscoverability is guaranteed [88].

Note that $\mathcal{L}(Q_{rd}) \subset \mathcal{L}(Q_{par})$. The IM algorithm selects the parallel cut ($\wedge, \{a, c\}, \{b, d\}$) and returns the more general Q_{par} that also allows for traces like $\langle a, b, c, a \rangle$ and $\langle a, c, a, b, d, b \rangle$ not possible in Q_{rd} .

Importantly, the IM algorithm *always produces a sound process model able to replay the whole event log*. Unlike many other algorithms, fitness is guaranteed. Since the models are block-structured and activities are not duplicated, the models tend to be simple and general (overfitting models are often the result of excessive label splitting). However, the fall-through in the IM algorithm may create underfitting models. This may occur in situations where there is no process tree without duplicate and silent activities generating the observed behavior.

Apart from the very special situation sketched in Fig. 7.28, any process tree without duplicate and silent activities is language-rediscoverable. When allowing for duplicate and silent activities such guarantees are more difficult to provide. The basic IM algorithm described in Sect. 7.5.1 never duplicates activities. Silent activities are only introduced for base cases and empty traces, e.g., $\times(\tau, a)$ if a can be skipped, $\bigcirc(a, \tau)$ if a can be repeated, and $\bigcirc(\tau, a)$ if a can be skipped and repeated. In the presence of duplicate and silent activities, the directly-follows graph provides insufficient information to ensure language-rediscoverability. However, weaker guarantees such as the ability to replay the event log without any problems still hold.

To get an understanding of the limitations of the basic IM algorithm, we consider the example logs used to introduce the α -algorithm in Chap. 6. As already shown, the process trees discovered by the IM algorithm for logs L_1 and L_2 are behaviorally equivalent to the WF-nets discovered by the α -algorithm (see Fig 7.20 and Fig 7.24). Some other logs used in Chap. 6 are considered next:

$$L_3 = [\langle a, b, c, d, e, f, b, d, c, e, g \rangle, \langle a, b, d, c, e, g \rangle^2,$$

$$\langle a, b, c, d, e, f, b, c, d, e, f, b, d, c, e, g \rangle]$$

$$L_4 = [\langle a, c, d \rangle^{45}, \langle b, c, d \rangle^{42}, \langle a, c, e \rangle^{38}, \langle b, c, e \rangle^{22}]$$

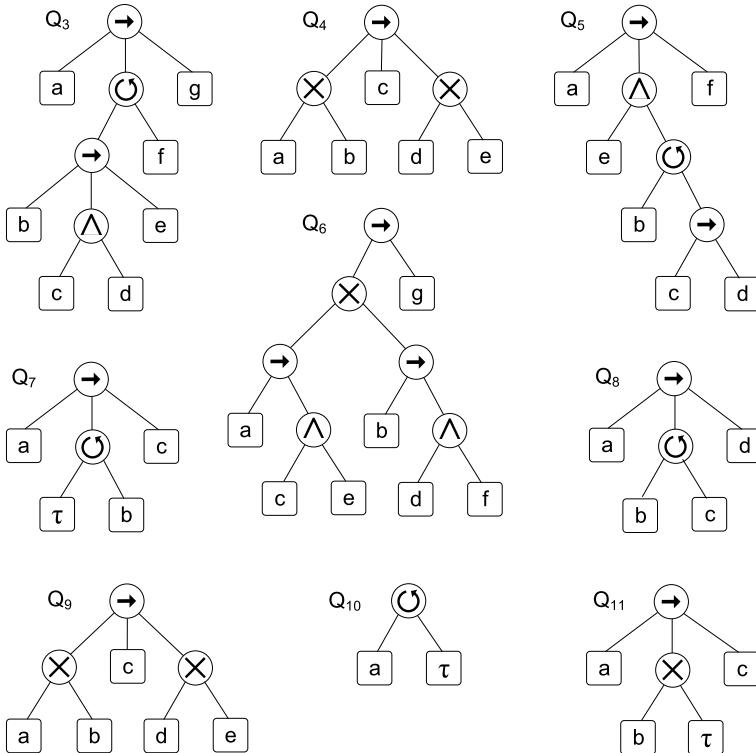


Fig. 7.29 Process trees learned for event logs L_3 – L_{11} introduced in Chap. 6

$$L_5 = [\langle a, b, e, f \rangle^2, \langle a, b, e, c, d, b, f \rangle^3, \langle a, b, c, e, d, b, f \rangle^2, \langle a, b, c, d, e, b, f \rangle^4, \langle a, e, b, c, d, b, f \rangle^3]$$

$$L_6 = [\langle a, c, e, g \rangle^2, \langle a, e, c, g \rangle^3, \langle b, d, f, g \rangle^2, \langle b, f, d, g \rangle^4]$$

$$L_7 = [\langle a, c \rangle^2, \langle a, b, c \rangle^3, \langle a, b, b, c \rangle^2, \langle a, b, b, b, b, c \rangle]$$

$$L_8 = [\langle a, b, d \rangle^3, \langle a, b, c, b, d \rangle^2, \langle a, b, c, b, c, b, d \rangle]$$

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

$$L_{10} = [\langle a, a \rangle^{55}]$$

$$L_{11} = [\langle a, b, c \rangle^{20}, \langle a, c \rangle^{30}]$$

These example logs were used to illustrate the characteristics and limitations of the α -algorithm. Figure 7.29 shows the process trees learned for these event logs using the basic IM algorithm described in Sect. 7.5.1: $Q_3 = \text{IM}(L_3)$, $Q_4 = \text{IM}(L_4)$, etc.

The α -algorithm was able to discover “correct” models for L_1-L_5 where correctness is defined as the ability to reproduce the event log. For L_6 the α -algorithm produces a Petri net with two redundant places, but the discovered model is trace equivalent to the desired model (see Fig. 6.9). The α -algorithm cannot handle the loop of length one required for L_7 . Also loops of length two cannot be discovered and hence event log L_8 is not handled well. The α -algorithm creates an underfitting model for L_9 . The α -algorithm is also unable to handle the repetition of a in L_{10} and the skipping of b in L_{11} .

Figure 7.29 shows the process trees generated by the IM algorithm for L_3-L_{11} . Unlike the α -algorithm, all models are by definition sound and can replay the respective event log (i.e., perfect fitness). Hence, Q_1-Q_{11} are “correct” in the sense mentioned before. Whereas the α -algorithm was unable to handle short loops (length one or length two), the IM algorithm creates Q_7 and Q_8 illustrating that loops pose no problem. Process tree Q_{11} shows that the skipping of activities can be handled by the IM algorithm. However, process trees Q_9 and Q_{10} also show that the discovered models may be underfitting.

In event log $L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$, activity a is eventually followed by d and b is eventually followed by e , but process tree $Q_9 = \rightarrow(\times(a, b), c, \times(d, e))$ does not capture this non-local dependency and allows a to be followed by e . This is not surprising since the process tree representation does not allow for such a non-local dependency (without label splitting). Process tree $Q'_9 = \times(\rightarrow(a, c, d), \rightarrow(b, c, e))$ better captures the behavior seen in L_9 , but requires the duplication of activity c . Label duplication would be the best choice here, but often label duplication leads to overfitting models simply enumerating parts of the event log.

The basic IM algorithm also cannot handle repetitions of a fixed length. Process tree $Q_{10} = \circlearrowleft(a, \tau)$ is discovered for event log $L_{10} = [\langle a, a \rangle^{55}]$. Hence, Q_{10} also allows for unobserved traces like $\langle a \rangle$ and $\langle a, a, a, a \rangle$. Process tree $Q'_{10} = \rightarrow(a, a)$ better captures the behavior seen in L_{10} , but requires the duplication of activity a .

The examples in Fig. 7.29 illustrate the characteristics of the IM algorithm. The algorithm *always* produces a process tree which is sound by construction and able to replay *all* behavior seen (perfect fitness). However, process trees constructed by the IM algorithm may be underfitting if the observed behavior requires a process tree with duplicate or silent activities. Different processes may have the same directly-follows graph, e.g., $\circlearrowleft(\wedge(a, b), \wedge(c, d))$ and $\wedge(\circlearrowleft(a, c), \circlearrowleft(b, d))$ (see Fig. 7.28). Also $\rightarrow(\times(a, b), c, \times(d, e))$ and $\times(\rightarrow(a, c, d), \rightarrow(b, c, e))$ have the same directly-follows graph. Such processes cannot be distinguished by the IM algorithm.

7.5.3 Extensions and Scalability

The basic IM algorithm described in Sect. 7.5.1 was introduced only recently (in 2013) [88]. Yet, several extensions and refinements have been proposed [89–91]. The basic algorithm cannot abstract from infrequent behavior and does not handle incompleteness well. The log is assumed to be directly-follows complete and frequencies are not taken into account. Fortunately, the IM framework is quite flexible.

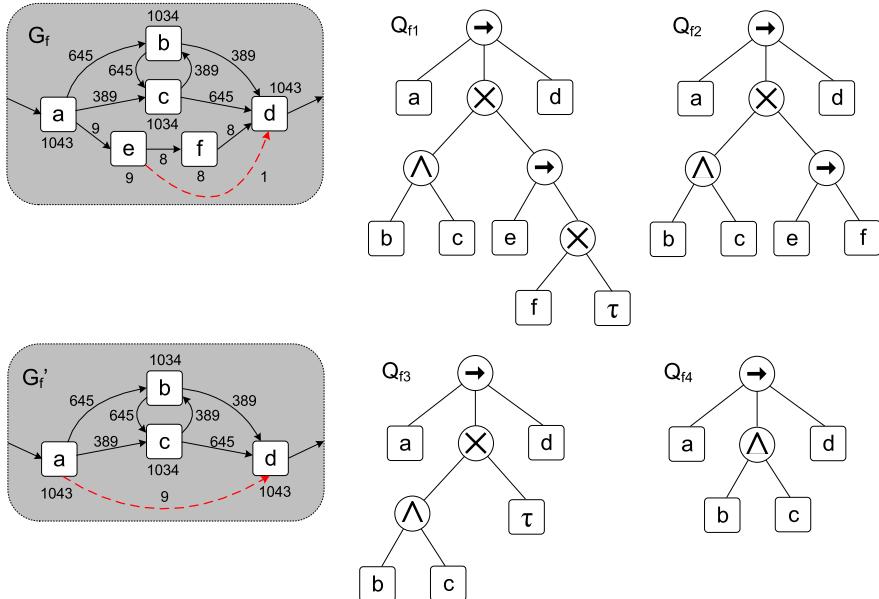


Fig. 7.30 Process trees $Q_{f1}-Q_{f4}$ learned for event log $L_f = [\langle a, b, c, d \rangle^{645}, \langle a, c, b, d \rangle^{389}, \langle a, e, f, d \rangle^8, \langle a, e, d \rangle]$ using frequency-based filtering. G_f is the directly-follows graph for the whole log. G'_f is the directly-follows graph after filtering based on activity frequency. The dashed lines indicate directly-follows relations that are less frequent and candidates for filtering

Using the basic ideas presented thus far a *family of inductive mining techniques* has been developed. Example members of this family of process discovery algorithms are:

- *Inductive Miner—infrequent* (IMF, [89]),
- *Inductive Miner—incompleteness* (IMC, [90]),
- *Inductive Miner—directly-follows based* (IMD, [91]),
- *Inductive Miner—infrequent—directly-follows based* (IMFD, [91]), and
- *Inductive Miner—incompleteness—directly-follows based* (IMCD, [91]).

To illustrate the IMF (Inductive Miner - infrequent) algorithm consider event log $L_f = [\langle a, b, c, d \rangle^{645}, \langle a, c, b, d \rangle^{389}, \langle a, e, f, d \rangle^8, \langle a, e, d \rangle]$. Figure 7.30 shows the directly-follows graph $G_f = G(L_f)$ based on event log L_f . The numbers indicate frequencies, e.g., activity b was executed 1034 times and was directly followed by activity c 645 times. The arc between e and d is dashed because this directly-follows relation is infrequent compared to all other arcs (the only “witness” is trace $\langle a, e, d \rangle$ that occurs once). Filtering out this arc would yield process tree Q_{f2} rather than Q_{f1} . In Q_{f2} , activity f cannot be skipped, this option was removed because it happens only once in L_f . Next to filtering arcs, it is also possible to filter activities. Activities e and f occur less frequent than the other activities. It is possible to remove these activities from the event log resulting in event log $L'_f = [\langle a, b, c, d \rangle^{645},$

$\langle a, c, b, d \rangle^{389}, \langle a, d \rangle^9]$. Figure 7.30 shows the directly-follows graph $G'_f = G(L'_f)$ based on the filtered event log. Process tree Q_{f3} is based on this directly-follows graph. The arc between a and d is dashed because this directly-follows relation is infrequent (only 9 times a is followed by d). Filtering out this arc would yield process tree Q_{f4} rather than Q_{f3} . The IMF algorithm uses various types of filtering with the goal to show the mainstream behavior. Figure 7.30 only sketches the basic principles. The IMF algorithm is more sophisticated and also uses the “eventually-follows graph” for filtering (next to the directly-follows graph). Note that there are some similarities with the heuristic miner. See [89] for details.

The IMC (Inductive Miner—incompleteness) algorithm [90] complements the IMF algorithm. Instead of removing exceptional behavior, the problem of missing behavior due to the incompleteness of the event log is addressed. The assumption that event logs are directly-follows complete is unrealistic for less structured processes and relatively small event logs. Consider $Q_c = \wedge(a_1, a_2, \dots, a_{10})$ and some event log L_c generated from this process tree. There are $10! = 3,628,800$ possible interleavings in Q_c . Suppose we have an event log with 500 cases. Obviously, this event log only shows a fraction of the possible interleavings (less than 1 out of 7000). Since the different interleavings have different probabilities (e.g., due to different delay distributions), it may also be the case that not all of the 90 possible directly-follows relations appear in L_c . Hence, arcs may be missing in the directly-follows graph. This makes it impossible to apply the parallel cut ($\wedge, \{a_1\}, \{a_2\}, \dots, \{a_{10}\}$). As a result the fall-through described before needs to be used ($\circlearrowleft(\tau, a_1, \dots)$), resulting in an underfitting model. The IMC algorithm uses so-called “probabilistic activity relations” [90] based on both the directly-follows graph and the eventually-follows graph. These are used to select the “most likely cut” even if the requirements stated in Definition 7.6 are not fully satisfied.

The IM, IMF, and IMC algorithms perform quite well compared to other algorithms (e.g., much faster than region-based techniques). However, the event log needs to be split recursively. This may create quite some overhead for larger event logs. Ideally, a single pass through the event log is preferable from a performance point of view. However, the IM, IMF, and IMC algorithms repeatedly traverse the event log to create smaller logs.

The *Inductive Miner—directly-follows based* (IMD) framework recurses on the directly-follows graph directly without creating sublogs [91]. This makes the framework extremely scalable. A single pass through the event log suffices and the work can be distributed easily. However, there are some limitations related to the accuracy of the results. There exist variants of the IM, IMF, and IMC algorithms using this framework. These are called the IMD (Inductive Miner—directly-follows based) algorithm, the IMFD (Inductive Miner—infrequent—directly-follows based) algorithm, and the IMCD (Inductive Miner—incompleteness—directly-follows based) algorithm. The cut detection works as before. However, the directly-follows graph is split into disjoint subgraphs (the graphs are not recomputed over sublogs).

The IMD algorithm runs in $O(n^3)$ where n is the number of activities in the directly-follows graph [91]. However, the guarantees provided by the IMD algorithm are similar to the basic IM algorithm. Still most process trees without dupli-

cate and silent activities are language-rediscoverable. If the event log is directly-follows complete and situations such as the one shown in Fig. 7.28 are excluded, then the IMD algorithm is able to rediscover the model used to generate the event log. This only holds under the assumption that there are no duplicate and silent activities.

The IMD framework also has some limitations. The model returned by the IMD algorithm is no longer fitness preserving. Consider directly-follows complete event logs L_1 and L_2 for process trees $Q_1 = \wedge(\rightarrow(a, b), c)$ and $Q_2 = \times(\rightarrow(a, c, b, c, a, b), c)$. The two logs have identical directly-follows graphs, $G(L_1) = G(L_2)$. The IMD algorithm returns Q_1 . However, Q_1 cannot reproduce any trace in L_2 . Hence, IMD is not fitness preserving for L_2 . If the IMD algorithm would not return Q_1 , then there would be no event log for which Q_1 could be constructed. This would be undesirable given the prevalence of Q_1 's behavior. Hence, fitness preservation is impossible in this setting (without using a fall-through).

The basic IM algorithm that recursively splits the event log is able to distinguish between L_1 and L_2 . The IM algorithm rediscovered Q_1 : $\text{IM}(L_1) = \wedge(\rightarrow(a, b), c)$. Q_2 is not rediscovered: $\text{IM}(L_2) = \wedge(\circlearrowleft(\tau, \rightarrow(a, b)), \circlearrowleft(c, \tau))$. However, unlike the IMD algorithm the log-splitting IM algorithm guarantees fitness preservation: $\mathcal{L}(Q_2) \subseteq \mathcal{L}(\text{IM}(L_2))$.

The limitations of the IMD framework are counterbalanced by its remarkable *scalability*. The IMD algorithm can handle event logs with billions of events while using only 2 GB of RAM [91]. It can be used to learn process models with over 10,000 activities. Moreover, computation can be easily distributed (e.g., using the Map-Reduce programming model and Hadoop-like infrastructures, see Chap. 12).

The family of inductive mining techniques also includes approaches that take into account transactional information (e.g., start and complete). The basic idea of all algorithms is to use a *divide-and-conquer* approach in combination with process trees that are *sound by construction*.

The different inductive mining algorithms (IM, IMF, IMC, IMD, IMFD, IMCD, etc.) combine interesting properties. The produced models are always sound. The algorithms are highly scalable, in particular IMD and IMFD. If desired, the algorithms are fitness-preserving (i.e., the log can be reproduced by models discovered using IM or IMC). Moreover, models can be seamlessly simplified by leaving out infrequent behavior (IMF and IMFD) and even event logs that are not directly-follows complete can be handled (IMC and IMCD). For particular classes of models even rediscoverability is guaranteed (IM and IMD). Trade-offs between scalability, accuracy, generalization, and precision are supported. These characteristics make inductive mining the current frontrunner in process discovery.

7.6 Historical Perspective

On the one hand, process mining is a relatively young field. All the process discovery techniques described in this chapter were developed in the last decade. More-

over, it is only recently that mature process discovery techniques and effective implementations have become available. On the other hand, process discovery has its roots in various established scientific disciplines ranging from concurrency theory, inductive inference and stochastics to data mining, machine learning and computational intelligence. It is impossible to do justice to the numerous contributions to process mining originating from different scientific domains. Hence, this section should be seen as a modest attempt to provide a historical perspective on the origins of process discovery.

In 1967 Mark Gold showed in his seminal paper “Language identification in limit” [61] that even regular languages cannot be exactly identified from positive examples only. In [61] Gold describes several inductive inference problems. The challenge is to guess a “rule” (e.g., a regular expression) based on an infinite stream of examples. An inductive inference method is able to “learn the rule in the limit” if after a finite number of examples the method is always able to guess the correct rule and does not need to revise its guess anymore based on new examples. A regular language is a language that can be accepted by a finite transition system (also referred to as a finite state machine). Regular languages can also be described in terms of regular expressions. For example, the regular expression $ab^*(c|d)$ denotes the set of traces starting with a , then zero or more b ’s and finally a c or d . Regular expressions were introduced by Stephen Cole Kleene [85] in 1956. In the Chomsky hierarchy of formal grammars, regular languages are the least expressive (i.e., Type-3 grammar). For example, it is impossible to express the language $\{a^n b^n \mid n \in \mathbb{N}\}$, i.e., the language containing traces that start with any number of a ’s followed by the same number of b ’s. Despite the limited expressiveness of regular expressions, Gold showed in [61] that they *cannot* be learned in the limit from positive examples only.

Many inductive inference problems have been studied since Gold’s paper (see the survey in [15]). For instance, subclasses of the class of regular languages have been identified that can be learned in the limit (e.g., the so-called k -reversible languages [15]). Moreover, if an “oracle” is used that can indicate whether particular examples are possible or not, a larger class of languages can be learned. This illustrates the importance of negative examples when learning. However, as indicated before, one will not find negative examples in an event log; *the fact that something did not happen provides no guarantee that it cannot happen*. Inductive inference focuses on learning a language perfectly. This is not the aim of process mining. Real-life event logs will contain noise and are far from complete. Therefore, the theoretical considerations in the context of inductive inference are less relevant for process mining.

Before the paper of Gold, there were already techniques to construct a finite state machine from a finite set of example traces. A naïve approach is to use the state representation function $I_1^{state}(\sigma, k) = hd^k(\sigma)$ described in Sect. 7.4.1 to construct a finite state machine. Such a finite state machine can be made smaller by using the classical Myhill–Nerode theorem [108]. Let L be a language over some alphabet \mathcal{A} and consider $\sigma_x, \sigma_y \in \mathcal{A}^*$. σ_x and σ_y are *equivalent* if there is no $\sigma_z \in \mathcal{A}^*$ such that $\sigma_x \oplus \sigma_z \in L$ while $\sigma_y \oplus \sigma_z \notin L$ or $\sigma_y \oplus \sigma_z \in L$ while $\sigma_x \oplus \sigma_z \notin L$. Hence, two traces

are equivalent if their “sets of possible futures” coincide. This equivalence notion divides the elements of L into equivalence classes. If L is a regular language, then there are finitely many equivalence classes. The Myhill–Nerode theorem states that if there are k such equivalence classes, then the smallest finite state machine accepting L has k states. Several approaches have been proposed to minimize finite state machines using these insights (basically folding equivalent states). In [21], a modification of the Myhill–Nerode equivalence relation is proposed for constructing a finite state machine based on a set of sample traces L with a parameter to balance precision and complexity. Here two states are considered equivalent if their k -tails are the same. In 1972, Alan Biermann also proposed an approach to “learn” a Turing machine from a set of sample computations [20].

In the mid 1990s, people like Rakesh Agrawal and others developed various data mining algorithms to find frequent patterns in large datasets. In [7], the Apriori algorithm for finding association rules was presented. These techniques were extended to sequences and episodes [69, 94, 131]. However, none of these techniques aimed at discovering end-to-end processes. More related is the work on hidden Markov models [9]. Here end-to-end processes can be considered. However, these models are sequential and cannot be easily converted into readable business process models.

In the second half of the 1990s, Cook and Wolf developed process discovery techniques in the context of software engineering processes. In [33], they described three methods for process discovery: one using neural networks, one using a purely algorithmic approach, and one Markovian approach. The authors considered the latter two to be the most promising approaches. The purely algorithmic approach builds a finite state machine in which states are fused if their futures (in terms of possible behavior in the next k steps) are identical. (Note that this is essentially the approach proposed by Biermann and Feldmann in [21].) The Markovian approach uses a mixture of algorithmic and statistical methods and is able to deal with noise. All approaches described in [33] are limited to sequential processes, i.e., no concurrency is discovered.

In 1998, two papers [8, 38] appeared that, independently of one another, proposed to apply process discovery in the context of business process management.

In [8], Agrawal, Gunopulos, and Leymann presented an approach to discover the so-called “conformal process graph” from event logs. This work was inspired by the process notation used by Flowmark and the presence of event logs in WFM systems. The approach discovers causal dependencies between activities, but is not able to find AND/XOR/OR-splits and joins, i.e., the process logic is implicit. Moreover, the approach has problems dealing with loops: a trace $\langle a, a, a \rangle$ is simply relabeled into $\langle a_1, a_2, a_3 \rangle$ to make the conformal process graph acyclic.

In the same year, Anindya Datta [38] proposed a technique to discover business process models by adapting the Biermann–Feldmann algorithm [21] for constructing finite state machines based on example traces. Datta added probabilistic elements to the original approach and embedded the work in the context of workflow management and business process redesign. The approach assumes that case identifiers are unknown, i.e., the setting is similar to the work in [53] where the challenge is to correlate events and discover cases. The resulting process model is again a sequential model.

Joachim Herbst [71, 72] was one of the first aiming at the discovery of more complicated process models. He proposed stochastic task graphs as an intermediate representation before constructing a workflow model in terms of the ADONIS modeling language. In the induction step, task nodes are merged and split in order to discover the underlying process. A notable difference with most approaches is that the same activity can appear multiple times in the process model, i.e., the approach allows for duplicate labels. The graph generation technique is similar to the approach of [8]. The nature of splits and joins (i.e., AND or OR) is discovered in the transformation step, i.e., the step in which the stochastic task graph is transformed into an ADONIS workflow model with block-structured splits and joins.

Most of the classical approaches have problems dealing with concurrency, i.e., either sequential models are assumed (e.g., transition systems, finite state machines, Markov chains, and hidden Markov models) or there is a post-processing step to discover concurrency. The first model to adequately capture concurrency was already introduced by Carl Adam Petri in 1962 [111]. (Note that the graphical notation as we know it today was introduced later.) However, classical process discovery techniques do not take concurrency into account. The α -algorithm [157] described in Sect. 6.2 and a predecessor of the heuristic miner [184] described in Sect. 7.2 were developed concurrently and share the same ideas when it comes to handling concurrency. These were the first process discovery techniques taking concurrency as a starting point (and not as an afterthought or post-optimization). The α -algorithm was used to explore the theoretical limits of process discovery [157]. Several variants of the α -algorithm have been proposed to lift some of its limitations [10, 11, 171, 174, 185]. The focus of heuristic mining was (and still is) on dealing with noise and incompleteness [183, 184].

Techniques such as the α -algorithm and heuristic mining do not guarantee that the model can replay all cases in the event log. In [171, 172], an approach is presented that guarantees a fitness of 1, i.e., all traces in the event log can be replayed in the discovered model. This is achieved by creatively using OR-splits and joins. As a result, the discovered model is typically underfitting. In [59, 60], artificially generated “negative events” are inserted to transform process discovery into a classification problem. The insertion of negative events corresponds to the completeness assumptions made by algorithms like the α -algorithm, e.g., “if a is never directly followed by b , then this is not possible”.

Region-based approaches are able to express more complex control-flow structures without underfitting. State-based regions were introduced by Ehrenfeucht and Rozenberg [51] in 1989 and generalized by Cortadella et al. [34]. In [165, 173], it is shown how these state-based regions can be applied to process mining. In parallel, several authors applied language-based regions to process mining [19, 170]. In [28], Joseph Carmona and Jordi Cortadella present an approach based on convex polyhedra. Here, the Parikh vector of each prefix in the log is seen as a polyhedron. By taking the convex hull of these convex polyhedra one obtains an over-approximation of the possible behavior. The resulting polyhedron can be converted into places using a construction similar to language-based regions. The synthesis/region-based approaches typically guarantee a fitness of 1. Unfortunately, these approaches also have problems dealing with noise.

For practical applications of process discovery it is essential that noise and incompleteness are handled well. Surprisingly, only few discovery algorithms described in literature focus on addressing these issues. Notable exceptions are heuristic mining [183, 184], fuzzy mining [66], genetic process mining [12, 26], and inductive mining [89–91]. Therefore, we put emphasis on these techniques in this chapter.

See [156, 160, 174] for additional pointers to earlier related work. These surveys do not include recent developments such as the family of inductive mining techniques presented in Sect. 7.5. The different inductive mining algorithms (IM, IMF, IMc, IMD, IMFD, IMCD, etc.) always produce sound models and are highly scalable. Moreover, these algorithms come with formal guarantees. For example, the IM algorithm is fitness-preserving and for particular classes of models even rediscoverability is guaranteed.

Part IV

Beyond Process Discovery

Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling
and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery:
An Introduction

Chapter 7
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance
Checking

Chapter 9
Mining Additional
Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining
Software

Chapter 12
Process Mining in the
Large

Chapter 13
Analyzing “Lasagna
Processes”

Chapter 14
Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15
Cartography and
Navigation

Chapter 16
Epilogue

In the previous part, the focus was on process discovery. However, in many situations, there is already a (partial) process model. Chapter 8 presents techniques for checking the quality of such models. Moreover, process models (discovered or made by hand) should not only describe control-flow: the other perspectives also need to be addressed. Chapter 9 exhibits techniques for mining additional perspectives involving resources, time, and data. Chapter 10 extends the scope even further and shows how process mining can be used to directly influence cases that are still running.

Chapter 8

Conformance Checking

After covering control-flow discovery in depth in Part III, this chapter looks at the situation in which both a process model and an event log are given. The model may have been constructed by hand or may have been discovered. Moreover, the model may be normative or descriptive. Conformance checking relates events in the event log to activities in the process model and compares both. The goal is to find commonalities and discrepancies between the modeled behavior and the observed behavior. Conformance checking is relevant for business alignment and auditing. For example, the event log can be replayed on top of the process model to find undesirable deviations suggesting fraud or inefficiencies. Moreover, conformance checking techniques can also be used for measuring the performance of process discovery algorithms and to repair models that are not aligned well with reality.

8.1 Business Alignment and Auditing

In Sect. 2.4, we introduced the terms Play-In, Play-Out, and Replay. *Play-Out* is the classical use of process models; the model generates behavior. For instance, by playing the “token game” in a WF-net, example behaviors can be generated. Simulation and workflow engines use Play-Out to analyze and enact process models. *Play-In* is the opposite of Play-Out, i.e., example behavior is taken as input and the goal is to construct a model. The discovery techniques presented in Chaps. 6 and 7 can be used for Play-In. *Replay* uses both an event log and a process model as input, i.e., history is replayed using the model to analyze various phenomena. For example, in Chap. 9 we will show that replay can be used for analyzing bottlenecks and decision analysis. In Chap. 10, replay will be used to predict the behavior of running cases and to recommend suitable actions. In this chapter, we focus on *conformance checking* using replay.

Figure 8.1 illustrates the main idea of conformance checking. The behavior of a process model and the behavior recorded in an event log are compared to find commonalities and discrepancies. Such analysis results in *global conformance measures*

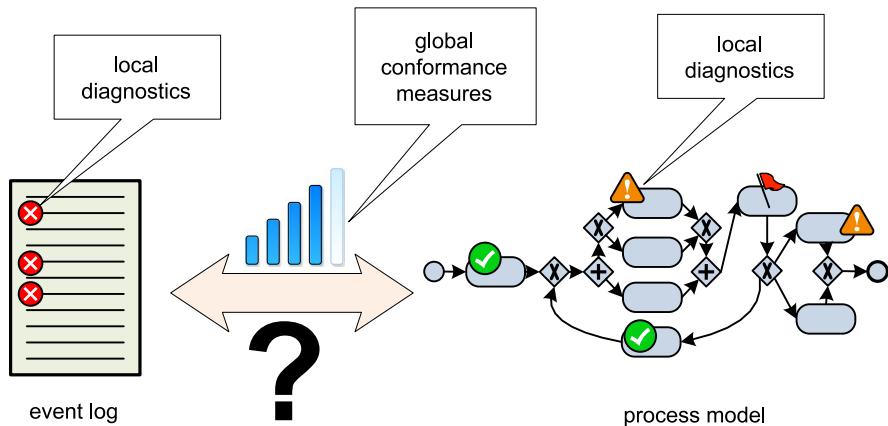


Fig. 8.1 Conformance checking: comparing observed behavior with modeled behavior. Global conformance measures quantify the overall conformance of the model and log. Local diagnostics are given by highlighting the nodes in the model where model and log disagree. Cases that do not fit are highlighted in the visualization of the log

(e.g., 85% of the cases in the event log can be replayed by the model) and *local diagnostics* (e.g., activity x was executed 15 times although this was not allowed according to the model). The interpretation of non-conformance depends on the purpose of the model. If the model is intended to be *descriptive*, then discrepancies between model and log indicate that the model needs to be improved to capture reality better. If the model is *normative*, then such discrepancies may be interpreted in two ways. Some of the discrepancies found may expose *undesirable deviations*, i.e., conformance checking signals the need for a better control of the process. Other discrepancies may reveal *desirable deviations*. For instance, workers may deviate to serve the customers better or to handle circumstances not foreseen by the process model. In fact, flexibility and non-conformance often correlate positively. For example, in some hospitals the phrase “breaking the glass” is used to refer to deviations that are recorded but that actually save lives. Nevertheless, even if most deviations are desired, it is important that stakeholders have insight into such discrepancies.

When checking conformance it is important to view deviations from two angles: (a) the model is “wrong” and does not reflect reality (“How to improve the model?”), and (b) cases deviate from the model and corrective actions are needed (“How to improve control to enforce a better conformance?”). Conformance checking techniques should support both viewpoints. Therefore, Fig. 8.1 shows deviations on both sides.

In Chap. 2, we related process mining to corporate governance, risk, compliance, and legislation such as the Sarbanes–Oxley Act (SOX) and the Basel II Accord. Corporate accounting scandals have triggered a series of new regulations. Although country-specific, there is a large degree of commonality between Sarbanes–Oxley (US), Basel II/III (EU), J-SOX (Japan), C-SOX (Canada), 8th EU Directive (EURO-SOX), BilMoG (Germany), MiFID (EU), Law 262/05 (Italy), Code Lippens

(Belgium), Code Tabaksblat (Netherlands), and others. These regulations require companies to identify the financial and operational risks inherent to their business processes, and establish the appropriate controls to address them. Although the focus of these regulations is on financial aspects, they illustrate the desire to make processes transparent and auditable. The ISO 9000 family of standards is another illustration of this trend. For instance, *ISO 9001:2008* requires organizations to model their operational processes. Currently, these standards do not force organizations to check conformance at the event level. For example, the real production process may be very different from the modeled production process. Nevertheless, the relation to conformance checking is evident. In this chapter, we take a more technological perspective and show concrete techniques for quantifying conformance and diagnosing non-conformance. However, before doing so, we briefly reflect on the relation between conformance checking, business alignment, and auditing.

The goal of *business alignment* is to make sure that the information systems and the real business processes are well aligned. People should be supported by the information system rather than work behind its back to get things done. Unfortunately, there is often a mismatch between the information system on the one hand and the actual processes and needs of workers and management on the other hand. There are various reasons for this. First of all, most organization use product software, i.e., generic software that was not developed for a specific organization. A typical example is the SAP system which is based on so-called “best practices”, i.e., typical processes and scenarios are implemented. Although such systems are configurable, the particular needs of an organization may be different from what was envisioned by the product software developer. Second, processes may change faster than the information system, because of external influences. Finally, there may be different stakeholders in the organization having conflicting requirements, e.g., a manager may want to enforce a fixed working procedure whereas an experienced worker prefers to have more flexibility to serve customers better.

Process mining can assist in improving the alignment of information systems, business processes, and the organization. By analyzing the real processes and diagnosing discrepancies, new insights can be gathered showing how to improve the support by information systems.

The term *auditing* refers to the evaluation of organizations and their processes. Audits are performed to ascertain the validity and reliability of information about these organizations and associated processes. This is done to check whether business processes are executed within certain boundaries set by managers, governments, and other stakeholders. For instance, specific rules may be enforced by law or company policies and the auditor should check whether these rules are followed or not. Violations of these rules may indicate fraud, malpractice, risks, and inefficiencies. Traditionally, auditors can only provide *reasonable assurance* that business processes are executed within the given set of boundaries. They check the operating effectiveness of controls that are designed to ensure reliable processing. When these controls are not in place, or otherwise not functioning as expected, they typically only check samples of factual data, often in the “paper world”.

However, today detailed information about processes is being recorded in the form of event logs, audit trails, transaction logs, databases, data warehouses, etc.

Therefore, it should no longer be acceptable to only check a small set of samples off-line. Instead, *all events in a business process can be evaluated and this can be done while the process is still running*. The availability of log data and advanced process mining techniques enables new forms of auditing [166]. Process mining in general, and conformance checking in particular, provide the means to do so.

8.2 Token Replay

In Sect. 6.4.3, we discussed four quality criteria: fitness, precision, generalization, and simplicity. These were illustrated using Fig. 6.24. In this figure one event log is given and four process models are shown. For each of these models, a subjective judgment is given with respect to the four quality criteria. As the models are rather extreme, the scores for the various quality criteria are evident. However, in a more realistic setting it is much more difficult to judge the quality of a model. This section shows how the notion of *fitness* can be quantified. Fitness measures “the proportion of behavior in the event log possible according to the model”. Of the four quality criteria, fitness is most related to conformance.

To explain the various fitness notions, we use the event log L_{full} described in Table 8.1. This is the same event log as the one used in Fig. 6.24. There are 1391 cases in L_{full} distributed over 21 different traces. For example, there are 455 cases following trace $\sigma_1 = \langle a, c, d, e, h \rangle$, 191 cases following trace $\sigma_2 = \langle a, b, d, e, g \rangle$, etc.

Figure 8.2 shows four models related to event log L_{full} . WF-net N_1 is the process model discovered when applying the α -algorithm to L_{full} . WF-net N_2 is a sequential model that, compared to N_1 , requires the examination (activity b or c) to take place before checking the ticket (activity d). Clearly, N_2 does not allow for all traces in Table 8.1. For example, $\sigma_3 = \langle a, d, c, e, h \rangle$ is not possible according to WF-net N_2 . WF-net N_3 has no choices, e.g., the request is always rejected. Many traces in Table 8.1 cannot be replayed by this model, e.g., $\sigma_2 = \langle a, b, d, e, g \rangle$ is not possible according to WF-net N_3 . WF-net N_4 is a variant of the “flower model”: the only requirement is that traces need to start with a and end with g or h . Clearly, all traces in Table 8.1 can be replayed by N_4 .

A naïve approach towards conformance checking would be to simply count the fraction of cases that can be “parsed completely” (i.e., the proportion of cases corresponding to firing sequences leading from $[start]$ to $[end]$). Using this approach the fitness of N_1 is $\frac{1391}{1391} = 1$, i.e., all 1391 cases in L_{full} correspond to a firing sequence of N_1 (“can be replayed”). The fitness of N_2 is $\frac{948}{1391} = 0.6815$ because 948 cases can be replayed correctly whereas 443 cases do not correspond to a firing sequence of N_2 . The fitness of N_3 is $\frac{632}{1391} = 0.4543$: only 632 cases have a trace corresponding to a firing sequence of N_2 . The fitness of N_4 is $\frac{1391}{1391} = 1$ because the “flower model” is able to replay all traces in Table 8.1. This naïve fitness metric is less suitable for more realistic processes. Consider for instance a variant of WF-net N_1 in which places $p1$ and $p2$ are merged into a single place. Such a model will have a

Table 8.1 Event log L_{full} : $a = \text{register request}$, $b = \text{examine thoroughly}$, $c = \text{examine casually}$, $d = \text{check ticket}$, $e = \text{decide}$, $f = \text{reinitiate request}$, $g = \text{pay compensation}$, and $h = \text{reject request}$

Frequency	Reference	Trace
455	σ_1	$\langle a, c, d, e, h \rangle$
191	σ_2	$\langle a, b, d, e, g \rangle$
177	σ_3	$\langle a, d, c, e, h \rangle$
144	σ_4	$\langle a, b, d, e, h \rangle$
111	σ_5	$\langle a, c, d, e, g \rangle$
82	σ_6	$\langle a, d, c, e, g \rangle$
56	σ_7	$\langle a, d, b, e, h \rangle$
47	σ_8	$\langle a, c, d, e, f, d, b, e, h \rangle$
38	σ_9	$\langle a, d, b, e, g \rangle$
33	σ_{10}	$\langle a, c, d, e, f, b, d, e, h \rangle$
14	σ_{11}	$\langle a, c, d, e, f, b, d, e, g \rangle$
11	σ_{12}	$\langle a, c, d, e, f, d, b, e, g \rangle$
9	σ_{13}	$\langle a, d, c, e, f, c, d, e, h \rangle$
8	σ_{14}	$\langle a, d, c, e, f, d, b, e, h \rangle$
5	σ_{15}	$\langle a, d, c, e, f, b, d, e, g \rangle$
3	σ_{16}	$\langle a, c, d, e, f, b, d, e, f, d, b, e, g \rangle$
2	σ_{17}	$\langle a, d, c, e, f, d, b, e, g \rangle$
2	σ_{18}	$\langle a, d, c, e, f, b, d, e, f, b, d, e, g \rangle$
1	σ_{19}	$\langle a, d, c, e, f, d, b, e, f, b, d, e, h \rangle$
1	σ_{20}	$\langle a, d, b, e, f, b, d, e, f, d, b, e, g \rangle$
1	σ_{21}	$\langle a, d, c, e, f, d, b, e, f, c, d, e, f, d, b, e, g \rangle$

fitness of $\frac{0}{1391} = 0$, because none of the traces can be replayed. This fitness notion seems to be too strict as most of the model seems to be consistent with the event log. This is especially the case for larger process models. Consider, for example, a trace $\sigma = \langle a_1, a_2, \dots, a_{100} \rangle$ in some log L . Now consider a model that cannot replay σ , but that can replay 99 of the 100 events in σ (i.e., the trace is “almost” fitting). Also consider another model that can only replay 10 of the 100 events in σ (i.e., the trace is not fitting at all). Using the naïve fitness metric, the trace would simply be classified as non-fitting for both models without acknowledging that σ was almost fitting in one model and in complete disagreement with the other model. Therefore, we use a fitness notion defined *at the level of events* rather than full traces.

In the naïve fitness computation just described, we stopped replaying a trace once we encounter a problem and mark it as non-fitting. Let us now just continue replaying the trace on the model but record all situations where a transition is forced to fire without being enabled, i.e., we count all missing tokens. Moreover, we record the tokens that remain at the end. To explain the idea, we first replay σ_1 on top of WF-net N_1 . Note that σ_1 can be replayed completely. However, we use this example to introduce the notation. Figure 8.3 shows the various stages of replay. Four counters are shown at each stage: p (produced tokens), c (consumed tokens),

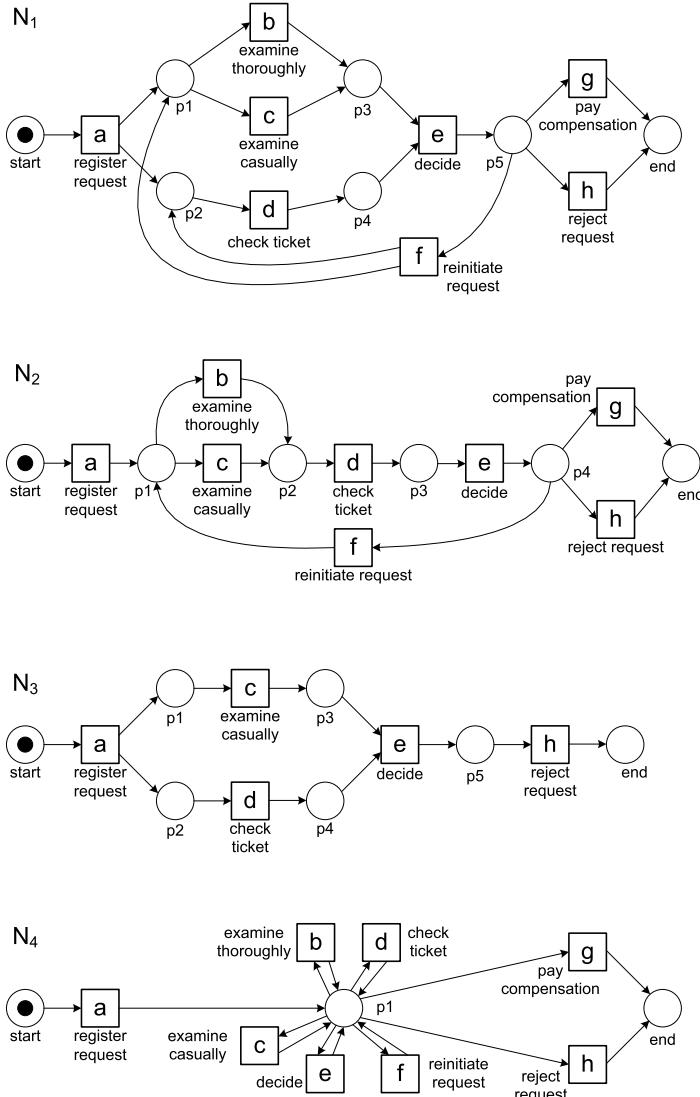
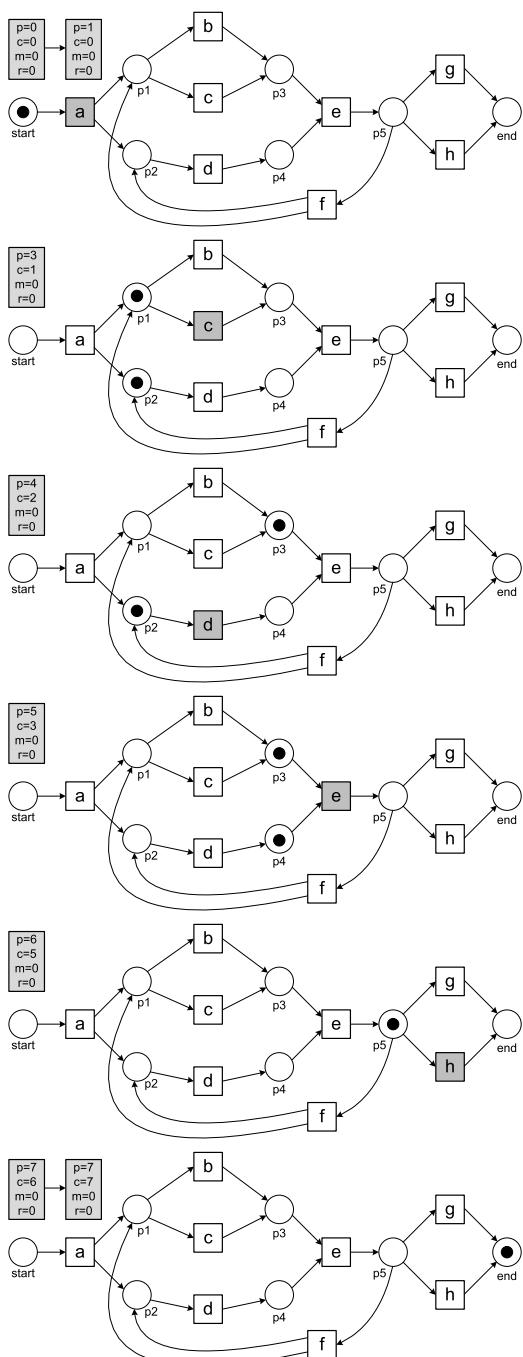


Fig. 8.2 Four WF-nets: N_1 , N_2 , N_3 and N_4

m (missing tokens), and r (remaining tokens). Let us first focus on p and c . Initially, $p = c = 0$ and all places are empty. Then the environment produces a token for place *start*. Therefore, the p counter is incremented: $p = 1$. Now we need to replay $\sigma_1 = \langle a, c, d, e, h \rangle$, i.e., we first fire transition a . This is possible. Since a consumes one token and produces two tokens, the c counter is incremented by 1 and the p counter is incremented by 2. Therefore, $p = 3$ and $c = 1$ after firing transition a . Then we replay the second event (c). Firing transition c results in $p = 4$

Fig. 8.3 Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$ on top of WF-net N_1 . There are four counters: p (produced tokens), c (consumed tokens), m (missing tokens), and r (remaining tokens)



and $c = 2$. After replaying the third event (i.e. d) $p = 5$ and $c = 3$. Then we replay e . Since e consumes two tokens and produces one, the result is $p = 6$ and $c = 5$. Then we replay the last event (h). Firing h results in $p = 7$ and $c = 6$. At the end, the environment consumes a token from place end . Hence the final result is $p = c = 7$ and $m = r = 0$. Clearly, there are no problems when replaying the σ_1 , i.e., there are no missing or remaining tokens ($m = r = 0$).

The fitness of a case with trace σ on WF-net N is defined as follows:

$$\text{fitness}(\sigma, N) = \frac{1}{2} \left(1 - \frac{m}{c} \right) + \frac{1}{2} \left(1 - \frac{r}{p} \right)$$

The first part computes the fraction of missing tokens relative to the number of consumed tokens. $1 - \frac{m}{c} = 1$ if there are no missing tokens ($m = 0$) and $1 - \frac{m}{c} = 0$ if all tokens to be consumed were missing ($m = c$). Similarly, $1 - \frac{r}{p} = 1$ if there are no remaining tokens and $1 - \frac{r}{p} = 0$ if none of the produced tokens was actually consumed. We use an equal penalty for missing and remaining tokens. By definition: $0 \leq \text{fitness}(\sigma, N) \leq 1$. In our example, $\text{fitness}(\sigma_1, N_1) = \frac{1}{2}(1 - \frac{0}{7}) + \frac{1}{2}(1 - \frac{0}{7}) = 1$ because there are no missing or remaining tokens.

Let us now consider a trace that cannot be replayed properly. Fig. 8.4 shows the process of replaying $\sigma_3 = \langle a, d, c, e, h \rangle$ on WF-net N_2 . Initially, $p = c = 0$ and all places are empty. Then the environment produces a token for place $start$ and the p counter is updated: $p = 1$. The first event (a) can be replayed. After firing a , we have $p = 2, c = 1, m = 0$, and $r = 0$. Now we try to replay the second event. This is not possible, because transition d is not enabled. To fire d , we need to add a token to place $p2$ and record the missing token, i.e., the m counter is incremented. The p and c counter are updated as usual. Therefore, after firing d , we have $p = 3, c = 2, m = 1$, and $r = 0$. We also tag place $p2$ to remember that a token was missing. Then we replay the next three events (c, e, h). The corresponding transitions are enabled. Therefore, we only need to update p and c counters. After replaying the last event, we have $p = 6, c = 5, m = 1$, and $r = 0$. In the final state $[p2, end]$ the environment consumes the token from place end . A token remains in place $p2$. Therefore, place $p2$ is tagged and the r counter is incremented. Hence the final result is $p = c = 6$ and $m = r = 1$. Figure 8.4 shows diagnostic information that helps to understand the nature of non-conformance. There was a situation in which d occurred but could not happen according to the model (m -tag) and there was a situation in which d was supposed to happen but did not occur according to the log (r -tag). Moreover, we can compute the fitness of trace σ_3 on WF-net N_2 based on the values of p, c, m , and r :

$$\text{fitness}(\sigma_3, N_2) = \frac{1}{2} \left(1 - \frac{1}{6} \right) + \frac{1}{2} \left(1 - \frac{1}{6} \right) = 0.8333$$

As a third example, we replay $\sigma_2 = \langle a, b, d, e, g \rangle$ on top of WF-net N_3 . Now the situation is slightly different because N_3 does not contain all activities appearing in the event log. In such a situation it seems reasonable to abstract from these events. Hence, we effectively replay $\sigma'_2 = \langle a, d, e \rangle$. Figure 8.5 shows the process of replaying these three events. The first problem surfaces when replaying e . Since c did not

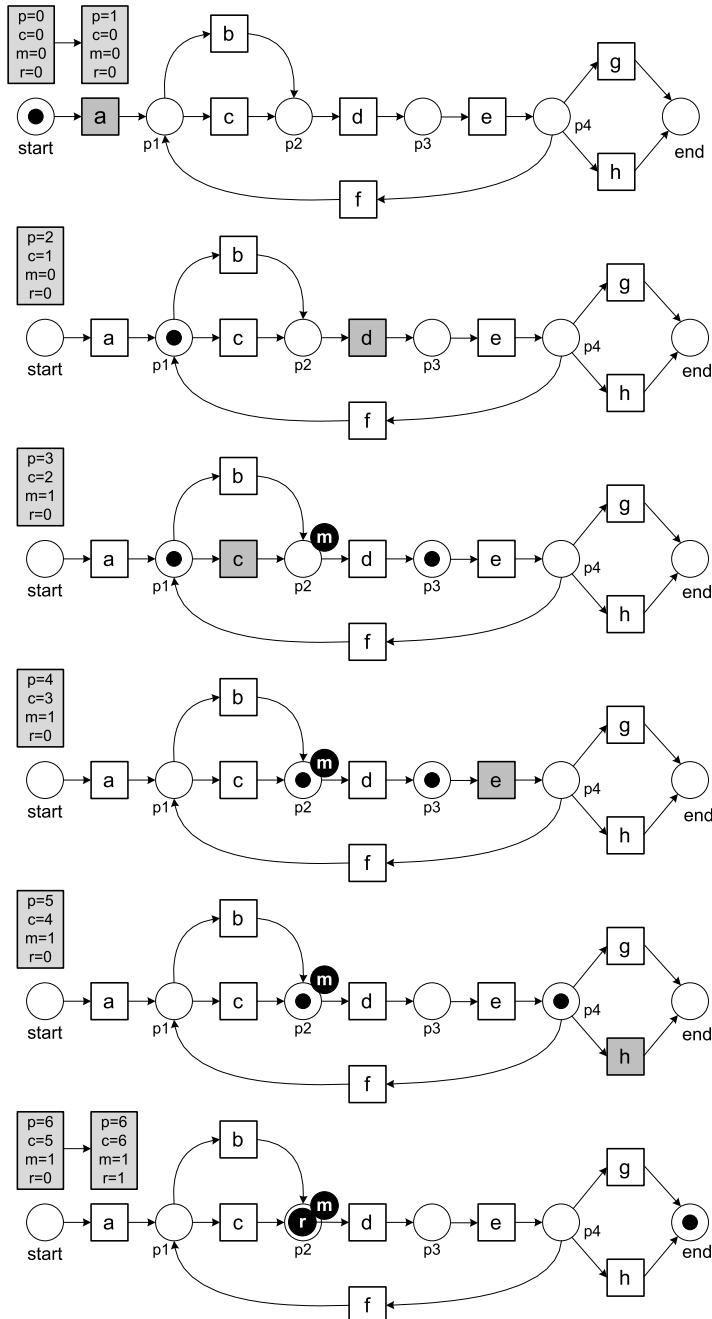


Fig. 8.4 Replayng $\sigma_3 = \langle a, d, c, e, h \rangle$ on top of WF-net N_2 : one token is missing ($m = 1$) and one token is remaining ($r = 1$). The r -tag and m -tag highlight the place where σ_3 and the model diverge

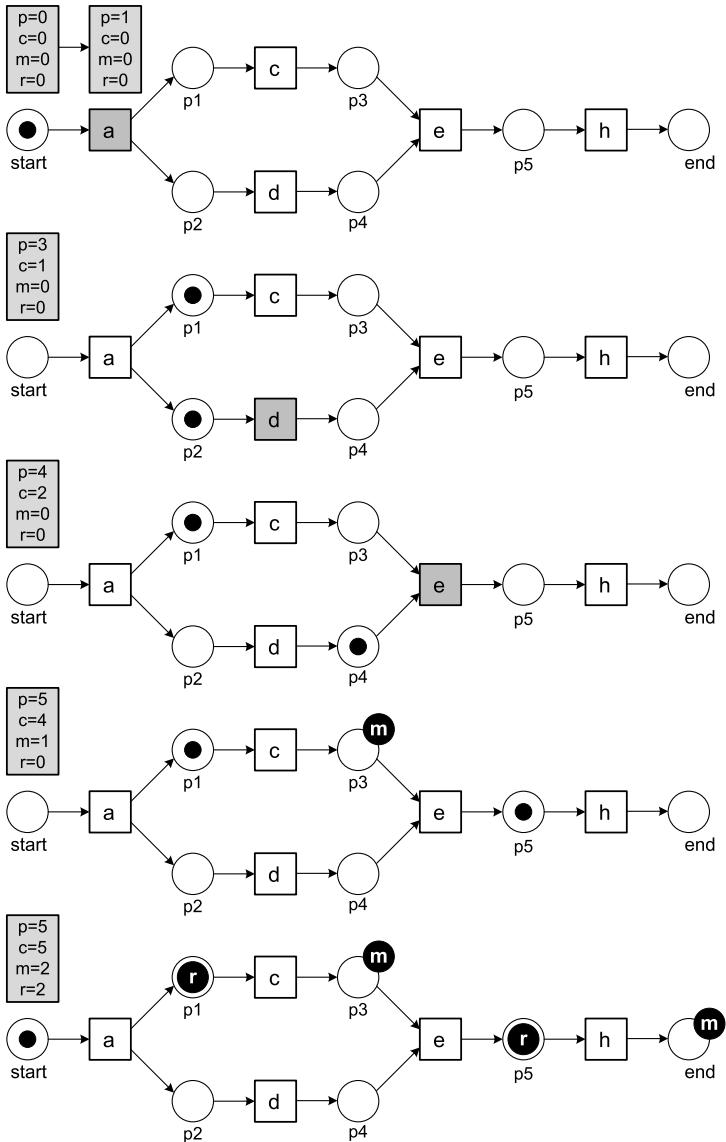


Fig. 8.5 To replay $\sigma_2 = \langle a, b, d, e, g \rangle$ on top of WF-net N_3 , all events not corresponding to activities in the model are removed first. Replaying $\sigma'_2 = \langle a, d, e \rangle$ shows that two tokens are missing ($m = 2$) and two tokens are remaining ($r = 2$) thus resulting in a fitness of 0.6

fire, place $p3$ is still empty and e is not enabled. The missing token is recorded ($m = 1$) and place $p3$ gets an m -tag. After replaying σ'_2 , the resulting marking is $[p1, p5]$. Now the environment needs to consume the token from place end . However, place end is not marked. Therefore, another missing token is recorded ($m = 2$)

and also place *end* gets an m -tag. Moreover, two tokens are remaining: one in place *p1* and one in place *p5*. The places are tagged with an r -tag, and the two remaining tokens are recorded $r = 2$. This way we find a fitness of 0.6 for trace σ_2 and WF-net N_3 based on the values $p = 5$, $c = 5$, $m = 2$, and $r = 2$:

$$\text{fitness}(\sigma_2, N_3) = \frac{1}{2} \left(1 - \frac{2}{5} \right) + \frac{1}{2} \left(1 - \frac{2}{5} \right) = 0.6$$

Moreover, Fig. 8.5 clearly shows the cause of this poor conformance: c was supposed to happen according to the model but did not happen, e happened but was not possible according to the model, and h was supposed to happen but did not happen.

Figures 8.3, 8.4, 8.5 illustrate how to analyze the fitness of a single case. The same approach can be used to analyze the fitness of a log consisting of many cases. Simply take the sums of all produced, consumed, missing, and remaining tokens, and apply the same formula. Let $p_{N,\sigma}$ denote the number of produced tokens when replaying σ on N . $c_{N,\sigma}$, $m_{N,\sigma}$, $r_{N,\sigma}$ are defined in a similar fashion, e.g., $m_{N,\sigma}$ is the number of missing tokens when replaying σ on N . Now we can define the fitness of an event log L on WF-net N :

$$\text{fitness}(L, N) = \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}} \right) + \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}} \right)$$

Note that $\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}$ is total number of missing tokens when replaying the entire event log, because $L(\sigma)$ is the frequency of trace σ and $m_{N,\sigma}$ is the number of missing tokens for a single instance of σ . The value of $\text{fitness}(L, N)$ is between 0 (very poor fitness; none of the produced tokens is consumed and all of the consumed tokens are missing) and 1 (perfect fitness; all cases can be replayed without any problems). Although $\text{fitness}(L, N)$ is a measure focusing on tokens in places, we will interpret it as a measure on events. The intuition of $\text{fitness}(L, N) = 0.9$ is that about 90% of the *events* can be replayed correctly.¹ This is only an informal characterization as fitness depends on missing and remaining tokens rather than events. For instance, a transition that is forced to fire during replay may have multiple empty input places. Note that if two subsequent events are swapped in a sequential process, this results in one missing and one remaining token. This seems reasonable, but also shows that the relation between the proportion of events that cannot be replayed correctly and the proportion of tokens that are missing or remaining is rather indirect.

By replaying the entire event log, we can now compute the fitness of event log L_{full} for the four models in Fig. 8.2:

$$\text{fitness}(L_{full}, N_1) = 1$$

¹In the remainder of this book, we often use this intuitive characterization of fitness, although from a technical point of view this is incorrect as $\text{fitness}(L, N)$ is only an indication of the fraction of events that can be replayed correctly.

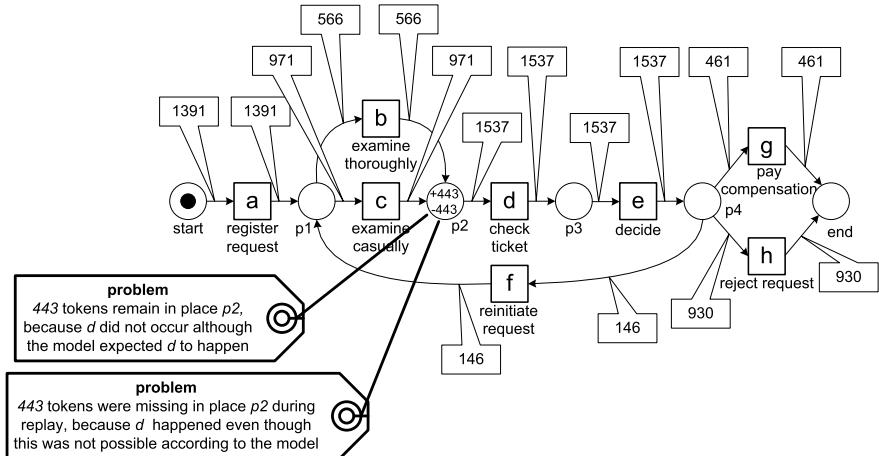


Fig. 8.6 Diagnostic information showing the deviations ($\text{fitness}(L_{full}, N_2) = 0.9504$)

$$\text{fitness}(L_{full}, N_2) = 0.9504$$

$$\text{fitness}(L_{full}, N_3) = 0.8797$$

$$\text{fitness}(L_{full}, N_4) = 1$$

This shows that, as expected, N_1 and N_4 can replay event log L_{full} without any problems (i.e., fitness 1). $\text{fitness}(L_{full}, N_2) = 0.9504$. Intuitively, this means that about 95% of the events in L_{full} can be replayed correctly on N_2 . As indicated earlier, this can be viewed in two ways:

- Event log L_{full} has a fitness of 0.9504, i.e., about 5% of the events deviate; and
- Process model N_2 has a fitness of 0.9504, i.e., the model is unable to explain 5% of the observed behavior.

The first view is used when the model is considered to be normative and correct (“the event log, i.e. reality, does not conform to the model”). The second view is used when the model should be descriptive (“the process model does not conform to reality”). $\text{fitness}(L_{full}, N_3) = 0.8797$, i.e., about 88% of the events in L_{full} can be replayed on N_3 . Hence, process model N_3 has the lowest fitness of the four models.

Typically, the event-based fitness is higher than the naïve case-based fitness. This is also the case here. WF-net N_2 can only replay 68% of the cases from start to end. However, about 95% of the individual events can be replayed.

Figure 8.6 shows some the diagnostics than can be generated based on replaying event log L_{full} on process model N_2 . The numbers on arcs indicate the flow of produced and consumed tokens. These show how cases flowed through the model, e.g., 146 times a request was reinitiated, 930 requests were rejected and 461 requests resulted in a payment. The places tagged during replay (i.e., the m and r -tags in Figs. 8.3, 8.4, and 8.5) can be aggregated to diagnose conformance problems and reveal their severity. As Fig. 8.6 shows, 443 times activity d happened although

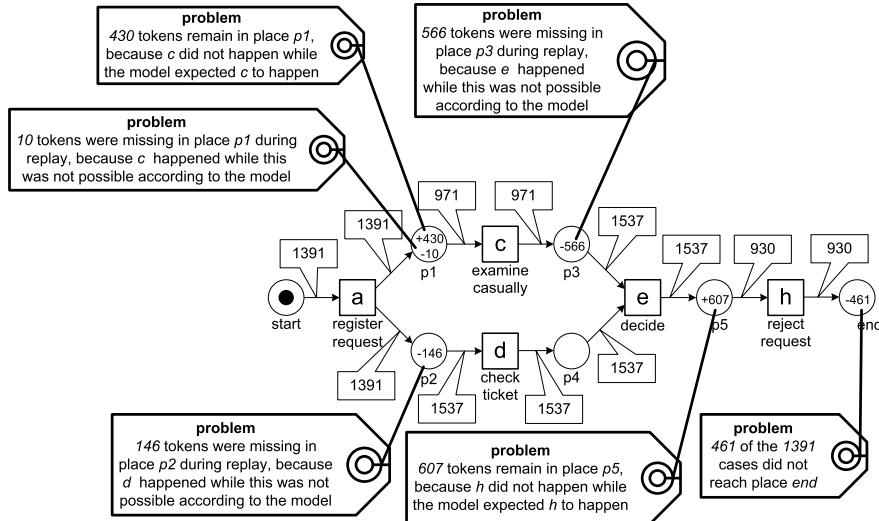


Fig. 8.7 Diagnostic information showing the deviations ($\text{fitness}(L_{\text{full}}, N_3) = 0.8797$)

it was not supposed to happen and 443 times activity *d* was supposed to happen but did not. The reason is that *d* was executed before *b* or *c*, which is not possible according to this sequential model.

Similarly, diagnostic information is shown for N_3 in Fig. 8.7. There the problems are more severe. For example, 566 times a decision was made (activity *e*) without being examined casually (activity *c*), and 461 cases did not reach the end because the request was not rejected.

As Fig. 8.8 shows, an event log can be split into two sublogs: *one event log containing only fitting cases and one event log containing only non-fitting cases*. Each of the event logs can be used for further analysis. For example, one could construct a process model for the event log containing only deviating cases. Also other data and process mining techniques can be used. For instance, it is interesting to know which people handled the deviating cases and whether these cases took longer or were more costly. In case fraud is suspected, one may create a social network based on the event log with deviating cases (see Sect. 9.3).

One could also use classification techniques to further investigate non-conformance. Recall that a decision tree can be learned from a table with one response variable and multiple predictor variables. Whether a case fits or not can be seen as the value of a response variable whereas characteristics of the case (e.g., case and event attributes) serve as predictor variables. The resulting decision tree attempts to explain conformance in terms of characteristics of the case. For example, one could find out that cases from gold customers handled by Pete tend to deviate. We will elaborate on this in Sect. 9.5.

See [14, 119, 121] for more information on token-based replay.

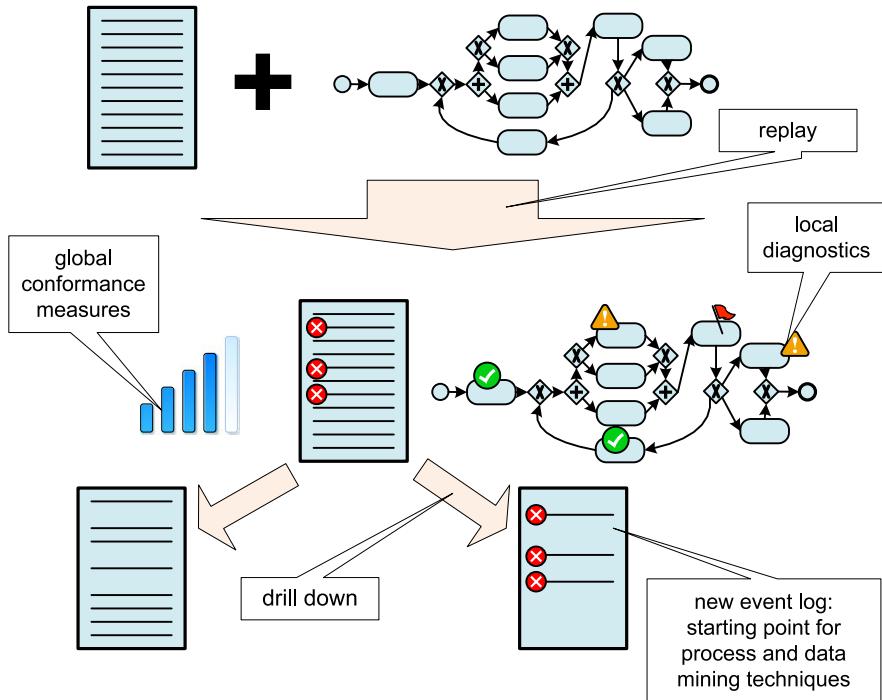


Fig. 8.8 Conformance checking provides global conformance measures like $\text{fitness}(L, N)$ and local diagnostics (e.g., showing activities that were executed although not allowed according to the process model). Moreover, the event log is partitioned into fitting and non-fitting cases. Both sublogs can be used for further analysis, e.g., discovering a process model for the deviating cases

8.3 Alignments

Using token-based replay we can differentiate between fitting and non-fitting cases (see Fig. 8.8). Moreover, the approach is easy to understand and can be implemented efficiently. However, the approach also has some drawbacks. Intuitively, fitness values tend to be too high for extremely problematic event logs. If there are many deviations, the Petri net gets “flooded with tokens” and subsequently allows for any behavior. The approach is also Petri-net specific and can only be applied to other representations after conversion. Moreover, if a case does not fit, the approach does not create a corresponding path through the model. We would like to map observed behavior onto modeled behavior to provide better diagnostics and to relate also non-fitting cases to the model. For example, to compute the mean waiting time between two activities, we cannot leave out all activities that do not fit perfectly. If we would do so, the results could be biased. *Alignments* were introduced to overcome these limitations [169].

To explain the notion of alignments informally, consider trace $\sigma = \langle a, d, b, e, h \rangle$ and the four models in Fig. 8.2. It is easy to see that σ fits perfectly in N_1 and N_4 ,

but not in N_2 and N_3 . A so-called *optimal alignment* is a best match given a trace and a model. Given σ and N_1 there is precisely one optimal alignment,

$$\gamma_1 = \begin{array}{|c|c|c|c|c|} \hline a & d & b & e & h \\ \hline a & d & b & e & h \\ \hline \end{array}$$

The top row corresponds to σ and the bottom row corresponds to a path from the initial marking to the final marking of N_1 .

Given σ and N_2 there are multiple optimal alignments:

$$\gamma_{2a} = \begin{array}{|c|>>|c|c|c|c|} \hline a & >> & d & b & e & h \\ \hline a & b & d & >> & e & h \\ \hline \end{array} \quad \gamma_{2b} = \begin{array}{|c|>>|c|c|c|c|} \hline a & >> & d & b & e & h \\ \hline a & c & d & >> & e & h \\ \hline \end{array} \quad \gamma_{2c} = \begin{array}{|c|c|c|>>|c|c|} \hline a & d & b & >> & e & h \\ \hline a & >> & b & d & e & h \\ \hline \end{array}$$

The “ $>>$ ” symbols denote misalignments. In γ_{2a} , the model makes a “ b move” before d may occur in both log (top row) and model (bottom row). Subsequently, the log makes a “ b move”, not possible anymore in the model. In γ_{2b} , the model makes a “ c move” (rather than a “ b move”) before d . In γ_{2c} , the log first makes a “ d move” not possible in the model, followed by b and a “ d move” made by the model. All three alignments have two $>>$ ’s (“no moves”).

Given σ and N_3 , there are also three optimal alignments:

$$\gamma_{3a} = \begin{array}{|c|>>|c|c|c|c|} \hline a & >> & d & b & e & h \\ \hline a & c & d & >> & e & h \\ \hline \end{array} \quad \gamma_{3b} = \begin{array}{|c|c|>>|c|c|c|} \hline a & d & >> & b & e & h \\ \hline a & d & c & >> & e & h \\ \hline \end{array} \quad \gamma_{3c} = \begin{array}{|c|c|c|>>|c|c|} \hline a & d & b & >> & e & h \\ \hline a & d & >> & c & e & h \\ \hline \end{array}$$

The model needs to make a “ c move” and the log needs to make “ b move” not possible in the model.

Given σ and N_4 , there is just one optimal alignment,

$$\gamma_4 = \begin{array}{|c|c|c|c|c|} \hline a & d & b & e & h \\ \hline a & d & b & e & h \\ \hline \end{array}$$

The alignment shows that σ perfectly fits N_4 : there are no $>>$ ’s signaling discrepancies between modeled and observed behavior.

The examples illustrate the usefulness of alignments. *Detailed diagnostics* can be given per case and these can be aggregated into *diagnostics at the process model level*. For example, we can indicate that a specific activity is often skipped or that some other activity occurs at times it is not supposed to happen. Moreover, observed behavior is related to modeled behavior in a precise manner.

Token-based conformance checking becomes more complicated when there are duplicate and silent activities, e.g., transitions with a τ label or two transitions with the same label. Alignments can be defined for any process notation, including Petri nets having duplicate and silent activities. To illustrate this, consider Fig. 8.9. The labeled Petri net N_5 is composed of 8 transitions and 7 places. Transition t_1 has label a modeling the initial registration step, transition t_2 has label b modeling an examination step, etc. There are two decision transitions (t_4 and t_5) having the same

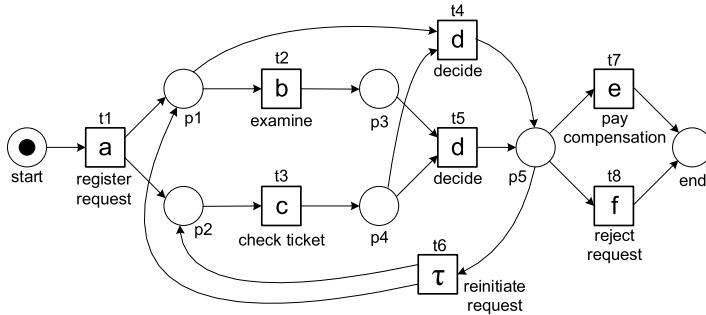


Fig. 8.9 A WF-net N_5 with duplicate and silent activities

label (d). There is one *silent* transition (t_6). This transition models the reinitiation step. This step is invisible as is reflected by the τ label. Given $\sigma_1 = \langle a, c, d, e \rangle$ and N_5 , there is precisely one optimal alignment,

$$\gamma_{5,1} = \begin{array}{|c|c|c|c|} \hline a & c & d & e \\ \hline a & c & d & e \\ \hline t_1 & t_3 & t_4 & t_7 \\ \hline \end{array}$$

The top row of the alignment corresponds to “moves in the log” and the bottom two rows correspond to “moves in the model”. Moves in the model are now represented by both the transition and its label. This is needed because there could be multiple transitions with the same label, e.g., in N_5 both t_4 and t_5 have a d label. The d event in trace $\sigma_1 = \langle a, c, d, e \rangle$ represents a decision and is not connected to a specific transition. However, during replay it becomes clear that d must refer to t_4 .

If a move in the model cannot be mimicked by a move in the log, then a “ \gg ” (“no move”) appears in the top row. Consider $\sigma_2 = \langle a, b, d, f \rangle$. There are two optimal alignments for σ_2 and N_5 :

$$\gamma_{5,2a} = \begin{array}{|c|c|c|c|c|} \hline a & b & \gg & d & f \\ \hline a & b & c & d & f \\ \hline t_1 & t_2 & t_3 & t_5 & t_8 \\ \hline \end{array} \quad \gamma_{5,2b} = \begin{array}{|c|c|c|c|c|} \hline a & \gg & b & d & f \\ \hline a & c & b & d & f \\ \hline t_1 & t_3 & t_2 & t_5 & t_8 \\ \hline \end{array}$$

If a move in the log cannot be mimicked by a move in the model, then a “ \gg ” (“no move”) appears in the bottom row. Consider $\sigma_3 = \langle a, c, d, e, f \rangle$. There are two optimal alignments for σ_3 and N_5 :

$$\gamma_{5,3a} = \begin{array}{|c|c|c|c|c|} \hline a & c & d & e & f \\ \hline a & c & d & e & \gg \\ \hline t_1 & t_3 & t_4 & t_7 & \\ \hline \end{array} \quad \gamma_{5,3b} = \begin{array}{|c|c|c|c|c|} \hline a & c & d & e & f \\ \hline a & c & d & \gg & f \\ \hline t_1 & t_3 & t_4 & & t_8 \\ \hline \end{array}$$

Silent transition t_6 leaves no trail in the event log. Given $\sigma_4 = \langle a, c, d, b, c, d, c, d, c, b, d, f \rangle$ and N_5 , there is precisely one optimal alignment,

$$\gamma_{5,4} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline & a & c & d & \gg & b & c & d & \gg & c & d & \gg & c & b & d & f \\ \hline & a & c & d & \tau & b & c & d & \tau & c & d & \tau & c & b & d & f \\ \hline t1 & t3 & t4 & t6 & t2 & t3 & t5 & t6 & t3 & t4 & t6 & t3 & t2 & t5 & t8 \\ \hline \end{array}$$

The alignment loops back three times. All \gg 's correspond to model moves of silent transition t_6 . These are considered harmless because these moves are invisible and cannot be observed in the log anyway. Hence, we consider $\sigma_4 = \langle a, c, d, b, c, d, c, b, d, f \rangle$ and N_5 to be perfectly fitting.

A *move* is a pair $(x, (y, t))$ where the first element refers to the log and the second element refers to the model. For example, $(a, (a, t_1))$ means that both log and model make an “ a move” and the move in the model is caused by the occurrence of transition t_1 . $(\gg, (c, t_3))$ means that the occurrence of transition t_3 with label c is not mimicked by a corresponding move of the log. (f, \gg) means that the log makes an “ f move” not followed by the model.

$(x, (y, t))$ is a *legal move* if one of the following four cases holds:

- $x = y$ and y is the visible label of transition t (*synchronous move*),
- $x = \gg$ and y is the visible label of transition t (*visible model move*),
- $x = \gg$, $y = \tau$ and transition t is silent (*invisible model move*), or
- $x \neq \gg$ and $(y, t) = \gg$ (*log move*).

Other moves such as (\gg, \gg) and $(x, (y, t))$ with $x \neq y$ are illegal moves.

An *alignment* is a sequence of legal moves such that after removing all \gg symbols, the top row corresponds to the trace in the log, and the bottom row corresponds to a firing sequence starting in some initial state of the process model and ending in some final state. Consider, for example, $\gamma_{5,2a}$. This is an alignment for σ_2 and N_5 because the top row $\langle a, b, \gg, d, f \rangle$ is indeed σ_2 after removing the \gg and the bottom row $\langle t_1, t_2, t_3, t_5, t_8 \rangle$ is indeed a firing sequence leading from [start] to [end].

Given a log trace and a process model, there may be many (if not infinitely many) alignments. For $\sigma_2 = \langle a, b, d, f \rangle$ and N_5 , there are additional alignments like:

$$\gamma_{5,2c} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline & \gg & \gg & \gg & \gg & a & b & c & d & f \\ \hline & \gg & \gg & \gg & \gg & t1 & t2 & t3 & t5 & t8 \\ \hline \end{array} \quad \gamma_{5,2d} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline & a & b & d & f & \gg & \gg & \gg & \gg \\ \hline & a & \gg & \gg & \gg & c & d & e \\ \hline t1 & & t3 & t4 & t7 \\ \hline \end{array}$$

Alignments $\gamma_{5,2a}$ and $\gamma_{5,2b}$ have just one \gg , alignment $\gamma_{5,2c}$ has nine \gg 's, and $\gamma_{5,2d}$ has six \gg 's. Clearly, $\gamma_{5,2a}$ (or $\gamma_{5,2b}$) describes the relation between σ_2 and N_5 better than the two longer alignments $\gamma_{5,2c}$ and $\gamma_{5,2d}$.

To select the most appropriate alignment, we associate *costs* to undesirable moves and select an alignment with the lowest total costs. Cost function δ assigns costs to legal moves. Moves where log and model agree have no costs, i.e., $\delta(x, (y, t)) = 0$ for *synchronous moves* (with $x = y$). Moves in model only have no costs if the transition is invisible, i.e., $\delta(\gg, (\tau, t)) = 0$ for *invisible model moves*.

$\delta(\gg, (y, t)) > 0$ is the cost when the model makes an “ y move” without a corresponding move of the log (*visible model move*). $\delta(x, \gg) > 0$ is the cost for an “ x move” in just the log (*log move*). These costs may depend on the nature of the activity, e.g., skipping a payment may be more severe than sending too many letters.

For simplicity we assume a fixed *standard cost function* which assigns cost 1 to all visible model moves and log moves ($\delta(\gg, (y, t)) = \delta(x, \gg) = 1$ with $y \neq \tau$). The cost of an alignment is simply the sum of the costs of all its moves. For example, $\delta(\gamma_{5,2a}) = \delta(\gamma_{5,2b}) = 1$, $\delta(\gamma_{5,2c}) = 9$, and $\delta(\gamma_{5,2d}) = 6$. $\delta(\gamma_{5,1}) = 0$ indicating that there are no misalignments. Also $\delta(\gamma_{5,4}) = 0$ because $\delta(\gg, (\tau, t6)) = 0$.

An alignment is *optimal* if there is no alternative alignment with lower costs. Obviously, $\gamma_{5,1}$ and $\gamma_{5,4}$ are optimal because the costs are 0 and cannot be lower. $\gamma_{5,2a}$ and $\gamma_{5,2b}$ are optimal alignments for trace σ_2 and model N_5 . Both $\gamma_{5,3a}$ and $\gamma_{5,3b}$ are optimal alignments for σ_3 and N_5 . These examples show that optimal alignments do not need to be unique. However, without loss of generality, we can assume a *deterministic* mapping that assigns any log trace σ to an optimal alignment $\lambda_{opt}^N(\sigma)$ in the context of a particular process model N . Such a mapping is sometimes referred to as an “Oracle”: for any observed behavior a suitably chosen path through the model is returned.

It is possible to convert misalignment costs into a fitness value between 0 (poor fitness, i.e., maximal costs) and 1 (perfect fitness, zero costs). The worst-case scenario is that there are no synchronous moves and only “moves in model only” and “moves in log only”. Note that we can always create an alignment where all events in trace σ are converted to log moves and a shortest path from an initial state to a final state of the model is added as a sequence of model moves. An example of a “worst-case alignment” for $\sigma_2 = \langle a, b, d, f \rangle$ and N_5 is

$$\gamma_{5,2w} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|} \hline & a & b & d & f & \gg & \gg & \gg & \gg \\ \hline \gg & \gg & \gg & \gg & \gg & a & c & d & f \\ \hline & t1 & t3 & t4 & t8 & & & & \\ \hline \end{array}$$

This alignment has “moves in log only” for the observed events in $\sigma_2 = \langle a, b, d, f \rangle$ and “moves in model only” for firing sequence $\langle t1, t3, t4, t8 \rangle$.

A worst-case alignment always yields a valid alignment and there cannot be optimal alignments with higher costs. Let us call this alignment $\lambda_{worst}^N(\sigma)$. Now the fitness of a trace σ can be defined as follows:

$$fitness(\sigma, N) = 1 - \frac{\delta(\lambda_{opt}^N(\sigma))}{\delta(\lambda_{worst}^N(\sigma))}$$

Assuming there is a path from some initial state to some final state, this always yields a value between 0 and 1. For σ_2 and model N_5 , $\delta(\lambda_{opt}^{N_5}(\sigma_2)) = 1$, $\delta(\lambda_{worst}^{N_5}(\sigma_2)) = 8$, and $fitness(\sigma_2, N_5) = 1 - \frac{1}{8} = 0.875$.

The fitness notion can be extended to event logs in a straightforward manner:

$$fitness(L, N) = 1 - \frac{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{opt}^N(\sigma))}{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{worst}^N(\sigma))}$$

Note that $\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda_{opt}^N(\sigma))$ is the sum of all costs when replaying the entire event log using optimal alignments. This is divided by the worst-case scenario to obtain a normalized overall fitness value.

To show alignment-based conformance checking in action, we revisit the event log L_{full} described in Table 8.1 and the four models in Fig. 8.2.

Let us first consider model N_1 in Fig. 8.2. For any trace in L_{full} , the optimal alignment has costs 0. There are 7539 synchronous moves for the 1391 cases. There are no separate log or model moves. Hence, $fitness(L_{full}, N_1) = 1$.

Next we consider model N_2 in Fig. 8.2. This model does not allow for concurrency and cannot handle traces where d occurs before b or c . There are 457 situations in event log L_{full} where d occurs before b or c . For some traces multiple optimal alignments are possible. This enables us to use a deterministic “Oracle” returning a particular optimal alignment. The choice of the Oracle does not influence the fitness computation. By definition these all yield the same fitness value. Consider, for example, $\sigma_8 = \langle a, c, d, e, f, d, b, e, h \rangle$ that occurred 47 times in L_{full} . Two examples of optimal alignments for this trace are (transition names are omitted):

$$\gamma_{8a} = \boxed{a|c|d|e|f| \gg |d| |b| |e|h} \quad \gamma_{8b} = \boxed{a|c|d|e|f| |d| |b| \gg |e|h}$$

For a particular collection of optimal alignments for L_{full} there are 7082 synchronous moves, 457 model moves, and 457 log moves. Hence, $\sum_{\sigma \in L_{full}} L_{full}(\sigma) \times \delta(\lambda_{opt}^{N_2}(\sigma)) = 457 + 457 = 914$. There are 7539 events in L_{full} and the shortest path from the initial marking to the final marking takes 5 model moves. Hence, $\sum_{\sigma \in L_{full}} L_{full}(\sigma) \times \delta(\lambda_{worst}^{N_2}(\sigma)) = 7539 + 1391 \times 5 = 14494$. This is the worst-case scenario. Therefore, $fitness(L_{full}, N_2) = 1 - \frac{914}{14494} = 0.936939$. Note that the fitness value is slightly lower than the fitness value using token based replay. This is caused by the cases where d occurs multiple times before b or c (within the same case). These are not sufficiently penalized using token based replay. A second or third misalignment in the same case is not detected due to a token remaining from the first misalignment.

Figure 8.10 shows the diagnostics based on a particular collection of optimal alignments. Compare the token replay diagnostics in Fig. 8.6 to the alignment-based diagnostics in this figure. Figure 8.6 suggests that d was executed 443 times before b or c . This is not the case as Fig. 8.10 clearly shows. b was executed 170 times after d . c was executed 287 times after d . d was executed $170 + 287 = 457$ times before b or c . The difference between 443 in Fig. 8.6 and the correct 457 in Fig. 8.10 is caused by tokens remaining in place $p2$ after the first iteration.

Next we consider model N_3 in Fig. 8.2. Several activities appearing in the event log do not appear in the model, thus causing unavoidable “moves in log only”. Taking a particular collection of optimal alignments, we note that there are 6064 synchronous moves, 891 model moves, and 1475 log moves. Hence, $\sum_{\sigma \in L_{full}} L_{full}(\sigma) \times \delta(\lambda_{opt}^{N_3}(\sigma)) = 891 + 1475 = 2366$. The worst-case scenario still has costs 14494 since there are 7539 events in L_{full} and the shortest path from the

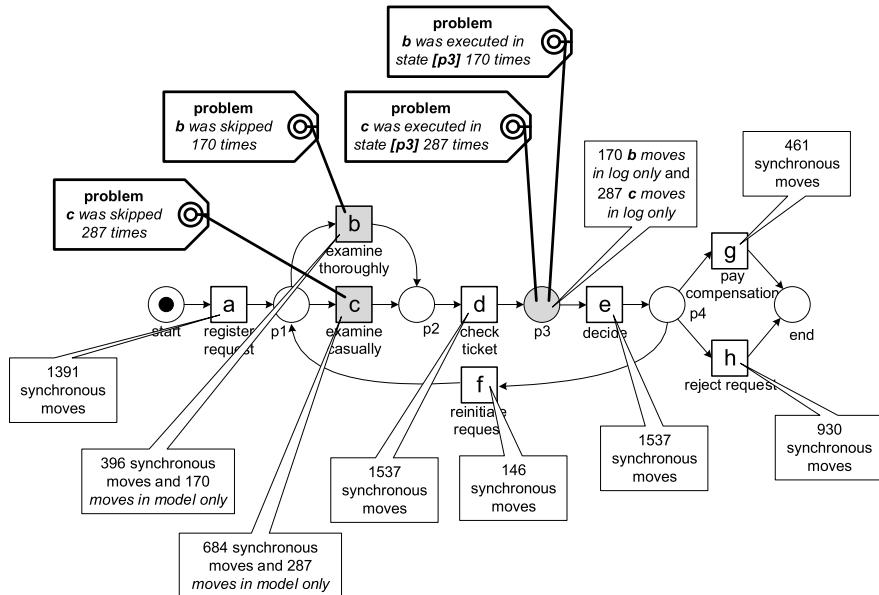


Fig. 8.10 Diagnostic information showing the deviations ($\text{fitness}(L_{full}, N_2) = 0.936939$)

initial marking to the final marking is still 5 steps. Therefore, $\text{fitness}(L_{full}, N_3) = 1 - \frac{2366}{14494} = 0.83676$.

Figure 8.11 shows the diagnostics based on this particular collection of optimal alignments. Activity c in the model was skipped 430 times in the event log and activity h was skipped 461 times. This explains the 891 “moves in model only”. The 1475 log moves are scattered over the different states of the model. Figure 8.7 shows that c was executed 971 times. However, as Fig. 8.11 shows, activity c was executed 961 times at a time allowed by the model. The $971 - 961 = 10$ occurrences of c happened in the second or third iteration which are non-existent in N_3 . Hence, the 971 in Fig. 8.7 is misleading. Unlike token-based replay, alignments map all traces in the log onto *actually existing* firing sequences from an initial state to a final state.

Finally, we align L_{full} with the “flower model” N_4 in Fig. 8.2. As expected, there are 7539 synchronous moves and no “moves in model only” and no “moves in log only”. Hence, $\text{fitness}(L_{full}, N_4) = 1$.

Based on the above examples, we conclude that the following differences exist between token-based and alignment-based conformance checking:

- Alignments provide more *detailed* but *easy to understand diagnostics*. Skipped and inserted events are easier to interpret than missing and remaining tokens.
- Alignments provide more *accurate diagnostics*. Token-based replay may provide misleading diagnostics due to remaining tokens (earlier deviations mask later deviations). As a result fitness values are generally too low.

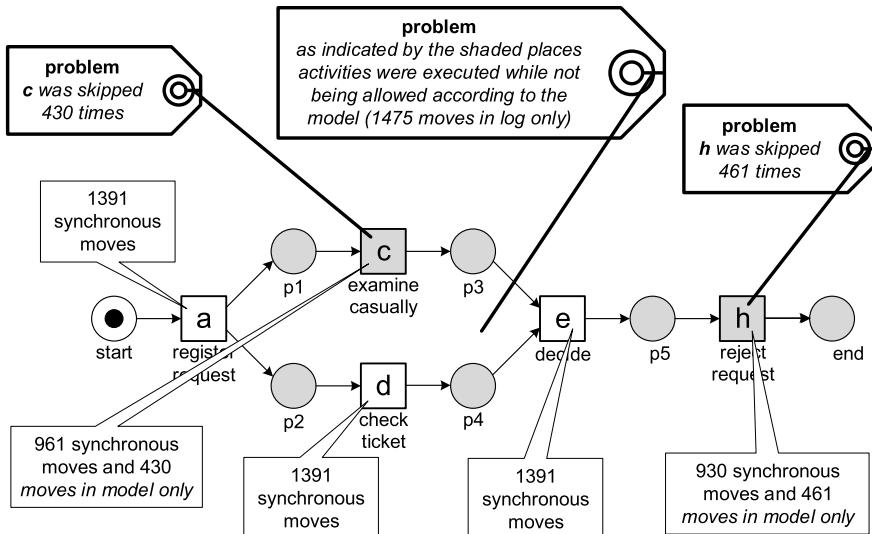


Fig. 8.11 Diagnostic information showing the deviations ($\text{fitness}(L_{full}, N_3) = 0.83676$)

- Alignments are *configurable* through the cost function. One can use multiple cost functions depending on the likelihood of a deviation and its severity [2].
- Alignments can be used to *map each case onto a feasible path in model*. This is important for projecting information (e.g., bottlenecks) on models. Moreover, the mapping ensures that non-fitting extra behavior is not causing misleading diagnostics. Token-based replay also relates observed and modeled behavior, but does not create the corresponding end-to-end execution sequences in the model.
- Alignments are *model independent*. Any process model with formal semantics and initial and final states can be used. Token-based replay assumes a Petri net, so conversions may be needed (e.g., from BPMN to Petri nets).
- Token-based replay provides *deterministic diagnostics* whereas multiple optimal alignments may exist for a trace. This can be addressed by deterministically picking one of possibly many optimal alignments. This does not influence the overall fitness value, but influences diagnostics based on alignments. Multiple optimal alignments can be returned for the same case, but this further complicates interpretation.

See [2–5, 169] for more precise alignment definitions and examples. As mentioned, the idea to align event logs and process models is not limited to Petri nets. Any process modeling notation with executable semantics can replay the event log in some way. See also the replay techniques used in [12, 66, 183, 184].

8.4 Comparing Footprints

In Sect. 6.2, we defined the notion of a *footprint*, i.e., a matrix showing causal dependencies. Such a matrix characterizes the event log. For instance, Table 8.2 shows the

Table 8.2 Footprint of L_{full} and N_1

	a	b	c	d	e	f	g	h
a	#	\rightarrow	\rightarrow	\rightarrow	#	#	#	#
b	\leftarrow	#	#	\parallel	\rightarrow	\leftarrow	#	#
c	\leftarrow	#	#	\parallel	\rightarrow	\leftarrow	#	#
d	\leftarrow	\parallel	\parallel	#	\rightarrow	\leftarrow	#	#
e	#	\leftarrow	\leftarrow	\leftarrow	#	\rightarrow	\rightarrow	\rightarrow
f	#	\rightarrow	\rightarrow	\rightarrow	\leftarrow	#	#	#
g	#	#	#	#	\leftarrow	#	#	#
h	#	#	#	#	\leftarrow	#	#	#

Table 8.3 Footprint of N_2 shown in Fig. 8.2

	a	b	c	d	e	f	g	h
a	#	\rightarrow	\rightarrow	#	#	#	#	#
b	\leftarrow	#	#	\rightarrow	#	\leftarrow	#	#
c	\leftarrow	#	#	\rightarrow	#	\leftarrow	#	#
d	#	\leftarrow	\leftarrow	#	\rightarrow	#	#	#
e	#	#	#	\leftarrow	#	\rightarrow	\rightarrow	\rightarrow
f	#	\rightarrow	\rightarrow	#	\leftarrow	#	#	#
g	#	#	#	#	\leftarrow	#	#	#
h	#	#	#	#	\leftarrow	#	#	#

Table 8.4 Differences between the footprints of L_{full} and N_2 . The event log and the model “disagree” on 12 of the 64 cells of the footprint matrix

	a	b	c	d	e	f	g	h
a					$\rightarrow : \#$			
b					$\parallel : \rightarrow$	$\rightarrow : \#$		
c					$\parallel : \rightarrow$	$\rightarrow : \#$		
d	$\leftarrow : \#$	$\parallel : \leftarrow$	$\parallel : \leftarrow$				$\leftarrow : \#$	
e		$\leftarrow : \#$	$\leftarrow : \#$					
f					$\rightarrow : \#$			
g								
h								

footprint matrix of L_{full} . This matrix is derived from the “directly follows” relation $>_{L_{full}}$. Clearly, process models also have a footprint: simply generate a complete event log, i.e., Play-Out the model and record execution sequences. From the viewpoint of a footprint matrix, an event log is complete if and only if all activities that can follow one another do so at least once in the log. Applying this to N_1 in Fig. 8.2 results in the same footprint matrix (i.e., Table 8.2). This suggests that the event log and the model “conform”.

Table 8.3 shows the footprint matrix generated for WF-net N_2 , i.e., Play-Out N_2 to record a complete log and derived its footprint. Comparing both footprint matrices (Tables 8.2 and 8.3) reveals several differences as shown in Table 8.4. For example, the relation between a and d changed from \rightarrow to $\#$. When comparing event log L_{full}

with WF-net N_2 it can indeed be seen that in L_{full} activity a is directly followed by d whereas this is not possible in N_2 . The relation between b and d changed from \parallel to \rightarrow . This reflects that in WF-net N_2 both activities are no longer parallel. Besides providing detailed diagnostics, Table 8.4 can also be used to quantify conformance. For instance, 12 of the 64 cells differ. Hence, one could say that the conformance based on the footprints is $1 - \frac{12}{64} = 0.8125$.

Conformance analysis based on footprints is only meaningful if the log is complete with respect to the “directly follows” relation $>_L$. This can be verified using k -fold cross-validation (see Sect. 4.6.2).

Interestingly, both models and event logs have footprints. This allows for *log-to-model* comparisons as just described, i.e., it can be checked whether a model and log “agree” on the ordering of activities. However, the same approach can be used for *log-to-log* and *model-to-model* comparisons. Comparing the footprints of two process models (model-to-model comparison) allows for the quantification of their similarity. Comparing the footprints of two event logs (log-to-log comparison) can, for example, be used for detecting *concept drift*. The term concept drift refers to the situation in which the process is changing while being analyzed. For instance, in the beginning of the event log two activities may be concurrent whereas later in the log these activities become sequential. This can be discovered by splitting the log into smaller logs and analyzing the footprints of the smaller logs. A log-to-log comparison of a sequence of event logs may reveal concept drift. Such a “second order process mining” requires lots of data because all the smaller logs are assumed to be complete with respect to $>_L$.

A topic typically neglected in literature and tools is the *cross-validation of conformance*. The event log is just a sample of behavior. This sample may be too small to make a reliable statement about conformance. Moreover, there may be additional complications like concept drift. For example, the average conformance over 2011 is 0.80, however, in the beginning of the year it was 0.90, but during the last two months conformance has been below 0.60. Most techniques provide a single conformance metric without stating anything about the reliability of the measure or concept drift. For instance, suppose we have a large event log L_1 and a small event log L_2 such that $L_2 \subset L_1$ and $|L_2| = 0.01 \times |L_1|$, i.e., L_2 contains 1% of the cases in L_1 . Suppose that $fitness(L_1, N) = 0.9$ and $fitness(L_2, N) = 0.6$. Clearly, the first value is much more reliable (as it is based on a log 100 times larger) but this is not expressed in the metric. If there is enough data to do cross-validation, the event log could be split randomly into k parts (see also Sect. 4.6.2). Then the fitness could be computed for all k parts. These k independent measures could then be used to create a confidence interval for the conformance of the underlying process, e.g., the fitness is, with 90% confidence, between 0.86 and 0.94. Some of the conformance measures have a tendency to go up or down when the event log is larger or smaller. Whereas token replay and alignments are insensitive to the size of the log, other measures like the ones based on the footprint matrix depend on the size and completeness of the event log. Consider, for example, an event log L split into two smaller logs L_1 and L_2 . Assuming that the process is in steady state, the expected value for $fitness(L, N)$ is identical to the expected value for $fitness(L_1, N)$

and $\text{fitness}(L_2, N)$. This does not hold for measures like the footprint matrix: relation $>_L$ can only grow if the log gets larger. Relative thresholds, as used for heuristic mining, may be used to reduce this effect.

The footprint is just one of many possible characterizations of event logs and models. In principle, any temporal property can be used. Instead of the “directly follows” relation also an “eventually follows” relation \gg_L can be used. $a \gg_L b$ means that there is at least one case for which a was eventually followed by b . This can also be combined with some time window, e.g., a was followed by b within four steps or a was followed by b within four hours. It is also possible to take frequencies into account (see for example measures such as $|a >_L b|$ and $|a \Rightarrow_L b|$ defined in the context of heuristic mining) and use thresholds. Clearly, the characterizations used to compare logs and models should match the notion of conformance one is interested in.

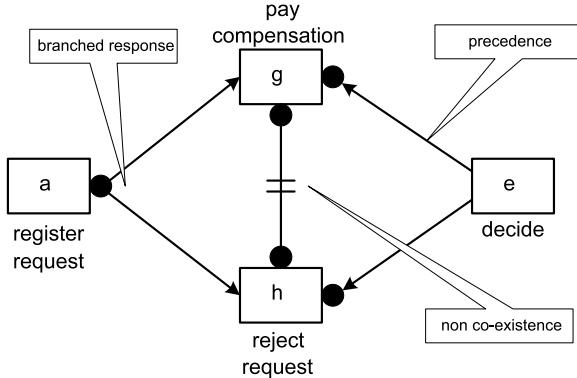
The token-based replay technique described in Sect. 8.2, the alignments in Sect. 8.3, and the comparison of footprint matrices can be used to check the conformance of an event log and a *whole* process model. It is of course also possible to directly check specific *constraints*, sometimes referred to as “business rules”. An example of a constraint is that activity a should always be followed by activity b . Another example is the so-called *4-eyes principle*: activities a and b should never be executed by the same person, e.g., to avoid fraud. In [158], it is shown how an LTL-based language can be used in the context of process mining. *Linear Temporal Logic* (LTL) is an example of a temporal logic that, in addition to classical logical operators, uses temporal operators such as: always (\square), eventually (\diamond), until (\sqcup), weak until (W), and next time (\circ) [30]. For instance, one could formulate the rule $\square(a \Rightarrow \diamond(g \vee h))$, i.e., if a occurs, it should eventually be followed by g or h . Another example is the rule $\diamond g \Leftrightarrow \neg(\diamond h)$ stating that eventually either g should happen or h should happen, but not both. The directly follows relation $a >_L b$ used earlier can be expressed in LTL: “ $\diamond(a \wedge \circ(b))$ ” (for at least one case). This illustrates that behavioral characterizations such as footprints can often be expressed in terms of LTL. LTL-based constraints as defined in [158] may also include explicit time and data. For example, it is possible to state that within two days after the occurrence of activity e , one of the activities g or h should have happened. Another example is that for gold customers the request should be handled in one week and for silver customers in two weeks.

Constraints may be used to split the log into two parts as described in Fig. 8.8. This way it is possible to further investigate cases that violate some business rule.

Declare: A constraint-based workflow language

In this book, we focus on mainstream process modeling languages like Petri nets, BPMN, EPCs, and YAWL. These languages are procedural and aim to describe end-to-end processes. In the context of conformance checking it is interesting to also consider declarative process modeling languages. *Declare* is such a language (in fact a family of languages) and a fully func-

Fig. 8.12 *Declare* specification consisting of four constraints: two precedence constraints, one non-coexistence constraint, and one branched response constraint



tional WFM system [103, 162]. Declare uses a graphical notation and semantics based on LTL. Figure 8.12 shows a Declare specification consisting of four constraints. The construct connecting activities *g* and *h* is a so-called *non-coexistence constraint*. In terms of LTL this constraint means “ $\neg((\Diamond g) \wedge (\Diamond h))$ ”; $\Diamond g$ and $\Diamond h$ cannot both be true, i.e., it cannot be the case that both *g* and *h* happen for the same case. There are two *precedence constraints*. The semantics of the precedence constraint connecting *e* to *g* can also be expressed in terms of LTL: “ $(\neg g) W e$ ”, i.e., *g* should not happen before *e* has happened. Since the weak until (*W*) is used in “ $(\neg g) W e$ ”, traces without any *g* and *e* events also satisfy the constraint. Similarly, *h* should not happen before *e* has happened: “ $(\neg h) W e$ ”. The constraint connecting *a* to *g* and *h* is a so-called branched constraint involving three activities. This *response constraint* states that every occurrence of *a* should eventually be followed by *g* or *h*: “ $\Box(a \Rightarrow (\Diamond(g \vee h)))$ ”. The latter constraint allows for $\langle a, a, a, g, h, a, a, h \rangle$ but not $\langle a, g, h, a \rangle$. Example traces that satisfy all four constraints are $\langle a, a, e, e, g \rangle$ and $\langle a, e, h, e \rangle$.

Procedural languages only allow for activities that are explicitly triggered through control-flow (token semantics). In a declarative language like Declare “*everything is possible unless explicitly forbidden*”.

The Declare language is supported by a WFM system that is much more flexible than traditional procedural WFM/BPM systems [162]. Moreover, it is possible to learn Declare models by analyzing event logs [86, 103]. The graphical constraint language is also suitable for conformance checking. Given an event log, it is possible to check all the constraints. Consider, for instance, Fig. 8.12. Given an event log one can show for each constraint the proportion of cases that respects this constraint [103, 162]. In case of conformance checking, complex time-based constraints may be used (e.g., after every occurrence of activity *a* for a gold customer, activity *g* or *h* should happen within 24 hours).

8.5 Other Applications of Conformance Checking

Conformance checking can be used for improving the alignment of business processes, organizations, and information systems. As shown, replay techniques and footprint analysis help to identify differences between a process model and the real process as recorded in the event log. The differences identified may lead to changes of the model or process. For example, exposing deviations between the model and process may lead to better work instructions or changes in management. Conformance checking is also a useful tool for auditors that need to make sure that processes are executed within the boundaries set by various stakeholders.

In this section, we show that conformance checking can be used for other purposes such as repairing models and evaluating process discovery algorithms. Moreover, through conformance checking event logs get connected to process models and thus provide a basis for all kinds of analysis.

8.5.1 Repairing Models

When a process model and an event log “disagree” on the process, this should lead to adaptations of the model or the process itself. Let us assume that we want to use conformance checking to *repair the model*, i.e., to align it with reality. The diagnostics provided in Figs. 8.6 and 8.7 can be used to (semi-)automatically repair the model. For instance, paths that are never taken can be removed from the model. Note that Figs. 8.6 and 8.7 show the frequency of activities and their causal dependencies. This may lead to the removal of activities that are never (or seldom) executed or the removal of choices. Token replay does not help to remove concurrency that is never used (e.g. activities modeled in parallel but executed in sequence). However, this can be seen in the footprint matrix. After removing unused parts of the model, the m and r -tags pointing to missing and remaining tokens can be used to repair the model. An m -tag points out activities that happened in the process but that were not possible according to the model. An r -tag points out activities that did not happen but that were supposed to happen according to the model. Comparing the footprint matrices of the log and model will show similar problems. Such information can be used by a designer to repair the model. In principle, it is possible to do this automatically. For example, given a set of edit operations on the model one could look for the model that is “closest” to the original model but that has a fitness of, say, more than 0.9. It is fairly straightforward to develop a genetic algorithm that minimizes the edit distance while ensuring a minimal fitness level: edit operations to repair a model are closely related to the genetic operations (mutation and crossover) described in Sect. 7.3. See [52] for a concrete approach to repair process models using event data.

8.5.2 Evaluating Process Discovery Algorithms

The focus of this chapter has been on conformance checking and quantifying fitness, i.e., measuring the ability to replay observed behavior on a predefined process model. However, fitness is just one of four quality dimensions and conformance checking is also related to the evaluation of process discovery techniques. Therefore, we provide a few pointers to literature.

In Sect. 6.4, we discussed the challenges that process discovery algorithms are facing: incompleteness, noise, etc. Process discovery is a complex task and many algorithms have been proposed in literature. As discussed in [122], it is not easy to compare the different algorithms. Compared to classical data mining challenges, there seem to be much more dimensions both in terms of *representation* (see, for instance, the more than 40 control-flow patterns gathered in the context of the Workflow Patterns Initiative [155, 191]) and *quality criteria* (even for the fitness notion several definitions exist). In [43], several process discovery techniques are evaluated using real-life event logs and multiple criteria. The study does not include recent approaches like inductive mining, but shows that automated comprehensive evaluations are possible.

In Sect. 6.4.3, we described four quality dimensions: *fitness*, *simplicity*, *precision*, and *generalization*. Obviously, conformance checking is closely related to measuring the fitness of a discovered model. Whether the model used for conformance checking is made by hand or discovered using some process mining algorithm is irrelevant for the techniques presented in this chapter. Hence, conformance checking, as described in this chapter, can also be used to evaluate and compare process discovery algorithms. However, the “flower model” N_4 in Fig. 8.2 illustrates that fitness covers just one dimension. Simplicity, precision, and generalization also need to be considered when evaluating a discovered model. Leaving out one dimension may lead to degenerate models as shown in [26, 169].

Obviously, a process discovery algorithm should aim to generate the *simplest model* possible that is able to explain the observed behavior. See [101] for an overview of metrics used to quantify the complexity and understandability of a process model. The metrics consider aspects such as the size of the model (e.g., the number of nodes and/or arcs) and the “structuredness” or “homogeneity” of the model [101].

A model with a severe lack of *precision* is “underfitting” and will, on average, have too many enabled transitions during replay. See [5, 26, 105, 121, 169] for approaches to qualify precision.

Precision: Avoid underfitting

Precision can be quantified in different ways. Here, we sketch the approach used in [105, 169]. Let $E \subseteq \mathcal{E}$ be the set of events in some event log L . M is the corresponding process model having a set of states S . Let A be the set of activities and assume the default classifier $\underline{e} = \#_{activity}(e) \in A$ is the *activity* associated to event $e \in E$ (see Sect. 5.2).

Assume the log has been aligned and was “squeezed” into model M using alignments (see Sect. 8.3). Hence, without loss of generality we may assume that each event e fits into the model and that we are able to compute $\#_{state}(e) \in S$. This is the *state* just *before* the occurrence of event e . Using optimal alignments, events (log moves) are synchronized with model moves, yielding a deterministic mapping from events to model states.

Let $en_M(e) \subseteq A$ be the set of activities enabled in the model in $\#_{state}(e)$. Let $\#_{hist}(e) \in A^*$ be the history of e , i.e., the sequence of activities executed for the same case until e . $\#_{hist}(e)$ does not include the latest activity corresponding to e , but the sequence of activities leading to e . $en_L(e) = \{\#_{activity}(e') \mid e' \in E \wedge \#_{hist}(e') = \#_{hist}(e)\} \subseteq A$ is the set of activities that were executed by events having the same history. We assume that events with the same history are mapped onto the same state, i.e., $\#_{hist}(e') = \#_{hist}(e)$ implies $\#_{state}(e') = \#_{state}(e)$. This is the case for most process modeling notations (BPMN, Petri nets, UML activity diagrams, etc.).

If precision is high, the model does not allow for much more behavior than observed. Hence, $|en_M(e)| \approx |en_L(e)|$. If precision is low, the model allows for much more behavior than observed. Hence, $|en_M(e)| \gg |en_L(e)|$. Precision can now be defined as follows:

$$precision(L, M) = \frac{1}{|E|} \sum_{e \in E} \frac{|en_L(e)|}{|en_M(e)|}$$

By definition, $en_L(e) \subseteq en_M(e)$ because the event log is perfectly fitting. Therefore, $0 < precision(L, M) \leq 1$ (assuming $E \neq \emptyset$). If all behavior allowed by the model is actually observed, then $precision(L, M) = 1$. If the model allows for much more behavior than observed, then $precision(L, M) \ll 1$. By taking the average over all events, we automatically take frequencies into account. If the model has an activity that is enabled on a frequent path but the activity is never executed, then this is more severe than an unused activity enabled along an infrequent path.

Figure 8.13 illustrates the precision computation. Three events (e_7 , e_8 , and e_9) share the same history and therefore also occur in the same state ($[p5, p5]$). In this state, four activities are possible (d , e , f , and g), but only two occur in the event log, d (events e_7 and e_9) and e (event e_8). There are no events having the same history corresponding to the occurrence of f or g . Therefore, $en_L(e_7) = en_L(e_8) = en_L(e_9) = \{d, e\}$ and $en_M(e_7) = en_M(e_8) = en_M(e_9) = \{d, e, f, g\}$. If the scope is limited to the three events in Fig. 8.13, then $precision(L, M) = \frac{1}{3}(\frac{2}{4} + \frac{2}{4} + \frac{2}{4}) = 0.5$. Of course, we would need to consider all events to compute the overall precision.

To illustrate the precision metric, consider the two perfectly fitting models in Fig. 8.2 (N_1 and N_4). For the other two models, alignment computations are needed to first “squeeze” the observed behavior into the model [5]. Using

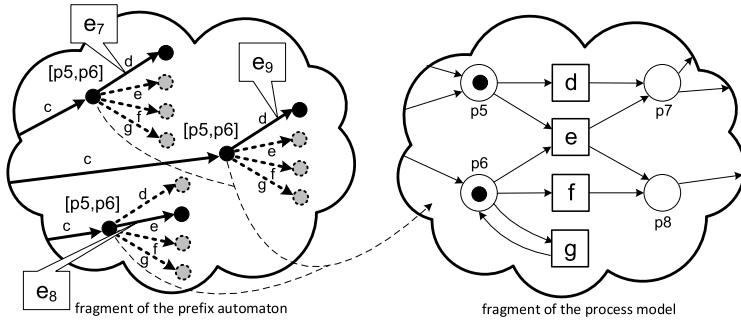


Fig. 8.13 Computing precision: $\#_{state}(e_7) = \#_{state}(e_8) = \#_{state}(e_9) = [p5, p6]$, $en_M(e_7) = en_M(e_8) = en_M(e_9) = \{d, e, f, g\}$, and $en_L(e_7) = en_L(e_8) = en_L(e_9) = \{d, e\}$

ProM, we find $precision(L, N_1) = 0.955$ and $precision(L, N_4) = 0.304$. This matches our intuition, N_1 is much more precise than the “flower model” N_4 that is indeed severely underfitting.

The precision computation just sketched can be applied to *any* type of process model for which optimal alignments can be computed (i.e., not just Petri nets). Using function $\#_{hist}$ to group events, we are implicitly creating a so-called “prefix automaton” (cf. Fig. 8.13). However, other abstractions (next to $\#_{hist}$) are possible as discussed in [2, 5, 169].

In general, a process model should not restrict behavior to just the examples seen in the log. A model that does not generalize is “overfitting”: Future observations are likely to deviate from the model. It is difficult to reason about generalization because this refers to unseen examples.

When evaluating a process discovery *algorithm* and not a specific model, one can use *cross-validation* (e.g., k -fold cross-validation or leave-one-out cross-validation) (see Sect. 6.4.2.3). For cross-validation first a process model is learned for a selected part of the event log (e.g., 80% of the cases), called the *training log*. The remaining 20% of the cases forms the *test log*. Then the test log is replayed or aligned using the model learned for the training log. Such a test can be repeated k times when k -folds are used. If the average fitness for the different test logs is good, the discovery technique is able to generalize. If the average fitness is poor, then the discovery technique is clearly overfitting. In the latter case, the discovery technique produces models that are unable to explain future observations.

Cross-validation *cannot* be used to evaluate a *specific* process model (see Sect. 6.4.2.3). When the model is already given, there is no point in creating a test and training log. In such situations, we need to resort to simple frequency-based metrics such as the one presented in [169]. Every event can be seen as an observation of an activity in some state $s \in S$. Suppose that state s is visited n times and that w is the number of different activities observed in this state. Suppose that n

is very large (say 985 visits to the same state) and w is very small (say 3 unique activities observed in the state), then it is unlikely that a new event visiting this state will correspond to an activity not seen before in this state. However, if n and w are of the same order of magnitude, then it is more likely that a new event visiting state s will correspond to an activity not seen before in this state. An estimator can be derived under the Bayesian assumption that there is an unknown number of possible activities in state s and that the probability distribution over these activities follows a multinomial distribution. It estimates the probability that a new observation will reveal a path not seen before. The weight of each state is based on the number of visits. The computed *generalization value* in [169] is close to 0 if it is likely that new events will exhibit behavior not seen before. The computed generalization value is close to 1 if it is unlikely that the next event will reveal new behavior.

Another approach to compute precision and generalization is to *project* process model and event log on smaller sets of activities and compare their behaviors with respect to these activities only. This can be viewed as a generalization of comparing footprints using $k \geq 1$ dimensions and not limited to the “directly follows” relation. Language inclusion on the projected models and logs can be used to compute precision and recall metrics.

Also see the precision and recall metrics in [14] used to compare two models in the context of an event log.

The examples and pointers in this section show that many approaches are available to quantify the four quality dimensions introduced in Sect. 6.4.3. These can be used to objectively assess the quality of process discovery results.

8.5.3 Connecting Event Log and Process Model

While replaying the event log on the process model, events in the log are related to activities in the model, i.e., the event log is *connected* to the process model. This may seem insignificant at first sight. However, this is of crucial importance for the subsequent chapters. *By relating events to activities, information extracted from the log can be used to enrich the process.* For example, timestamps in the event log can be used to make statements about the duration of some modeled activity.

Figure 8.14 shows the class model presented in Sect. 5.4.1 (cf. Fig. 5.9) without case attributes. The process model level and the event level may *exist independent of one another*. People make process models without relating them to (raw) data in the information system and processes generate data while being unaware of the process models that may exist. Notable exceptions are WFM and BPM systems for which such a connection already exists. This is why process-aware systems are not just important for process automation and also serve as a powerful enabler for process analysis. However, for the majority of processes, there is no supporting WFM or BPM system. As a result, process models (if they exist) and event data are at best loosely coupled. Fortunately, both token replay and alignments *can be used to establish a tight coupling between the process model level and the event level*.

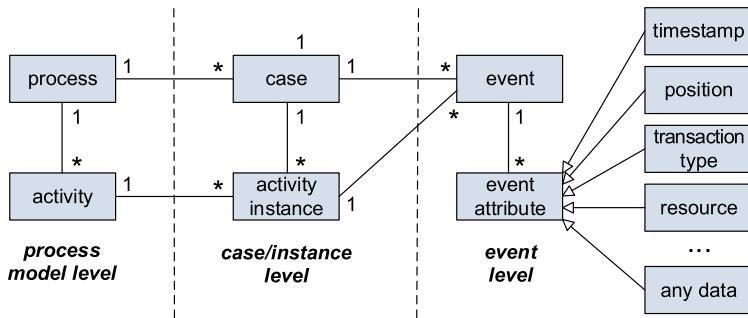


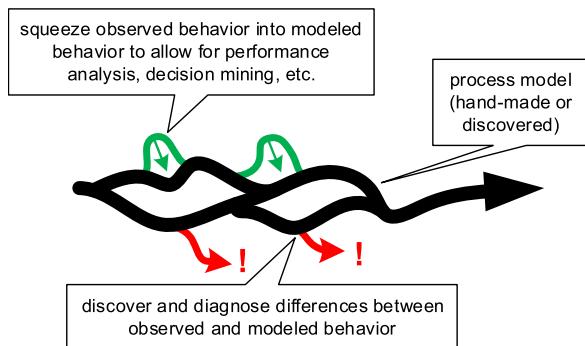
Fig. 8.14 Observed behavior (events) is related to modeled behavior (activities)

The case/instance level shown in Fig. 8.14 consists of *cases* and *activity instances* that connect *processes* and *activities* in the model to *events* in the event log. Within the same case (i.e., process instance) there may be multiple instances of the same activity. For instance, some check activity may be performed multiple times for the same customer request (cf. loops).

Typical event data exist in the form of collections of data records possibly distributed over multiple database tables. A record in such a table may correspond to one or more events while listing certain properties (attributes), e.g., date information, some amount, and credit rating. As discussed in Chap. 5, one of the main challenges is to locate these events and to correlate them. Each event needs to be related to a particular case. When replaying the event log on a model, each event that “fits into the model” is connected to an activity instance. Note that there can be multiple instances of the same activity for each case. Moreover, a single activity instance may correspond to multiple events. Consider a case c with a loop involving activity a . Two instances of a are executed for c , $a_{c,1}$ and $a_{c,2}$. For each of these two activity instances there may be multiple events. For example, the first activity is offered, started, and aborted (three events corresponding to $a_{c,1}$) whereas the second activity is assigned, started, suspended, resumed, and completed (five events corresponding to $a_{c,2}$). See also Sect. 5.2 where the transactional life-cycle is discussed in detail. When describing token replay and alignment computations we did not elaborate on the different *event types* (start, complete, abort, etc.). However, such transactional information can be taken into account when replaying the event log.

In Sect. 8.2, we showed that *token-based replay* can be used to relate observed behavior to modeled behavior. Section 8.3 introduced the notion of *alignments* as an even more direct way of relating observed behavior and modeled behavior. Both approaches can be used to detect and diagnose deviations as sketched in Fig. 8.15. Moreover, alignments can also be used to “squeeze” reality into the model for further analysis. Even if a case does not fit completely, we can find a corresponding path in the model. If we leave out non-fitting cases, the remaining set of cases is no longer representative for the whole. Therefore, it is important to “squeeze” reality into the model even when there are (minor) discrepancies. Subsequently, event data can be used to “breathe life” into otherwise static process models. As a result, all

Fig. 8.15 The ability to replay event data on a process model helps to detect and diagnose deviations and to “squeeze” reality into the model for further analysis



kinds of information extracted from the event log can be projected onto the model, e.g., showing bottlenecks and highlighting frequent paths. The event attributes in Fig. 8.14 provide valuable information that can be aggregated and mapped onto activities and resources. For instance, timestamps can be used to visualize bottlenecks, waiting times, etc. Resource data attached to events can be used to learn working patterns and allocation rules. Cost information can be projected onto process models to see inefficiencies. The next chapter will elaborate on this.

Chapter 9

Mining Additional Perspectives

Whereas the main focus of process discovery is on the control-flow perspective, event logs may contain a wealth of information relating to other perspectives such as the organizational perspective, the case perspective, and the time perspective. Therefore, we now shift our attention to these other perspectives. Organizational mining can be used to get insight into typical work patterns, organizational structures, and social networks. Timestamps and frequencies of activities can be used to identify bottlenecks and diagnose other performance related problems. Case data can be used to better understand decision-making and analyze differences among cases. Moreover, the different perspectives can be merged into a single model providing an integrated view on the process. Such an integrated model can be used for “what if” analysis using simulation.

9.1 Perspectives

Thus far the focus of this book was on control-flow, i.e., the ordering of activities. The chapters on process discovery and conformance checking often used a so-called “simple event log” as a starting point (see Definition 5.4). However, as discussed in Chap. 5, event logs typically contain much more information. Events and cases can have any number of attributes (see Definitions 5.1 and 5.3). The extension mechanism of XES illustrates how such attributes can be structured and stored. Moreover, as stressed in Sect. 2.2, process mining is not limited to the control-flow perspective. Therefore, we now focus on adding some of the other perspectives.

Figure 9.1 shows a typical scenario. The starting point is an event log and some initial process model. Note that the process model may have been constructed manually or discovered through process mining. Important is that the process model and event log are connected. In Sect. 8.5.3, we showed that the replay approaches used in the context of conformance checking can be used to tightly couple model and log. As discussed using Fig. 8.14, activity instances discovered during replay connect modeled activities to recorded events. This way attributes of events (resources,

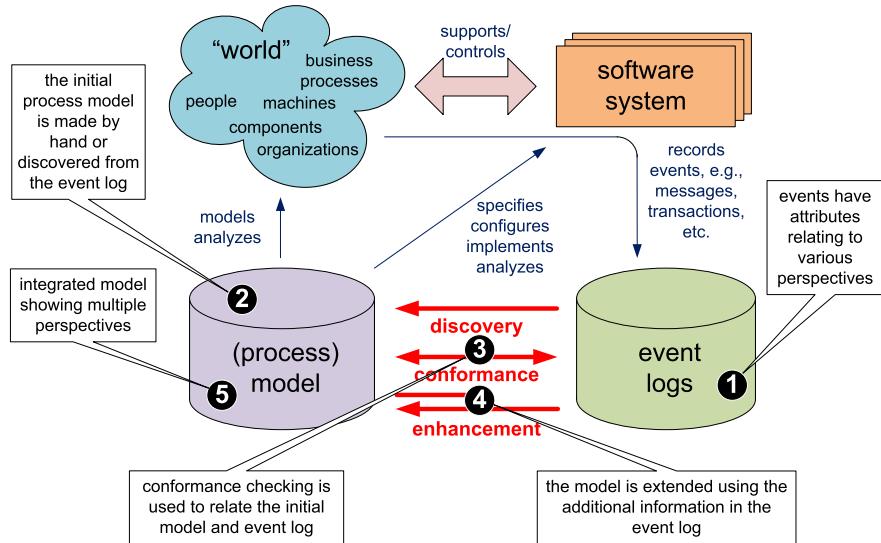


Fig. 9.1 The organizational, case, and time perspectives can be added to the original control-flow model using attributes from the event log

timestamps, costs, etc.) can be used to extend the initial model. For example, information about service or waiting times extracted from the event log can be added to the model. After adding the different perspectives, an *integrated* process model is obtained.

Figure 9.1 lists the three main types of process mining: *discovery*, *conformance*, and *enhancement*. Let us focus on the third type of process mining. Enhancement aims to extend or improve an existing process model using information about the actual process recorded in some event log. One type of enhancement is *repair* as discussed in Sect. 8.5.1. Here, we devote our attention to other type of enhancement: *extension*. Through extension we add a new perspective to the process model by cross-correlating it with the log.

In the remainder, we show some examples of log-based model extension. Section 9.3 discusses various process mining techniques related to the organizational perspective. Here, information about resources is used to analyze working patterns and to see how work “flows” through an organization. Extensions based on the time perspective are discussed in Sect. 9.4. When events bear timestamps it is possible to discover bottlenecks, measure service levels, monitor the utilization of resources, and predict the remaining processing time of running cases. Section 9.5 focuses on other attributes and their effects on decision making. This section illustrates that classical data mining techniques such as decision tree learning can be used to extend a process model with the case perspective. The different perspectives can be merged into a single integrated process model. Section 9.6 shows how such an integrated model can be constructed and used. For instance, a complete simulation model can be mined and subsequently used for “what if” analysis.

Table 9.1 A fragment of some event log: each line corresponds to an event

Case id	Event id	Properties				
		Time	Activity	Trans	Resource	Cost
1	35654423	30-12-2010:11.02	register request	start	Pete	
	35654424	30-12-2010:11.08	register request	complete	Pete	50
	35654425	31-12-2010:10.06	examine thoroughly	start	Sue	
	35654427	31-12-2010:10.08	check ticket	start	Mike	
	35654428	31-12-2010:10.12	examine thoroughly	complete	Sue	400
	35654429	31-12-2010:10.20	check ticket	complete	Mike	100
	35654430	06-01-2011:11.18	decide	start	Sara	
	35654431	06-01-2011:11.22	decide	complete	Sara	200
	35654432	07-01-2011:14.24	reject request	start	Pete	
	35654433	07-01-2011:14.32	reject request	complete	Pete	200
2	35654483	30-12-2010:11.32	register request	start	Mike	
	35654484	30-12-2010:11.40	register request	complete	Mike	50
	35654485	30-12-2010:12.12	check ticket	start	Mike	
	35654486	30-12-2010:12.24	check ticket	complete	Mike	100
	35654487	30-12-2010:14.16	examine casually	start	Pete	
	35654488	30-12-2010:14.22	examine casually	complete	Pete	400
	35654489	05-01-2011:11.22	decide	start	Sara	
	35654490	05-01-2011:11.29	decide	complete	Sara	200
	35654491	08-01-2011:12.05	pay compensation	start	Ellen	
	35654492	08-01-2011:12.15	pay compensation	complete	Ellen	200
...						

9.2 Attributes: A Helicopter View

Before discussing approaches to discover the resource, time, and case perspectives, we provide another example showing the kind of information one can find in a typical event log. Table 9.1 shows a small fragment of a larger event log. Compared to earlier examples, each event now also has a *transaction type*. Consider, for example, the first two events in Table 9.1. The first event refers to the *start* of an activity instance, whereas the second event refers to the *completion* of this instance. By taking the difference between the timestamps of both events, it can be derived that Pete worked for six minutes on case 1 when registering the request of the customer. Only events with transaction type *complete* have a cost attribute. Note that Sue and Mike are both working on the same case at the same time, because activities *examine thoroughly* and *check ticket* for case 1 are overlapping.

Table 9.2 shows the *case attributes* stored in the event log. These are attributes that refer to the case as a whole rather than an individual event (see Definition 5.3). Case 1 is a request initiated by customer *Smith*. This customer has an identification

Table 9.2 Attributes of cases

Case id	Custid	Name	Type	Region	Amount
1	9911	Smith	gold	south	989.50
2	9915	Jones	silver	west	546.00
3	9912	Anderson	silver	north	763.20
4	9904	Thompson	silver	west	911.70
5	9911	Smith	gold	south	812.10
6	9944	Baker	silver	east	788.00
7	9944	Baker	silver	east	792.80
8	9911	Smith	gold	south	544.70
...

number 9911. Customer Smith is a *gold* customer in region *south*. The amount of compensation requested is € 989.50. Cases 5 and 8 are also initiated by the same customer. Case 2 is initiated by silver customer *Jones* from region *west*. This customer claimed an amount of € 546.00.

Each of the events implicitly refers to attributes of the corresponding case. For instance, event 35654483 implicitly refers to silver customer Jones because the event is executed for case 2. In Chap. 5, we formalized the notion of an event log and event attributes. Consider for example $e = 35654431$ and some of its attributes: $\#_{case}(e) = 1$, $\#_{activity}(e) = \text{decide}$, $\#_{time}(e) = 06-01-2011:11.22$, $\#_{resource}(e) = \text{Sara}$, $\#_{trans}(e) = \text{complete}$, $\#_{cost}(e) = 200$, $\#_{custid}(e) = 9911$, $\#_{name}(e) = \text{Smith}$, $\#_{type}(e) = \text{gold}$, $\#_{region}(e) = \text{south}$, and $\#_{amount}(e) = 989.50$. For process discovery, we ignored most of these attributes. This chapter will show how to use these attributes to create an integrated model covering different perspectives.

A first step in any process mining project is to get a feeling for the process and the data in the event log. The so-called *dotted chart* provides a helicopter view of the process [129]. In a dotted chart, each event is depicted as a dot in a two dimensional plane as shown in Fig. 9.2. The horizontal axis represents the *time* of the event. The vertical axis represents the *class* of the event. To determine the class of an event we use a *classifier* as described in Definition 5.2. A classifier is a function that maps the attributes of an event onto a label, \underline{e} is the *class* of the event. An example of a classifier is $\underline{e} = \#_{case}(e)$, i.e., the case id of the event. Other examples are $\underline{e} = \#_{activity}(e)$ (the name of the activity being executed) and $\underline{e} = \#_{resource}(e)$ (the resource triggering the event). In this particular example, $\underline{e} = \#_{region}(e)$ would be a classifier mapping the event onto the region of the customer.

Every line in the dotted chart shown in Fig. 9.2 refers to a class, e.g., if the classifier $\underline{e} = \#_{resource}(e)$ is used, then every line corresponds to a resource. The dots on such a line describe the events belonging to this class, e.g., all events executed by a particular resource. The time dimension can be *absolute* or *relative*. If time is relative, the first event of each case takes place at time zero. Hence, the horizontal position of the dot depends on the time passed since the first event for the same case. The time dimension can be *real* or *logical*. For real time, the actual timestamp is used. For logical time, events are simply enumerated without considering the actual timestamps: only their ordering is taken into account. The first event has time 0, the

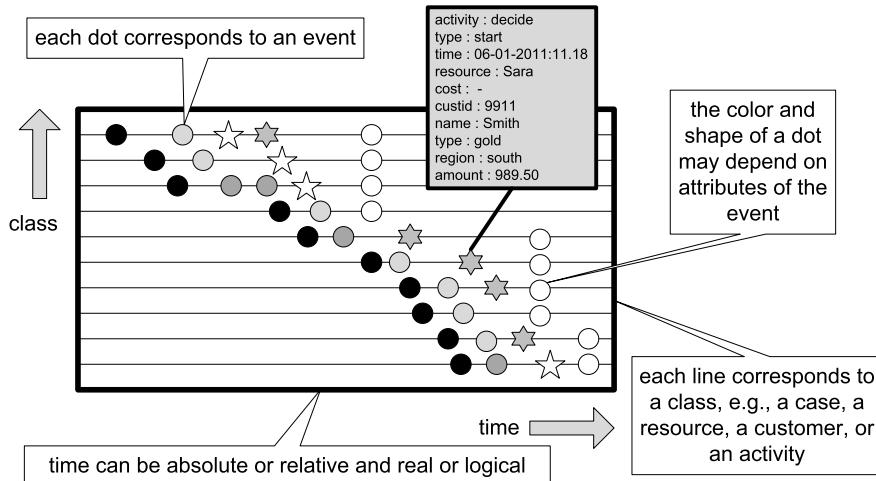


Fig. 9.2 Dotted chart: events are visualized as *dots*. Their position, color, and shape depend on the attributes of the corresponding event

second event has time 1, etc. Also logical time can be absolute (global numbering) or relative (each case starts at time 0).

As Fig. 9.2 shows, the shape and color of a dot may depend on other attributes, i.e., there is also a classifier for the shape of the dot and a classifier for the color of the dot. For instance, if the classifier $e = \#_{case}(e)$ is used, then every line corresponds to a case. The shape of the dot may depend on the resource triggering the corresponding event and the color of the dot may depend on the name of the corresponding activity. In our example, the shape of the dot can also depend on the type of customer (silver or gold) and the color of the dot may depend on the region (north, east, south, or west).

Figure 9.2 shows only a schematic view of the dotted chart. Figures 9.3 and 9.4 show two dotted charts based on a real event log. The event log was extracted from the database of a large Dutch housing agency. The event log contains 5987 events relating to 208 cases and 74 activity names. Each case refers to a housing unit (e.g., an apartment). The case starts when the tenant wants to terminate the current lease and ends when the new tenant has moved into the unit. Both figures show 5987 dots. The classifier $e = \#_{case}(e)$ is used, i.e., every line corresponds to a unit. The color of the dot depends on the name of the corresponding activity. There are 74 colors: one for each of the possible activities. Figure 9.3 uses absolute/real times for the horizontal dimension. Cases are sorted by the time of the first event. These initial events do not form a straight line. If the arrival rate of new cases would be constant, the frontier formed by initial events would resemble a straight line rather than the curve shown in Fig. 9.3. The curved frontier line shows that the arrival process increases in intensity towards the middle of the time window visualized in Fig. 9.3. Moreover, it seems that events are not evenly spread over the time window. There are periods with little activity. Figure 9.4 uses relative/real times. This figure shows

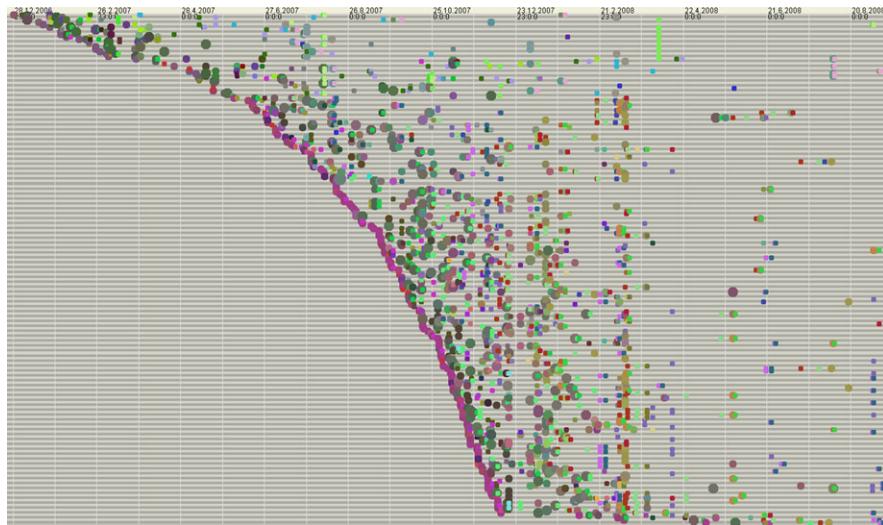


Fig. 9.3 Dotted chart for a process of a housing agency using absolute time. The influx of new cases increases over time. Moreover, several periods with little activity can be identified



Fig. 9.4 Dotted chart for the process of the housing agency using relative time, i.e., all cases start at time zero. The chart reveals a large variation in flow times: some cases are handled in a few days whereas others take more than a year

that there is a huge variation in flow time. About 45% of the cases are handled in less than 150 days whereas about 10% of the cases take more than one year.

The dotted chart is a very powerful tool to view a process from different angles. One can see all events in one glance while potentially showing different perspectives at the same time (class, color, shape, and time). Moreover, by zooming in one can investigate particular patterns. For example, when classifier $e = \#\text{resource}(e)$ is used one can immediately see when a resource has been inactive for a longer period.

In the dotted charts shown in this section, timestamps are used to align events in the horizontal dimension. As shown in [79], it is also possible to align events based on their context rather than time. As a result repeating patterns in the event log are aligned so that it becomes easy to see common behavior and deviations without constructing a process model. The identification of such patterns helps understanding the “raw” behavior captured in the event log. As indicated in Chap. 5, event logs often contain low-level events that are of little interest to management. The challenge is to aggregate low-level events into events that are meaningful for stakeholders. Therefore, the event log is often preprocessed after a visual inspection of the log using dotted charts. There are several approaches to preprocess low-level event logs. For example, frequently appearing low-level patterns can be abstracted into events representing activities at the business level [77]. Also activity-based filtering can be used to preprocess the log. We elaborate on this in Chaps. 14 and 15.

The dotted chart can be seen as an example of a *visual analytics* technique. Visual analytics leverages on the remarkable capabilities of humans to visually identify patterns and trends in large datasets. Even though Fig. 9.3 shows almost six thousand events, people involved in this process can see patterns, trends, and irregularities in one glance.

9.3 Organizational Mining

Organizational mining focuses on the *organizational perspective* [130, 159]. Starting point for organizational mining is typically the $\#\text{resource}(e)$ attribute present in most event logs. Table 9.3 shows a fragment of a larger event log in which each event has a resource attribute; all complete events have been projected onto their resource and activity attributes. This event log is based on the process model from Chap. 2. Using such information, there are techniques to learn more about people, machines, organizational structures (roles and departments), work distribution, and work patterns.

By analyzing an event log as shown in Table 9.3 it is possible to analyze the relation between resources and activities. Table 9.4 shows the mean number of times a resource performs an activity per case. For instance, activity a is executed exactly once for each case (take the sum of the first column). Pete, Mike, and Ellen are the only ones executing this activity. In 30% of the cases, a is executed by Pete, 50% is executed by Pete, and 20% is executed by Ellen. Activities e and f are always executed by Sara. Activity e is executed, on average, 2.3 times per case. The event

Table 9.3 Compact representation of the event log highlighting the resource attribute of each event ($a = \text{register request}$, $b = \text{examine thoroughly}$, $c = \text{examine casually}$, $d = \text{check ticket}$, $e = \text{decide}$, $f = \text{reinitiate request}$, $g = \text{pay compensation}$, and $h = \text{reject request}$)

Case id	Trace
1	$\langle a^{\text{Pete}}, b^{\text{Sue}}, d^{\text{Mike}}, e^{\text{Sara}}, h^{\text{Pete}} \rangle$
2	$\langle a^{\text{Mike}}, d^{\text{Mike}}, c^{\text{Pete}}, e^{\text{Sara}}, g^{\text{Ellen}} \rangle$
3	$\langle a^{\text{Pete}}, c^{\text{Mike}}, d^{\text{Ellen}}, e^{\text{Sara}}, f^{\text{Sara}}, b^{\text{Sean}}, d^{\text{Pete}}, e^{\text{Sara}}, g^{\text{Ellen}} \rangle$
4	$\langle a^{\text{Pete}}, d^{\text{Mike}}, b^{\text{Sean}}, e^{\text{Sara}}, h^{\text{Ellen}} \rangle$
5	$\langle a^{\text{Ellen}}, c^{\text{Mike}}, d^{\text{Pete}}, e^{\text{Sara}}, f^{\text{Sara}}, d^{\text{Ellen}}, c^{\text{Mike}}, e^{\text{Sara}}, f^{\text{Sara}}, b^{\text{Sue}}, d^{\text{Pete}}, e^{\text{Sara}}, h^{\text{Mike}} \rangle$
6	$\langle a^{\text{Mike}}, c^{\text{Ellen}}, d^{\text{Mike}}, e^{\text{Sara}}, g^{\text{Mike}} \rangle$
...	...

Table 9.4 Resource-activity matrix showing the mean number of times a person performed an activity per case

	a	b	c	d	e	f	g	h
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

log conforms to the process model shown in Fig. 2.2. Hence, for some cases e is executed only once whereas for other cases e is executed repeatedly (2.3 times on average). On average, activity f is executed 1.3 times. This suggests that the middle part of the process (composed of activities b , c , d , e , and f) needs to be redone for the majority of cases. Consider for example case 5 in Table 9.3; e is executed three times and f is executed twice for this case.

9.3.1 Social Network Analysis

Sociometry, also referred to as *sociography*, refers to methods that present data on interpersonal relationships in graph or matrix form [182]. The term sociometry was coined by Jacob Levy Moreno who already used such techniques in the 1930s to better assign students to residential cottages in a training facility. Until recently, the input data for sociometry consisted mainly of interviews and questionnaires. However, with the availability of vast amounts of electronic data, new ways of gathering input data are possible.

Here we restrict ourselves to *social networks* as shown in Fig. 9.5. The nodes in a social network correspond to organizational entities. Often, but not always, there is a one-to-one correspondence between the resources found in the log and organizational entities (i.e., nodes). In Fig. 9.5 nodes x , y , and z could refer to persons.

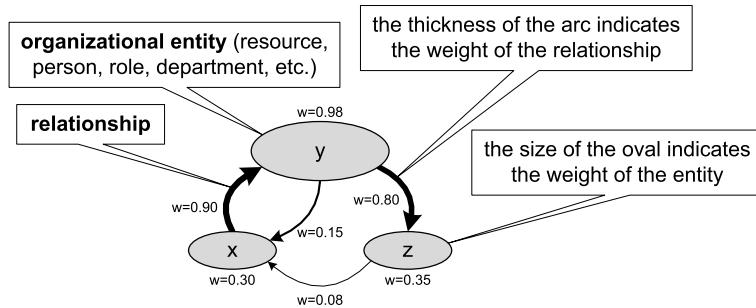


Fig. 9.5 A social network consists of nodes representing organizational entities and arcs representing relationships. Both nodes and arcs can have weights indicated by “*w = …*” and the size of the shape

The nodes in a social network may also correspond to aggregate organizational entities such as roles, groups, and departments. The arcs in a social network correspond to relationships between such organizational entities. Arcs and nodes may have *weights*. The weight of an arc or node indicates its *importance*. For instance, node *y* is more important than *x* and *z* as is indicated by its size. The relationship between *x* and *y* is much stronger than the relationship between *z* and *x* as shown by the thickness of the arc. The interpretation of “importance” depends on the social network. Later, we will give some examples to illustrate the concept.

Sometimes the term *distance* is used to refer to the inverse of the weight of an arc. An arc connecting two organizational entities has a high weight if the distance between both entities is small. If the distance from node *x* to node *y* is large, then the weight of the corresponding arc is small (or the arc is not present in the social network).

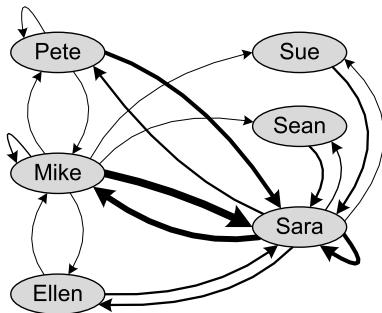
A wide variety of metrics have been defined to analyze social networks and to characterize the role of individual nodes in such a diagram [182]. For example, if all other nodes are in short distance to a given node and all geodesic paths (i.e., shortest paths in the graph) visit this node, then clearly the node is very central (like a spider in the web). There are different metrics for this intuitive notion of *centrality*. The Bavelas–Leavitt index of centrality is a well-known example that is based on the geodesic paths in the graph. Let *i* be a node and let $D_{j,k}$ be the geodesic distance from node *j* to node *k*. The Bavelas–Leavitt index of centrality is defined as $BL(i) = (\sum_{j,k} D_{j,k}) / (\sum_{j,k} D_{j,i} + D_{i,k})$. The index divides the sum of all geodesic distances by the sum of all geodesic distances from and to node *i*. Other related metrics are *closeness* (1 divided by the sum of all geodesic distances to a given node) and *betweenness* (a ratio based on the number of geodesic paths visiting a given node) [159, 182]. Recall that distance can be seen as the inverse of arc weight.

Notions such as centrality analyze the position of one organizational entity, say a person, in the whole social network. There are also metrics making statements about the network as a whole, e.g., the degree of connectedness. Moreover, there are also

Table 9.5 Handover of work matrix showing the mean number of handovers from one person to another per case

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

Fig. 9.6 Social network based on handover of work at the level of individual resources using a threshold of 0.1. The thickness of the arcs is based on the frequency of handovers from one person to another



techniques to identify *cliques* (groups of entities that are strongly connected to each other while having fewer connections to entities outside the clique).

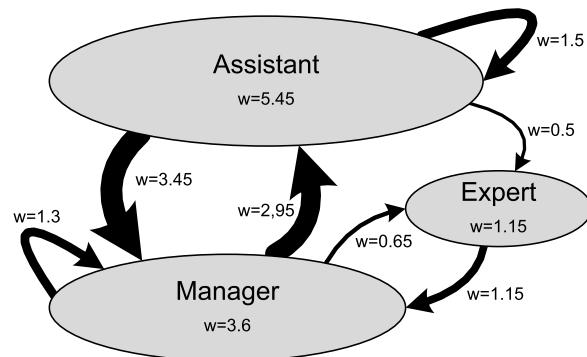
Clearly event logs with $\#_{\text{resource}}(e)$ attributes provide an excellent source of information for social network analysis. For instance, based on the event log one can count the number of times *work is handed over from one resource to another*. Consider for example case 1 having the following trace: $(a^{\text{Pete}}, b^{\text{Sue}}, d^{\text{Mike}}, e^{\text{Sara}}, h^{\text{Pete}})$. Clearly, there is a handover of work from Pete to Sue and Mike after the completion of a . Note that Sue does not hand over work to Mike, because b and d are concurrent. However, both Sue and Mike hand over work to Sara, because activity e requires input from both b and d . Finally, Sara hands over work to Pete. Hence, in total there are five handovers: $(a^{\text{Pete}}, b^{\text{Sue}})$, $(a^{\text{Pete}}, d^{\text{Mike}})$, $(b^{\text{Sue}}, e^{\text{Sara}})$, $(d^{\text{Mike}}, e^{\text{Sara}})$, and $(e^{\text{Sara}}, h^{\text{Pete}})$. Table 9.5 shows the average number of handovers from one resource to another. For instance, Mike frequently hands over work to Sara: on average 1.725 times per case. Sue and Sean only hand over work to Sara as they only execute activity b . It is important to note that the discovered process model is exploited when constructing the social network. The causal dependencies in the process model are used to count handovers in the event log. This way only “real” handovers of work are counted, e.g., concurrent activities may follow one another but do not contribute the number of handovers.

Table 9.5 encodes a social network. All non-zero cells represent “handover of work” relationships. When visualizing a social network typically a threshold is used. If we set the threshold to 0.1, we obtain the social network shown in Fig. 9.6. All cells with a value of at least 0.1 are turned into arcs in the social network. To keep the diagram simple, we only assigned weights to arcs and not to nodes. As Fig. 9.6

Table 9.6 Handover of work matrix at the role level

	Assistant	Expert	Manager
Assistant	1.5	0.5	3.45
Expert	0	0	1.15
Manager	2.95	0.65	1.3

Fig. 9.7 Social network based on handover of work at the level of roles. The weights of nodes are based on the number of times a resource having the role performs an activity. The weights of the arcs are based of the average number of times a handover takes place from one role to another per case



shows there is a strong connection between Mike and Sara. On average, there are 1.725 handovers from Mike to Sara and 1.475 handovers from Sara to Mike. The social network clearly shows the flow of work in the organization and can be used to compute metrics such as the Bavelas–Leavitt index of centrality. Such analysis shows that Sara and Mike are most central in the social network.

The nodes in a social network correspond to organizational entities. In Fig. 9.6 the entities are individual resources. However, it is also possible to construct social networks at the level of departments, teams, or roles. Assume for example that there are three roles: *Assistant*, *Expert*, and *Manager*. Pete, Mike, and Ellen have the role *Assistant*, Sue and Sean have the role *Expert*, and Sara is the only one having the role *Manager*. Later, we will show that such roles can be discovered from frequent patterns in the event log. Moreover, such information is typically available in the information system. Now we can count the number of handovers at the role level. Consider again case 1: $\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete} \rangle$. Using the information about roles we can rewrite this trace to $\langle a^{Assistant}, b^{Expert}, d^{Assistant}, e^{Manager}, h^{Assistant} \rangle$. Again we find five handovers: one from role *Assistant* to role *Expert* ($a^{Assistant}, b^{Expert}$), one from role *Assistant* to role *Assistant* ($a^{Assistant}, d^{Assistant}$), one from role *Expert* to role *Manager* ($b^{Expert}, e^{Manager}$), one from role *Assistant* to role *Manager* ($d^{Assistant}, e^{Manager}$), and one from role *Manager* to role *Assistant* ($e^{Manager}, h^{Assistant}$). Table 9.6 shows the average frequency of such handovers per case. This matrix containing sociometric information can be converted into a social network as shown in Fig. 9.7.

The social network in Fig. 9.7 has weighted nodes and arcs. The weights are visualized graphically. For instance, the biggest node is role *Assistant* with a weight of 5.45. This weight indicates the average number of activities executed by this role.

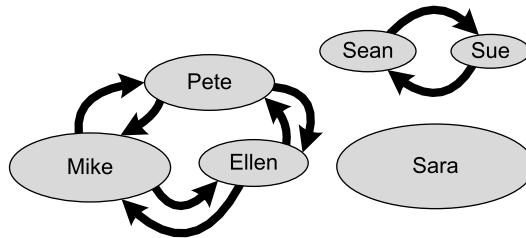


Fig. 9.8 Social network based on similarity of profiles. Resources that execute similar collections of activities are related. Sara is the only resource executing e and f . Therefore, she is not connected to other resources. Self-loops are suppressed as they contain no information (self-similarity)

The weight of role *Expert* is only 1.15, because the two experts (Sue and Sean) only execute activity b which, on average, is executed 1.15 times per case. The weights of the arcs are directly taken from Table 9.6. Clearly, handovers among the roles *Assistant* and *Manager* are most frequent.

Counting handovers of work is just one of many ways of constructing a social network from an event log. In [159] various types of social networks are presented. For example, one can simply count how many times two resources have worked on the same case, i.e., two nodes have a strong relationship when they frequently work together on common cases. One can also use Table 9.4 to quantify the *similarity* of two resources. Every row in the resource-activity matrix can be seen as the *profile* of a resource. Such a vector describes the relevant features of a resource. For example, Pete has profile $P_{Pete} = (0.30, 0.0, 0.345, 0.69, 0.0, 0.0, 0.135, 0.165)$, Mike has profile $P_{Mike} = (0.5, 0.0, 0.575, 1.15, 0.0, 0.0, 0.225, 0.275)$, and Sara has profile $P_{Sara} = (0.0, 0.0, 0.0, 0.0, 2.3, 1.3, 0.0, 0.0)$. Clearly, P_{Pete} and P_{Mike} are very similar whereas P_{Pete} and P_{Sara} are not. The distance between two profiles can be quantified using well-known distance measures such as the *Minkowski distance*, *Hamming distance*, and *Pearson's correlation coefficient*. Moreover, clustering techniques such as *k-means clustering* and *agglomerative hierarchical clustering* can be used to group similar resources together based on their profile (see Sect. 4.3). Two resources in the same cluster (or in close proximity according to the distance metric) are strongly related whereas resources in different clusters (or far away from each other) have no significant relationship in the social network.

For the resource-activity matrix shown in Table 9.4 it does not matter which distance metric or clustering technique is used. All will come to the conclusion that Pete, Mike, and Ellen are very similar and thus have a strong relationship in the social network based on similarity. Similarly, Sue and Sean have a strong relationship in the social network based on similarity. Sara is clearly different from the resources in the two other groups. Figure 9.8 shows the social network based on similarity. Here one can clearly see the roles *Assistant*, *Expert*, and *Manager* mentioned before. However, now the roles are discovered based on the profiles of the resources.

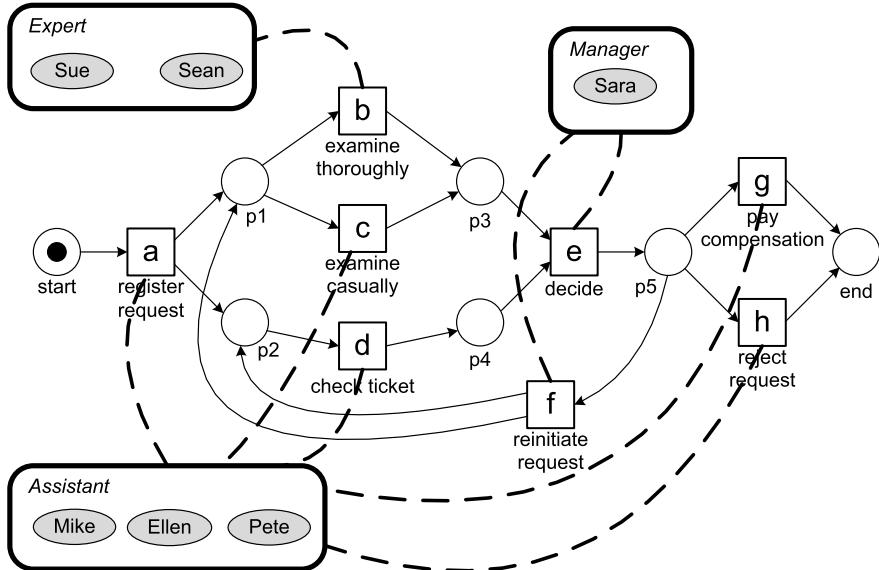


Fig. 9.9 Organizational model discovered based on the event log

9.3.2 Discovering Organizational Structures

The behavior of a resource can be characterized by a *profile*, i.e., a vector indicating how frequent each activity has been executed by the resource. By using such profiles, various clustering techniques can be used to discover similar resources. Figure 9.8 showed an example in which three roles are discovered based on similarities of the profiles of the six resources. In Sect. 4.3 we introduced k -means clustering and agglomerative hierarchical clustering. For k -means clustering the number of clusters is decided upfront. Agglomerative hierarchical clustering produces a dendrogram allowing for a variable number of clusters depending on the desired granularity. Additional relevant features of resources (authorizations, salary, age, etc.) can be added to the profile before clustering. This all depends on the information available. After clustering the resources into groups, these groups can be related to activities in the process. Figure 9.9 shows the end result using the roles discovered earlier.

The three roles *Assistant*, *Expert*, and *Manager* in Fig. 9.9 have the property that they partition the set of resources. In general this will not be the case, e.g., a resource can have multiple roles (e.g., a consultant that is also team leader). Moreover, each activity corresponds to precisely one role. Also this does not always need to be the case. Figure 9.10 sketches a more general situation.

The hypothetical organizational model in Fig. 9.10 connects the process model and the resources seen in the event log. There are eight organizational entities: oe_1, \dots, oe_8 . The model is hierarchical, e.g., oe_4 contains resource r_5 and all resources of oe_6, oe_7 , and oe_8 . Hence five resources belong to organizational entity

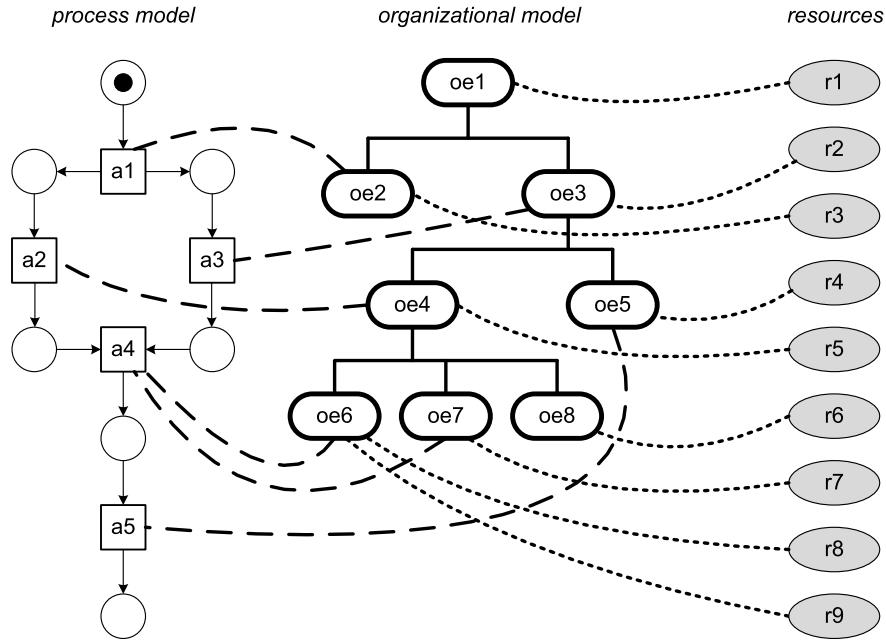


Fig. 9.10 The organizational entities discovered connect activities in the process model to sets resources

oe4: r5, r6, r7, r8, and r9. Organizational entity $oe1$, i.e., the root node, contains all nine resources. If agglomerative hierarchical clustering is used to cluster resources, one automatically gets such a hierarchical structure. Figure 4.7 in Sect. 4.3 shows how agglomerative hierarchical clustering creates a *dendrogram* and Fig. 4.8 shows how any horizontal line defines a level in the hierarchy. The translation from a dendrogram to a hierarchical structure as shown in Fig. 9.10 is straightforward.

Activity $a1$ in Fig. 9.10 can only be performed by resource $r3$ whereas activity $a2$ can be executed by $r5, r6, r7, r8$, or $r9$. For more information about organizational mining we refer to [130].

9.3.3 Analyzing Resource Behavior

Figure 9.10 shows how activities, organizational entities, and resources can be related. Since events in the log refer to activities and resources (and indirectly also to organizational entities), performance measures extracted from the event log can be projected onto such models. For instance, frequencies can be projected onto activities, organizational entities, and resources. It could be shown that resource $r5$ performed 150 activities in the last month: 100 times $a2$ and 50 times $a3$. By aggregating such information it could be deduced that organizational entity $oe4$ was used 300 times in the same period.

In Table 9.3, we abstracted from transaction types, i.e., we did not consider the start and completion of an activity instance. Most logs will contain such information. For example, Table 9.1 shows the start and completion of each activity instance. Some logs will even show when a workitem is offered to a resource or when it is assigned. If such events are recorded, then a diagram such as Fig. 9.10 can also show detailed time related information. For example, the utilization and response times of resources can be shown.

Assuming that the event log contains high quality information including precise timestamps and transaction types, the behavior of resources can be analyzed in detail [139]. Of course privacy issues play an important role here. However, the event log can be anonymized prior to analysis. Moreover, in most organizations one would like to do such analysis at an aggregate level rather than at the level of individuals. For instance, in Sect. 3.1, we mentioned the Yerkes-Dodson law of arousal which describes the relation between workload and performance of people. This law hypothesizes that people work faster when the workload increases. If the event log contains precise timestamps and transaction types, then it is easy to empirically investigate this phenomenon. For any activity instance, one knows its duration and by scanning the log it is also easy to see what the workload was when the activity instance was being performed by some resource. Using supervised learning (e.g., regression analysis or decision tree analysis) the effects of different workloads on service and response times can be measured. See [139] for more examples.

Privacy and anonymization

Event logs may contain sensitive or private data. Events refer to actions and properties of customers, employees, etc. For instance, when applying process mining in a hospital it is important to ensure data privacy. It would be unacceptable that data about patients would be used by unauthorized persons or that event data about treatments would be used in a way not intended when releasing the data. The challenge in process mining is to use event logs to improve processes and information systems while protecting personally identifiable information and not revealing sensitive data. Therefore, most event logs contain *anonymized* attribute values. For example, the name of the customer or employee is often irrelevant for questions that need to be answered. To make an attribute anonymous, the original value is mapped onto a new value in a deterministic manner. This ensures that one can correlate attributes in one event to attributes in another event without knowing the actual values. For instance, all occurrences of the name “Wil van der Aalst” are mapped onto “Q2T4R5R7X1Y9Z”. The mapping of the original value onto the anonymized value should be such that it is not easy (or even impossible) to compute the inverse of the mapping. Anonymous data can sometimes be de-anonymized by combining different data sources. For example, it is often possible to trace back an individual based on her birth date and the birth dates of her children. Therefore, even “anonymous data” should be handled carefully.

Table 9.7 Compact representation of the event log highlighting timestamps; artificial timestamps are used to simplify the presentation of the time-based replay approach

Case id	Trace
1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{50}, e_{complete}^{59}, g_{start}^{70}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, d_{start}^{62}, b_{complete}^{65}, d_{complete}^{67}, e_{start}^{80}, e_{complete}^{87}, g_{start}^{90}, g_{complete}^{98} \rangle$
...	...

Note that process mining techniques do not create *new* data. The information stored in event logs originates from other databases and audit trails. Therefore, privacy and security issues already exist before applying process mining. Nevertheless, the active use of data and process mining techniques increases the risk of data misuse. Organizations should therefore continuously balance the benefits of creating and using event data against potential privacy and security problems.

9.4 Time and Probabilities

The *time perspective* is concerned with the timing and frequency of events. In most event logs, events have a timestamp ($\#_{time}(e)$). The granularity of these timestamps may vary. In some logs only date information is given, e.g., “30-12-2010”. Other event logs have timestamps with millisecond precision. The presence of timestamps enables the discovery of bottlenecks, the analysis of service levels, the monitoring of resource utilization, and the prediction of remaining processing times of running cases. In this section we focus on *replaying event logs with timestamps*. A small modification of the replay approach presented in Sect. 8.2 suffices to include the time perspective in process models.

Table 9.7 shows a fragment of some larger event log highlighting the role of timestamps. To simplify the presentation, we use fictive two-digit timestamps rather than verbose timestamps like “30-12-2010:11.02”. Moreover, we assume that each event has a start event and a complete event. Obviously, the replay approach does not depend on these simplifying assumptions.

Figure 9.11 shows some raw diagnostic information after replaying the three cases shown in Table 9.7. Activity a has three activity instances; one for each case. The first instance of a runs from time 12 to time 19. Hence, the duration of this activity instance is 7 time units. Activity d has four activity instances. For case 3 there are two instances of d ; one running from time 35 to time 40 and one running from time 62 to time 67. The durations of all activity instances are shown. Also places are annotated to indicate how long tokens remained there. For example, there

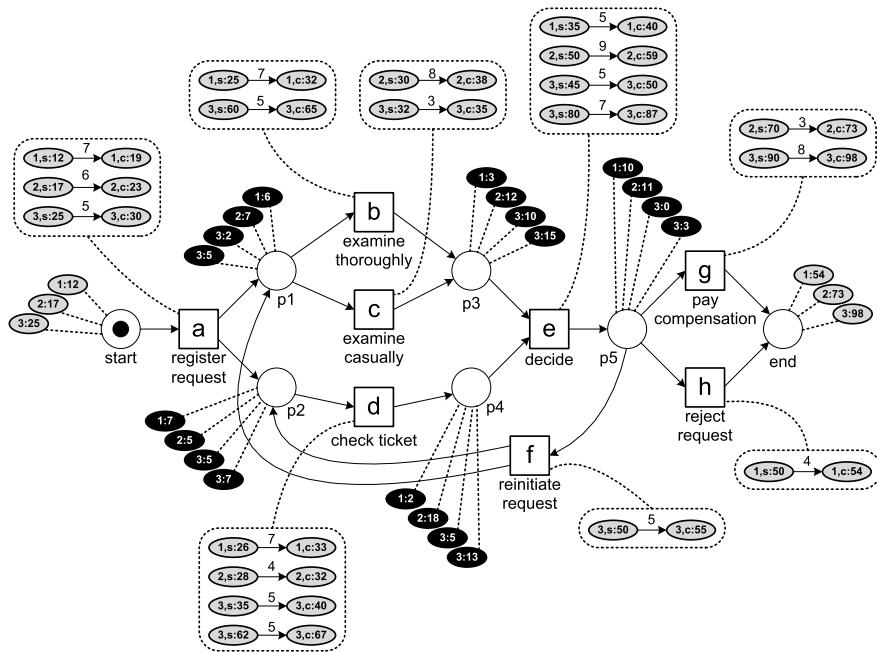


Fig. 9.11 Timed replay of the first three cases in the event log: case 1 starts at time 12 and ends at time 54, case 2 starts at time 17 and ends at time 73, case 3 starts at time 25 and ends at time 98

were four periods in which a token resided in place p_1 : one token corresponding to case 1 resided in p_1 for 6 time units (from time 19 until time 25), one token corresponding to case 2 resided in p_1 for 7 time units (from time 23 until time 30), and two tokens corresponding to case 3 resided in this place (one for $32 - 30 = 2$ time units and one for $60 - 55 = 5$ time units). These times can be found using the approach presented in Sect. 8.2. The only modifications are that now tokens bear timestamps and statistics are collected during replay. In this example, all three cases fit perfectly (i.e., no missing or remaining tokens). One needs to ignore non-fitting events or cases to deal with logs that do not have a conformance of 100%. Heuristics are needed to deal with such situations, but here we assume perfect fitness.

Figure 9.12 shows another view on the information gathered while replaying the three cases. Consider for instance case 3. For this case, an instance of activity a was running from time 25 until time 30. At time 30, c and d became enabled. However, as shown, c started at time 32 and d started at time 35. This implies that there was a waiting time of 2 before c started and a waiting time of 5 before d started. After completing c and d , i.e., at time 40, the first instance of e became enabled. Since the first instance of activity e ran from time 45 until 50, the waiting time for this instance of e was $45 - 40 = 5$ time units. Note that from time 35 until time 45 there was a token in place p_3 (because c completed at time 35 and e started at time 45). However, only half of this period should be considered as waiting time for e , because e only got enabled at time 40 when d completed. As discussed in Sect. 8.5.3, such

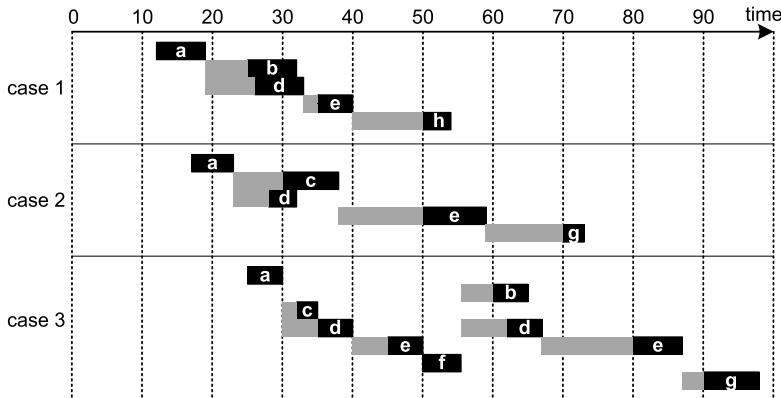


Fig. 9.12 Timeline showing the activity instances of the first three activities

diagnostics are *only possible because events in the log have been coupled to model elements through replay*.

After replay, for each place a collection of “token visits” has been recorded. Each token visit has a start and end time. Hence, a multi-set of durations can be derived. In the example, place p_1 has the multi-set $[6, 7, 2, 5, \dots]$ of durations. For a large event log such a multi-set will contain thousands of elements. Hence, it is possible to fit a distribution and to compute standard statistics such as mean, standard deviation, minimum, and maximum. The same holds for activity instances. Every activity instance has a start and end time. Hence, a multi-set of service times can be derived. For example, activity e in the example has the multi-set $[5, 9, 5, 7, \dots]$ of activity durations. Also here standard statistics can be computed. These can also be computed for waiting times. It is also possible to compute confidence intervals to derive statements such as “the 90% confidence interval for the mean waiting time for activity x is between 40 and 50 minutes”.

Figures 9.11 and 9.12 demonstrate that replay can be used to provide various kinds of performance related information:

- *Visualization of waiting and service times.* Statistics such as the average waiting time for an activity can be projected onto the process model. Activities with a high variation in service time could be highlighted in the model, etc.
- *Bottleneck detection and analysis.* The multi-set of durations attached to each place can be used to discover and analyze bottlenecks. The places where most time is spent can be highlighted. Moreover, cases that spend a long time in a particular place can be further investigated. This is similar to the selection of non-conforming cases described earlier (cf. Fig. 8.8), i.e., the sublog of delayed cases can be analyzed separately to find root causes for the delays.
- *Flow time and SLA analysis.* Fig. 9.11 also shows that the overall flow time can be computed. (In fact, no process model is needed for this.) One can also point to two arbitrary points in the process, say x and y , and compute how many times

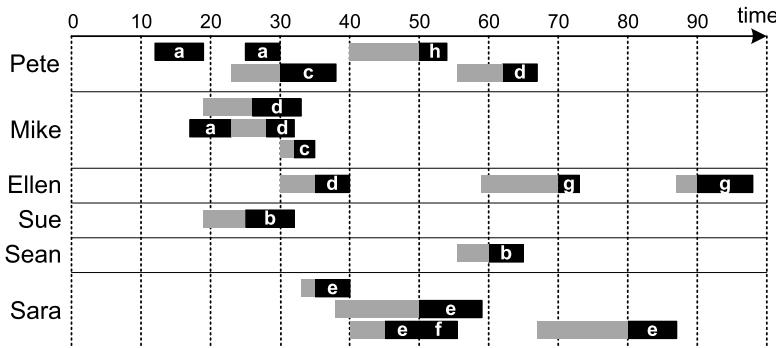


Fig. 9.13 Timeline showing the activity instances projected onto resources. Such projections of the event log allow for the analysis of resource behavior and their utilization

a case flows from x to y . The multi-set of durations to go from x to y can be used to compute all kinds of statistics, e.g., the average flow time between x and y or the fraction of cases taking more than some preset norm. This can be used to monitor Service Level Agreements (SLAs). For instance, it could be that there is a contractual agreement that for 90% of the cases y should be executed within 48 hours after the completion of x . Non-conformance with respect to such an SLA can be highlighted in the model.

- *Analysis of frequencies and utilization.* While replaying the model, times and frequencies are collected. These can be used to show routing probabilities in the model. For example, after e there is a choice to do f , g or h . By analyzing frequencies, one can indicate in the model that in 56% of choices, e is followed by f , in 20% g is chosen, and in 24% h is chosen. By combining frequencies and average service times one can also compute the utilization of resources. Figure 9.13 shows all activity instances and their waiting times *projected onto the resources* executing them. This illustrates that replay can be used to analyze resource behavior.

The event log shown in Table 9.7 only contains *start* and *complete* events. In Chap. 5 we identified additional event types such as *assign*, *schedule*, *suspend*, *resume*, *manualskip*, *abort_case*, and *withdraw*. If an event log contains such events, more statistics can be collected during replay. For instance, if *start* events are preceded by *assign* events, it is possible to analyze how long it takes to start executing the activity instance after being assigned to a specific resource. The transactional life-cycle shown in Fig. 5.3 can be used when replaying the event log.

It can also be the case that the event log does not contain transactional information, i.e., the $\#_{trans}(e)$ attribute is missing in the log. In this case activities are assumed to be atomic. Nevertheless, it is still possible to analyze the time that passes in-between such atomic activities. In addition, heuristics can be applied to “guess” the duration of activity instances.

9.5 Decision Mining

The *case perspective* focuses on properties of cases. Each case is characterized by its case attributes, the attributes of its events, the path taken, and performance information (e.g., flow times).

First, we focus on the influence of case and event attributes on the routing of cases. In Fig. 9.9 there are two *decision points*:

- after registering the request (activity *a*) either a thorough examination (activity *b*) or a casual examination (activity *c*) follows; and
- after making a decision (activity *e*), activity *g* (pay compensation), activity *h* (reject request), or activity *f* (reinitiate request) follows.

Both decision points are of type XOR-split: precisely one of several alternatives is chosen. *Decision mining* aims to find rules explaining such choices in terms of characteristics of the case [120]. For example, by analyzing the event log used to discover Fig. 9.9 one could find that customers from the southern region are always checked thoroughly and that requests by silver customers always get rejected. Clearly, a *classification technique like decision tree learning can be used to find such rules* (see Sect. 4.2). Recall that the input for decision tree learning is a table where every row lists one categorical response variable (e.g., the chosen activity) and multiple predictor variables (e.g., properties of the customer). The decision tree aims to explain the response variable in terms of the predictor variables.

Consider, for example, the situation shown in Fig. 9.14. Using three different notations (YAWL, BPMN, and Petri nets) a choice is depicted: activity *x* is followed by either activity *y* or activity *z*. The table in Fig. 9.14 shows different cases for which this choice needs to be made. There are three predictor variables (*type*, *region*, and *amount*) and one response variable (*activity*). Variables *type*, *region*, and *activity* are categorical and variable *amount* is numerical. The predictor variables correspond to knowledge known about the case at the point in time when the decision was made. The response variable *activity* is determined based on a scan of the event log. The event log will reveal whether *x* was followed by *y* or *z*. The table in Fig. 9.14 serves as input for some decision tree learning algorithm as explained in Sect. 4.2. The resulting decision tree can be rewritten into a rule. Based on the example table, classification will show that the value of the response variable is *y* if the customer is a gold customer and the amount is lower than € 500. Otherwise, the value of the response variable is *z* as shown in Fig. 9.14.

Petri nets cannot express OR-splits and joins directly. However, in higher-level languages like BPMN and YAWL one can express such behavior. Figure 9.15 shows an OR-split using the YAWL and BPMN notation: activity *x* is followed by *y*, or *z*, or *y* and *z*. Note that the response variable *activity* is still categorical and can be determined by scanning the log. The table in Fig. 9.15 can be analyzed using a decision tree learner and the result can be transformed into one rule for each of the output arcs. The response variable is “just *y*” if the customer is a gold customer and the amount is less than € 500, the response variable is “just *z*” if the customer is a silver customer and the amount is at least € 500, and the response variable is

type	region	amount	activity
gold	south	987.30	z
silver	north	178.70	z
gold	south	211.50	y
silver	west	587.70	z
silver	east	224.70	z
silver	south	278.50	z
gold	north	488.50	y
silver	west	443.20	z
silver	south	673.70	z
gold	west	413.50	y
silver	south	687.70	z
gold	south	987.30	z
silver	north	378.80	z
gold	south	314.50	y
silver	north	537.70	z
silver	west	158.70	z
gold	east	344.50	y
...

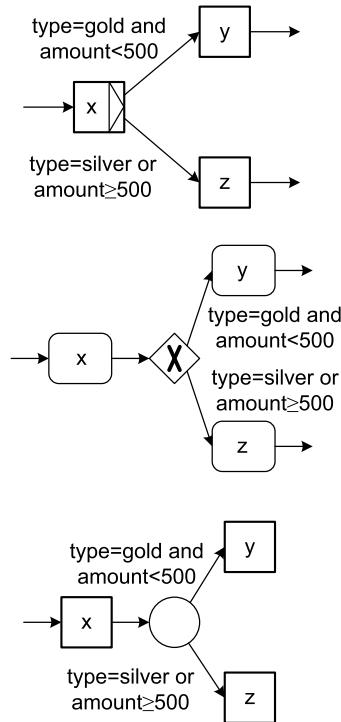


Fig. 9.14 Decision mining: using case and event attributes, a rule is learned for the XOR-split. The result is shown using different notations: YAWL (top), BPMN (middle), and Petri nets (bottom)

type	region	amount	activity
gold	south	987.30	y and z
silver	north	178.70	y and z
gold	south	211.50	just y
silver	west	587.70	just z
silver	east	224.70	y and z
silver	south	278.50	y and z
gold	north	488.50	just y
silver	west	443.20	y and z
silver	south	673.70	just z
...

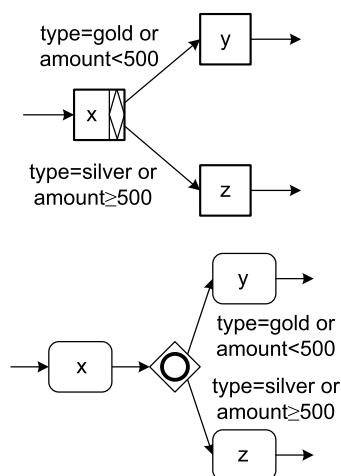


Fig. 9.15 Decision mining: using case and event attributes, a rule is learned for the OR-split. The response variable has three possible values: just y, just z, and both y and z

“y and z” in all other cases. Based on this classification the conditions shown in the YAWL and BPMN models can be derived.

For the predictor variables all case and event attributes can be used. Consider for instance the decision point following activity e and event 35654431 in Table 9.1. The case and event attributes of this event are shown in Tables 9.1 and 9.2. Hence, predictor variables for event 35654431 are: $case = 1$, $activity = decide$, $time = 06-01-2011:11.22$, $resource = Sara$, $trans = complete$, $cost = 200$, $custid = 9911$, $name = Smith$, $type = gold$, $region = south$, and $amount = 989.50$. As described in [120] also the attributes of earlier and later events can be taken into account. For example, all attributes of all events in the trace up to the decision moment can be used. In the process shown in Fig. 9.9 one could find the rule that all cases that involve Sean get rejected in the decision point following activity e .

There may be loops in the model. Hence, the same decision point may be visited multiple times for the same case. Each visit corresponds to a new row in the table used by the decision tree algorithm. For example, in the process shown in Fig. 9.9, there may be cases for which e is executed four times. The first three times e is followed by f and the fourth time e is followed by g or h . Each of the four decisions corresponds to a row in the table used for classification. Using replay, the outcome of the decision (i.e., the response variable) can be identified for each row. Also note that the values of the predictor variables for these four rows may be different.

In some cases, it may be impossible to derive a reasonable decision rule. The reason may be that there is too little data or that decisions are seemingly random or based on considerations not in the event log. In such cases, replay can be used to provide a probability for each branch. Hence, such a decision point is characterized by probabilities rather than data dependent decision rules.

The procedure can be repeated for all decision points in a process model. The results can be used to extend the process model, thus incorporating the case perspective.

Classification in process mining

The application of classification techniques like decision tree learning is *not limited to decision mining* as illustrated by Figs. 9.14 and 9.15: *additional predictor variables* may be used and *alternative response variables* can be analyzed.

In Figs. 9.14 and 9.15 only attributes of events and cases are used as predictor variables. However, also behavioral information can be used. For instance, in Fig. 9.9 it would be interesting to count the number of times that f has been executed. This may influence the decision point following activity e . For example, it could be the case that a request is never initiated more than two times. It may also be that timing information is used as a predictor variable. For instance, if the time taken to check the ticket is less than five minutes, then it is more likely that the request is rejected. It is also possible to use *contextual* information as a predictor variable. Contextual information is in-

formation that is not in the event log and that is not necessarily related to a particular case. For example, the weather may influence a decision. This can only be discovered if the weather condition is taken into account as a predictor variable. Decisions may also depend on the volume of work in the pipeline. One can imagine that the choice between b and c in Fig. 9.9 depends on the workload of the two experts Sue and Sean. When they are overloaded, it may be less likely that b is selected. These examples illustrate that predictor variables are not limited to case and event attributes. However, note the “curse of dimensionality” discussed in Sect. 4.6.3. Analyzing decision points with many predictor variables may be computationally intractable.

In Figs. 9.14 and 9.15 we used classification to learn decision rules. The predictor variables can also be used to learn *other properties of the process*. For instance, one may be interested in characterizing cases for which a particular activity is executed. Classification can also be used to uncover reasons for non-conformance. As shown in Fig. 8.8, the event log can be split into two sublogs: one event log containing only fitting cases and one event log containing only non-fitting cases. The observation whether a case fits or not, can be seen as a response variable. Hence, classification techniques like decision tree learning can be used to characterize cases that deviate. For example, one could learn the rule that cases of gold customers from the southern region tend to deviate from the normative model. Similarly, one could learn rules related to the lateness of cases. For instance, one could find out that cases involving Ellen tend to be delayed.

These examples show that established classification techniques can be combined with process mining once the process model and the event log are connected through replay techniques.

9.6 Bringing It All Together

In this chapter we showed that a control-flow model can be extended with additional perspectives extracted from the event log. Figure 9.16 sketches the approach to obtain a fully integrated model covering all relevant aspects of the process at hand. The approach consists of five steps. For each step we provide pointers to chapters and sections in this book

- *Step 1: obtain an event log.* Chapter 5 showed how to extract event data from a variety of systems. As explained using Fig. 5.1, this is an iterative process. The dotted chart described in Sect. 9.2 helps to explore the event log and guide the filtering process.
- *Step 2: create or discover a process model.* Chapters 6 and 7 focus on techniques for process discovery. Techniques such as heuristic mining and genetic mining can be used to obtain a process model. However, also existing hand-made models can be used.

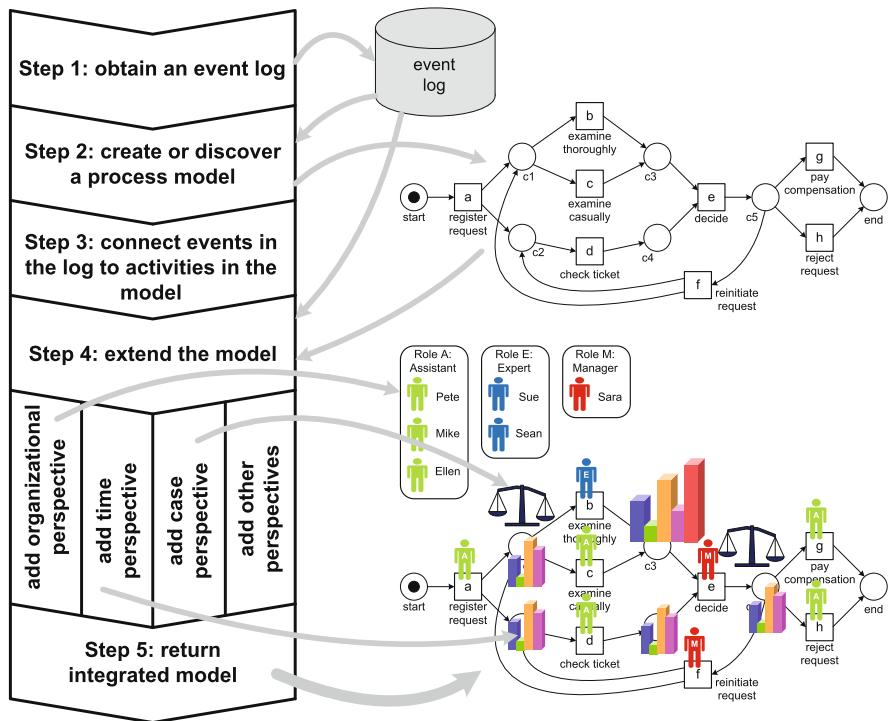


Fig. 9.16 Approach to come to a fully integrated model covering the organizational, time, and case perspectives

- **Step 3: connect events in the log to activities in the model.** As discussed in Sect. 8.5.3, this step is essential for projecting information onto models and to add perspectives. Using the replay technique described in Sect. 8.2, events in the log and activities in the model get connected.
- **Step 4: extend the model.** This is the topic of the current chapter.
 - **Step 4a: add the organizational perspective.** As shown in Sect. 9.3, it is possible to analyze the social network and subsequently identify organizational entities that connect activities to groups of resources.
 - **Step 4b: add the time perspective.** Timestamps and frequencies can be used to learn probability distributions that adequately describe waiting and service times and routing probabilities. Section 9.4 demonstrates that the replay techniques used for conformance checking can be modified to add the time perspective to process models.
 - **Step 4c: add the case perspective.** Section 9.5 showed how to use attributes in the log for decision mining. This shows which data is relevant and should be included in the model.
 - **Step 4d: add other perspectives.** Depending on the information in the log other perspectives may be added to model. For example, information on risks and

costs can be added to the model. Existing risk analysis techniques and costing approaches such as Activity Based Costing (ABC) and Resource Consumption Accounting (RCA) can be used to extend the model [31].

- *Step 5: return the integrated model.*

In Chaps. 13 and 14, we provide an overall life-cycle describing a process mining project (L^* life-cycle model). This more elaborate life-cycle incorporates Fig. 9.16.

The integrated model resulting from the steps in Fig. 9.16, can be used for various purposes. First of all, it provides a *holistic view* on the process. This provides new insights and may generate various ideas for process improvement. Moreover, the integrated model can be used as input for other tools and approaches. For instance, it can be used as a starting point for configuring a WFM or BPM system. During the configuration of such a system for a specific process, one needs to provide a model for the control-flow and the other perspectives. The integrated model can also be used to generate a simulation model covering all perspectives. For example, in [124] it is shown that the techniques described in this chapter can be used to generate a simulation model in *CPN Tools*. CPN Tools is a powerful simulation environment based on *colored Petri nets* [82, 149] (see www.cpntools.org).

The resulting simulation model closely follows reality as it is based on event logs rather than human modeling. The colored Petri net models control-flow, data flow, decisions, resources, allocation rules, service times, routing probabilities, arrival processes, etc., thus capturing all aspects relevant for simulation. The integrated simulation model can be used for “what if” analysis to explore different redesigns and control strategies.

Short-term simulation

As stressed earlier, it is essential that events in the log are connected to model elements. This allows for the projection of dynamic information onto models: the event log “breathes life” into otherwise static process models. Moreover, the merging of the various perspectives into a single model depends on this. Establishing a good connection between an event log and model may be difficult and require several iterations. However, when using a BPM system, this connection already exists; BPM systems are driven by explicit workflow models and provide excellent event logs. Moreover, internally such systems also have an explicit representation of the state of each running case. This enables a new type of simulation called *short-term simulation* [125, 139]. The key idea is to start all simulation runs from the current state and focus the analysis of the transient behavior. This way a “*fast forward button*” into the future is provided.

Figure 9.16 sketches how a simulation model could be obtained that closely matches reality. The current state obtained from the BPM system, i.e., the markings of all cases and related data elements, can be loaded into the simulation as a realistic initial state.

To understand the importance of short-term simulation, we elaborate on the difference between *transient analysis* and *steady-state analysis*. The key idea of simulation is to execute a model repeatedly. The reason for doing the experiments repeatedly, is to not come up with just a single value (e.g., “the average response time is 10.36 minutes”) but to provide confidence intervals (e.g., “the average response time is with 90 percent certainty between 10 and 11 minutes”). For transient analysis the focus is on the initial part of future behavior, i.e., starting from the initial state the “near future” is explored. For transient analysis the initial state is very important. If the simulation starts in a state with long queues of work, then in the near future flow times will be long and it may take some time to get rid of the backlog. For steady-state analysis the initial state is irrelevant. Typically, the simulation is started “empty” (i.e., without any cases in progress) and only when the system is filled with cases the measurements start.

Steady-state analysis is most relevant for answering strategic and tactical questions. Transient analysis is most relevant for operational decision making. Lion’s share of contemporary simulation support aims at steady-state analysis and, hence, is limited to strategic and tactical decision making. Short-term simulation focuses on operational decision making; starting from the current state—loaded from the BPM system—the “near future” is explored repeatedly [139]. This shows what will happen if no corrective actions are taken. Moreover, “what if” analysis can be used to explore the effects of different actions (e.g., adding resources and reconfiguring the process).

In [125] it is shown how this approach can be realized using the BPM system *YAWL*, the process mining tool *ProM*, and the simulation tool *CPN Tools*. This illustrates the potentially spectacular synergetic effects that can be achieved by combining workflow automation, process mining, and simulation.

Chapter 10

Operational Support

Most process-mining techniques work on “post mortem” event data, i.e., they analyze events that belong to cases that have already completed. Obviously, it is not possible to influence the execution of “post mortem” cases. Moreover, cases that are still in the pipeline cannot be guided on the basis of “post mortem” event data only. Today, however, many data sources are updated in (near) real-time and sufficient computing power is available to analyze events when they occur. Therefore, process mining should not be restricted to off-line analysis and can also be used for online operational support. This chapter broadens the scope of process mining to include online decision support. For example, for a running case the remaining flow time can be predicted and suitable actions can be recommended to minimize costs.

10.1 Refined Process Mining Framework

Thus far we identified three main types of process mining: *discovery*, *conformance*, and *enhancement* (cf. Figs. 2.5 and 9.1). Orthogonal to these types of process mining we identified several perspectives including: the *control-flow perspective* (“How?”), the *organizational perspective* (“Who?”), and the *case/data perspective* (“What?”). The classification of process mining techniques into discovery, conformance, and enhancement does reflect that analysis can be done *online* or *off-line*. Moreover, Figs. 2.5 and 9.1 do not acknowledge that there are essentially two types of models (“*de jure models*” and “*de facto models*”) and two types of data (“*pre mortem*” and “*post mortem*” event data) [139].

Figure 10.1 shows our *refined process mining framework*. As before, we assume some external “world” consisting of business processes, people, organizations, etc. and supported by some information system. The information system records information about this “world” in such a way that event logs can be extracted as described in Chap. 5.

Figure 10.1 emphasizes the systematic, reliable, and trustworthy recording of events by using the term *provenance*. This term originates from scientific comput-

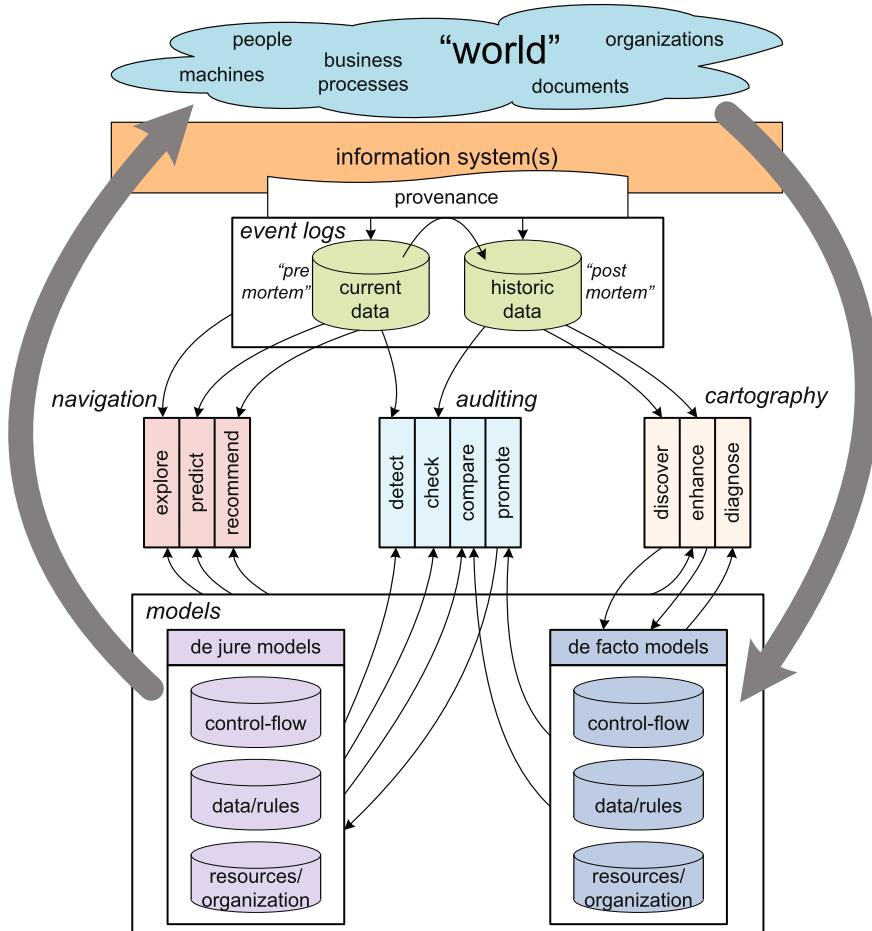


Fig. 10.1 Refined process mining framework

ing, where it refers to the data that is needed to be able to reproduce an experiment [39]. *Business process provenance* aims to systematically collect the information needed to reconstruct what has actually happened in a process or organization. When organizations base their decisions on event data it is essential to make sure that these describe history well. Moreover, from an auditing point of view it is necessary to ensure that event logs cannot be tampered with. Business process provenance refers to the set of activities needed to ensure that history, as captured in event logs, “cannot be rewritten or obscured” such that it can serve as a reliable basis for process improvement and auditing.

Data in event logs are partitioned into “*pre mortem*” and “*post mortem*” event data in the refined process mining framework depicted in Fig. 10.1. “Post mortem” event data refer to information about cases that have completed, i.e., these data can

be used for process improvement and auditing, but not for influencing the cases they refer to. Most event logs considered thus far contained only historic, i.e., “post mortem”, event data. “Pre mortem” event data refer to cases that have not yet completed. If a case is still running, i.e., the case is still “alive” (pre mortem), then it may be possible that information in the event log about this case (i.e., current data) can be exploited to ensure the correct or efficient handling of this case.

“Post mortem” event data is most relevant for *off-line process mining*, e.g., discovering the control-flow of a process based on one year of event data. For *online process mining* a mixture of “pre mortem” (current) and “post mortem” (historic) data is needed. For example, historic information can be used to learn a predictive model. Subsequently, information about a running case is combined with the predictive model to provide an estimate for the remaining flow time of the case.

The refined process mining framework also distinguishes between two types of models: “*de jure models*” and “*de facto models*”. A *de jure model* is normative, i.e., it specifies how things should be done or handled. For example, a process model used to configure a BPM system is normative and forces people to work in a particular way. A *de facto model* is descriptive and its goal is not to steer or control reality. Instead, de facto models aim to capture reality. The techniques presented in Chaps. 6 and 7 aim to produce de facto models. Figure 10.1 also highlights that models can cover different perspectives, i.e., process mining is not limited to control-flow and is also concerned with resources, data, organizational entities, decision points, costs, etc. The two large arrows in Fig. 10.1 illustrate that de facto models are derived from reality (right downward arrow) and that de jure models aim to influence reality (left upward arrow).

After refining event logs into “pre mortem” and “post mortem” and partitioning models into “*de jure*” and “*de facto*”, we can identify ten process mining related activities as shown in Fig. 10.1. These ten activities are grouped into three categories: *cartography*, *auditing*, and *navigation*.

10.1.1 Cartography

Process models can be seen as the “maps” describing the operational processes of organizations, i.e., just like geographic maps, process models aim to describe reality. In order to do this, *abstractions* are needed. For example, on a roadmap a highway may be denoted by an orange line having a thickness of four millimeters. In reality the highway will not be orange; the orange coloring is just used to emphasize the importance of highways. If the scale of the map is 1 : 500000, then the thickness of the line corresponds to a highway of 2 kilometers wide. In reality, the highway will not be so broad. If the thickness of the line would correspond to reality (assuming the same scale), it would be approximately 0.05 millimeter (for a highway of 25 meters wide). Hence, the highway would be (close to) invisible. Therefore, the scale is modified to make the map more readable and useful. When making process models, we need to use similar abstractions. In Chap. 15, we will elaborate on the relationships between process maps and geographic maps. Also note that in Sect. 6.4.4 we

already used the metaphor of a “process view” to argue that a discovered process model views reality from a particular “angle”, is “framed”, and shown using a particular “resolution”. Metaphors such as “maps” and “views” help in understanding the role of process models in BPM.

Figure 10.1 shows that three activities are grouped under cartography: *discover*, *enhance*, and *diagnose*.

- *Discover*. This activity is concerned with the extraction of (process) models as discussed in Chaps. 6 and 7.
- *Enhance*. When existing process models (either discovered or hand-made) can be related to event logs, it is possible to enhance these models. The connection can be used to repair models or to extend them. In Sect. 8.5.1, we showed that models can be made more faithful using the diagnostics provided by conformance checking techniques. Chap. 9 illustrated how attributes in event logs can be used to add additional perspectives to a model.
- *Diagnose*. This activity does not directly use event logs and focuses on classical model-based process analysis as discussed in Sect. 3.3, e.g., process models can be checked for the absence of deadlocks or alternative models can be simulated to estimate the effect of various redesigns on average cycle times.

10.1.2 Auditing

In Sect. 8.1, we defined *auditing* as the set of activities used to check whether business processes are executed within certain boundaries set by managers, governments, and other stakeholders [166]. In Fig. 10.1, the auditing category groups all activities that are concerned with the comparison of behaviors, e.g., two process models or a process model and an event log are put side by side.

- *Detect*. This activity compares de jure models with current “pre mortem” data (events of running process instances) with the goal to detect deviations at run-time. The moment a predefined rule is violated, an alert is generated.
- *Check*. As demonstrated in Chap. 8, historic “post mortem” data can be cross-checked with de jure models. The goal of this activity is to pinpoint deviations and quantify the level of compliance.
- *Compare*. De facto models can be compared with de jure models to see in what way reality deviates from what was planned or expected. Unlike for the previous two activities, no event log is used directly. However, the de facto model may have been discovered using historic data; this way event data are used indirectly for the comparison. In Sect. 8.4, we showed that footprints can be used for model-to-model (and log-to-model) comparisons.
- *Promote*. Based on an analysis of the differences between a de facto model and a de jure model, it is possible to promote parts of the de facto model to a new de jure model. By promoting proven “best practices” to the de jure model, existing processes can be improved.

Note that the *detect* and *check* activities are similar except for the event data used. The former activity uses “pre mortem” data and aims at online analysis to be able to react immediately when a discrepancy is detected. The latter activity uses “post mortem” data and is done off-line.

10.1.3 Navigation

The last category of process mining activities aim at business process *navigation*. Unlike the cartography and auditing activities, navigation activities are forward-looking. For example, process mining techniques can be used to make predictions about the future of a particular case and guide the user in selecting suitable actions. When comparing this with a car navigation system from TomTom or Garmin, this corresponds to functionalities such predicting the arrival time and guiding the driver using spoken instructions. In Chap. 15, we elaborate on the similarities between car navigation and process mining.

Figure 10.1 lists three navigation activities: *explore*, *predict*, and *recommend*.

- *Explore*. The combination of event data and models can be used to explore business processes at run-time. Running cases can be visualized and compared with similar cases that were handled earlier.
- *Predict*. By combining information about running cases with models (discovered or hand-made), it is possible to make predictions about the future, e.g., the remaining flow time and the probability of success.
- *Recommend*. The information used for predicting the future can also be used to recommend suitable actions (e.g. to minimize costs or time). The goal is to enable functionality similar to the guidance given by car navigation systems.

In earlier chapters, we focused on activities using historic (“post mortem”) data only, i.e., activities *discover*, *enhance*, and *check* in Fig. 10.1. In the remainder of this chapter, we shift our attention to online analysis also using “pre mortem” data.

10.2 Online Process Mining

Traditionally, process mining has been used in an off-line fashion using only “post mortem” data. This means that only completed cases are being considered, i.e., the traces in the event log are *complete* traces corresponding to cases that were fully handled in the past. For *operational support* we also consider “pre mortem” event data and respond to such data in an online fashion. Now only running cases are considered as these can, potentially, still be influenced. A running case may still generate events. Therefore, it is described by a *partial* trace.

Figure 10.2 shows the essence of operational support. Consider a case for which activities a and b have been executed. Partial trace $\sigma_p = \langle a, b \rangle$ describes the known

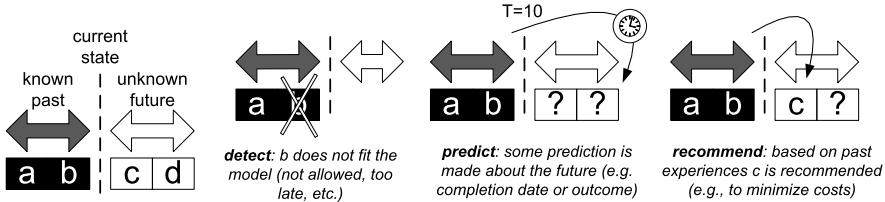


Fig. 10.2 Three process mining activities related to operational support: *detect*, *predict*, and *recommend*

Table 10.1 Fragment of event log with timestamps and transactional information. For instance, event a_{start}^{12} denotes the start of activity a at time 12

Case id	Trace
1	$\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$
2	$\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28}, c_{start}^{30}, d_{complete}^{32}, c_{complete}^{38}, e_{start}^{50}, e_{complete}^{59}, g_{start}^{70}, g_{complete}^{73} \rangle$
3	$\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, d_{start}^{62}, b_{complete}^{65}, d_{complete}^{67}, e_{start}^{80}, e_{complete}^{87}, g_{start}^{90}, g_{complete}^{98} \rangle$
...	...

past of the case. Note that the two events may have all kinds of attributes (e.g., timestamps and associated resources), but these are not shown here. In the state after observing σ_p , the future of the case is not known yet. One possible future could be that c and d will be executed resulting in a complete trace $\sigma_c = \langle a, b, c, d \rangle$. Figure 10.2 shows three operational support activities: *detect*, *predict*, and *recommend*. These correspond to the activities already mentioned in the context of Fig. 10.1.

- **Detect.** This activity compares the partial trace σ_p with some normative model, e.g., a process model or an LTL constraint. Such a check could reveal a violation as shown in Fig. 10.2. If b was not allowed after a , an alert would be generated.
- **Predict.** This activity makes statements about the events following σ_p . For example, the expected completion time could be predicted by comparing the current case to similar cases that were handled in the past.
- **Recommend.** Recommendations guide the user in selecting the next activity after σ_p . For example, it could be that, based on historic information, it is recommended to execute activity c next (e.g., to minimize costs or flow time).

Note that all three activities assume some model, e.g., predictions and recommendations could be based on a regression model or obtained using simulation. Besides the three operational support activities illustrated by Fig. 10.2, it is also possible to simply explore partial traces. For example, dotted chart visualization and other visual analytics techniques can also be applied to running cases.

In the remainder, we show how some of the process mining techniques presented earlier can be modified to provide operational support. In order to do this, we use the event log shown in Table 10.1. This log was also used in earlier chapters and is based

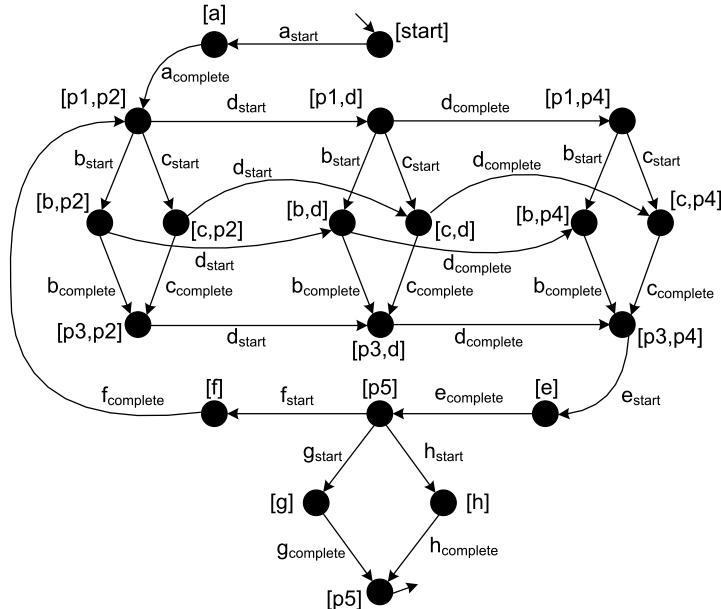


Fig. 10.3 Transition system modeling the process that generated the event log shown in Table 10.1. This process was already modeled in terms of a WF-net (Fig. 2.2) and in terms of BPMN (Fig. 2.3). However, in this transition system we model the start and completion of an activity explicitly. In terms of the WF-net in Fig. 2.2, this means that transition a is split into transitions a_{start} and $a_{complete}$ connected by a place named a ; etc.

on the running example introduced in Chap. 2. The WF-net shown in Fig. 2.2 models the process for which events have been recorded in Table 10.1. Figure 2.3 models the same process in terms of BPMN. Independent of the notation used, we can also derive a transition system modeling the same process as is shown in Fig. 10.3. The transition system labels the nodes with markings of the corresponding Petri net in which each activity is modeled by a start and complete transition. Transition a_{start} consumes a token from place *start* and produces a token for place *a*. The token in place *a* models that activity *a* is being executed. Transition $a_{complete}$ consumes a token from place *a* and produces a token for each of the places *p1* and *p2* (state $[p1, p2]$ in Fig. 10.3). The state labeled $[b, d]$ in Fig. 10.3 corresponds to the marking with tokens in *b* and *d*, i.e., activities *b* and *d* are being executed in parallel. The state space of the BPMN model shown in Fig. 2.3 is isomorphic to the transition system shown in Fig. 10.3.

10.3 Detect

The first operational support activity we elaborate on is *detecting* deviations at run-time. This can be seen as conformance checking “on-the-fly”. Compared to confor-

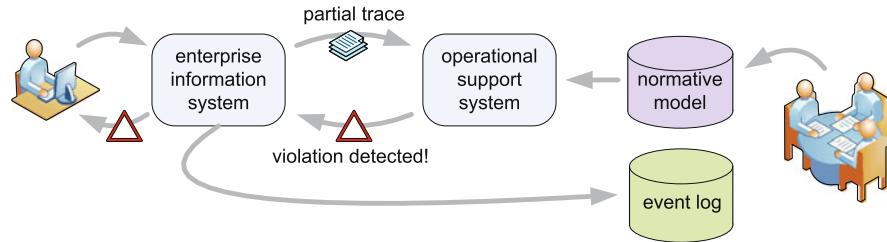


Fig. 10.4 Detecting violations at run-time: the moment a deviation is detected, an alert is generated

mance checking as described in Chap. 8 there are two important differences: (a) we do no consider the log as a whole but focus on the *partial trace of a particular case*, and (b) in case of a deviation there should be an *immediate response* when the deviation occurs. Figure 10.4 illustrates this type of operational support. Users are interacting with some enterprise information system. Based on their actions, events are recorded. The partial trace of each case is continuously checked by the operational support system, i.e., each time an event occurs, the partial trace of the corresponding case is sent to the operational support system. The operational support system immediately generates an alert if a deviation is detected. The enterprise information system and its users can take appropriate actions based on this alert, e.g., a manager is notified such that corrective actions can be taken.

All cases in the event log shown in Table 10.1 conform to the transition system of Fig. 10.3, the WF-net shown in Fig. 2.2, and the BPMN model shown in Fig. 2.3. Therefore, when these cases were executing, no deviations could be detected with respect to these models. Assume now that the more restrictive WF-net shown in Fig. 10.5 describes the desired normative behavior. Compared to the original model activity d (i.e., checking the ticket) should occur after b or c (i.e., one of the examinations).¹

Let us now consider the first case, $\sigma_1 = \langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$. After each event it is checked whether there is a deviation or not. At time 12, after executing the first event a_{start}^{12} no deviation is found, because trace $\langle a_{start}^{12} \rangle$ can be replayed in Fig. 10.5 without missing tokens.² The next two events can also be replayed, i.e., $\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25} \rangle$ is a possible firing sequence of the WF-net in which each activity is refined into a start

¹Note that this diagram can be simplified by removing place $c2$, the arc from $c3$ to e , and the arc from d to $c3$ (i.e., N_2 in Fig. 8.2). The simplified model has the same behavior, i.e., both are bisimilar.

²The WF-net Fig. 10.5 has only one transition per activity while the log contains start and complete events. As described in Sect. 6.2.4, each activity can be described by a small subprocess. Assume that all transitions in Fig. 10.5 are split into a start transition and complete transition connected through a place named after the activity. For example, transition a is refined into transitions a_{start} and $a_{complete}$ connected by a place a . Note that the transition system in Fig. 10.3 used the same naming convention.

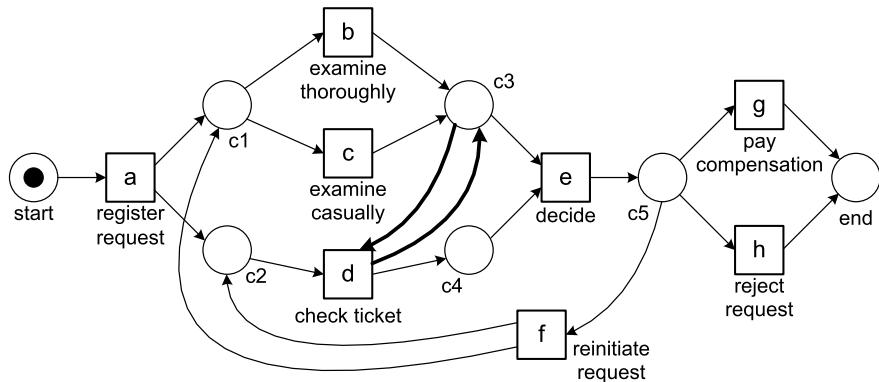
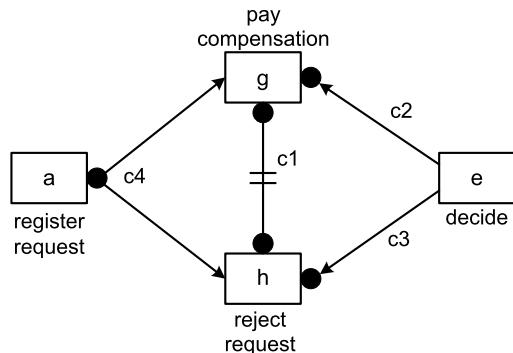


Fig. 10.5 WF-net modeling an additional constraint: d can only be started once b or c has completed

Fig. 10.6 Declare specification composed of four constraints: $c1$, $c2$, $c3$, and $c4$



and complete transition. The state after replaying the three events is $[c2, b]$. The next event, i.e., d_{start}^{26} is not possible in this state. Hence, an *alert* is generated at time 26 based on partial trace $\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26} \rangle$. The alert signals that activity d was started without being enabled. For the second case a deviation is detected at time 28; based on the partial trace $\langle a_{start}^{17}, a_{complete}^{23}, d_{start}^{28} \rangle$ an alert is generated stating that d was started before it was enabled. For the third case a deviation is detected at time 62. The prefix $\langle a_{start}^{25}, a_{complete}^{30}, c_{start}^{32}, c_{complete}^{35}, d_{start}^{35}, d_{complete}^{40}, e_{start}^{45}, e_{complete}^{50}, f_{start}^{50}, f_{complete}^{55}, b_{start}^{60}, d_{start}^{62} \rangle$ cannot be replayed properly because the second instance of d is started without being enabled. These examples show that the replay approach from Chap. 8 can also be used at run-time for detecting deviations the moment they happen.

In Chap. 8, we introduced *Declare* as an example of a constraint-based language. We used the Declare model shown in Fig. 10.6 to explain some of basic concepts. Each of the four constraints shown can be specified in terms of LTL. Constraint $c1$ is a non-coexistence constraint stating that g and h should not both happen. The LTL expression for this constraint is $!((\Diamond g) \wedge (\Diamond h))$. Constraint $c2$ is a precedence

constraint $((!g) W e)$ modeling the requirement that g should not happen before e has happened. Constraint $c3$ is a similar precedence constraint, but now referring to h rather than g . Constraint $c4$ is a branched response constraint stating that every occurrence of a should eventually be followed by g or h , i.e., $\square(a \Rightarrow (\Diamond(g \vee h)))$.

Consider some case having a partial trace σ_p listing the events that have happened thus far. Each constraint c in Fig. 10.6 is in one of the following states for partial trace σ_p :

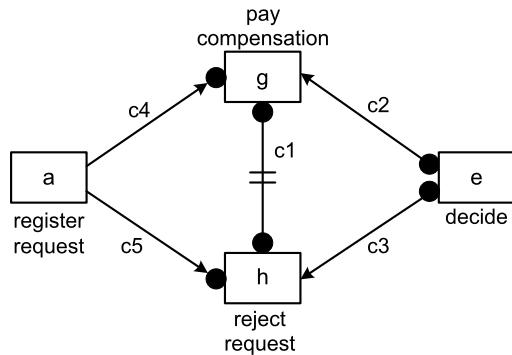
- *Satisfied*. The LTL formula corresponding to c evaluates to true for the partial trace σ_p .
- *Temporarily violated*. The LTL formula corresponding to c evaluates to false for σ_p , however, there is a longer trace σ'_p that has σ_p as a prefix and for which the LTL formula corresponding to c evaluates to true.
- *Permanently violated*. The LTL formula corresponding to c evaluates to false for σ_p and all its extensions, i.e., there is no σ'_p that has σ_p as a prefix and for which the LTL formula evaluates to true.

These three notions can be lifted from the level of a *single constraint* to the level of a *complete Declare specification*. A Declare specification is *satisfied* for a case if all of its constraints are satisfied. A Declare specification is *temporarily violated* by a case if for the current partial trace at least one of the constraints is violated, however, there is a possible future in which all constraints are satisfied. A Declare specification is *permanently violated* by a case if no such future exists.

None of the cases shown in Table 10.1 violates any of the constraints shown in Fig. 10.6, i.e., for each trace, at the *end*, all constraints are satisfied. Let us now consider a scenario in which for a case the trace $\sigma = \langle a, b, d, g \rangle$ is executed. For simplicity, we removed timestamps and transactional information. Initially, i.e., for trace $\sigma_0 = \langle \rangle$, all constraints are satisfied. After executing a , i.e., for prefix $\sigma_1 = \langle a \rangle$, constraint $c4$ is temporarily violated. Because there is a possible future in which all constraints are satisfied, there is no need to generate an alert. However, diagnostic information stating that constraint $c4$ is temporarily violated could be provided. Executing b and d does not change the situation, i.e., both partial traces $\sigma_2 = \langle a, b \rangle$ and $\sigma_3 = \langle a, b, d \rangle$ temporarily violate $c4$. However, after executing g the situation changes. Partial trace $\sigma_4 = \langle a, b, d, g \rangle$ satisfies constraint $c4$. However, constraint $c2$ is permanently violated by σ_4 as there is no “possible future” in which e occurs before g . Therefore, a deviation is detected and reported.

Figure 10.7 shows another Declare model. Constraint $c1$ is the same non-coexistence constraint as before. Constraint $c2$ is a response constraint stating that every occurrence of activity e should eventually be followed by g , i.e., $\square(e \Rightarrow (\Diamond g))$ in LTL terms. Constraint $c3$ is a similar response constraint (every occurrence of activity e should eventually be followed by h). Constraint $c4$ is a precedence constraint $((!g) W a)$ modeling the requirement that g should not happen before a has happened. Constraint $c5$ is also a precedence constraint $((!h) W a)$. Assume that Fig. 10.7 is the normative model. Let us first consider a scenario in which for a case the trace $\sigma = \langle a, b, d, g \rangle$ is executed. For all prefixes, all of the constraints are satisfied, i.e., no alerts need to be executed during the lifetime of this case. Let us

Fig. 10.7 Another Declare specification. Note that c_1 , c_2 , and c_3 imply that e cannot be executed without permanently violating the specification



now consider the scenario $\sigma = \langle a, b, d, e, g \rangle$. No alerts need to be generated for the first three events. In fact at any stage all five constraints are satisfied. However, after executing e constraints c_2 and c_3 are temporarily violated. To remove these temporary violations, both g and h need to be executed after $\langle a, b, d, e \rangle$. However, the execution of both g and h results in a permanent violation of c_1 . Because there is no possible future in which *all* constraints are satisfied, the Declare specification is permanently violated by prefix $\langle a, b, d, e \rangle$ and an alert is generated directly after e occurs. Note that in the latter scenario, there are only temporarily violated constraints whereas the whole specification is permanently violated. Therefore, advanced reasoning is required to determine whether an event signifies a deviation or not. As shown in [103, 162], one can use model checking or abductive logic programming to detect such deviations and provide informative alerts.

10.4 Predict

The second operational support activity we consider is *prediction*. As shown in Fig. 10.8, we again consider the setting in which users are interacting with some enterprise information system. The events recorded for cases can be sent to the operational support system in the form of partial traces. Based on such a partial trace and some predictive model, a prediction is generated. Examples of predictions are:

- The predicted remaining flow time is 14 days;
- The predicted probability of meeting the legal deadline is 0.72;
- The predicted total cost of this case is € 4500;
- The predicted probability that activity a will occur is 0.34;
- The predicted probability that person r will work on this case is 0.57;
- The predicted probability that a case will be rejected is 0.67; and
- The predicted total service time is 98 minutes.

In the fictive example shown in Fig. 10.8, the operational support system predicts that the completion date will be April 25th, 2011.

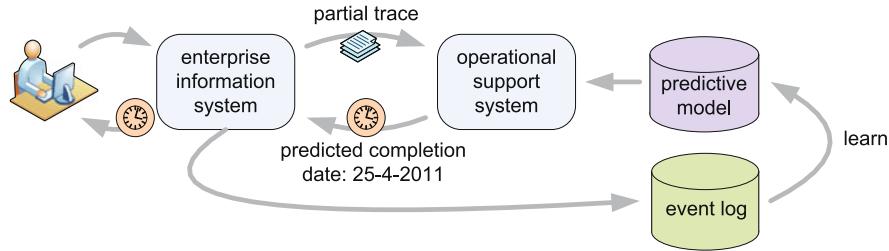


Fig. 10.8 Both the partial trace of a running case and some predictive model are used to provide a prediction (e.g., remaining flow time, expected total costs, or probability of success)

Various techniques can be used to generate predictions. For example, the supervised learning techniques discussed in Sect. 4.1.2 can be used to answer some of these questions. Using feature extraction, relevant properties of the partial trace need to be mapped onto predictor variables. Moreover, the feature we would like to predict is mapped onto a response variable. The response variable is often a performance indicator, e.g., remaining flow time or total costs. If the response variable is numeric, typically regression analysis is used. For a categorical response variable, classification techniques such as decision tree learning can be used. The predictive model is based on historic “post mortem” event data, but can be used to make predictions for the cases that are still running.

Given the variety of approaches and the broad spectrum of possible questions, we cannot provide a comprehensive overview of prediction techniques. Therefore, as an example, we select one particular technique answering a specific question. In the remainder, we show how to *predict the remaining flow time using an annotated transition system* [164, 167]. Starting point for this approach is an event log with timestamps as shown in Table 10.1 and a transition system such as the one shown in Fig. 10.3. The transition system can be obtained by computing the state-space of a process model expressed in another language (WF-nets, BPMN, YAWL, EPCs, etc.). For example, the transition system in Fig. 10.3 can be obtained from the WF-net in Fig. 2.2 or the BPMN model in Fig. 2.3. The transition system can also be obtained using the technique described in Sect. 7.4.1, i.e., using an event log L and a state representation function $l^{state}()$, one can automatically generate a transition system able to replay the event log.

Assuming that the event log fits the transition system, one can replay the events on the model and collect timing information. Non-fitting events and/or cases can be simply ignored or handled as described in Sect. 8.2. Figure 10.9 shows the timed replay of the first two traces in Table 10.1.

Let us consider the first case, $\langle a_{start}^{12}, a_{complete}^{19}, b_{start}^{25}, d_{start}^{26}, b_{complete}^{32}, d_{complete}^{33}, e_{start}^{35}, e_{complete}^{40}, h_{start}^{50}, h_{complete}^{54} \rangle$. This case started at time 12 and ended at time 54. Hence, its flow time was 42 time units. States visited by this case are annotated with a tag (t, e, r, s) where t is the *time* the state is visited, e is the *elapsed time* since the start when visiting the state, r is the *remaining flow time*, and s is the *sojourn time*. State $[a]$ is tagged with the annotation $(t = 12, e = 0, r = 42, s = 7)$ because

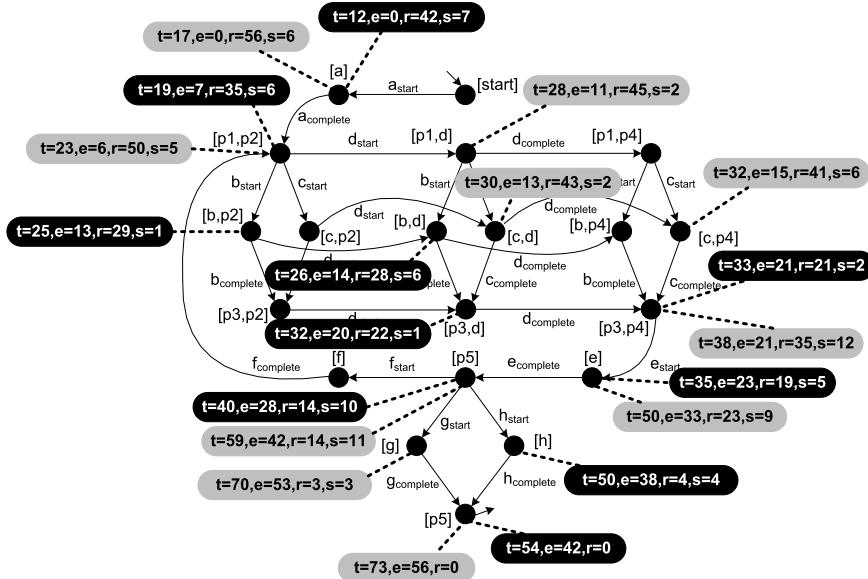


Fig. 10.9 Statistics collected while replaying the first two cases: t is the time the state is visited, e is the elapsed time since the start when visiting the state, r is the remaining flow time, and s is the sojourn time

this state was visited by the case directly after the first event a_{start}^{12} occurred. $t = 12$ because event a_{start}^{12} occurred at time 12. $e = 12 - 12 = 0$ because no time elapsed after executing just one event. $r = 54 - 12 = 42$ is the remaining time until the end of the case after a was started at time 12. $s = 19 - 12 = 7$ because the next event occurred 7 time units later. State $[p1, p2]$ is tagged with annotation ($t = 19, e = 7, r = 35, s = 6$) because a completed at time $t = 19$. $e = 19 - 12 = 7$ because a completed 7 time units after the case started. $r = 54 - 19 = 35$ because the case ended at time 54. $s = 25 - 19 = 6$ because the next event occurred 6 time units later. Figure 10.9 shows all annotations related to the first two cases. For example, state $[p3, p4]$ was visited once by each of the two cases resulting in annotations ($t = 33, e = 21, r = 21, s = 2$) and ($t = 38, e = 21, r = 35, s = 12$). The initial state $[start]$ has no annotations since no events have occurred when visiting this state. The final state $[p5]$ has no sojourn time because there is no next event when visiting this state.

Table 10.1 shows only a fragment of the whole event log. However, it is obvious that the other cases in the log can be replayed in a similar fashion to gather more annotations. For example, the third case visited state $[p3, p4]$ twice, after event $d_{complete}^{40}$ and after event $d_{complete}^{67}$. The first visit resulted in annotation ($t = 40, e = 15, r = 58, s = 5$) and the second visit resulted in annotation ($t = 67, e = 42, r = 31, s = 13$). Assuming a large event log, there may be hundreds or even thousands of annotations per state. For each state x it is possible to create

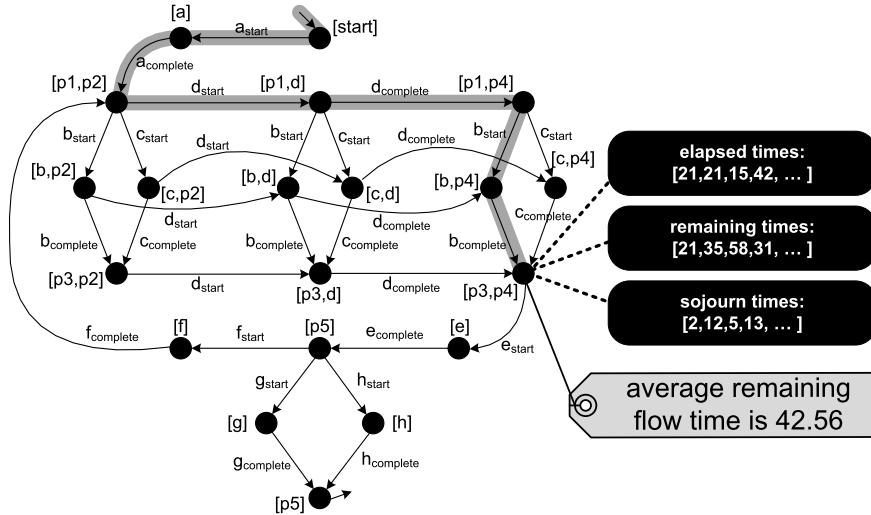


Fig. 10.10 Each state has a multi-set of remaining flow times (one element for each visit). This is the basis for predicting the remaining flow time of future cases. For a case with partial trace $\langle a_{start}^{512}, a_{complete}^{518}, d_{start}^{525}, d_{complete}^{526}, b_{start}^{532}, b_{complete}^{533} \rangle$, the predicted remaining flow time is 42.56. This is the mean remaining flow time of cases in state $[p3, p4]$

a multi-set $Q_x^{remaining}$ of remaining flow times based on these annotations. For state $[p3, p4]$ this multi-set is $Q_{[p3, p4]}^{remaining} = [21, 35, 58, 31, \dots]$: the first case visited state $[p3, p4]$ once (21 time units before completion), the second case visited $[p3, p4]$ once (35 time units before completion), the third case visited $[p3, p4]$ twice (58 and 31 time units before completion), etc. Similar multi-sets exist for elapsed times ($Q_{[p3, p4]}^{elapsed} = [21, 21, 15, 42, \dots]$) and sojourn times ($Q_{[p3, p4]}^{sojourn} = [2, 12, 5, 13, \dots]$). Based on these multi-sets all kinds of statistics can be computed. For example, the *mean remaining flow time* in state $[p3, p4]$ is $\sum_{q \in Q} \frac{Q(q) \times q}{|Q|}$ with $Q = Q_{[p3, p4]}^{remaining}$. Like in Sect. 9.4, it is possible to compute other standard statistics such as standard deviation, minimum, and maximum. One can also fit a distribution on the sample data using standard statistical software. For example, based on the samples $Q_{[p3, p4]}^{remaining} = [21, 35, 58, 31, \dots]$ one could find that these remaining flow times are best described by a Gamma distribution with parameters $r = 8.0502$ and $\lambda = 0.18915$. This distribution has a mean of 42.56 and a standard deviation of 15.0. As shown in Chap. 9, such insights can be used to extend models with the time information. Moreover, the annotated transition system can also be used actively, and predict the remaining time for a running case.

Figure 10.10 shows the transition system with annotations for state $[p3, p4]$. Moreover, the path of a partial trace of a case that is still running is highlighted in the figure. The partial trace of this case is $\langle a_{start}^{512}, a_{complete}^{518}, d_{start}^{525}, d_{complete}^{526}, b_{start}^{532}, b_{complete}^{533} \rangle$. At time 533 we are interested in the remaining flow time of this case. An obvious predictor for the remaining flow time of the running case is the mean

remaining flow time of all earlier cases in the same state, i.e., 42.56. Hence, the case is expected to complete around time 575.56. This illustrates that for any running case, at any point in time, one can predict the remaining flow time.

The annotated transition system can be used to make more refined statements about the predicted remaining flow time. For example, it is clear that the size of multi-set $Q_{[p3, p4]}^{\text{remaining}}$ and the standard deviation of the historic samples in this multi-set have impact on the reliability of the prediction. Rather than giving a single prediction value, it is also possible to produce predictions like “With 90% confidence the remaining flow time is predicted to be between 40 and 45 days” or “78% of similar cases were handled within 50 days”. Moreover, as shown in [167], it is possible to use cross-validation to determine the quality of predictions.

The approach based on an annotated transition system is not restricted to predicting the remaining flow time. Obviously, one could predict the sojourn time in a similar fashion. Moreover, also non-time related predictions can be made using the same approach. For example, suppose that we are interested in whether the request is accepted (activity g occurs) or rejected (activity h occurs). To make such predictions, we annotate states with information about known outcomes for “post mortem” cases. For example, $Q_{[p3, p4]}^{\text{accepted}} = [0, 1, 1, 1, \dots]$. For state $[p3, p4]$, a “0” is added to this multi-set for each visit of a case that will be rejected and “1” is added for each visit of a case that will be accepted. The average value of $Q_{[p3, p4]}^{\text{accepted}}$ is a predictor for the probability that a case visiting state $[p3, p4]$ will be accepted. This example shows that a *wide variety of predictions* can be generated using a suitable annotated transition system. It is important to note that process-related information is taken into account, i.e., the prediction is based on the *state* of the running case rather than some static attribute. Classical data mining approaches (e.g., based on regression or decision trees) typically use static attributes of a case rather than state information.

The transition system shown in Fig. 10.10 happens to coincide with the states of the WF-net and BPMN model provided earlier. However, as discussed in Sect. 7.4.1, different transition systems can be constructed based on an event log. The event log L and the state representation function $I^{\text{state}}()$ determine the level of detail and the aspects considered. For example, it is possible to abstract from irrelevant activities resulting in a more coarse-grained transition system. However, it is also possible to include information about resources and data in the state, thus resulting in a more fine-grained transition system. There should be sufficient visits to all states to make reliable predictions. The transition system is too fine-grained if many states are rarely visited when replaying log L . The level of abstraction should be consistent with the size of the log and the response variable that needs to be predicted. For supervised learning this is generally referred to as the problem of feature extraction, i.e., determining the predictor variables that are most relevant for predicting the response variable. See [167] for more details and examples.

The approach based on annotated transition systems is just one of many approaches that could be used for prediction. For example, *short-term simulation* could be used to explore the possible futures of a particular case in a particular state (see Sect. 9.6). The simulation model learned based on historic data is initialized with the

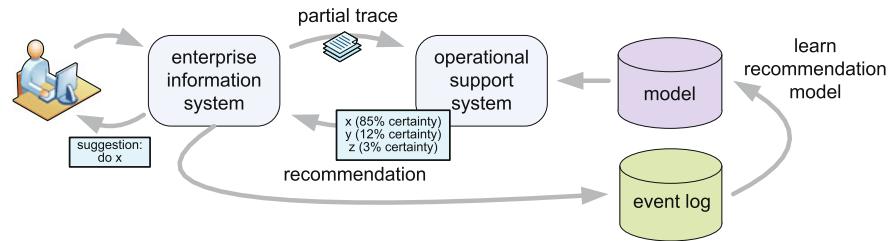


Fig. 10.11 A model based on historic data is used to provide recommendations for running cases. Recommendations are not enforced and may have quality attributes attached, e.g., in 85% of similar cases, x is the activity that minimizes flow time

current state of the running case. Subsequently, the remaining lifetime of the case is simulated repeatedly to obtain sample measurements for the performance indicator to be predicted.

10.5 Recommend

The third operational support activity we consider in this chapter is *recommendation*. As Fig. 10.11 shows, the setting is similar to prediction, i.e., a partial trace is sent to the operational support system followed by a response. However, the response is not a prediction but a recommendation about what to do next. To provide such a recommendation, a model is learned from “post mortem” event data. Moreover, the operational support system should know what the *decision space* is, i.e., what are the possible actions from which to choose one. Based on the recommendation model these actions are ordered. For example, in Fig. 10.11 the operational support system recommends to do action x with 85% certainty. The other two possible actions have a “lower” recommendation: y is recommended with 12% certainty and z is recommended with 3% certainty. In most cases it is impossible to give a recommendation that is guaranteed to be optimal; the best choice for the next step may depend on the occurrence of unknown external events in the future. For example, in Fig. 10.11 there may be cases for which z turns out to be the best choice.

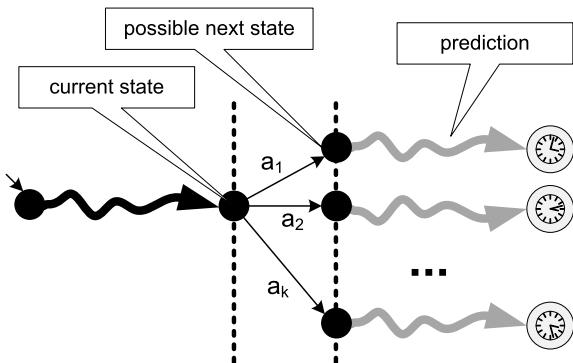
A recommendation is always given *with respect to a specific goal*. Examples of goals are:

- Minimize the remaining flow time;
- Minimize the total costs;
- Maximize the fraction of cases handled within 4 weeks;
- Maximize the fraction of cases that is accepted; and
- Minimize resource usage.

These goals can also be aggregated and combined, e.g., to balance between cost reduction and flow time reduction. To operationalize such a goal, a performance indicator needs to be defined, e.g., remaining flow time or total costs. This performance indicator corresponds to the response variable in supervised learning.

Fig. 10.12

Recommendations can be based on predictions. For every possible choice, simply predict the performance indicator of interest. Then, recommend the best one(s)



A recommendation makes statements about a set of possible actions, i.e., the decision space. The decision space may be a set of activities, e.g., $\{f, g, h\}$. This means that in the current state activities f , g , and h are possible candidates and the question to be answered by the operational support system is “Which candidate is best given the goal selected?”. However, the decision space may also consist of a set of resources and the goal is then to recommend the best resource to execute a given activity. For example, the operational support system could recommend allocating activity h to Mike to minimize the flow time. This example shows that recommendations are not limited to control-flow and can also refer to other perspectives. Therefore, we use the term “action” rather than activity. The decision space for a running case may be part of the message sent from the enterprise information system to the operational support system. Otherwise, the recommendation model should be able to derive the decision space based on the partial trace.

As shown in Fig. 10.12, recommending an action to achieve a goal is closely related to predicting the corresponding performance indicator. Suppose that for a case having a partial trace σ_p we need to recommend some action from a set of possible actions $\{a_1, a_2, \dots, a_k\}$. The existing partial trace can be extended by assuming that action a_1 is selected (although it did not happen yet). σ_1 is the resulting extended partial trace, i.e., $\sigma_1 = \sigma_p \oplus a_1$. (Here we assume that a_1 is an activity and we use simple traces.) The same can be done for all other actions resulting in a set of partial traces $D = \{\sigma_1, \sigma_2, \dots, \sigma_k\}$. Now a prediction is made for the selected performance indicator and each element of D . The resulting predictions are compared and ranked. If σ_2 has the best predicted value (e.g., shortest remaining flow time), then a_2 is recommended first.

Depending on the prediction technique used, the recommendation can also include information about its reliability/quality, e.g., the confidence or certainty that a particular selection is optimal with respect to the goal. For example, in Fig. 10.11 the recommendation attaches a confidence to each of the three possible actions. How to interpret such confidence values depends on the underlying prediction method. For example, if short-term simulation is used, then the 85% certainty of x mentioned in Fig. 10.11 (i.e., the confidence attached to recommendation x) would mean that in 85% of the simulation experiments action x resulted in the shortest remaining flow time.

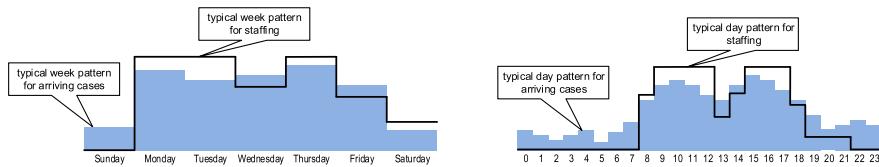


Fig. 10.13 Recurring weekly and daily patterns that are highly predictable but seldom used in analysis

10.6 Processes Are Not in Steady State!

In Sect. 3.1, we argued that models are often an idealized view on reality. As an example, we mentioned the so-called Yerkes–Dodson law suggesting that a person’s speed of working increases when workload increases (until a point where performance degrades due to stress). Such phenomena are typically not captured in simulation models [139, 163]. Obviously, this limits the predictive value of such models. Also models learned from event data may be blind to such phenomena. The main complication is that *processes are not in steady state*. Processes are influenced by working hours, weekends, contextual factors, and drifts. Understanding these phenomena is important, not only for operational support, but for interpreting process mining results in real-life settings.

10.6.1 Daily, Weekly and Seasonal Patterns in Processes

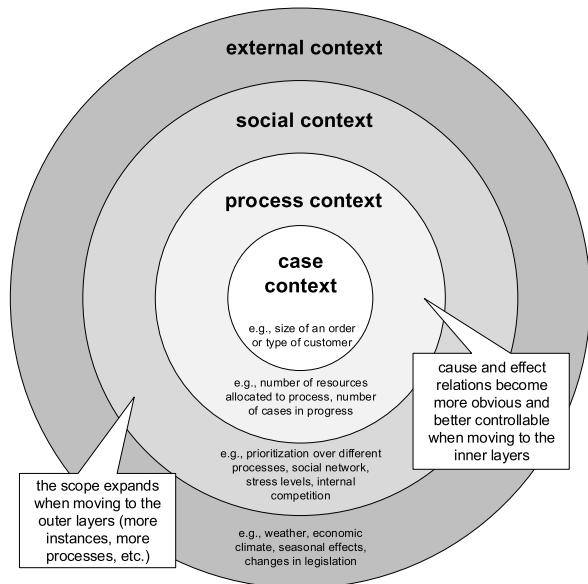
Figure 10.13 shows examples of typical weekly and daily patterns. Less work arrives during weekends and at night. Resource availability also fluctuates in a predictable manner due to office hours, lunch breaks and weekends. Staffing may be adapted to such patterns. However, there may still be congestion causing delays (often at foreseeable points in times).

Daily, weekly and seasonal patterns can be learned from event data. This is highly relevant for operational support. Assume that a person works only on Fridays or that she is only in the office in the morning. Using averages over historic data, we may predict that an activity will be completed by this person on Wednesday in the late afternoon. Such predictions are wrong and can easily be detected using cross-validation, however, the corresponding insights easily get lost in standard performance measures. The challenge is to incorporate recurring patterns when making predictions and suggesting improvements.

10.6.2 Contextual Factors

Processes are executed in a particular *context*, but this context is often neglected during analysis [118, 148]. The approach based on an annotated transition system

Fig. 10.14 The *context* in which events occur and processes unfold should be taken into account



presented in Sect. 10.4 reduces the context to the average remaining flow time of cases that visited the same state before (see Fig. 10.10). Let us compare predictions in operational processes to predicting driving times by a navigation system. Suppose we would like to know the time it takes to drive from Eindhoven to Amsterdam. Such a prediction could be based on properties of the driver and car. How long did people of the same age category driving the same brand of car take in the past? Such type of prediction is comparable to the analysis of the remaining flow time of cases in state $[p3, p4]$ in Fig. 10.10. However, such analysis neglects *important contextual factors*. The driving time may depend on the weather and the time of the day. More important, the driving time strongly depends on the other cars currently driving in the same direction! This illustrates the relevance of context in process mining.

In Fig. 10.14, we distinguish four types of context relevant for process mining [148]:

- **Case Context.** Process instances (i.e., cases) may have various properties that influence their execution. Consider, for example, the way a customer order is handled. The type of customer placing the order may influence the path the case follows in the process. The size of the order may influence the type of shipping selected or may influence the transportation time. These properties can be directly related to the individual process instance and we refer to them as the *case context*. Typically, it is not difficult to discover relationships between the case context and the observed behavior of the case. For example, one could discover that an activity is typically skipped for gold customers.
- **Process Context.** A process may be instantiated many times, e.g., thousands of customer orders are handled by the same process per year. Yet, the corresponding

process model typically describes the life-cycle of one order in isolation. Although interactions among process instances are not made explicit in such models, cases may influence each other. For example, instances may compete for the same resources. An order may be delayed by too much work-in-progress. Looking at one instance in isolation like in Fig. 10.10 is not sufficient for understanding the observed behavior. Process mining techniques should also consider the *process context*, e.g., the number of instances being handled and the number of resources available for the process. For example, when predicting the expected remaining flow time for a particular case one should not only consider the case context (e.g., the status of the order) but also the process context (e.g., workload and resource availability).

- *Social Context.* The process context considers all factors that can be directly related to a process and its instances. However, people and organizations are typically not allocated to a single process and may be involved in many additional processes. Moreover, activities are executed by people that operate in a social network. Friction between individuals may delay process instances and the speed at which people work may vary due to circumstances that cannot be fully attributed to the process being analyzed. All of these factors are referred to as the *social context*. This context characterizes the way in which people work together *within a particular organization*. Today's process mining techniques tend to neglect the social context even though it is clear that this context directly impacts the way that cases are handled.
- *External Context.* The external context captures all factors that are part of an even wider ecosystem that extends beyond the control sphere of the organization. For example, the weather, the economic climate, and changing regulations may influence the way that cases are being handled. The weather may influence the workload, e.g., a storm or flooding may lead to increased volumes of insurance claims. Changing oil prices may influence the number of customer orders (e.g., the demand for heating oil increases when prices drop). More stringent identity checks may influence the order in which social security related activities are being executed. Although the external context can have a dramatic impact on the process being analyzed, it is difficult to select the relevant variables.

The factors closely related to a case are often easy to identify. However the social and external contexts are more difficult to capture in a few variables that can be used by process mining algorithms. Moreover, analysis (e.g., predictions) may suffer from the so-called “curse of dimensionality” (see Sect. 4.6.3). In high-dimensional feature spaces, enormous amounts of event data are required to reliably learn the effect of contextual factors.

10.6.3 Concept Drift in Processes

The term *concept drift* refers to the situation in which the process is changing while being analyzed [81]. For instance, in the beginning of the event log two activities

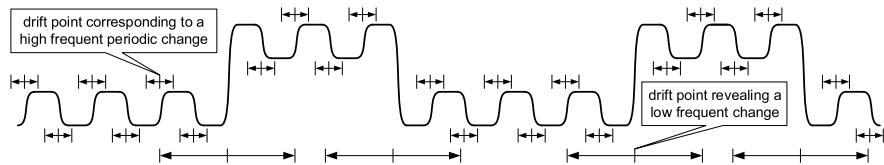


Fig. 10.15 A periodically changing process with two types of drift at different time scales. Shorter sliding windows are used to detect short-term drifts. The left half of the sliding window is compared with right half to spot statistically relevant changes. The larger sliding windows at the bottom are used for long-term drifts

may be concurrent whereas later in the log these activities become sequential. Processes may change due to periodic/seasonal changes (e.g., “in December there is more demand” or “on Friday afternoon there are fewer employees available”) or due to changing conditions (e.g., “the market is getting more competitive”). Such changes impact processes and it is vital to detect and analyze them.

There are three challenges when dealing with concept drift [81]:

- *Change point detection.* Did the process change? If so, when did it change?
- *Change localization and characterization.* What has changed?
- *Change process discovery.* How to capture and predict “second-order” dynamics?

Once a point of change has been identified, the next step is to characterize the nature of change, and to identify the region(s) of change in a process. Concept drift is challenging because of the dynamic nature of processes. Processes in steady-state are not static (“first-order” dynamics), making the detection of irregular behavior (“second-order” dynamics) challenging. Most approaches use a sliding window approach [81, 177, 188]. Different windows with events are compared using statistical methods to detect significant changes. To precisely localize change points and to detect drift at different time scales, the lengths of such windows need to be varied. Different types of drifts may be intertwined as illustrated by Fig. 10.15.

10.7 Process Mining Spectrum

The refined process mining framework shown in Fig. 10.1 illustrates the broadness of the *process mining spectrum*. We identified 10 process mining activities ranging from discovery and conformance checking to the three operational support activities described in this chapter. These activities may be concerned with “de jure” or “de facto” models, “pre mortem” or “post mortem” event data, and one or more perspectives (control-flow perspective, organizational perspective, case/data perspective, etc.). In this chapter, we showed that process mining techniques originally intended for off-line analysis can be adapted for operational support. For example, replay techniques originally developed for conformance checking can be used to *detect* policy violations, *predict* remaining flow times, and *recommend* activities in an online setting.

Part V

Putting Process Mining to Work

Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
**Process Mining:
The Missing Link**

Part II: Preliminaries

Chapter 3
**Process Modeling
and Analysis**

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
**Process Discovery:
An Introduction**

Chapter 7
**Advanced Process
Discovery Techniques**

Part IV: Beyond Process Discovery

Chapter 8
**Conformance
Checking**

Chapter 9
**Mining Additional
Perspectives**

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
**Process Mining
Software**

Chapter 12
**Process Mining in the
Large**

Chapter 13
**Analyzing “Lasagna
Processes”**

Chapter 14
**Analyzing “Spaghetti
Processes”**

Part VI: Reflection

Chapter 15
**Cartography and
Navigation**

Chapter 16
Epilogue

This part reviews available tool support for process mining. ProM and various commercial tools are introduced. Given the size of today's event logs, “Big Data” approaches to improve scalability are described. Next to tools, a methodology for process mining is provided. The methodology uses the L^* life-cycle model and identifies two types of processes (Lasagna and Spaghetti processes). Moreover, practical guidelines and real-life examples are provided.

Chapter 11

Process Mining Software

The successful application of process mining relies on good tool support. Traditional Business Intelligence (BI) tools are data-centric and focus on rather simplistic forms of analysis. Mainstream data mining and machine learning tools provide more sophisticated forms of analysis, but are also not tailored towards the analysis and improvement of processes. Fortunately, there are dedicated process mining tools able to transform event data into actionable process-related insights. For example, ProM is an open-source process mining tool supporting all of the techniques mentioned in this book. Process discovery, conformance checking, social network analysis, organizational mining, clustering, decision mining, prediction, and recommendation are all supported by ProM plug-ins. However, the usability of the hundreds of available plug-ins varies and the complexity of the tool may be overwhelming for end-users. In recent years, several vendors released dedicated process mining tools (e.g., Celonis, Disco, EDS, Fujitsu, Minit, myInvenio, Perceptive, PPM, QPR, Rialto, and SNP). These tools typically provide less functionality than ProM, but are easier to use while focusing on data extraction, performance analysis and scalability. This chapter provides an overview of available tools and trends.

11.1 Process Mining Not Included!

This book revolves around the analysis of behavior based on event data. Fueled by the growing availability of data (“Big Data”), data science emerged as a new discipline. As discussed in Sect. 1.3, data science approaches tend to be process agnostic. Process mining aims for duality (yin and yang) between data-driven forces and process-centric forces (see Fig. 2.1). The process mining spectrum is broad and, as shown in the previous chapters, extends far beyond process discovery and conformance checking. Process mining connects data science and process science (see Fig. 1.7). Hence, it is inevitable that process mining objectives are overlapping with those of other approaches, methodologies, principles, methods, tools, and paradigms. In Sect. 2.5, we discussed the relation to BPM, BPR, BI, Big Data, data

mining, Lean Six Sigma, etc. We posed questions like: “How does process mining compare to data mining?” (Sect. 2.5.2) and “How does process mining compare to Business Intelligence?” (Sect. 2.5.5). Books on data mining and BI seldom cover process mining techniques. The same holds for data mining and BI software. *Defining process mining as a particular type of machine learning, data mining or BI technique, will not extend the actual capabilities of (machine learning, data mining or BI) tools.* Software packages for machine learning and data mining *cannot* deal with process models (i.e., BPMN, EPC, UML, Petri nets, etc.) and do *not* support tasks like conformance checking. One needs dedicated process mining software for this: It is not included!

In the remainder of this chapter, we describe the capabilities of ProM and various commercial process mining tools. However, before doing so, we briefly discuss the market for BI products.

Forrester defines *Business Intelligence* (BI) in two ways. The broad definition provided by Forrester is “BI is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making” [55]. Forrester also provides a second, more narrow, definition: “BI is a set of methodologies, processes, architectures, and technologies that leverage the output of information management processes for analysis, reporting, performance management, and information delivery” [55].

Some of the most widely used BI products are [56]: *IBM Cognos Business Intelligence* (IBM), *Oracle Business Intelligence* (Oracle), *SAP BusinessObjects* (SAP), *MS SQL Server/Power BI* (Microsoft), *MicroStrategy* (MicroStrategy), *QlikView* (QlikTech), *SAS Business Intelligence* (SAS), *TIBCO Spotfire Analytics* (TIBCO), *Jaspersoft BI Enterprise* (Jaspersoft), and *Pentaho BI Platform* (Pentaho). The typical functionality provided by these products includes:

- *ETL* (Extract, Transform, and Load). All products support the extraction of data from various sources. The extracted data is then transformed into a standard data format (typically a multidimensional table) and loaded into the BI system.
- *Ad-hoc querying*. Users can explore the data in an ad-hoc manner (e.g., drilling down and “slicing and dicing”).
- *Reporting*. All BI products allow for the definition of standard reports. Users without any knowledge of the underlying data structures can simply generate such predefined reports. A report may contain various tables, graphs, and scorecards.
- *Interactive dashboards*. All BI products allow for the definition of dashboards consisting of tabular data and a variety of graphs. These dashboards are interactive, e.g., the user can change, refine, aggregate, and filter the current view using predefined controls.
- *Alert generation*. It is possible to define events and conditions that need to trigger an alert, e.g., when sales drop below a predefined threshold an e-mail is sent to the sales manager.

The mainstream BI products from vendors such as IBM, Oracle, SAP, and Microsoft do *not* support process mining. All of the systems mentioned earlier are

data-centric and are *unaware* of the processes the data refers to. The focus is on fancy-looking dashboards and rather simple reports, instead of a deeper analysis of the data collected. This is surprising as the “I” in BI refers to “intelligence”. Unfortunately, the business ~~un~~intelligence market is dominated by large vendors that focus on monitoring and reporting rather than analytics. Data mining or statistical analysis are often added as an afterthought.

Most BI tools provide interfaces to data mining tools. For example, open-source BI products from organizations like Jaspersoft and Pentaho can connect to open-source data mining tools such as *WEKA* (Waikato Environment for Knowledge Analysis, weka.wikispaces.com), *RapidMiner* (www.rapidminer.com), *KNIME* (Konstanz Information Miner, www.knime.org), and *R* (www.r-project.org). These provide more “intelligence” than mainstream BI tools.

WEKA is a widely-used prototypical example of a data mining tool [190]. *WEKA* supports classification (e.g., decision tree learning), clustering (e.g., *k*-means clustering), and association rule learning (e.g., the Apriori algorithm). *WEKA* expects so-called “arff” files as input. Such a file stores tabular data such as shown in Tables 4.1, 4.2, and 4.3. It is impossible to directly load an event log into *WEKA*. However, it is possible to convert XES or MXML data into tabular data that can be analyzed by *WEKA* [42]. After conversion each row either corresponds to an event or a case. For example, it is possible to extract variables like flow time and the frequency of some activity for each case. Similarly, it is possible to create a table where each row lists the attributes of some event. However, either way, the original event notion is lost. This illustrates that data mining tools, like the mainstream BI products, are data-centric rather than process-centric.

Tools such as *RapidMiner*, *KNIME*, and *R* are extendible. For example, *RapidMiner* provides a marketplace where users can acquire additional building blocks (e.g., for text mining). *RapidProM* (www.rapidprom.org), available through the *RapidMiner* marketplace, provides a collection of process mining building blocks based on ProM plug-ins (see Sect. 11.3.3). This way users of *RapidMiner* are able to use process mining techniques without installing a separate process mining tool [23, 97].

In general, one cannot assume that BI and data mining tools provide any process mining capabilities. Fortunately, plenty of dedicated process mining tools are available. These are discussed in the remainder of this chapter.

11.2 Different Types of Process Mining Tools

Before describing concrete process mining tools, we first discuss different ways of characterizing process mining software.

Potentially, there may be very different groups of users interacting with process mining software. On the one hand, there may be experts that need to be able to answer “one of a kind” questions requiring ad-hoc data extractions, complex data transformations, and sophisticated analysis techniques. On the other hand, there can

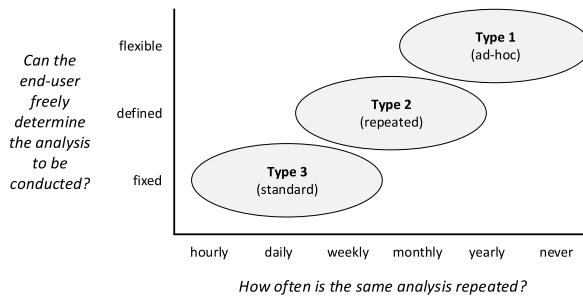


Fig. 11.1 Three types of use cases: *Type 1* (for ad-hoc questions requiring data exploration/extraction and problem-driven selection of analysis techniques), *Type 2* (for repeated questions in a known setting but possibly still requiring configuration), and *Type 3* (for standard questions in a fixed pre-configured setting)

be end-users that just want to look at standard overviews (“process-centric dashboards”) generated using process mining.

The spectrum of process mining use cases can be characterized through the following two questions:

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

Fig. 11.1 defines three types of use cases based on answers to these two questions.

Use cases of *Type 1* (ad-hoc) require a spreadsheet-like tool: questions are ad-hoc and the user needs to have complete freedom to perform analysis. The analysis process is iterative and undefined. The results of one analysis step may lead to unanticipated additional data extractions (or transformations) to enable the next analysis step. Analysis workflows are unique and seldom repeated.

Use cases of *Type 2* (repeated) involve questions that are recurring, but at a lower frequency. Analysis workflows may be predefined but not completely fixed. Customization may be needed and the interpretation of the results requires knowledge of process mining and understanding of the data.

Use cases of *Type 3* (standard) involve routine questions that are recurring frequently. The different analysis views are fixed and no customization is possible. The user only needs to understand predefined dashboard-like views.

The three types are on the diagonal in Fig. 11.1. Use cases not on the diagonal do not make much sense. For example, we cannot provide a predefined dashboard for “one of a kind” questions (corresponding to the combination of “never” and “fixed” in Fig. 11.1). Moreover, if there is a continuous need to answer the same question based on the latest data, then there is no need to explore the data in an ad-hoc manner (i.e., also the combination of “hourly” and “flexible” in Fig. 11.1 makes no sense).

Process mining tools may be tailored to one of the three types in Fig. 11.1. For example, a tool like Disco (Fluxicon) is comparable to a spreadsheet program (but for “behavior” rather than “numbers”, see Sect. 1.3). The user can load the data of interest, pick a particular view, and get immediate results without any system configuration. Such style of interaction is good for exploration and fast results (*Type 1*),

but less suitable for end-users that do not understand the underlying data and analysis techniques (*Type 3*).

The initial investment for a *Type 1* analysis is low, but less suitable for situations where many users repeatedly need to do the same type of analysis. The initial investment for a *Type 3* analysis is much higher. An expert needs to configure the way data is extracted and define the views on the data provided to end users. However, after the initial investment, analysis is easier and highly repeatable. *Type 2* analysis is in-between *Type 1* and *Type 3*. Use cases of *Type 2* benefit from analysis workflows that are (partly) predefined but not completely fixed.

Another way to categorize process mining software is based on the way it is bundled:

- *Dedicated* process mining software—pure play process mining tools devoted to the analysis of event data and processes.
- *Embedded* process mining software—tools that provide process mining functionality, but that are embedded in a larger suite.

Most of the tools discussed in this chapter fall in the first category. However, process mining functionality may also be embedded in a larger BPM, ERP, BI or data mining product as an add-on. RapidProM, an extension of RapidMiner, is an example of embedded process mining software [97].

Process mining tools can also be classified based on their “openness”:

- *Open-source* process mining software—the source code is publicly available. Depending on the license other parties can extend, change, or redistribute the software.
- *Closed-source* process mining software—proprietary software whose source code is not published and cannot be changed or extended.

The commercial process mining tools described in Sect. 11.4 are closed-source. ProM is an example of an open-source tool.

All process mining tools are able to discover process models, but the types of models learned from event data vary. We distinguish three classes of models:

- *Informal* process models—“boxes and arrows” diagrams not having a formal interpretation that can be related to traces in the event log.
- *Formal low-level* process models—transition systems, Markov chains, episodes, sequences, etc.
- *Formal high-level* process models—end-to-end models allowing for choices, concurrency, loops, etc. This includes BPMN models, EPC models, UML activity diagrams, Petri nets, process trees, etc.

Formal models have executable semantics. Informal models are drawings composed of boxes and arrows without a clear relation to the traces in the event log. Such informal diagrams do not distinguish between choice and concurrency (there are no AND/XOR/OR-gateways/connectors/operators). A model is formal if, given a sequence of events, one can determine whether it fits or not. Process mining tools are characterized by the process models they support. Most commercial process

mining tools use a mixture of informal and low-level models (see Sect. 11.4.2). The fact that a discovered model can be saved in BPMN format (or any other format with AND/XOR/OR-gateways/connectors/operators) does not imply that the model can be interpreted as such.

Process mining starts from event data. Process mining tools may have different mechanisms to get event data:

- *File*. Events are stored in a XES, MXML, Excel, or CSV file.
- *Database*. Events are loaded from a database system, for example via a JDBC connection. Several tools support incremental event loading, i.e., periodically the database is inspected for new data.
- *Adapter*. Events are loaded from a particular application (e.g., SAP, Sharepoint, or SalesForce) through a dedicated piece of software. In most cases events can be loaded incrementally.
- *Streaming*. The process mining tool works on a stream of events emitted through an event bus or web service. Events are captured as they occur and not retrieved from a file, database, or application at a later point in time.

The process mining software may run locally or remotely. The event data typically resides at the same location. We distinguish three types of deployments:

- *Stand-alone*. The software runs locally, e.g., on the laptop used for analysis.
- *On premise*. The back-end of the software does not run locally, but on a server inside the organization.
- *Cloud*. The software runs on a server outside the organization.

Some products offer multiple forms of deployment. This is not only a technological decision, but also related to privacy laws, security, and ethics. For example, the cloud provider may store event data on a server in a different country.

The refined process mining framework (Sect. 10.1) identifies the following activities:

- *Discover*—learning (process) models from event data;
- *Enhance*—repair or extend models (adding additional perspectives to a model, e.g., to show bottlenecks);
- *Diagnose*—model-based process analysis;
- *Detect*—comparing de jure models with current “pre mortem” data (events of running process instances) to detect deviations at runtime;
- *Check*—checking conformance by comparing historic “post mortem” data with de jure models (e.g., to pinpoint deviations and quantify compliance);
- *Compare*—comparing de jure models with de facto models to see whether reality deviates from what was planned or expected;
- *Promote*—transferring “best practices” (learned from event data) to the de jure model;
- *Explore*—exploring business processes at run-time using a combination of event data and models;
- *Predict*—making process-related predictions, e.g., the remaining flow time and the probability of non-compliance;

- *Recommend*—supporting operational processes by recommending suitable actions (e.g. to minimize costs or time).

Process mining software can be characterized by the activities supported. For example, all tools support activity *discover*, but only few support activities like *predict* and *recommend*.

In the remainder, we first describe ProM and then provide an overview of other process mining tools, including 11 commercial products.

11.3 ProM: An Open-Source Process Mining Platform

ProM is the leading open-source process mining tool. The lion’s share of academic research is conducted by using and extending ProM. Moreover, the commercial process mining tools discussed in Sect. 11.4 are based on ideas first developed in the context of ProM. Therefore, this section first introduces ProM which is tailored towards use cases of *Type 1* (see Fig. 11.1).

11.3.1 Historical Context

In 2002, there were several, rather simple, stand-alone process mining tools available. Examples of tools developed around the turn of the century include: *MiMo* (α -miner based on ExSpect), *EMiT* (α -miner taking transactional information into account), *Little Thumb* (predecessor of the heuristic miner), *InWolvE* (miner based on stochastic activity graphs), and *Process Miner* (miner assuming structured models) [156]. At this time, several researchers were building simple prototypes to experiment with process discovery techniques. However, these tools were based on rather naïve assumptions (simple process models and small but complete data sets) and provided hardly any support for real-life process mining projects (scalability, intuitive user interface, etc.). Clearly, it did not make any sense to build a dedicated process mining tool for every newly conceived process discovery technique. This observation triggered the development of the *ProM framework*, a “plug-able” environment for process mining using MXML as input format. The goal of the first version of this framework was to provide a *common basis* for all kinds of process mining techniques, e.g., supporting the loading and filtering of event logs and the visualization of results. This way people developing new process discovery algorithms did not have to worry about extracting, converting, and loading event data. Moreover, for standard model types such as Petri nets, EPCs, and social networks, default visualizations were provided by the framework.

In 2004, the first fully functional version of the ProM framework (*ProM 1.1*) was released. This version contained 29 plug-ins: 6 mining plug-ins (the classic α -miner, the Tshinghua α miner, the genetic miner, the multi-phase miner, the social network miner, and the case data extraction miner), 7 analysis plug-ins (e.g., the

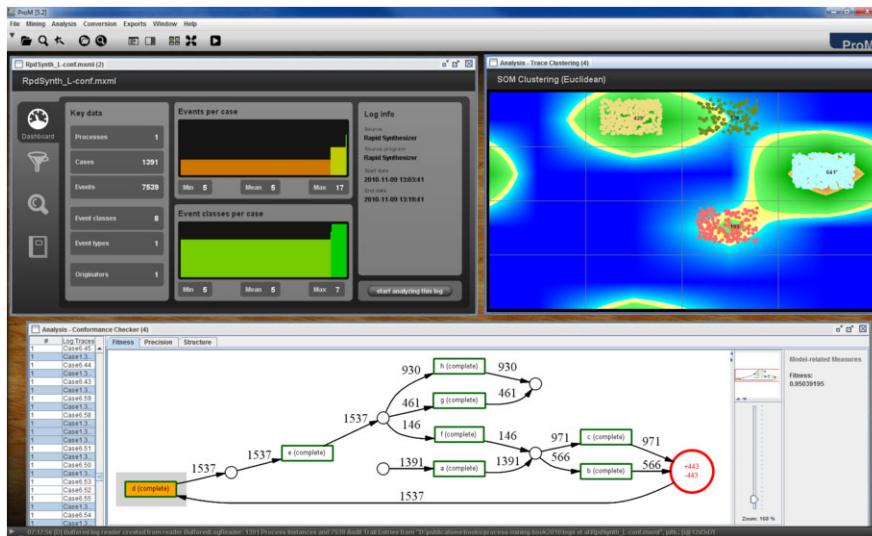


Fig. 11.2 Screenshot of ProM 5.2 showing two of the 286 plug-ins. The *bottom window* shows the *conformance checker* plug-in while checking the fitness of event log L_{full} described in Table 8.1 and WF-net N_2 depicted in Fig. 8.2. The plug-in identifies the conformance problem (the log and model disagree on the position of d) and returns a fitness value computed using the approach presented in Sect. 8.2, $\text{fitness}(L_{full}, N_2) = 0.95039195$. The *right window* shows the trace clustering plug-in using Self Organizing Maps (SOM) to find homogeneous groups of cases. The largest cluster contains 641 cases. These are the cases that were rejected without a thorough examination (i.e., traces $\sigma_1, \sigma_3, \sigma_{13}$ in Table 8.1)

LTL checker), 4 import plug-ins (e.g., plug-ins to load Petri nets and EPCs), 9 export plug-ins, and 3 conversion plug-ins (e.g., a plug-in to convert EPCs into Petri nets). Over time more plug-ins were added. For instance, ProM 4.0 (released in 2006) contained already 142 plug-ins. The 27 mining plug-ins of ProM 4.0 included also the heuristic miner and a region-based miner using Petrify. Moreover, ProM 4.0 contained a first version of the conformance checker described in [121]. ProM 5.2 was released in 2009. This version contained 286 plug-ins: 47 mining plug-ins, 96 analysis plug-ins, 22 import plug-ins, 45 export plug-ins, 44 conversion plug-ins, and 32 filter plug-ins. Figure 11.2 shows two plug-ins of ProM 5.2. This version already supported most of the process mining techniques presented in this book. For example, the 47 mining plug-ins of ProM 5.2 include most of the discovery algorithms presented in Chap. 7 (genetic mining, heuristic mining, fuzzy mining, etc.). The replay approach presented in Sect. 8.2 was supported by the conformance checker plug-in of ProM 5.2 [121].

The spectacular growth of the number of plug-ins in the period from 2004 to 2009 illustrates that ProM realized its initial goal to provide a platform for the development of new process mining techniques. ProM had become the de facto standard for process mining. Research groups from all over the globe contributed to the development of ProM and people from tens of thousands of organizations down-

loaded ProM (the ProM framework has been downloaded over 130.000 times). In the same period, we applied ProM at numerous organizations, e.g., in the context of joint research projects, Master projects, and consultancy projects. The large number of plug-ins and the many practical applications also revealed some problems. For example, ProM 5.2 can be quite confusing for the inexperienced user who is confronted with almost 300 plug-ins. Moreover, in ProM 5.2 (and earlier versions) the user interface and the underlying analysis techniques are tightly coupled, i.e., most plug-ins require user interaction. It was impossible to embed ProM functionality in data mining tools such as RapidMiner, KNIME, etc. due to this tight coupling.

To be able to run ProM remotely and to embed process mining functionality in other systems, we decided to completely re-implement ProM from scratch. This allowed us to learn from earlier experiences and to develop a completely new architecture based on an improved plug-in infrastructure.

ProM 6 was released in November 2010. This was the first version based on the new architecture and XES rather than MXML. XES, described in Sect. 5.3, is the process mining standard adopted by the IEEE Task Force on Process Mining. Although ProM 5.2 was already able to load enormous event logs, scalability and efficiency were further improved by using OpenXES [64, 65]. Not all plug-ins of ProM 5.2 have been re-implemented in ProM 6. Nevertheless, most of the process mining techniques described in this book are supported by plug-ins developed for ProM 6.

ProM 6 can distribute the execution of plug-ins over multiple computers. This can be used to improve performance (e.g., using grid computing) and to offer ProM as a service. For instance, at TU/e (Eindhoven University of Technology) we use a dedicated process mining grid to handle huge data sets and to conduct large-scale experiments. The user interface has been re-implemented to be able to deal with many plug-ins, logs, and models at the same time. Plug-ins are now distributed over so-called *packages* and can be chained into composite plug-ins. Packages contain related sets of plug-ins. ProM 6 provides a so-called package manager to add, remove, and update packages. Users should only load packages that are relevant for the tasks they want to perform. This way it is possible to avoid overloading the user with irrelevant functionality. Moreover, ProM 6 can be customized for domain-specific or even organization-specific applications.

Figures 11.3 and 11.4 show the selection of the ILP miner plug-in (based on language-based regions, see Sect. 7.4.3) and the resulting process model discovered by ProM 6. ProM 6.5.1a (SilvR+) was released in October 2015. There is also a “ProM Lite” version providing only the most used functionality.

11.3.2 Example ProM Plug-Ins

ProM is open-source software¹ and can be freely downloaded from www.promtools.org or www.processmining.org. Plug-ins can be installed via ProM’s package man-

¹ProM framework is released under the GNU Lesser General Public License (L-GPL).



Fig. 11.3 Screenshot of ProM 6.5. After loading an event log, a list of applicable plug-ins is shown and the plug-in implementing discovery using language-based regions (ILP miner) is selected

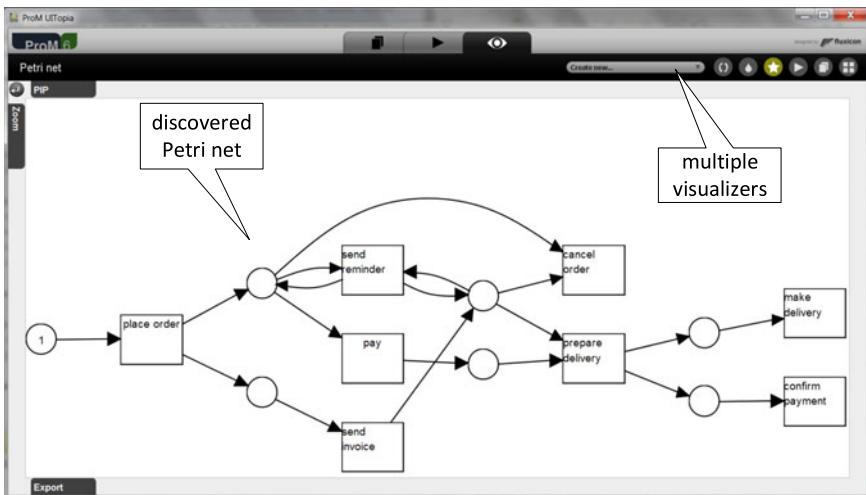


Fig. 11.4 Screenshot of ProM 6.5 showing the Petri net discovered using language-based regions after starting the plug-in selected in Fig. 11.3

ager. Currently, there are over 1500 plug-ins available (including deprecated plug-ins that are no longer supported). The ILP miner plug-in depicted in Fig. 11.4 is just one of these 1500 plug-ins. Hence, it is impossible to provide a complete overview of the functionality of ProM. The reader is encouraged to visit www.processmining.org to learn more about ProM's functionality and available plug-ins. Here, we only show a few examples.

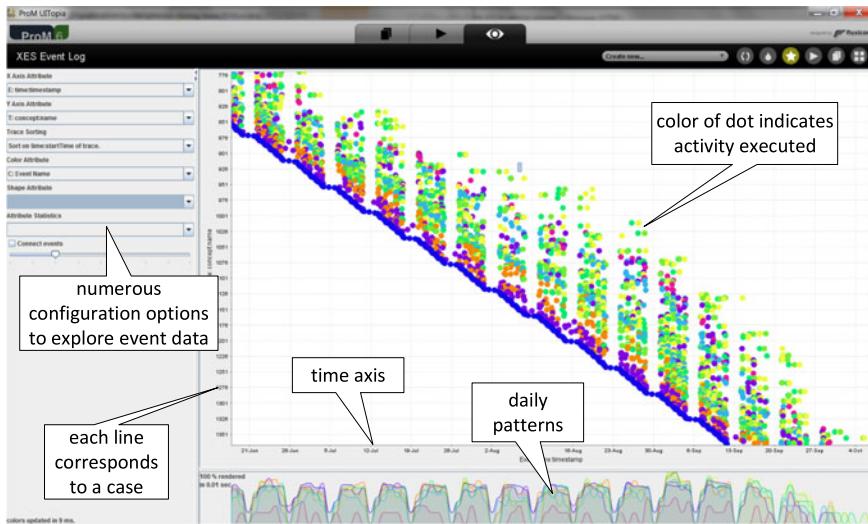


Fig. 11.5 ProM’s dotted chart can be used to explore the event data from different angles

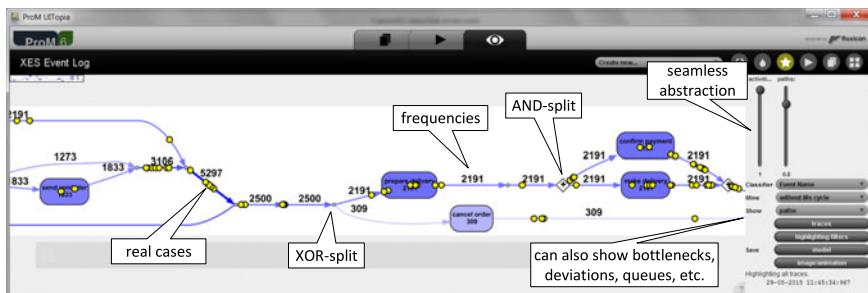


Fig. 11.6 Visual inductive miner replaying the event log on the discovered process model

ProM can load XES, MXML, and CSV files. To extract files from other data sources, tools such as XESame and ProMimport can be used (cf. Sect. 5.3). Figure 11.5 shows a *dotted chart* (see Sect. 9.2). The user can control both axes completely and influence the coloring and shape of the dots.

ProM supports dozens of process discovery algorithms. Next to the ILP miner shown in Fig. 11.4 and the α -algorithm [157], also heuristic mining [183, 184], fuzzy mining [66], genetic process mining [12, 26], and various forms of inductive mining [89–91] are supported. Figure 11.6 shows the *visual inductive miner*. This miner always returns a sound process model and is able to handle large and noisy event logs. Nevertheless, the miner can ensure (if desired) perfects fitness. Results can be converted to Petri nets, EPCs, statecharts and BPMN models. Moreover, the visual inductive miner supports bottleneck analysis and outlier detection.

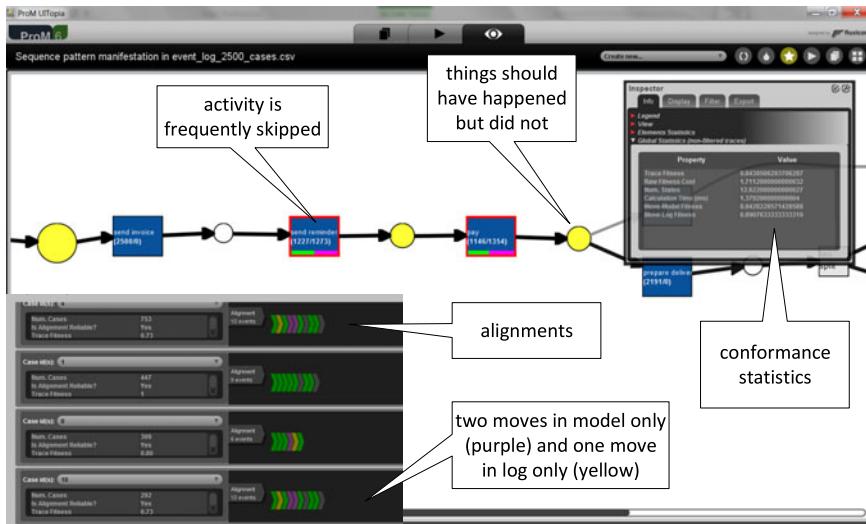


Fig. 11.7 Conformance checking based on alignments (cf. Sect. 8.3)

Conformance checking and performance analysis heavily depend on reliable replay algorithms [169]. Newer plug-ins in ProM rely on alignments as described in Sect. 8.3. Figure 11.7 shows deviations from both the log and model perspective. Next to fitness also notions such as precision are computed [5]. Similar views are provided for performance diagnostics based on alignments.

ProM also supports trace clustering [78], trace alignment [79], and model repair [52]. Next to a variety of procedural models (Petri nets, BPMN, YAWL, EPCs, etc.), ProM also support declarative models. Declare models can be discovered and the conformance of declarative models can be checked.

ProM is not limited to the control-flow and time perspectives. There are plug-ins to create social networks and to discover roles (organizational perspective). There are also plug-ins for decision mining [40, 120]. These plug-ins enhance control-flow models with guards based on the data perspective. Plug-ins can discover so-called *data-aware Petri nets* and check the conformance of such models [40]. Many of these plug-ins create classification problems. See [42] for a ProM plug-in that supports the interaction between process mining and data mining in a generic manner.

The plug-ins mentioned thus far are all related to process mining. However, it should be noted that ProM (both version 5.2 and 6.X) supports process analysis in the broadest sense, e.g., also the analysis techniques mentioned in Sect. 3.3 are supported by ProM or the tools that ProM interfaces with (e.g., CPN Tools). For example, the plug-in “Analyze structural properties of a Petri net” computes transition invariants, place invariants, S-components, T-components, traps, siphons, TP- and PT-handles, etc. The plug-in “Analyze behavioral properties of a Petri net” computes unbounded places, dead transitions, dead markings, home markings, coverability graphs, etc. The “Woflan” plug-in checks the soundness of WF-nets

(cf. Sect. 3.2.3) [179]. Moreover, powerful Petri-net-based analysis tools such as *LoLa*, *Wendy*, *Uma*, and *Petrify* are embedded in ProM as plug-ins.

The hundreds of ProM plug-ins implementing all of the techniques described in this book (and many more) illustrate the applicability and broadness of process mining.

11.3.3 Other Non-commercial Tools

Due to the success of ProM in the academic community, there are only a few other non-commercial process mining tools. Research groups all over the world have contributed to the 1500 plug-ins in ProM (also see the list of organizations mentioned in the Acknowledgements). Next to ProM, most other tools are commercial (cf. Sect. 11.4). A few notable exceptions are described next.

11.3.3.1 PMLAB

PMLAB is a scripting environment for process mining developed by the group of Josep Carmona at Universitat Politècnica de Catalunya in Barcelona. PMLAB can load XES, MXML, and CSV files and different analysis steps can be chained together in scripts. A variety of process discovery techniques based on the theory of regions and satisfiability modulo theories are supported. Tools such as *Genet*, *Petrify*, *Rbminer*, and *Dbminer* can be invoked from PMLAB. These tools are mostly based on state-based regions [34]. As shown in Sect. 7.4, an event log can be converted into a transition system and subsequently synthesized into a Petri net. Classical region theory needs to be extended/relaxed to make it more applicable for process discovery, e.g., Rbminer adapts the classical theory to provide more compact and readable process models [128]. PMLAB uses iPython, a framework for scripting in Python. PMLAB can also invoke ProM plugins through PMLAB scripts. The scripting tool was inspired by MATLAB and Mathematica. However, there are also similarities with RapidMiner, KNIME, and R. PMLAB can be downloaded from <https://www.cs.upc.edu/~jcarmona/PMLAB/>.

11.3.3.2 CoBeFra

CoBeFra is a benchmarking framework for conformance checking developed at the department of Management Informatics at KU Leuven in Belgium. It is mostly used for the systematic evaluation of process discovery techniques. CoBeFra reads event logs in XES or MXML file format and process models in PNML format. Given an event log and a process model, the tool evaluates the model with respect to dozens of metrics (e.g., various notions of fitness, precision, and simplicity). For large scale experiments (e.g. varying parameters of discovery algorithms to analyze the effects on fitness and precision), the metrics are automatically collected for sets of models and logs. CoBeFra can be downloaded from <http://www.processmining.be/cobefra>.

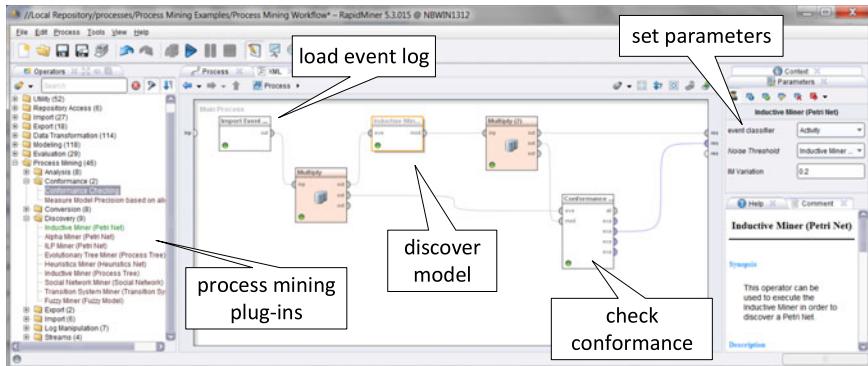


Fig. 11.8 A process mining workflow where an event log is loaded, a model is discovered using the inductive miner, and the result is checked using the conformance checker based on alignments

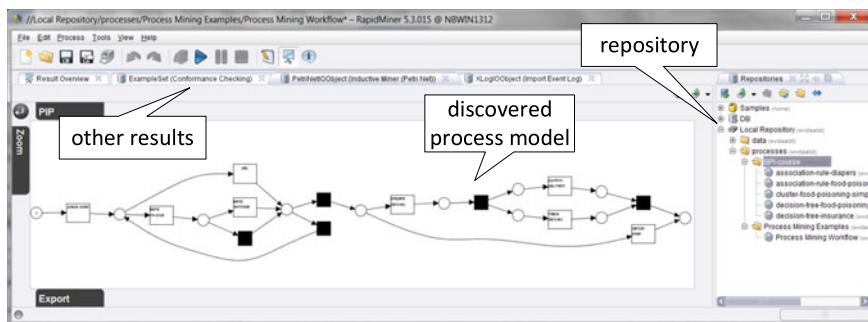


Fig. 11.9 One of the outcomes of the analysis workflow in Fig. 11.8

11.3.3.3 RapidProM

Many tools for data analysis support the definition and execution of *analysis workflows*, sometimes also called *scientific workflows*. For example, widely used tools like RapidMiner, KNIME, and R can chain together building blocks to form such workflows. However, these tools do not support process mining natively. Conversely, ProM does not provide such workflow support. Therefore, *RapidMiner* was extended with process mining plug-ins from ProM. The resulting tool is called *RapidProM* (www.rapidprom.org).

Figure 11.8 shows a process mining workflow created using RapidProM. First, a XES log is loaded. Second, a process model is discovered using the *Inductive Miner*—*infrequent* (IMF, [89]). The quality of this model is checked using alignments using another building block. The workflow in Fig. 11.8 can be stored and applied to any event log. Figure 11.9 shows one of the output objects produced by the workflow (the discovered process tree was automatically converted to a Petri net).

RapidProM can be used to do large scale experiments [23, 97]. For example, the workflow in Fig. 11.8 can be applied to thousands of event logs without any manual intervention. RapidProM can also be applied to answer recurring questions in a business setting, e.g., to create a report at the end of every week.

RapidProM is available via the *RapidMiner Marketplace*. Many other types of analysis are available in the RapidMiner ecosystem. This facilitates the combination of process mining, text mining, machine learning, data mining, and statistics. For example, cases can be grouped into clusters using standard data mining techniques followed by the application of process mining techniques on each of these clusters.

Whereas ProM is most suitable for use cases of *Type 1*, RapidProM, CoBeFra, and PMLAB are tailored towards use cases of *Type 2* (see Fig. 11.1).

11.4 Commercial Software

Several commercial process mining tools emerged on the market in recent years. Compared to ProM these tools are easier to use, but provide less functionality than the 1500 plug-ins in ProM. This lowers the threshold for using process mining significantly.

This section provides an overview of the commercial process mining tools currently on the market. *The goal is not to give detailed information on specific tools or to provide checklists.* The market and tools change rapidly. Most of the tools described did not exist when the first version of this book was published in 2011 [140]. Moreover, the capabilities of tools change with every release and usability and scalability cannot be expressed in simple checklists. *Hence, organizations that are selecting a commercial process mining tool are urged to evaluate the tools based on concrete questions and datasets.*

After illustrating some of the commercial tools in Sect. 11.4.1, we share a few general insights based on experiences with currently available process mining tools in Sect. 11.4.2.

11.4.1 Available Products

Table 11.1 lists 11 process mining tools in alphabetical order: *Celonis Process Mining* (Celonis), *Disco* (Disco), *Enterprise Discovery Suite* (EDS), *Interstage Business Process Manager Analytics* (Fujitsu), *Minit* (Minit), *myInvenio* (myInvenio), *Perceptive Process Mining* (Perceptive), *QPR ProcessAnalyzer* (QPR), *Rialto Process* (Rialto), *SNP Business Process Analysis* (SNP), and *webMethods Process Performance Manager* (PPM). For tools with a longer name, the shorter name between brackets is used. For example, “*webMethods Process Performance Manager*” is abbreviated to PPM.

Tools like Disco, Fujitsu, QPR, and PPM have been around for a few years. Minit, myInvenio, and Rialto emerged very recently (in 2015). Tools like *Process*

Table 11.1 Overview of commercial process mining tools

Short name	Full name of tool	Version	Vendor	Webpage
Celonis	Celonis Process Mining	4.0	Celonis GmbH	www.celonis.de
Disco	Disco	1.9	Fluxicon	www.fluxicon.com
EDS	Enterprise Discovery Suite	4	StereoLOGIC Ltd	www.stereologic.com
Fujitsu	Interstage Business Process Manager Analytics	12.2	Fujitsu Ltd	www.fujitsu.com
Minit	Minit	1.0	Gradient ECM	www.minitlabs.com
myInvenio	myInvenio	1.0	Cognitive Technology	www.my-invenio.com
Perceptive	Perceptive Process Mining	2.7	Lexmark	www.lexmark.com
QPR	QPR ProcessAnalyzer	2015.5	QPR	www.qpr.com
Rialto	Rialto Process	1.5	Exeura	www.exeura.eu
SNP	SNP Business Process Analysis	15.27	SNP Schneider-Neureither & Partner AG	www.snp-bpa.com
PPM	webMethods Process Performance Manager	9.9	Software AG	www.softwareag.com

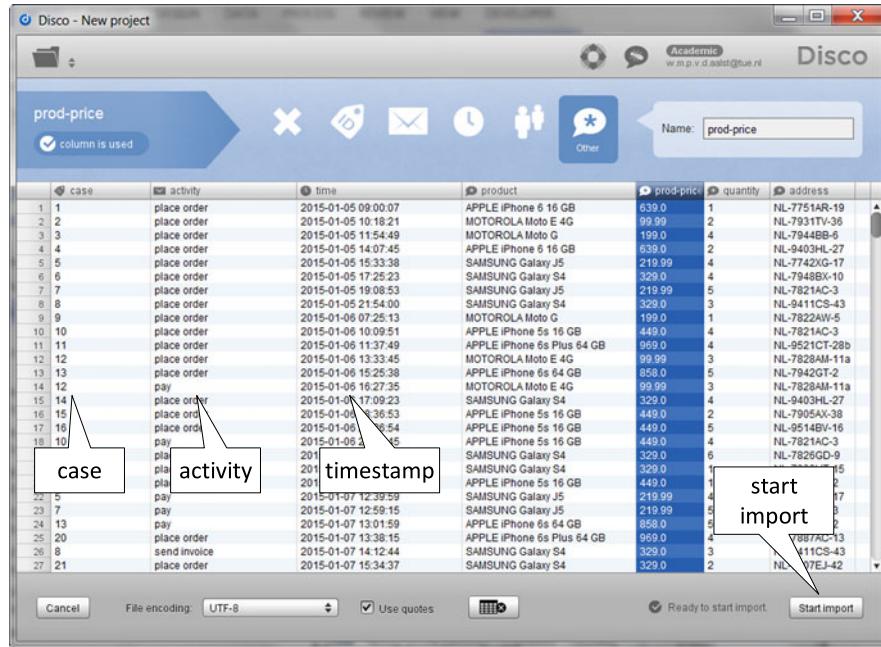


Fig. 11.10 Disco allows for the easy import of CSV files and supports process mining formats such as XES and MXML

Discovery Focus (Iontas/Verint Systems) and *Enterprise Visualization Suite* (Businesscape) are no longer available. Earlier products such as *Reflect|one* by Pallas Athena and *Reflect* by Futura Process Intelligence were further developed as part of the Perceptive suite of BPM tools. It is interesting to note that both Pallas Athena and Futura Process Intelligence were selected as “Cool Vendor” by Gartner in 2009 because of their process mining capabilities. Reflect was the first dedicated commercial process mining tool. The *ARIS Process Performance Manager* (PPM) was initially developed by IDS Scheer. Process mining capabilities were added later and PPM is now part of Software AG’s webMethods Operational Intelligence Platform.

As mentioned, it is not our goal to discuss particular tools in detail. However, we show a few screenshots to provide an impression of typical capabilities of available tools. Figure 11.10 shows a screenshot of Disco while loading a CSV file. The columns can be mapped onto process mining concepts such as case, activity, timestamp, and resource. Disco automatically suggests an initial mapping (including the format to be used for timestamps) that can be adapted. Disco can also load and save event logs in XES and MXML format. Researchers often use Disco for an initial analysis of the data (involving filtering, exploration, and bottlenecks analysis) after which XES files are saved for further analysis using ProM. The discovery algorithm used by Disco can be viewed as an improved and further developed version of ProM’s Fuzzy Miner [66]. The scalability and robustness of Disco are much better than the original Fuzzy Miner. Disco is easy to use and learn, and lowers the barrier

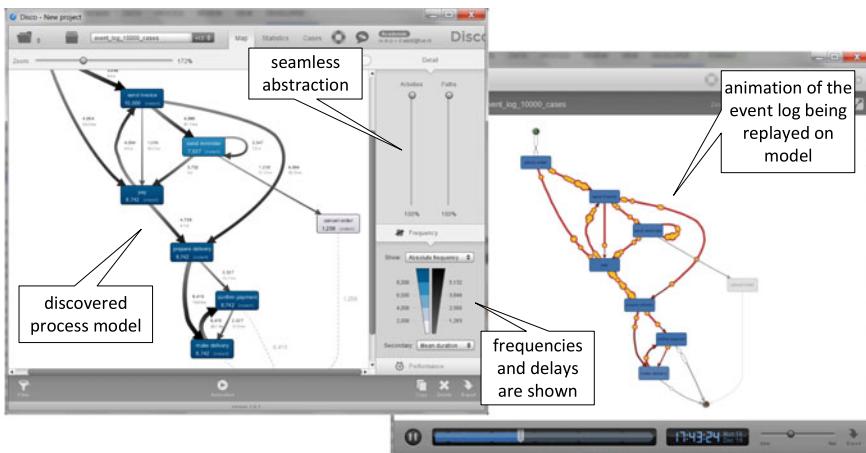


Fig. 11.11 The discovery algorithm of Disco is a further development of the Fuzzy miner and event data can be replayed at the selected abstraction level

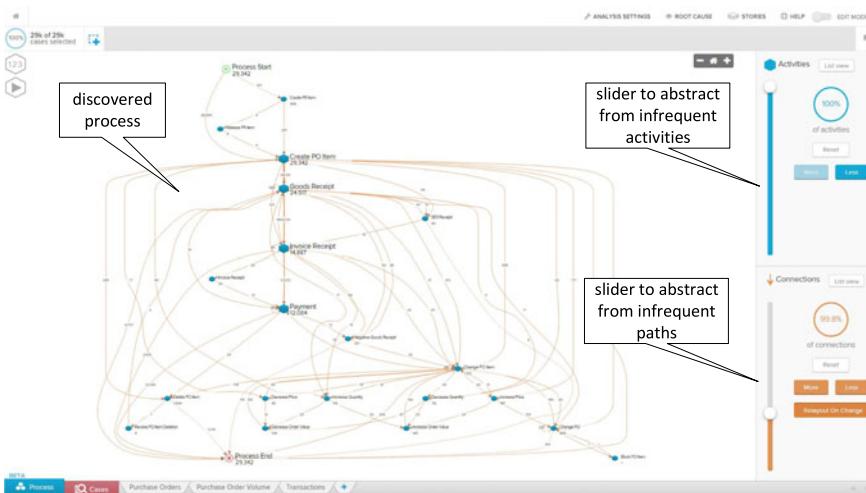


Fig. 11.12 A process model discovered using Celonis showing all activities in the event log

to get started with process mining significantly. Figure 11.11 shows a discovered process model and an animation based on the underlying event data. Animations can be saved as movies and show behavior that changes over time.

Figure 11.12 shows a process model discovered using Celonis. Celonis can load event data from CSV and XES files or database management systems such as SAP HANA, Oracle DB, MSSQL, MYSQL, PostgreSQL and IBM DB2. It is often used in conjunction with SAP. Events are stored in an OLAP-like data structure. Like Disco, Celonis provides sliders to seamlessly simplify models (if desired). In



Fig. 11.13 Animation of the process obtained by replaying the event data on a simplified process model and three charts showing trends in the event data

Fig. 11.14 A process model discovered using Minit

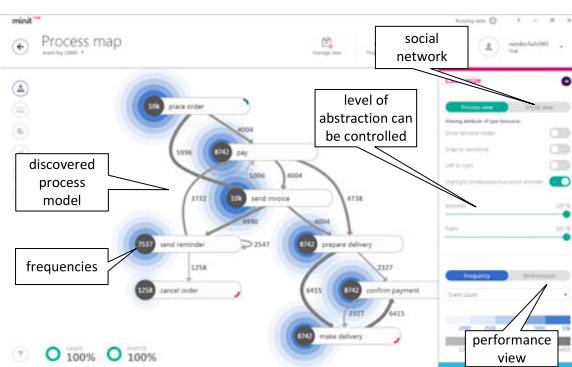


Fig. 11.13, a simplified model is used to show an animation of the process. Process related information can also be summarized in column-, line-, area-, pie-charts or tables. This is illustrated by the three charts in Fig. 11.13.

Figure 11.14 shows a screenshot of a model discovered using Minit. Minit also supports XES and uses a discovery algorithm similar to ProM's Fuzzy Miner. Like Disco and Celonis, Minit is able to handle large event logs efficiently.

Like Minit, myInvenio became available in 2015. These tools illustrate the growing interest in process mining. Figure 11.15 shows a process model discovered using myInvenio. Conformance checking is supported by comparing a reference model (e.g., specified in BPMN or XPDL) with the discovered process model. The differences can be highlighted as shown Fig. 11.15.

Figure 11.16 shows a process model discovered using Perceptive Process Mining. Performance-related information is mapped onto the process model (durations and

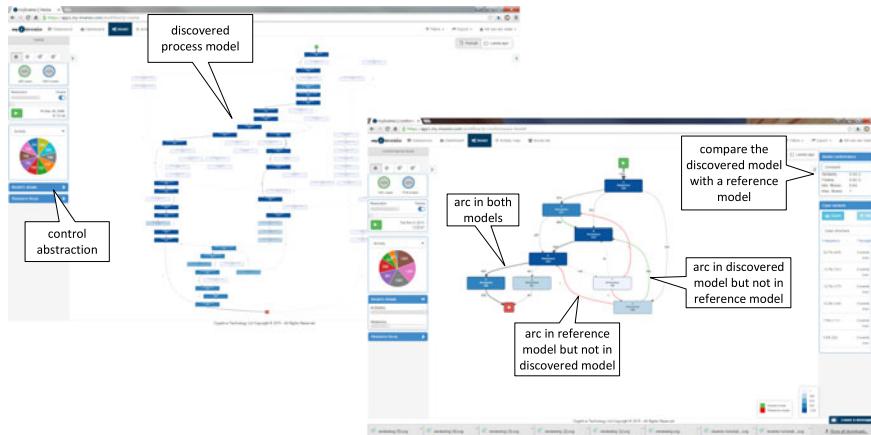


Fig. 11.15 A process model discovered using myInvenio (*left*) and the comparison of a reference model and a discovered model (*right*)

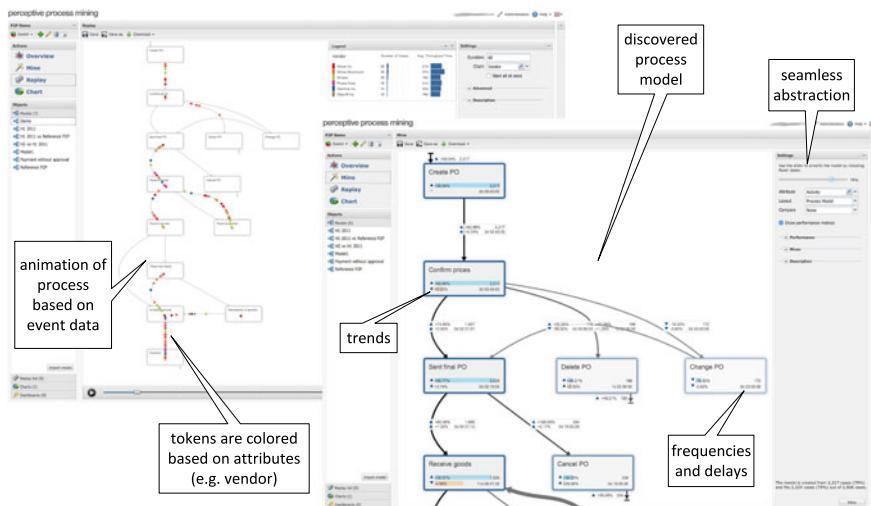


Fig. 11.16 The process model discovered using Perceptive is used to signal trends in performance (*right*) and to animate the process using “colored tokens” (*left*)

frequencies). Perceptive also shows trends, e.g., bottlenecks that are growing over time. The animation in Fig. 11.16 uses colored tokens. The coloring can be based on any case attribute (e.g., the vendor). This helps to spot differences between distinct groups of cases.

Figure 11.17 shows screenshots of four other process mining tools. Each process shown was discovered using event data.

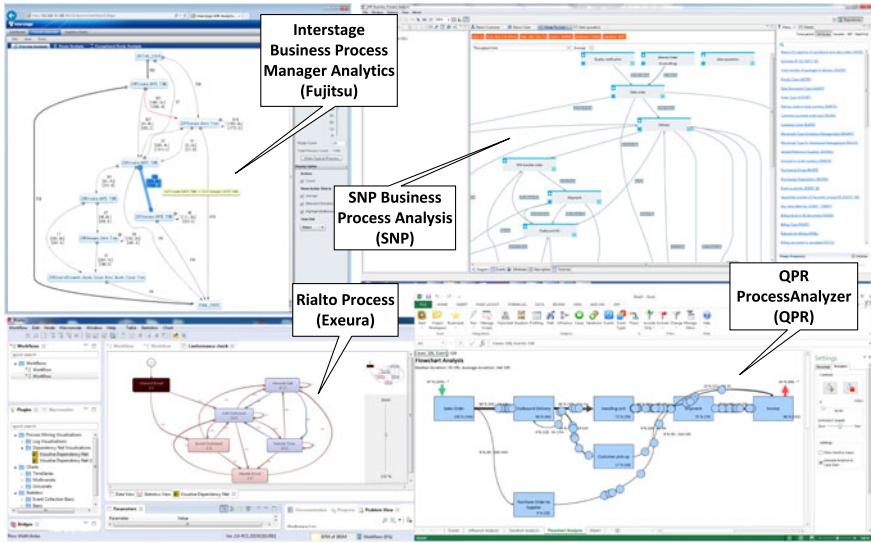


Fig. 11.17 Screenshots of four additional process mining tools: Fujitsu Interstage Business Process Manager Analytics (Fujitsu), SNP Business Process Analysis (SNP), QPR ProcessAnalyzer (QPR), and Exeura Rialto Process (Rialto)

11.4.2 Strengths and Weaknesses

The screenshots in Figs. 11.10–11.17 show that process mining capabilities are readily available in commercial tools. None of the products covers the range of process mining capabilities supported by the hundreds of available ProM plug-ins. However, ProM requires process mining expertise and is not supported by a commercial organization. Hence, it has the advantages and disadvantages common for open-source software. Fortunately, the 11 process mining tools listed in Table 11.1 nicely complement ProM.

On the one hand, there are quite some commonalities among the commercial tools (as illustrated by the screenshots in Figs. 11.10–11.17). On the other hand, there are major differences in usability and scalability. Some focus more on use cases of *Type 1* (e.g., Disco, Minit, and myInvenio) whereas other tools focus more on use cases of *Type 3* (e.g., Celonis and PPM). Organizations selecting a commercial process mining product are urged to do a pilot project where a few products are applied to organization-specific data and questions.

Despite the differences between the tools, we can make some general observations.

11.4.2.1 Limited Support for Concurrency

If a group of activities is not always executed in the same order, we would like to avoid the situation where all activities are connected to one another. Yet, process

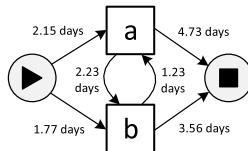


Fig. 11.18 Most tools that do not allow for concurrency have difficulties handling event logs like $L_{par} = [\langle a, b \rangle^{100}, \langle b, a \rangle^{100}]$: Due to the introduction of loops, the model allows for non-observed traces like $\langle a, b, a \rangle$, $\langle a \rangle$, $\langle b, a, b, a, b, a, b \rangle$, $\langle b \rangle$, etc.

discovery techniques that do not support concurrency do exactly that. In Sect. 11.2, these purely sequential models were called low-level models. The models discovered by such techniques tend to be very Spaghetti-like. Moreover, sequential models where every activity appears only once tend to be severely underfitting (e.g., parallel activities are turned into loops).

To illustrate the problem consider the artificial event log $L_{par} = [\langle a, b \rangle^{100}, \langle b, a \rangle^{100}]$. Clearly, there is no loop and one would expect that the discovered model shows that a and b both happen once per case. However, if a and b cannot be concurrent and the tool has one node per activity, then the tool is forced to introduce loops allowing for traces like $\langle a, b, a, b, a \rangle$ (see Fig. 11.18). Clearly, this model is underfitting and not adequately reflecting the observed behavior.

Some of the commercial tools do not support concurrency at all (e.g., SNP). Perceptive Process Mining offers two mining algorithms: a genetic algorithm able to discover concurrency (but time-consuming and not scalable) and a simpler, better performing, algorithm based on the directly follows relation having the problem mentioned above.

Also Disco deals with concurrency different from the algorithms described in Chaps. 6 and 7. Parallelism in Disco is discovered only if two activity instances for the same case overlap. This implies that concurrency cannot be discovered in event logs without explicit transactional information (e.g., when there are just complete events). If activities are interleaved (i.e., not overlapping), then the arrows are suppressed, unless the slider is moved up to ensure perfect fitness. Using the terminology introduced in Sect. 11.2, Disco shifts from an informal model with concurrency to a formal low-level model without concurrency to ensure correctness.

Other tools have similar issues and are often less clear about this. They operate in the space between informal models and low-level models, thus making interpretation tricky. Consider the delays in Fig. 11.18. When does the process end—4.73 days after the (last) completion of activity a or $2.23 + 3.56 = 5.79$ days after the (last) completion of activity a ? Probably none of the two answers is right, thus illustrating the confusion.

To summarize—*None of the commercial tools handles concurrency adequately*. There are at least two reasons for this. First of all, simple algorithms are used to ensure scalability and transparency. Second, the models learned have informal semantics. The latter is interesting because several tools claim to support BPMN and can export models to BPM systems. This may lead to misleading results. Most

tools do not show explicit AND/XOR-splits/joins. Adding logic when saving models will lead to confusion and may result in models that are not sound (e.g., having deadlocks).

As long as process models are interpreted as “pictures” this is not a problem. However, the way that models need to be interpreted *also influences frequencies and performance results*. For example, if it is unclear whether things need to be synchronized or not, computed waiting times cannot be trusted. The inductive mining techniques presented in Sect. 7.5 show that it is possible to discover concurrency without creating unsound or imprecise models. The different inductive mining algorithms (IM, IMF, IMc, IMD, IMFD, IMCD, etc.) always produce sound models and are highly scalable. Some of these algorithms even provide formal guarantees (e.g., perfect fitness).

11.4.2.2 Limited Support for Conformance Checking

Informal models that can only be interpreted as “pictures” cannot be used for conformance checking. Currently, there is no commercial tool that computes alignments or that is able to apply some other replay algorithm to precisely diagnose deviations in the presence of concurrency. The reasons are the same as before: scalability (computing alignments may be too time consuming) and informal semantics (e.g., not being able to distinguish between AND-joins and XOR-joins).

Conformance checking is not handled by replaying the event log on a precise end-to-end process model. Instead one or more of the following approaches are used:

- *Rule based.* The user can specify rules for filtering. For example, Disco and Celonis can be used to define a wide variety of rules (e.g., “ a is followed by b and not c and should be executed by a resource not involved in d ”). By applying such rules, the event log can be split into conforming and non-conforming cases.
- *Outlier based.* Infrequent paths that deviate from mainstream behavior are manually inspected. By classifying certain paths as deviating, the corresponding cases are tagged as non-conforming.
- *Side-by-side.* The discovered process model and the normative reference model are depicted next to each other. Users need to visually compare models to see deviations.
- *Overlay.* The discovered model and the reference model are stacked on top of each other and differences are highlighted. Figure 11.15 illustrates that myInvenio supports this type of comparison.

Comparing a discovered model and a reference model may lead to incorrect conclusions. Note that the discovered model generalizes over the data, i.e., paths possible in the model may never have happened. This may trigger the detection of deviations that never occurred in reality. The discovered model may also abstract from infrequent behavior. Therefore, rare (but possibly harmful) deviations may remain undetected. However, such peculiar deviations tend to be highly relevant for conformance checking.

Assume that the informal model in Fig. 11.18 was discovered for L_{par} and subsequently compared with a normative model putting a and b in parallel. A visual comparison of the two models would suggest non-existing deviations.

The techniques in Chap. 8 and the plug-ins of ProM show that conformance checking is possible. However, conformance checking can only be supported if the informal models are replaced by formal models (e.g., process trees or Petri nets with a defined initial and final marking). *As long as this functionality is not present, users are forced to capture the real semantics of the normative model in terms of a collection of rules used for filtering.*

11.4.2.3 Performance Perspective is Well Supported

The primary focus of commercial process mining tools is on performance. Each of the tools can visualize bottlenecks in the process. Tools such as Celonis, Perceptive, QPR, Minit, PPM, etc. provide a range of charts. Most of the commercial tools make it possible to quickly find bottlenecks, unnecessary rework and delays.

Note that the problems mentioned earlier may endanger the correctness of performance results. If misalignments and concurrency are not handled well, then the reported results may be incorrect. For example, tools may report negative waiting times if events are reversed or excessive times if events are missing.

11.4.2.4 Data Perspective Not in Models

None of the commercial process mining tools is able to discover data-aware process models. For example, it is impossible to learn guards or perform any other form of decision mining as described in Chap. 9. Conformance checking of models with data is also not supported.

Additional data in the event log can be used in rules for filtering. Moreover, some tools can show the distribution of values for particular groups of cases. However, data are not explicitly related to the process model.

11.4.2.5 Organizational Perspective

Most tools are able to construct a social network (see Chap. 9). It is typically also possible to see the utilization of resources. Nearly all tools consider resource information, roles, and other organizational entities as plain data elements. Hence, the organizational perspective can be handled in the same way as data (e.g., using filtering). Separation of duties (4-eyes principle) can be checked in this way. myInvenio also creates an activity map (a simplified RACI matrix). Normally, a RACI matrix shows the people Responsible, Accountable, Consulted, and Informed for each activity. Event logs need to be enriched to provide this type of information. Sophisticated analysis techniques to optimize work distribution and social network analysis are still missing.

11.4.2.6 Growing Support for XES

Next to tools like ProM, RapidProM, PMLAB, and CoBeFra, the XES standard is supported by a growing number of commercial tools. Currently, Disco, Celonis, Minit, Rialto, and SNP support XES. QPR and myInvenio have announced XES support for the next release. Perceptive, PPM, EDS, and Fujitsu do not (yet) support XES.

XES makes it easier to combine different tools, e.g., using a commercial tool in conjunction with ProM, RapidProM, PMLAB, or CoBeFra.

11.4.2.7 Getting Event Data from Other Sources

Vendors of commercial tools realize that substantial time is spent on extracting data from information systems. In Sect. 11.2, we listed four mechanisms to get event data: file, database, adapter, and streaming. Next to file-based imports of XES, MXML, and CSV, most tools support the extraction of data from JDBC databases. Events can often be loaded from systems such as MySQL, IBM DB2, Oracle DB, SQL Server, PostgreSQL, and SAP HANA.

Often datasets can also be *incrementally* updated (importing only changes since the last import). For example, Disco can retrieve data from a server with the so-called Airlift interface. On the server side of the Airlift connection, arbitrary databases and production systems can be connected.

Systems like Celonis provide additional support to obtain data from SAP systems. Due to the partnership between SAP and Celonis, integration with SAP products like SAP HANA is safeguarded. In fact, most process mining tools support application specific adapters, but the range of systems covered and the quality of these adapters varies per tool.

11.4.2.8 Filtering

Filtering plays a crucial role in most commercial systems. Figure 11.19 shows six types of filtering supported by Disco. Filters can be used to remove individual events or complete cases. For example, one can remove all slow cases, all exceptional cases, etc. One can also specify LTL or Declare-like rules, e.g., activity a should be eventually followed by b (the response constraint in Declare and “ $\square(a \Rightarrow (\Diamond b))$ ” in LTL). Filtering can be used for ad-hoc conformance checking and plays an important role in root-cause analysis.

Filtering is related to OLAP (see Sect. 12.4). The dimensions in an OLAP cube also split the data based on different criteria. Process mining tools like Celonis store events in multidimensional cubes to facilitate the selection and comparison of particular groups of cases.

11.4.2.9 No Automatic Clustering

Filtering and the selection of dimensions in an OLAP cube are based on user-defined criteria. However, one may also use clustering techniques that automatically group

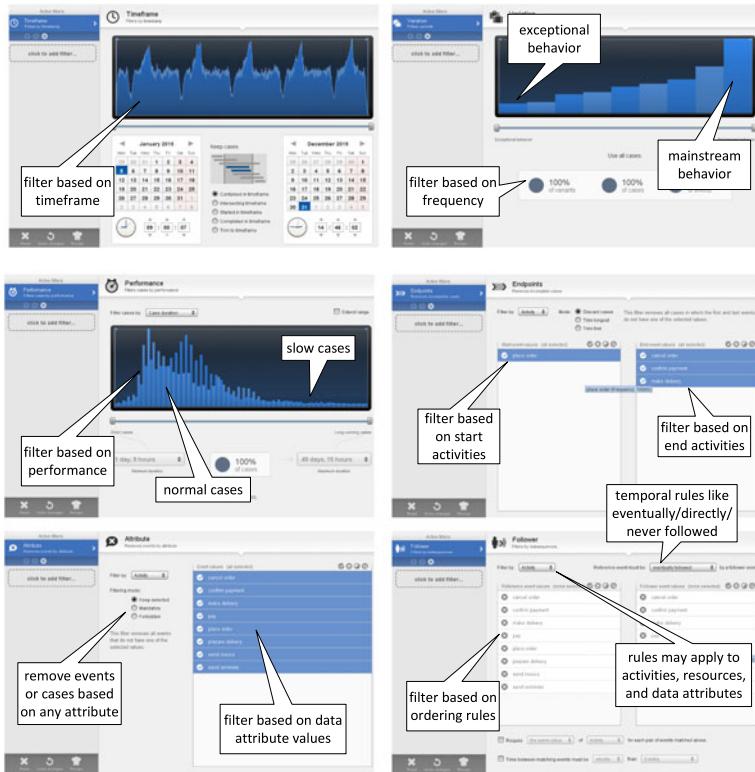


Fig. 11.19 Illustration of the extensive filtering capabilities in commercial systems like Disco

cases that are similar. ProM provides several ways of clustering similar cases based on selected features.

Standard techniques like k -means clustering (see Sect. 4.3 and Chap. 9) can be used as a preprocessing step for process mining [13, 62, 78]. The clusters themselves may already provide novel insights. Moreover, the clusters can often be used to discover multiple simple process models instead of one complex process model.

Surprisingly, clustering is not supported by the current generation of commercial process mining tools.

11.4.2.10 Reporting and Animation

Process mining results need to be communicated. Most tools provide means to create reports, for example, by storing artifacts such as charts, tables, and models. Compared to BI tools, the reporting facilities of process mining tools are often limited.

Disco, Celonis, Perceptive, Minit, QPR, and myInvenio support data-driven process animations. Figures 11.6, 11.11, 11.13, and 11.16 show screenshots where

event logs are animated by replaying them on discovered models. Such animations are instrumental when convincing management. Animation is also a means to support change management: It can be used to create a sense of urgency and to build consensus on root causes.

11.4.2.11 Links to Other Tools

Some of the process mining tools are part of a bigger suite. For example, PPM is part of the webMethods suite and the ARIS family of tools. Perceptive Process Mining was developed as part of Perceptive's BPM suite. It is expected that in time most BPM systems will provide a process mining component (similar to the simulation components in today's BPM systems).

Most process mining tools are able to export process models in a format that can be read by other tools (e.g., BPMN or XPDL). This way the results from process mining can be used as a starting point for modeling, simulation, and documentation.

As mentioned in the context of RapidProM, the interplay between process mining and data mining is extremely valuable. Hence, some process mining tools can export data in a form that can be analyzed by standard data mining tools. Other crossovers of tools are possible. For example, loading process mining results into Excel to create a chart or to compute some statistic.

11.4.2.12 Operational Support

Disco, Celonis, Perceptive, QPR, PPM, and Fujitsu can upload data periodically or incrementally. Analysis views are “refreshed” based on the new data. However, true predictive analyses, as described in Chap. 10, are seldom supported. QPR, PPM, and Rialto report integration efforts with dedicated prediction tools. However, these approaches do not seem to be process specific (i.e., the discovered model is not leveraged for prediction).

11.4.2.13 Scalability

Most of the commercial process mining tools have a good performance in terms of scalability and responsiveness. Some tools can even *handle event logs with billions of events, millions of cases, and hundreds of activities*. Loading such event logs may be time consuming (say up to an hour), but once the log is loaded analysis can be done within a few seconds.

Scalability depends on many different factors and not only the size of the event log. Some types of analysis are sensitive to the average trace length of cases, the number of distinct activities, or the number of attributes per event. Section 12.1.3 describes the key characteristics of logs relevant for scalability.

Organizations selecting a process mining tool are advised to test the scalability of tools on their own data using standard hardware. This is the only way to compare performance in a meaningful way (be aware of indexing and special hardware). See Chap. 12 for techniques to handle even larger event logs.

11.5 Outlook

It is impossible to give a complete overview of all products supporting process mining. Just ProM, the leading open-source process mining framework, already provides more than 1500 plug-ins. These plug-ins cover a wide range of analysis techniques. For example, all process discovery approaches described in this book are supported through ProM plug-ins. Moreover, ProM is not limited to process discovery and also supports conformance checking, social network analysis, bottleneck analysis, decision mining, operational support, verification, model conversion, etc. Most ProM plug-ins aim at use cases of *Type 1*. RapidProM, CoBeFra, and PMLAB support use cases of *Type 2*.

The 11 commercial process mining tools described in this chapter help to lower the threshold for process mining. Next to use cases of *Type 1*, also use cases of *Type 3* are supported using pre-configured dashboards and automated data extraction. Each of the eleven tools aims at supporting less experienced users. Sometimes process mining capabilities are embedded in larger software products. The scalability and usability of most commercial systems is good. Several tools can handle event logs with billions of events. However, compared to ProM there are also typical weaknesses such as the inability to discover concurrency well and the limited support for conformance checking. The focus is on performance analysis rather than conformance checking and precise models.

Since the process mining market is developing fast, readers are advised to test tools using their own event data. Even when tools look similar, differences in terms of practical usability and scalability may be significant.

Chapter 12

Process Mining in the Large

Process mining provides the technology to leverage the ever-increasing amounts of event data in modern organizations and societies. Despite the growing capabilities of modern computing infrastructures, event logs may be too large or too complex to be handled using conventional approaches. This chapter focuses on handling “Big Event Data” and relates process mining to Big Data technologies. Moreover, it is shown that process mining problems can be decomposed in two ways, *case-based decomposition* and *activity-based decomposition*. Many of the analysis techniques described can be made scalable using such decompositions. Also other performance-related topics such as streaming process mining and process cubes are discussed. The chapter shows that the lion’s share of process mining techniques can be “applied in the large” by using the right infrastructure and approach.

12.1 Big Event Data

Some of the process mining tools described in Chap. 11 can discover process models for logs with billions of events. However, performance highly depends on the characteristics of the event log (e.g., number of distinct activities and redundancy) and the questions asked (e.g., conformance checking is often more time consuming than discovery). Moreover, for some applications event logs may be even larger or results need to be provided instantly.

In Chap. 1, we listed the “four V’s of Big Data”: Volume, Velocity, Variety, and Veracity (Fig. 1.4). The term “Internet of Events”, introduced in Sect. 1.1, refers to the growing availability of event data. These data are omnipresent and an enabler for process mining. This chapter will focus on event data and the first two V’s. The first ‘V’ (Volume) refers to the size of some data sets, in our case event logs. We will discuss various *decomposition* and *distribution* strategies to turn large process mining problems into multiple smaller ones. The second ‘V’ (Velocity) refers to the speed of the incoming events that need to be processed. It may be impossible or undesirable to store all data. Therefore, we will also introduce the topic of *streaming* process mining.

Big Data is not limited to process-related data. However, Big Data infrastructures enable us to collect, store, and process huge event logs (see Sect. 2.5.9). Process mining tools can exploit such infrastructures. Therefore, we describe current trends in hardware and software (Sect. 12.1.2), before describing the characteristic features of event logs (Sect. 12.1.3). However, first we briefly discuss the possibilities and risks when going from sampled “small data” to “all data”.

12.1.1 $N = All$

In the past, conclusions were often based on human judgment or analysis of sample data. Either data were not available, unreliable, or it was impossible to process all data. In many businesses, we now witness a change from collecting *some data* to collecting *all data* [100]: “ $N = All$ ” where N refers to the sample size. As described in Sect. 1.1, the digital universe and physical universe are becoming more aligned. Money has become a predominantly digital entity. Queries on the availability of products are answered based on data in some database rather than a visit to the warehouse. The direct coupling between data and reality combined with our improved abilities to store and process data forms the playground of data science as described in Chap. 1.

Sampling was needed in the “analog era” characterized by information *scarcity*. Due to sampling error and sampling bias, it may be risky to extrapolate conclusions from sample data. Moreover, the granularity of analysis using sampled data is often too coarse making it impossible to draw conclusions for smaller subcategories and submarkets. Hence, the concept of sampling makes no sense if *all* data are available and we have the computing power to analyze all events. Consider, for example, conformance checking. Why just check the conformance of a few cases if we can check all cases and detect all deviations? Clearly, “ $N = All$ ” requires a new way of thinking. For example, auditors and accountants may be afraid of uncovering all deviations. Also the work of marketers and social scientists is changing: large-scale data analysis is replacing sampling and questionnaires.

Having all data ($N = All$) may also create problems:

- Hardware, software and analysis techniques need to be able to cope with the associated volumes.
- Overfitting the data may lead to “bogus conclusions” (cf. Bonferroni’s principle).

This chapter will focus on the first problem. However, to illustrate the second problem we consider the following example inspired by a similar example in [114].

Suppose some Dutch government agency is searching for terrorists by examining hotel visits of all of its 18 million citizens (18×10^6). The hypothesis is that terrorists meet multiple times at some hotel to plan an attack. Hence, the agency looks for suspicious “events” $\{p_1, p_2\} \uparrow \{d_1, d_2\}$ where persons p_1 and p_2 meet on days d_1 and d_2 in some hotel. How many of such suspicious events will the agency find if the behavior of people is completely random? To estimate this number, we make some

additional assumptions. On average, Dutch people go to a hotel every 100 days and a hotel can accommodate 100 people at the same time. We further assume that there are $\frac{18 \times 10^6}{100 \times 100} = 1800$ Dutch hotels where potential terrorists can meet.

The probability that two persons (p_1 and p_2) visit a hotel on a given day d is $\frac{1}{100} \times \frac{1}{100} = 10^{-4}$. The probability that p_1 and p_2 visit the *same* hotel on day d is $10^{-4} \times \frac{1}{1800} = 5.55 \times 10^{-8}$. The probability that p_1 and p_2 visit the same hotel on two different days d_1 and d_2 is $(5.55 \times 10^{-8})^2 = 3.086 \times 10^{-15}$. Note that different hotels may be used on both days. Hence, the probability of suspicious event $\{p_1, p_2\} \dagger \{d_1, d_2\}$ is 3.086×10^{-15} .

How many candidate events are there? Assume an observation period of 1000 days. Hence, there are $\binom{1000}{2} = \frac{1000 \times (1000 - 1)}{2} = 499,500$ combinations of days d_1 and d_2 . Note that the order of days does not matter, but the days need to be different. There are $\binom{18 \times 10^6}{2} = \frac{18 \times 10^6 \times (18 \times 10^6 - 1)}{2} = 1.62 \times 10^{14}$ combinations of persons p_1 and p_2 . Again the ordering of p_1 and p_2 does not matter, but $p_1 \neq p_2$. Hence, there are $499,500 \times 1.62 \times 10^{14} = 8.09 \times 10^{19}$ candidate events $\{p_1, p_2\} \dagger \{d_1, d_2\}$.

The expected number of suspicious events is equal to the product of the number of candidate events $\{p_1, p_2\} \dagger \{d_1, d_2\}$ and the probability of such events (assuming independence), $8.09 \times 10^{19} \times 3.086 \times 10^{-15} = 249,749$. Hence, there will be around a quarter million observed suspicious events $\{p_1, p_2\} \dagger \{d_1, d_2\}$ in a 1000 day period!

Suppose that there are only a handful of terrorists and related meetings in hotels. The Dutch government agency will need to investigate around a quarter million suspicious events involving hundreds of thousands innocent citizens. This is an illustration of Bonferroni's principle.

Bonferroni's principle

In statistics, the *Bonferroni's correction* is a method (named after the Italian mathematician Carlo Emilio Bonferroni) to compensate for the problem of multiple comparisons. Normally, one rejects the null hypothesis if the likelihood of the observed data under the null hypothesis is low. If we test many hypotheses, we also increase the likelihood of a rare event. Hence, the likelihood of incorrectly rejecting a null hypothesis increases. If the desired significance level for the whole collection of null hypotheses is α , then the Bonferroni correction suggests that one should test each individual hypothesis at a significance level of $\frac{\alpha}{k}$ where k is the number of null hypotheses. For example, if $\alpha = 0.05$ and $k = 20$, then $\frac{\alpha}{k} = 0.0025$ is the required significance level for testing the individual hypotheses.

Bonferroni's principle aims to avoid treating random observations as if they are real and significant [114]. To apply the principle, compute the number of observations of some phenomena one is interested in under the assumption that things occur at random. If this number is significantly larger than the real number of instances one expects, then most of the findings will be false positives.

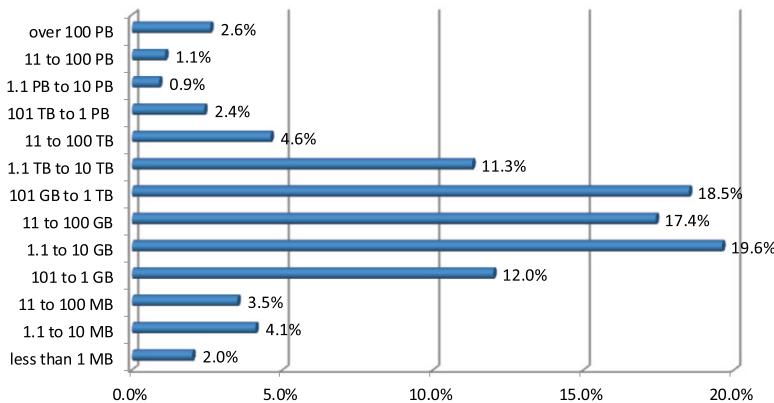


Fig. 12.1 Results of the KDnuggets poll (August 2015): “What was the largest data set you analyzed/data mined?” (1 Gigabyte (GB) equals 1000 MB, 1 Terabyte (TB) equals 1000 GB, 1 Petabyte (PB) equals 1000 TB)

Bonferroni’s principle is highly relevant for large data sets with many instances. The number of rare events of a certain type will increase as the volume of data grows even if there is no pattern and behavior is completely random.

If we are looking for terrorists and expect only a few terrorists to be active, then any hypothesis that points to hundreds of thousands of citizens behaving randomly is pointless. Bonferroni’s principle states that one can only find terrorists by looking for events that are so rare that they are unlikely to occur in random data.

Figure 12.1 is another illustration of the challenges posed by today’s data sets. In a recent KDnuggets poll, over 78% of respondents reported to have analyzed data sets of more than 1 Gigabyte and over 22% of respondents reported to have analyzed data sets of more than 1 Terabyte. Before discussing process-mining specific ways of dealing with large event logs, we first discuss some general technological developments relevant for mining massive data sets.

12.1.2 Hardware and Software Developments

In 1965, Gordon Moore predicted that the number of transistors would double every year until 1975 [104]. In 1975, Moore revised his prediction to a doubling of components every two years. This prediction turned out to be remarkably accurate, as shown in Fig. 12.2. The diagonal line shows *Moore’s law* predicting that the number of transistors on a chip is $1000 \times 2^{(y-1970)/2}$. Here, 1970 is used as basis.

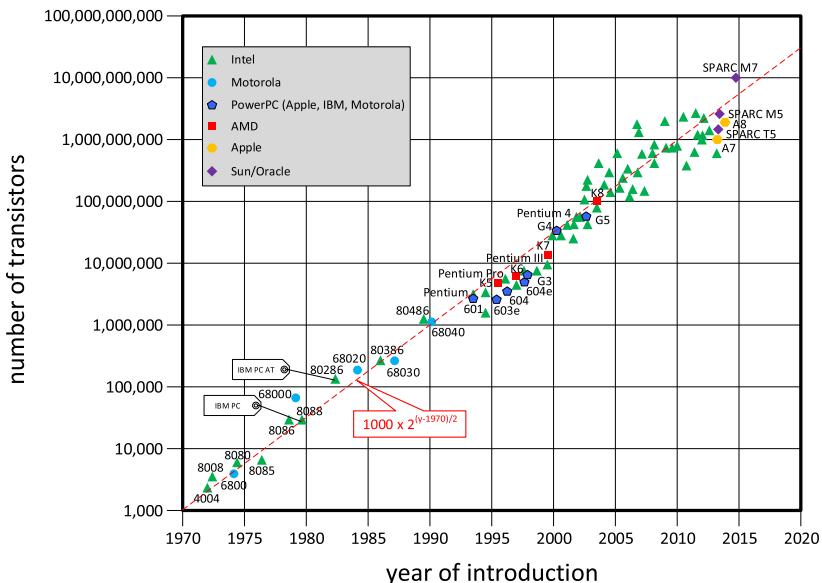


Fig. 12.2 Moore's law : The number of transistors on a chip has been growing exponentially since the early 1970s

For 2016 this formula predicts $1000 \times 2^{(2016-1970)/2} = 8.4 \times 10^9$ transistors which seems consistent with reality, e.g., the Intel Xeon E7-8890 v3 processor released in May 2015 has a total of 5.6 billion transistors.

The exponential growth is not limited to the number of transistors per chip. Performance of CPUs has been growing at a similar pace. Although clock speeds leveled off around 2004, multicore architectures continued boosting performance. Similar developments can be seen in memory and storage (e.g., size of hard disks and flash drives), graphics (e.g., pixels on a screen), and networking (e.g., the capacity of wireless networks). Also costs have been decreasing exponentially, e.g., the price of storing one Gigabyte of data or making a particular computation.

It is difficult to grasp the incredible developments in IT as reflected by Moore's law. In 1970, it took 1.5 hours to travel by train from Eindhoven to Amsterdam. If transportation would have followed the same developments, then the train trip would now take only $(1.5 \times 60 \times 60)/2^{(2016-1970)/2} = 0.00064$ seconds (i.e., less than a millisecond). Flying from Amsterdam to New York would only take 3.4 milliseconds ($(8 \times 60 \times 60)/2^{(2016-1970)/2} = 0.0034$ seconds). In 1970, a car would have consumed approximately 5000 liter of fuel to drive around the world (40,075 km). If fuel efficiency would have followed similar developments, less than 1 milliliter of fuel would be needed now ($5000/2^{(2016-1970)/2} = 0.00059$ liter). These examples illustrate that our ability to process data has developed at a spectacular pace. Although this development has been ongoing since 1970, data analysis has now reached a "tipping point" thus explaining the current "Big Data" hype.

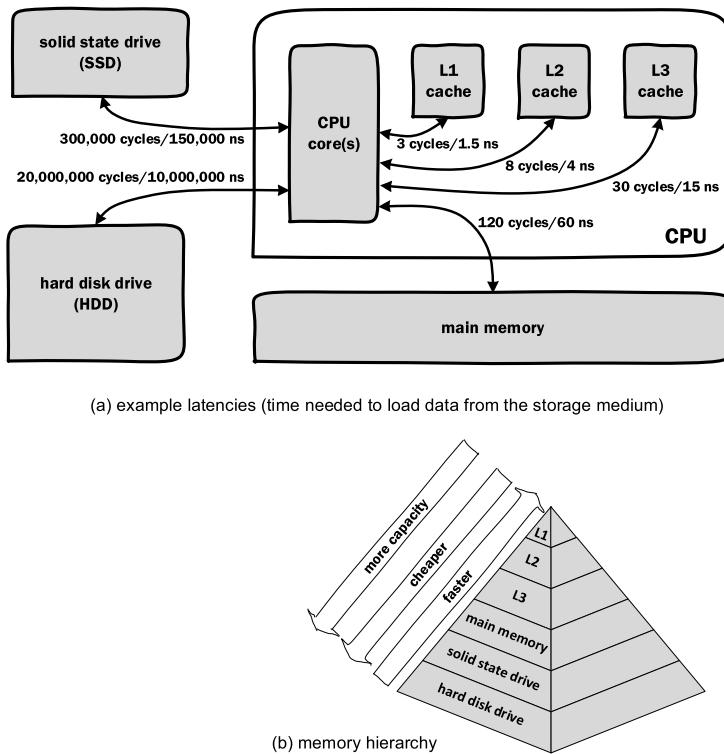


Fig. 12.3 Conceptual view of a typical computer’s memory hierarchy going from slow, spacious and cheap (HDD) to fast, small and expensive (L1 cache)

Let us not take a look inside a computer (see Fig. 12.3) and focus on the different types of storage and their latencies. The hard disk drive (HDD) is the cheapest, but also slowest form of storage. HDDs offer large amounts of storage. The solid state drive (SSD) is more expensive and has less capacity, but is considerably faster. Main memory is even more expensive, but orders of magnitude faster than drives. Fastest and most expensive are the different caches (L1, L2, and L3).

Figure 12.3 shows some typical latencies which come into play when the CPU needs to access data. Latency is the time delay experienced by the computer to load data from the storage medium until it is available in the CPU register. The L1 cache latency in Fig. 12.3 is 1.5 nanoseconds corresponding to 3 cycles of the processor (1 nanosecond is 10^{-9} second). Loading data from the L2 and L3 cache takes a bit more time, but is still faster than loading data from main memory. The main memory latency in Fig. 12.3 is 60 nanoseconds (120 cycles). The latencies for the different types of drives are much larger: 150,000 nanoseconds for SSD and 10,000,000 nanoseconds for HDD. Note that the numbers in Fig. 12.3 are just examples. They are used to exemplify the spectacular differences between the different types of storage media. The prices per byte of storage are inverse proportional to the latencies.

To illustrate the differences in Fig. 12.3, we assume that loading data corresponds to fetching a cup of coffee. Getting a data element from the L1 cache corresponds to getting a Nespresso coffee from the kitchen. For the analogy, let us assume that 1.5 nanoseconds corresponds to 7.5 meters (distance from desk to kitchen). Getting a data element from main memory then corresponds to getting a coffee from the Starbucks around the corner (300 meters equals 60 ns). Getting a data element from SSD then corresponds to flying from Eindhoven to Aosta (Italy) to get a really good cappuccino (750 kilometers equals 150,000 ns). Getting a data element from hard disk then corresponds to flying to Colombia, Ethiopia, and Kenya to hand-pick the best beans, process them in Amsterdam, take the beans to Rome and ask a barista to make a ristretto from the ground coffee beans (50,000 kilometers equals 10 milliseconds).

When implementing a process mining technique, it is important to be aware of these differences in latency. Whether an event is on disk or in memory makes a huge difference.

Figures 12.2 and 12.3 provide the context for a discussion on the rapidly changing IT landscape for Big Data analytics. A lot has changed in database technology over the last decade. Edgar Codd defined the relational model in 1970 [32]. For many years Relational Database Management Systems (RDBMS) and the Structured Query language (SQL) were the de facto standard. Database systems like Oracle V2, IBM's DB2, dBase, Sybase, Ingres, Informix, Access, Postgres, and MySQL released in the period 1975–2005 were all relational. However, due to the challenges of extremely large data sets (at the “scale of Google”), the dominance of relational databases and SQL ended. Since 2005 many new non-relational database systems have been released and the emergence of big data technologies like Hadoop caused a revolution in the way IT systems are organized. We sketch some of these developments in the remainder.

12.1.2.1 In-Memory Databases and Analytics

The numbers in Fig. 12.3 show that getting data from a hard disk is like traveling around the world to get a cup of coffee. An *in-memory database management system* primarily relies on main memory for data storage. This requires computer systems with a large main memory (e.g., a terabyte of main memory). Since the size of main memory is still limited compared to disk storage, these in-memory database management systems compress the data using a variety of compression mechanisms. When the limit of available main memory is reached, larger chunks of data are unloaded from main memory based on the characteristics of the application and are reloaded into main memory when they are required again.

Since data is stored in volatile memory, all stored information is lost when the device loses power or is reset. Durability, one of the standard ACID (atomicity, consistency, isolation, durability) properties, does not hold without special provisions. Solutions are snapshot files, checkpoint images, and transaction logging to recover an in-memory database after system failure.

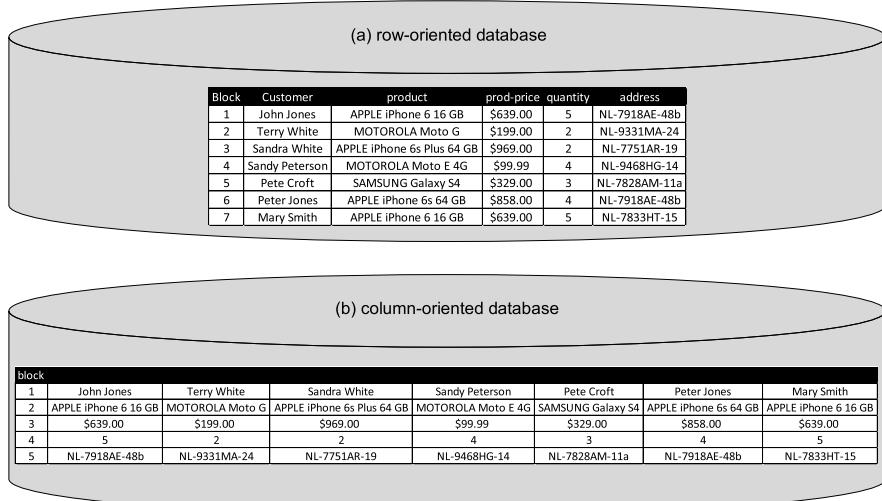


Fig. 12.4 Row-oriented versus column-oriented databases

Next to using an in-memory platform or database management system, the analytics tool itself can also manage data directly. The challenge is to keep the “hot data” in main memory and the “cold data” on disk.

OLAP applications can benefit from in-memory analytics. Users want to “slice and dice” data to look at the data from different angles (see Sect. 12.4). The effect of an OLAP operation needs to be instantaneous to fully support the analyst.

Although the costs of in-memory technology are improving, it is still quite expensive and hence only economic in situations where responsiveness is imperative.

12.1.2.2 Columnar Databases

In a traditional relational database, data is organized in horizontal rows and vertical columns. The rows correspond to instances (e.g., sales transactions) and the columns refer to attributes of these instances. This is very natural and corresponds to the way we plot data in spreadsheets or on paper. However, when analyzing data it is rare to analyze all the columns in a single row. During analysis we tend to do operations on all the rows in a single column, e.g., taking the sum or computing the average. In a *column store*, columns tend to be stored together rather than rows.

Figure 12.4 sketches the idea of organizing data by columns (b) rather than rows (a). *Columnar databases* use the row orientation and store all elements related to a particular attribute together in one block, e.g., all quantities, prices, product names, etc. are stored together. Aggregate operations such as sum and average can be done faster, since all the data needed is grouped together. Also compression can be improved, since there tends to be more repetition in the column-oriented blocks. For example, customer and product names are shared by multiple sales transactions whereas within a sales transaction the attributes will be different.

SAP HANA is a well-known example of an application server that includes an in-memory, column-oriented database management system. HANA employs a mix of columnar and row-based storage. MonetDB is a column-oriented database management system released under an open-source license. Other column-oriented systems include Apache Cassandra, HBase, Accumulo, Druid, and Vertica.

Column-oriented systems are part of a larger class of *NoSQL database management systems*. A NoSQL database provides a mechanism for the storage and retrieval of data that is no longer based on the traditional tabular relations from relational databases. Next to column-oriented systems, there are other classes of NoSQL systems such as document-oriented databases, graph databases, object databases, and key-value databases. Most of these systems were developed to cope with specific scalability challenges. For the ranking of web pages, we may need to perform iterated matrix–vector multiplications with billions of rows and columns. For searches in social networks, we need to analyze graphs with billions of nodes and edges. Traditional database systems cannot handle such problems. In fact, the NoSQL systems illustrate the end of the “one size fits all” approach promoted by relational databases.

12.1.2.3 Large-Scale Distributed File Systems

When datasets are small, analysis can be done using a single computer having its own memory, disk and CPU. However, if datasets get larger, parallel processing is needed using a network of parallel computers. In the past, large scientific computations were done using special-purpose parallel computers using specialized hardware. However, the need for cheap large-scale data processing triggered a trend towards the use of thousands of compute nodes operating more or less independently and using commodity hardware.

Scalability at reasonable costs is key for achieving a competitive advantage. For example, Google was forced to create its own hardware and software stack to realize a scalable commercial solution. In this context, Google developed a distributed file system, *Google File System* (GFS) [58]. GFS supports massive numbers of commodity servers with directly attached storage to be exposed as a single logical file system.

In such a distributed file system, files can be enormous (e.g., terabytes), but are rarely updated. Typically data is appended rather than changed. The files are split into chunks that are replicated at two or more compute nodes.

Compute nodes are typically stored in racks. The nodes are connected by fast networks for inter-rack and intra-rack communication. The replicated chunks of data are stored in different racks to be able to handle rack failure.

Replication is essential when the number of compute nodes increases. Recall that commodity hardware is used for compute nodes to lower costs. Assume that the Mean Time Between Failures (MTBF) for a given compute node is three years. This implies that the MTBF for a distributed system composed of 1000 nodes is approximately one day. Therefore, fault-tolerance needs to be built into the system: the replicated chunks of data are used to automatically recover from hardware failures.

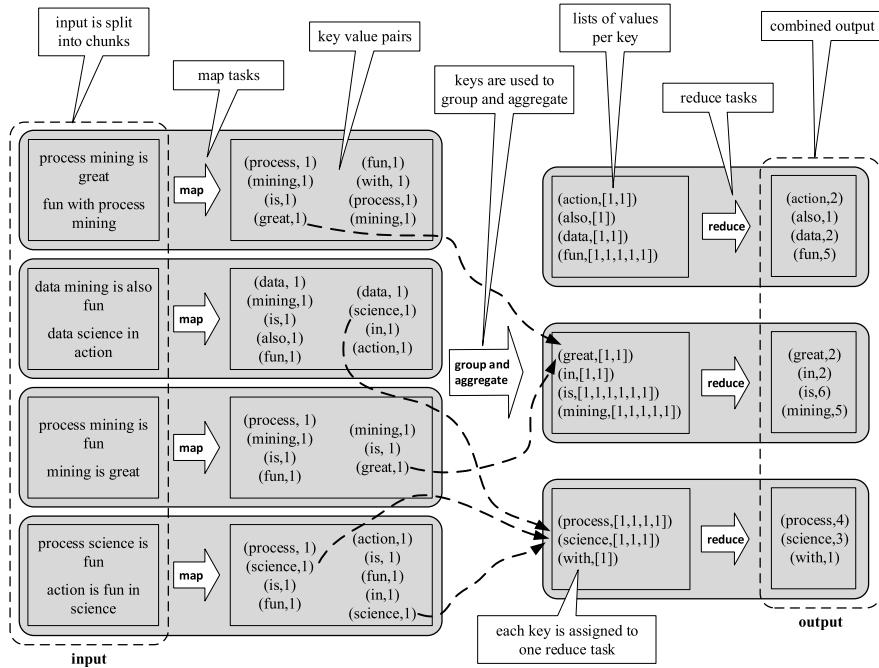


Fig. 12.5 Counting words using MapReduce

The *Hadoop Distributed File System* (HDFS) is an open-source distributed file system inspired by GFS. HDFS is the core of *Apache Hadoop*, an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Hadoop is based on the assumption that hardware failures are common and should be automatically handled by the framework.

Apache Hadoop evolved into an extensive ecosystem with many components that go far beyond this book. For example, *Hadoop YARN* is a resource-management platform responsible for managing and scheduling computing resources. *Hadoop MapReduce* is also part of the base framework and provides an implementation of the *MapReduce* programming model for large-scale data processing. MapReduce was originally developed within Google [44].

MapReduce is a programming model that allows complex problems to be broken up into simple map functions that can be executed concurrently together with reduce functions that combine the outputs from each parallel stream.

MapReduce

MapReduce is a style of distributed computing where the user only needs to provide two functions, called *Map* and *Reduce*. The system takes care of the

rest: managing the parallel execution, coordinating the concurrent tasks, and handling failures. There are two types of tasks, *Map* tasks and *Reduce* tasks.

Each *Map* task gets one or more chunks of data from the distributed file systems (e.g., GFS or HDFS). The *Map* task produces a sequence of *key–value pairs*, e.g., the output of task i could be of the form $\langle (k_{i,1}, v_{i,1}), (k_{i,2}, v_{i,2}), \dots, (k_{i,n}, v_{i,n}) \rangle$. The way the key value pairs are produced from the input data is specified by the *Map* function.

The key–value pairs produced by all *Map* tasks are redistributed by the system as preparation for the *Reduce* tasks. The keys are divided among the *Reduce* tasks, e.g., using hashing. A particular key k_j is assigned to one of the *Reduce* tasks and all values for this key are put into a list. Key and list of values (i.e., $(k_j, \langle v_{j,1}, v_{j,2}, \dots, v_{j,m} \rangle)$) are input for the *Reduce* task. Function *Reduce* specifies what the output is based on input $(k_j, \langle v_{j,1}, v_{j,2}, \dots, v_{j,m} \rangle)$. For example, the sum or average could be computed over the list of values for a particular key.

Figure 12.5 shows an example. The input consists of 8 sentences split into four chunks each containing two sentences. Each *Map* task emits one key–value pair per word in its input chunk. In this simple example, the value is always 1. For example, the key value pair (*process*, 1) refers to the first word, (*mining*, 1) refers to the second word, etc. Most keys appear in multiple *Map* tasks. The emitted key–value pairs are grouped using the keys and aggregated into lists. The infrastructure ensures that this is done efficiently. Each key is assigned to a particular *Reduce* task. In the example, there are three *Reduce* tasks each responsible for a few keys. Note that (*great*, (1, 1)) is based on key–value pairs emitted by the first and third *Map* tasks. (*science*, (1, 1, 1)) in the last *Reduce* task is based on three key–value pairs: one emitted by the second *Map* task and two emitted by the last *Map* task. The *Reduce* function takes as input pairs consisting of a key and its list of associated values and combines these values in some aggregated result. Here, the sum is taken. For example, (*science*, (1, 1, 1)) is reduced to (*science*, 3) indicating that the word “*science*” appears three times in the whole text.

Often the *Reduce* function is commutative and associative. In this case, part of the aggregation can be moved to the *Map* function. For example, the last *Map* task in Fig. 12.5 could have emitted a single key–value pair (*science*, 2) rather than two key–value pairs (*science*, 1). Typically, there are more *Map* tasks than *Reduce* tasks to limit communication.

Map and *Reduce* tasks can be done concurrently using possibly thousands of compute nodes. MapReduce is particularly useful in situations where even linear or quadratic algorithms are not fast enough. If an event log is so large that just scanning the log takes too long, the problem needs to be distributed to solve smaller problems in parallel. In the ideal scenario MapReduce scales linearly with the number of compute nodes, i.e., if the number of compute nodes is doubled, the problem is solved in half the time.

MapReduce was first implemented within Google [44]. Later it became a core ingredient of the Apache Hadoop ecosystem. See [114] for an illustration of the broad range of problems that can be tackled using MapReduce.

GFS and HDFS are used in conjunction with commodity hardware distributed over hundreds or thousands of nodes. Since the bandwidth for communication is limited, it is desirable to push computation to the data. Moreover, programming distributed systems is notoriously hard. This explains the relevance of the MapReduce programming model. The user only needs to write “map” and “reduce” functions, and the rest is left to the framework that handles work distribution and faults. Many analysis questions can be translated to “map” and “reduce” functions. In [114], several examples are given. Later, in Sect. 12.2, we will demonstrate that also process discovery can benefit from the MapReduce programming model.

Next to Apache Hadoop, a variety of alternative computing frameworks have been proposed, sometimes also installed on top of or alongside Hadoop. An example is Apache Spark which provides multi-stage in-memory primitives and is particularly suited for machine learning algorithms.

12.1.3 Characterizing Event Logs

This book focuses on a particular type of data, *event data*. The complexity and size of event logs are determined by different factors. This section defines the key characteristics of logs.

Consider, for example, event log $L_1 = [\langle a, b, c, f \rangle^2, \langle a, c, b, f \rangle^2, \langle a, d, e, f \rangle^2]$ in Fig. 12.6. This event log is composed of six cases. The average trace length of a case is four. In total there are 24 events in L_1 . There are three distinct traces, each of which appears twice. There are six distinct activities ($a-f$).

Event log L_2 is twice the size of L_1 in terms of cases and events. However, the number of distinct traces and distinct activities did not change.

Event log L_3 is also twice the size of L_1 in terms of events. However, the average trace length of a case has doubled. Event logs L_1 , L_2 , and L_3 have the same number of distinct traces and distinct activities.

The size of event log L_4 is the same as the size of L_1 (both in terms of cases and events). However, the number of distinct activities has doubled.

The size of event log L_5 is also the same as the size of L_1 (both in terms of cases and events). However, now all cases are distinct and the traces vary in length.

Figure 12.6 shows that the complexity of an event log does not just depend on the number of events. A perfectly fitting process model learned for L_4 will have twice the number of activities compared to the model learned for L_1 . The variety in L_5 seems higher than in L_1 and L_2 which are of the same size. Therefore, we require multiple *event log metrics* to adequately characterize the input of a process mining task.

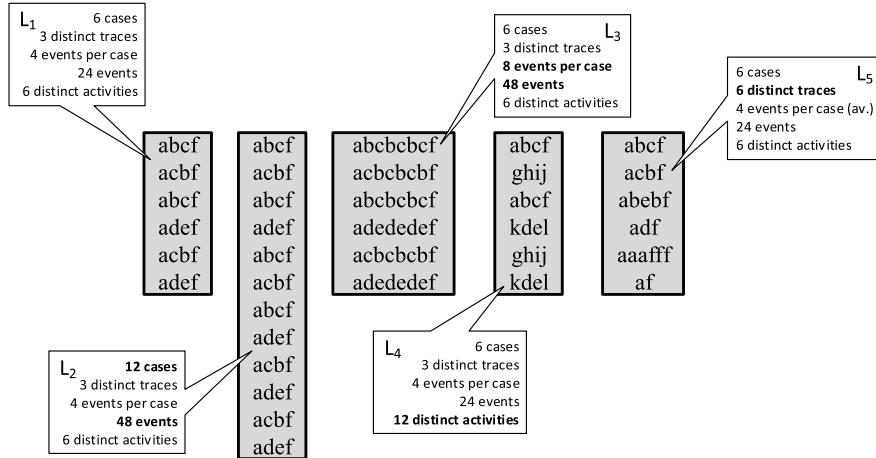


Fig. 12.6 Illustrating some of the key characteristics of an event log

Let us first focus on *simple event logs* without event attributes other than the activity name, i.e., $L \in \mathbb{B}(\mathcal{A}^*)$ is a multi-set of traces. Let us recall some notations introduced before. These will be used to define several event log metrics. In Definition 7.5, we denoted $G(L) = (A_L, \rightarrow_L, A_L^{start}, A_L^{end})$ as the *directly-follows graph* of L with $A_L = \{a \in \sigma \mid \sigma \in L\}$ as the set of observed activities, $\rightarrow_L = \{(a, b) \in A \times A \mid a >_L b\}$ as the directly follows relation, $A_L^{start} = \{a \in A \mid \exists_{\sigma \in L} a = \text{first}(\sigma)\}$ as the set of start activities, and $A_L^{end} = \{a \in A \mid \exists_{\sigma \in L} a = \text{last}(\sigma)\}$ as the set of end activities. Recall that $\partial_{set}(\sigma) = \{a_1, a_2, \dots, a_n\}$ for any $\sigma = \langle a_1, a_2, \dots, a_n \rangle$.

Definition 12.1 (Event log metrics) Let $L \in \mathbb{B}(\mathcal{A}^*)$ be an event log and $G(L) = (A_L, \rightarrow_L, A_L^{start}, A_L^{end})$. We define the following *event log metrics* for L :

- Number of cases,

$$\#cases(L) = |L|$$

- Average trace length of cases,

$$av_{lloc}(L) = \frac{\sum_{\sigma \in L} L(\sigma) \times |\sigma|}{|L|}$$

- Alternatively one can compute the minimal trace length, the maximal trace length, and the standard deviation of trace lengths.
- Number of distinct activities,

$$\#acts(L) = |A_L|$$

- Average number of distinct activities per case,

$$av_{dapc}(L) = \frac{\sum_{\sigma \in L} L(\sigma) \times |\partial_{set}(\sigma)|}{|L|}$$

Alternatively one can compute the minimal or maximal number of distinct activities or the standard deviation.

- Average set-based non-overlap of traces,

$$av_{sbnot}(L) = 1 - \frac{\sum_{\sigma_1, \sigma_2 \in L} L(\sigma_1) \times L(\sigma_2) \times \frac{|\partial_{set}(\sigma_1) \cap \partial_{set}(\sigma_2)|}{|\partial_{set}(\sigma_1) \cup \partial_{set}(\sigma_2)|}}{|L|^2}$$

$av_{sbnot}(L)$ compares pairs of traces in terms of overlap. If all traces refer to the same set of activities, then $av_{sbnot}(L) = 0$. If traces tend to refer to disjoint sets of activities, then $av_{sbnot}(L)$ will be closer to 1. Other distance measures could be used taking the cardinalities into account (Euclidean or Jaccard distance).

- Number of distinct cases,

$$\#dc(L) = |\{\sigma \in L\}|$$

- Number of events,

$$\#events(L) = \#cases(L) \times av_{tloc}(L) = \sum_{\sigma \in L} L(\sigma) \times |\sigma|$$

- Number of direct successions,

$$\#ds(L) = |\mapsto_L|$$

$\#ds(L)$ counts the number of arcs in the directly-follows graph.

- Number of start activities,

$$\#sa(L) = |A_L^{start}|$$

- Number of end activities,

$$\#ea(L) = |A_L^{end}|$$

Consider, for example, $L_1 = [\langle a, b, c, f \rangle^2, \langle a, c, b, f \rangle^2, \langle a, d, e, f \rangle^2]$ in Fig. 12.6. Here $\#cases(L_1) = 6$ (number of cases), $av_{tloc}(L_1) = 4$ (average trace length of cases), $\#acts(L_1) = 6$ (number of distinct activities), $av_{dapc}(L_1) = 4$ (average number of distinct activities per case), $av_{sbnot}(L_1) = 0.296$ (average set-based non-overlap of traces), $\#dc(L_1) = 3$ (number of distinct cases), $\#events(L_1) = 24$ (number of events), $\#ds(L_1) = 9$ (number of direct successions), $\#sa(L_1) = 1$ (number of start activities), and $\#ea(L_1) = 1$ (number of end activities). Table 12.1 also shows the event log metrics for the other event logs in Fig. 12.6.

Event logs are considered more challenging if the values for these metrics are higher. L_2 is most challenging (of the five toy logs) in terms of the number of cases ($\#cases(L_2) = 12$). L_3 is most challenging in terms of the average trace

Table 12.1 Event log metrics for the five event logs in Fig. 12.6

Event log metric		L_1	L_2	L_3	L_4	L_5
Number of cases	$\#_{cases}(L_i)$	6	12	6	6	6
Average trace length of cases	$av_{tloc}(L_i)$	4	4	8	4	4
Number of distinct activities	$\#_{acts}(L_i)$	6	6	6	12	6
Average number of dist. act. per case	$av_{dapc}(L_i)$	4	4	4	4	3.166
Average set-based non-overlap of traces	$av_{sbnor}(L_i)$	0.296	0.296	0.296	0.667	0.348
Number of distinct cases	$\#_{dc}(L_i)$	3	3	3	3	6
Number of events	$\#_{events}(L_i)$	24	48	48	24	24
Number of direct successions	$\#_{ds}(L_i)$	9	9	10	9	13
Number of start activities	$\#_{sa}(L_i)$	1	1	1	3	1
Number of end activities	$\#_{ea}(L_i)$	1	1	1	3	1

length ($av_{tloc}(L_3) = 8$). L_4 is most challenging in terms of the number of distinct activities ($\#_{acts}(L_4) = 12$), the least overlap of activities in pairwise comparison of traces ($av_{sbnor}(L_4) = 0.667$), and the number of start and end activities ($\#_{sa}(L_4) = \#_{ea}(L_4) = 3$). L_5 is most challenging in terms of the number of distinct cases ($\#_{dc}(L_5) = 6$) and the number of direct successions ($\#_{ds}(L_5) = 13$).

Events may have any number of attributes (see Definition 5.1). Some of the attributes are standard: timestamp, resource, and transaction type. Other attributes may be domain-specific, e.g., blood pressure, sales region, age, voltage, and grade. Attributes may be sparse or not. For example, every event is expected to have a timestamp. However, there may also be attributes like blood pressure that are attached to only a small fraction of all events. Such an attribute is sparse as only few events in the event log have it. Next to the control-flow oriented metrics in Definition 12.1, one can define additional event log metrics such as:

- The number of distinct attributes in an event log $\#_{ndael}(L)$ and
- The average number of attributes per event $av_{nape}(L)$.

The event logs in Fig. 12.6 are, of course, not representative for real-life event logs. In this chapter, we are particularly interested in challenging event logs, e.g., event logs with tens of thousands of cases ($\#_{cases}(L) \gg 10,000$), millions of events ($\#_{events}(L) \gg 1,000,000$), and hundreds of different activities ($\#_{acts}(L) \gg 100$).

For the α -algorithm, the heuristic miner, the fuzzy miner, and the inductive miner based on the directly-follows graph (IMD and IMFD algorithms), a single pass through the entire event log suffices. Even if such algorithms are exponential in the number of different activities, the pass through the event log typically remains most time consuming. Note that often the number of distinct activities ($\#_{acts}(L)$) is orders of magnitude smaller than the number of cases ($\#_{cases}(L)$) or the number of events ($\#_{events}(L)$). In later sections, we will show that it is possible to decompose computation based on structures like the directly-follows graph.

Conformance checking based on alignments and discovery based on language-based regions require solving optimization problems (e.g., an ILP problem). These

problems are more challenging and cannot be decomposed as easily. Here the number of distinct activities ($\#_{acts}(L)$) and the average trace length of cases ($av_{tloc}(L)$) are important.

As mentioned in Chap. 11, some of today’s process mining tools can handle logs with billions of events and millions of cases. However, this only holds for single-pass algorithms based on counting and does *not* apply to algorithms doing some form of optimization or state exploration (e.g., conformance checking or region-based discovery). Moreover, scalability depends on many factors. An event log with a lot of redundancy ($\#_{cases}(L) \gg \#_{dc}(L)$) is easier to analyze than a log where most cases are unique. Some systems have problems with large logs having many attributes per event (e.g., $av_{nape}(L) \gg 100$) even when these attributes are not used for analysis.

In the remainder, we will focus on process discovery and conformance checking. These are most challenging from a performance point of view and form the backbone for subsequent analyses. Note that after computing alignments, it is easy to extend the model with performance information and other aspects learned from the event log.

12.2 Case-Based Decomposition

Process mining is motivated by the availability of event data. However, as event logs become larger (say gigabytes or terabytes), performance becomes a concern. The only way to handle larger applications while ensuring acceptable response times, is to distribute analysis over a network of computers (e.g., multicore systems, grids, and clouds). This ultimately requires splitting the event log: each compute node becomes responsible for a part of the event data. We consider two types of decomposition, *case-based decomposition* and *activity-based decomposition*.

Case-based decomposition (called “vertical partitioning of the event log” in [141]), distributes events based on the case they belong to. Event logs may be composed of millions of cases. These can be distributed over the compute nodes such that each case is assigned to one node. Hence, each compute node works on a subset of the whole log after which the results need to be merged.

Activity-based decomposition (also called “horizontal partitioning of the event log” [141]), distributes events based on the activities they refer to. The number of unique activities is typically much smaller than the number of cases or events. However, many process mining algorithms are non-linear in the number of activities. Creating subproblems working on smaller groups of activities may therefore provide super-linear speedups. Hence, activity-based decomposition can also be used on a single computer solving the subproblems in sequence. For activity-based decomposition cases are split up in smaller traces working on subsets of activities. In principle, each compute node needs to consider all cases. Typically, the activity sets are partly overlapping as will be explained in Sect. 12.3.

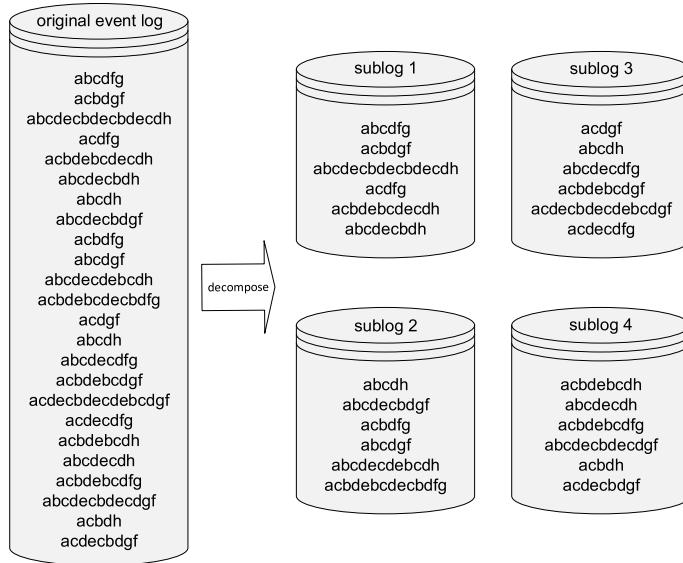


Fig. 12.7 Case-based decomposition of a small event log: The 24 cases are split into four groups of 6 cases

In this section, we focus on case-based decomposition. As shown in Fig. 12.7, the basic idea is very simple. The cases are simply distributed over the n sublogs. Each of the n sublogs can then be analyzed in parallel.

12.2.1 Conformance Checking Using Case-Based Decomposition

In Chap. 8, we introduced various conformance checking techniques. Token-based replay (Sect. 8.2) and alignment-based conformance checking (Sect. 8.3) are most interesting because they directly relate cases in the event log to the model. Alignment-based conformance checking provides better diagnostics (more detailed, accurate, and easy to understand) than token-based replay, but is more time consuming. To compute optimal alignments, optimization problems need to be solved which can become intractable if model and log are huge.

Case-based decomposition is a straightforward way to achieve a linear speedup for both token-based replay and alignment-based conformance checking. If there are n compute nodes each handling $\frac{1}{n}$ of the cases, then conformance can be checked in $\frac{t_s}{n}$ time where t_s is the time to check conformance using a single node. Such a linear speedup can only be achieved if the overhead is negligible compared to the time to do the actual conformance computations.

Consider the WF-net in Fig. 12.8. Suppose we would like to check conformance of the event log of Fig. 12.7 with respect to this process model. We can sequentially

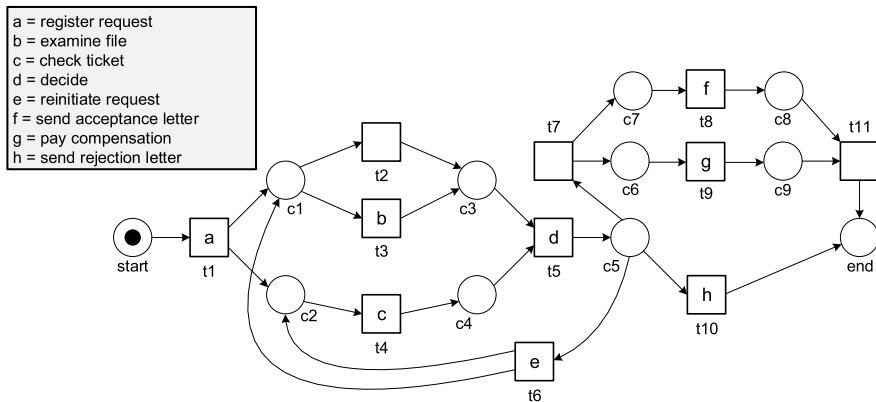


Fig. 12.8 Process model used to illustrate different decomposition approaches

check all 24 cases on a single compute node or we can split the event log into four parts as shown in Fig. 12.7. This means that 4×6 cases are checked in parallel. It is easy to combine the diagnostics for the four sublogs. For token-based replay (Sect. 8.2), we can simply *add up the consumed, produced, missing, and remaining tokens* (Sect. 8.2). For alignment-based conformance checking, we can *add up all misalignment costs* (Sect. 8.3). Also metrics such as the percentage of (non-)perfectly fitting cases, moves in model only, and moves in log only can be easily computed based on the intermediate results for the four sublogs.

When splitting the cases, we may exploit redundancy in the event log. By putting similar cases in the same sublog, further speedups are possible. If two cases have the same trace, then conformance only needs to be checked for one of them. Moreover, sophisticated alignment-based conformance checking techniques cache intermediate results and therefore handle similar cases faster. Obviously, there is a trade-off since it may take time to cluster similar cases. Also a simple hashing function can be used to group cases.

12.2.2 Process Discovery Using Case-Based Decomposition

In Chap. 6, we introduced the α -algorithm. More advanced process discovery algorithms were introduced in Chap. 7: heuristic mining (Sect. 7.2), genetic process mining (Sect. 7.3), region-based mining (Sect. 7.4), and inductive mining (Sect. 7.5). These algorithms have very different performance characteristics.

Generic process mining (Sect. 7.3) can be trivially parallelized in a number of ways. We can decompose the event log as shown in Fig. 12.7. However, we can also replicate the whole event log at each compute node. Per node there are separate generations of individuals (candidate models). Selection (including fitness computations) and reproduction (e.g., crossover and mutation) are done per node. Periodically, individuals are exchanged between compute nodes. By sharing the best

individuals between the parallel nodes, the evolutionary process typically converges faster.

Region-based mining techniques (Sect. 7.4) are difficult to parallelize. If region-based techniques are applied to sublogs, there is no easy way to merge the resulting models. For example, a place that can be added according to one sublog may not be feasible according to another sublog.

The *inductive mining techniques* (Sect. 7.5) that actually split the event log (i.e., IM, IMF, and IMC) are also difficult to decompose. The different sublogs can be analyzed separately, however, finding the initial exclusive-choice, sequence, parallel, or redo-loop cut is most time consuming. Only after splitting the event logs based on activities, the work can be distributed. This does not correspond to the notion of case-based decomposition. Learning a model per sublog is possible, but merging these models is problematic. Most likely the models disagree to certain ordering relations. Fortunately, inductive mining based on the directly-follows graph (without log splitting) can be decomposed, as will be explained next.

In the remainder, we focus on the α -algorithm, the heuristic miner, the fuzzy miner, and the directly-follows based inductive miners (IMD, IMFD, and IMCD). These have in common that they are based on the *directly-follows graph* or a similar internal aggregate structure. Key for the α -algorithm is relation $>_L$ that contains all pairs of activities in a “directly follows” relation, and the sets T_I and T_O (cf. Definition 6.4). The dependency graph used by the heuristic miner is similar to the “directly follows” relation, but incorporates frequencies. Miners such as IMD, IMFD, and IMCD start from a directly-follows graph $G(L) = (A_L, \mapsto_L, A_L^{start}, A_L^{end})$ (see Definition 7.5). Despite subtle differences the basis is the same—*counting local patterns in cases*.

Any process discovery technique that is based on counting local patterns in individual cases can benefit from case-based decomposition. Consider Fig. 12.7 and the directly follows relation between b and c . In the first sublog, b is four times directly followed by c , in the second sublog six times, in the third sublog four times, and in the fourth sublog also four times. Hence, the frequency of the directly follows relation between b and c is $4 + 6 + 4 + 4 = 18$. The same can be done with start and end activities. Frequencies of local patterns can be counted per case and therefore simply added per sublog. These frequencies can be aggregated over all sublogs, making decomposition easy. Note that case-based decomposition does not influence the outcome: The same process model is discovered.

Discovery approaches that count patterns in cases (e.g., the frequency of the directly follows relation) can exploit the *MapReduce* programming model introduced in Sect. 12.1.2. Consider the example in Fig. 12.9. An event log consisting of 16 cases is split into four chunks. The *Map* function emits a key–value pair for every direct succession. To keep track of initial and final activities, dummy starts and ends have been added denoted by \triangleright (start) and \square (end). The first trace $\langle a, b, c, f \rangle$ handled by the first *Map* task results in five key–value pairs: $(\triangleright a, 1)$, $(ab, 1)$, $(bc, 1)$, $(cf, 1)$, and $(f\square, 1)$. The *MapReduce* framework takes care of the grouping and aggregation of these key–value pairs. The first *Reduce* task is responsible for four keys ($\triangleright a$, ab , ac , and ad), and simply adds up the values in the lists. The output of this

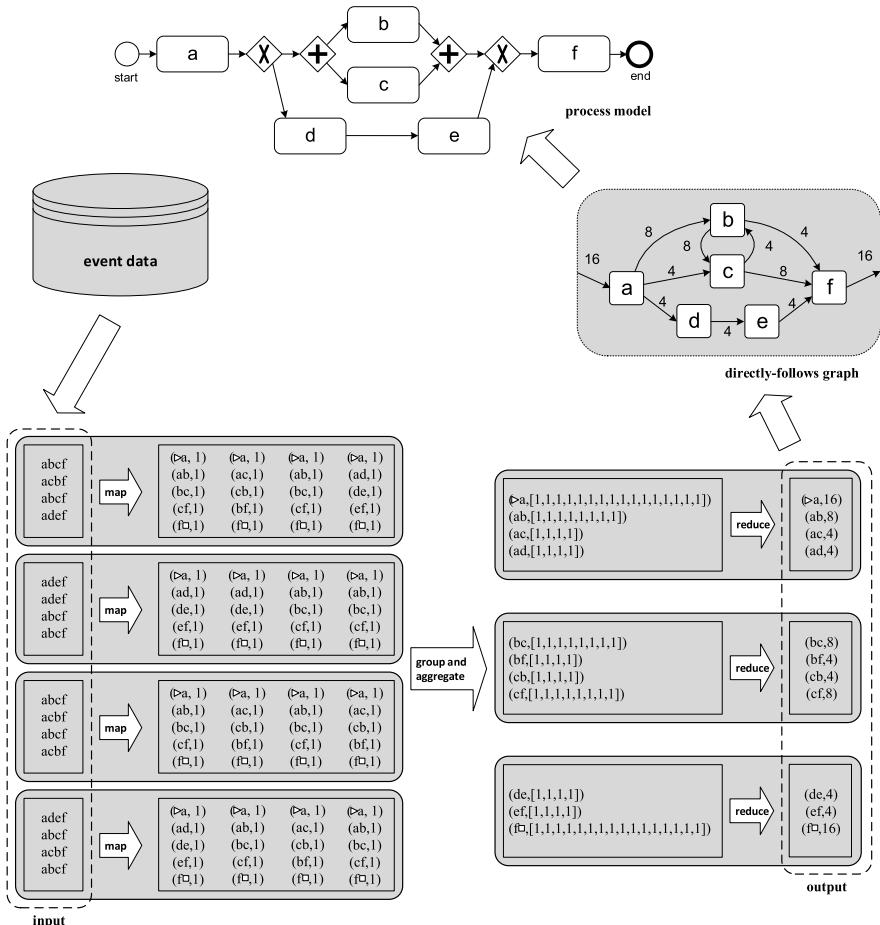


Fig. 12.9 Computing the directly-follows graph using MapReduce

Reduce task is $(\triangleright a, 16)$, $(ab, 8)$, $(ac, 4)$, and $(ad, 4)$. The combined output of all Reduce tasks provides all information needed to produce the directly-follows graph shown in Fig. 12.9 (including frequencies). Based on this graph we can apply the α -algorithm and the directly-follows based inductive miner (IMD). Other process discovery algorithms like the heuristic miner, the fuzzy miner, and other variants of the α -algorithm and inductive miner use similar inputs. Figure 12.9 shows the process model returned by both the α -algorithm and the IMD inductive miner (using the directly-follows graph).

The example logs in Figs. 12.7 and 12.9 are too small to really illustrate case-based decomposition. For small examples the overhead caused by decomposition will only slow down analysis. One needs to imagine that event logs contain millions

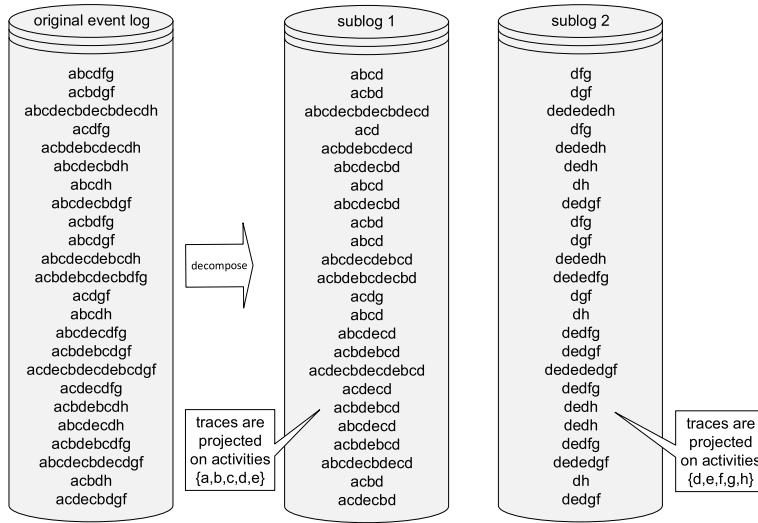


Fig. 12.10 Activity-based decomposition of a small event log: The two sublogs are projections of the original event log

or even billions of cases referring to hundreds or thousands of different activities. In such cases, it is vital to be able to distribute analysis. Figure 12.9 shows that it is fairly easy to formulate process discovery in terms of the MapReduce programming model and use a Hadoop-like infrastructure.

12.3 Activity-Based Decomposition

Case-based decomposition is most suitable when using techniques that ultimately count local patterns, e.g., to construct a directly-follows or dependency graph. However, discovery techniques that do not use such patterns (e.g., region-based discovery or inductive mining based on log-splitting) cannot use case-based decomposition. Conformance techniques can also use case-based decomposition, but the complexity is in the number of activities and the average trace length. Therefore, a linear speedup may not be enough. *If computing an alignment for a single trace takes too long or is even infeasible, case-based decomposition is not a viable solution.*

For situations where case-based decomposition is not good enough, one can consider *activity-based decomposition* as an alternative decomposition approach. Figure 12.10 sketches the idea. Sublogs are created by projecting cases onto subsets of activities. Each sublog is responsible for a particular subset of activities. The first sublog in Fig. 12.10 is responsible for subset $\{a, b, c, d, e\}$ and the second sublog is responsible for subset $\{d, e, f, g, h\}$. Note that the two activity sets are overlapping. Strictly speaking, this is not a requirement. However, later we will see that some overlap is desirable.

We use the projection operator introduced in Chap. 5 to explain activity-based decomposition. $\sigma \uparrow X$ is the projection of σ onto some subset $X \subseteq A$, e.g., $\langle a, b, c, a, b, c, d \rangle \uparrow \{a, b\} = \langle a, b, a, b \rangle$. In Fig. 12.10, the set of all activities is $A = \{a, b, c, d, e, f, g, h\}$. The two sublogs are based on subsets $A_1 = \{a, b, c, d, e\}$ and $A_2 = \{d, e, f, g, h\}$. Consider, for example, the first trace in the original log $\sigma = \langle a, b, c, d, f, g \rangle$. This corresponds to trace $\sigma \uparrow A_1 = \langle a, b, c, d \rangle$ in the first sublog and trace $\sigma \uparrow A_2 = \langle d, f, g \rangle$ in the second sublog.

We can also apply the projection operator to event logs. $L = [\langle a, b, c, d, f, g \rangle, \langle a, c, b, d, g, f \rangle, \dots, \langle a, c, d, e, c, b, d, g, f \rangle]$ is the original event log. $L_1 = L \uparrow A_1 = [\langle a, b, c, d \rangle, \langle a, c, b, d \rangle, \dots, \langle a, c, d, e, c, b, d \rangle]$ is first sublog, and $L_2 = L \uparrow A_2 = [\langle d, f, g \rangle, \langle d, g, f \rangle, \dots, \langle d, e, d, g, f \rangle]$ is the second sublog.

In the general setting, we can have any event log L with activities A decomposed in k subsets A_1, A_2, \dots, A_k such that $A = A_1 \cup A_2 \cup \dots \cup A_k$. The corresponding k sublogs are $L \uparrow A_1, L \uparrow A_2, \dots, L \uparrow A_k$. All sublogs have the same number of cases as the original log, but traces are much shorter if the subsets are relatively small. Process mining algorithms can be applied to each of the k sublogs, after which the partial results need to be merged. This is often surprisingly easy, as is shown next.

12.3.1 Conformance Checking Using Activity-Based Decomposition

In case of decomposed conformance checking, we have a model N and log L with activities A . Although the decomposition approach does not depend on Petri nets (see [142]), let us assume that N is a Petri net with an initial and final marking. We allow for duplicate and silent activities, e.g., transitions with a τ label or multiple transitions with the same label.

Assume we would like to check the conformance of $L = [\langle a, b, c, d, f, g \rangle, \langle a, c, b, d, g, f \rangle, \dots, \langle a, c, d, e, c, b, d, g, f \rangle]$ with respect to the Petri net in Fig. 12.8. *How to decompose the event log?* We need to identify subsets of activities and decompose model and log based on this. The basic idea is that we must cut the Petri net in Fig. 12.8 into fragments such that we only cut through transitions that are visible and unique. We cannot cut through places or silent transitions. We can also not cut through transitions having a label appearing at multiple places. In fact, such transitions should also not become separated. The exact requirements are given in [144].

Using the approach just described we can split the Petri net N in two parts as shown in Fig. 12.11. This way we find activity sets $A_1 = \{a, b, c, d, e\}$ and $A_2 = \{d, e, f, g, h\}$. Next we create two sublogs based on these activity sets, $L_1 = L \uparrow A_1$ and $L_2 = L \uparrow A_2$. These are the two sublogs shown in Fig. 12.10. The fragments also form two smaller process models. N_1 is the Petri net on the left-hand side of the dotted line (including the transitions labeled d and e). N_2 is the Petri net on the right-hand side of the dotted line (also including the boundary transitions). N_1 has

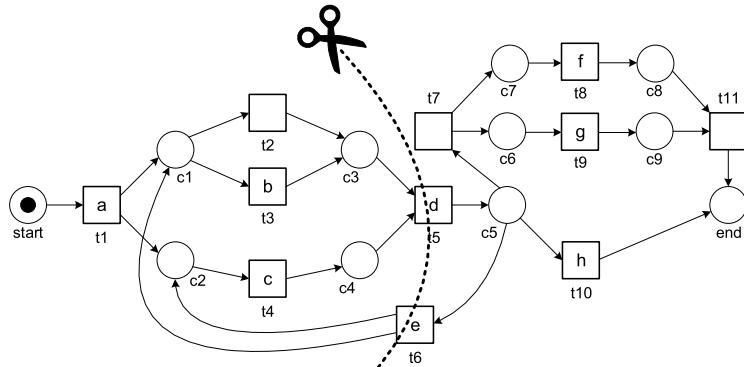


Fig. 12.11 By cutting through transitions that are visible and unique, we find activity sets $A_1 = \{a, b, c, d, e\}$ and $A_2 = \{d, e, f, g, h\}$

initial marking [*start*] and the empty final marking ([]). N_2 has initial marking [] and final marking [*end*].

Next we check conformance of $L_1 = L \uparrow A_1$ and $L_2 = L \uparrow A_2$ with respect to the two Petri net fragments N_1 and N_2 . In [144], it is shown that L is perfectly fitting N if and only if L_1 is perfectly fitting N_1 and L_2 is perfectly fitting N_2 . Hence, we can translate a larger conformance checking problem into two smaller ones without losing accuracy. Since L_1 is perfectly fitting N_1 and L_2 is perfectly fitting N_2 , we conclude that L is perfectly fitting N .

The Petri net can be decomposed into more parts using the rule explained before: The net can be further decomposed by cutting through transitions that are visible and unique. Figure 12.12 shows that the Petri net can also be split into three fragments. Again we have the same guarantee: the overall log is perfectly fitting if and only if each of the sublogs is perfectly fitting. The partitioning of the net into fragments following the rules described in [144] is called a *valid decomposition*. There is always a unique *maximal* decomposition having fragments which cannot be split further. The maximal decomposition of the Petri net in Fig. 12.8 has six fragments resulting in activity sets: $\{a\}$ (transitions and arcs connected to *start*), $\{a, b, d, e\}$ (transitions and arcs connected to c_1 and c_3), $\{a, c, e\}$ (transitions and arcs connected to c_2), $\{c, d\}$ (transitions and arcs connected to c_4), $\{d, e, f, g, h\}$ (transitions and arcs connected to c_5, c_6 and c_7), and $\{f, g, h\}$ (transitions and arcs connected to c_8, c_9 and *end*). Note that each place and each arc of the original Petri net appears in precisely one of the fragments.

Given a valid decomposition (maximal or not) of an arbitrary Petri net N into fragments N_1, N_2, \dots, N_k with activity sets A_1, A_2, \dots, A_k : L is perfectly fitting N if and only if for all $i \in \{1, \dots, k\}$ L_i is perfectly fitting N_i . If a deviation is found in one of the sublogs, then there is a deviation in the overall log. If no deviation is found in any of the sublogs, then there are no deviations in the overall log.

The approach in [144] is very general and can be applied to other process models [142] and combined with a variety of decomposition strategies [106]. For larger, relatively structured process models, the time for conformance checking may go from

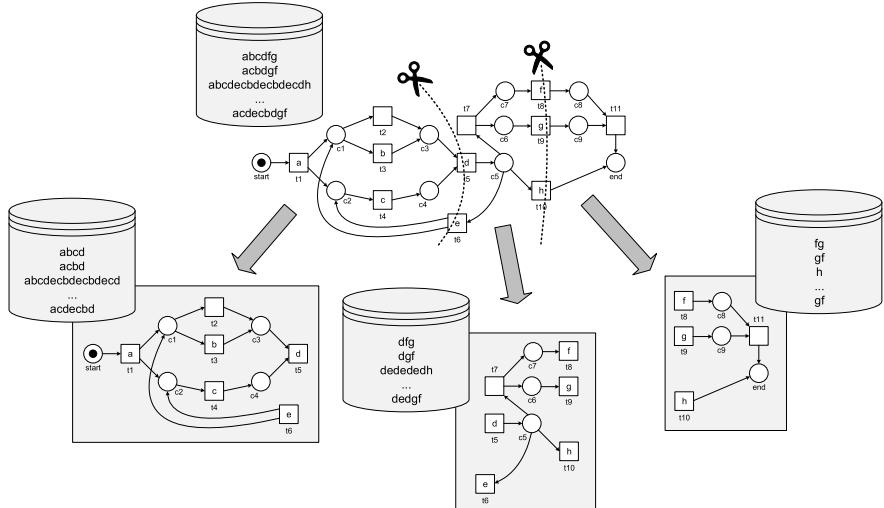


Fig. 12.12 Another valid decomposition of the process model resulting in three sublogs: $L_1 = L \uparrow \{a, b, c, d, e\}$, $L_2 = L \uparrow \{d, e, f, g, h\}$ and $L_3 = L \uparrow \{f, g, h\}$

hours or days to seconds or minutes (on a single computer). *Even when computation is not distributed over multiple computing nodes, decomposition can help to speed up conformance checking.* This can be explained by the fact that some algorithms are exponential in the number of different activities or the average length of the traces.

12.3.2 Process Discovery Using Activity-Based Decomposition

Assume we have an “Oracle” that, given an event log L over A , provides activity sets A_1, A_2, \dots, A_k with $A = A_1 \cup A_2 \cup \dots \cup A_k$. If we discover a perfectly fitting model N_i for each of the sublogs $L_i = L \uparrow A_i$, then L is also perfectly fitting the composed model $N = N_1 \oplus N_2 \oplus \dots \oplus N_k$. The composition synchronizes the different submodels based on shared activity labels. In terms of a Petri net with unique labels, this means that the composition is the net obtained by fusing transitions having the same label (see [144]). This is the reverse of the decompositions in Fig. 12.11 and Fig. 12.12.

We can apply any existing discovery technique to the sublogs $L_i = L \uparrow A_i$. It is even possible to apply different discovery techniques to different sublogs. This can also be done in parallel. Hence, given a suitable Oracle, we get a highly configurable approach to discover process models in a decomposed or distributed manner.

Figure 12.13 sketches the approach using an example. Based on the original log $L = [\langle a, b, c, d, e, f, g, i \rangle, \langle a, c, d, f, h, e, i \rangle, \langle a, c, b, d, g, f, e, i \rangle, \langle a, c, d, h, e, f, i \rangle, \dots]$, the Oracle suggests: $A_1 = \{a, b, c, d\}$, $A_2 = \{d, g, h, i\}$, and $A_3 =$

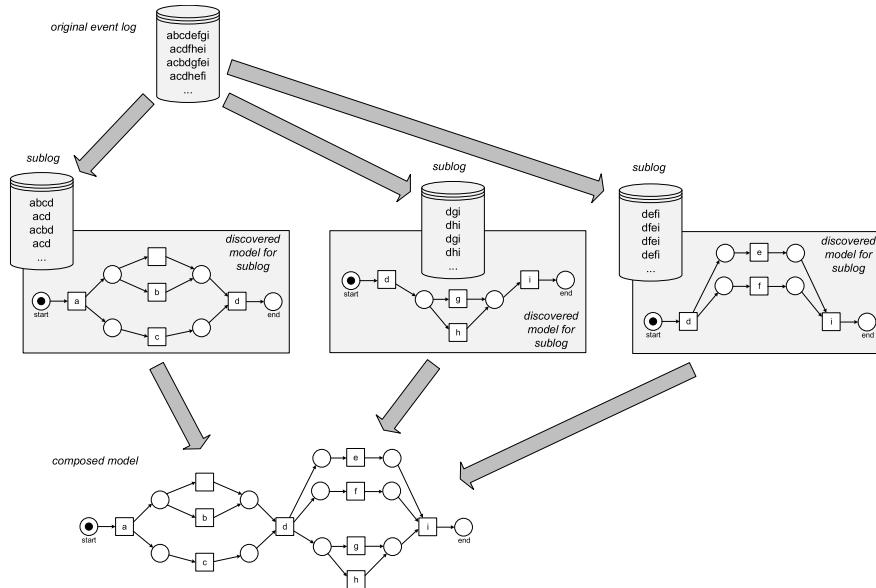


Fig. 12.13 Decomposed discovery: The event log is projected onto subsets of activities and the resulting sublogs are input for a standard discovery technique. The resulting process models are merged into an overall model by synchronizing overlapping activities

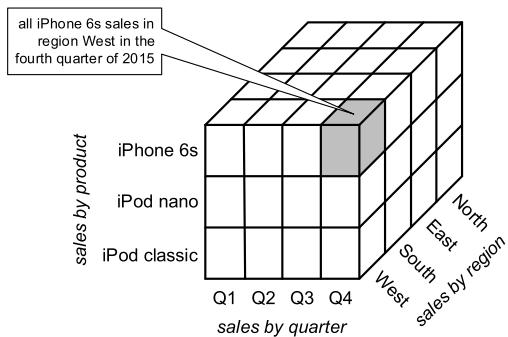
$\{d, e, f, i\}$. The three sublogs $L_1 = L \uparrow A_1$, $L_2 = L \uparrow A_2$, and $L_3 = L \uparrow A_3$ serve as input for conventional discovery approaches. This step can be distributed. The resulting three models are shown in Fig. 12.13 and can be merged into an overall model. Superfluous start/end places are removed while composing the models. Since each of the intermediate models is perfectly fitting, the overall model is also perfectly fitting [144].

All of this is done under the assumption of some Oracle providing A_1, A_2, \dots, A_k . If these sets are poorly chosen (e.g., no overlap between activity sets), then the resulting model may be severely underfitting. The Oracle may be based on *domain knowledge*, e.g., exploiting the natural hierarchy of a system or organization. Note that a software architecture provides information on which components can interact. Such information can be exploited to select partly overlapping activity sets.

We can also use sampling to quickly build a directly-follows graph (or use the MapReduce approach described before) and then use heuristics to decompose the graph [74]. There are various other approaches to quickly decide on activity sets without trying to discover an overall process model first. After projecting the event log on sublogs, more time-consuming approaches can be used.

A well-chosen collection A_1, A_2, \dots, A_k may also help in the balance between overfitting and underfitting and thus lead to better results in less time. For activities in different subsets we do not need to observe all interleavings to derive concurrency: We only need to see the “local” interleavings.

Fig. 12.14 Three dimensional OLAP cube containing sales data. Each cell refers to all sales of a particular product in a particular region and in a particular period. For each cell we can compute metrics such as the number of items sold or the total value



The goal of this section was to illustrate the different ways in which logs can be decomposed and used for both conformance checking and discovery. The goal was not to describe a concrete algorithm. Therefore, we skipped many of the details. However, the examples show that there are many ways to decompose process mining problems when event logs get extremely large.

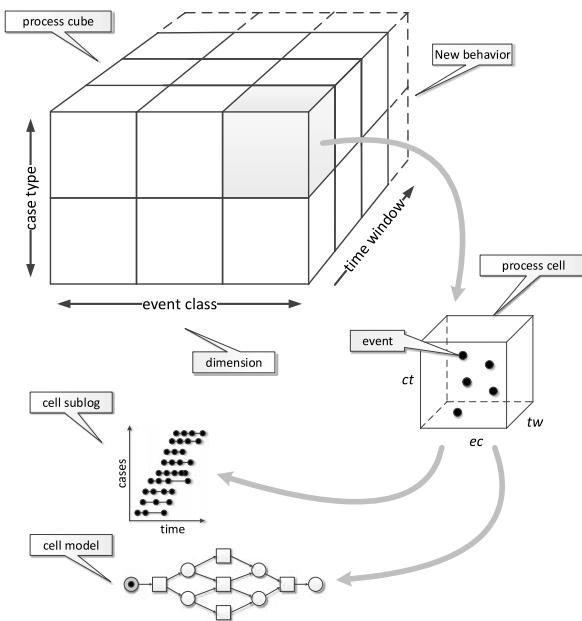
12.4 Process Cubes

Case-based and activity-based decomposition distribute events over sublogs. For case-based decomposition, each event was assigned to a particular sublog based on the case it belongs to. In case of activity-based decomposition an event may be assigned to multiple sublogs (activity sets may overlap). Decomposition was done for performance reasons. However, there may be more reasons for grouping events. These are discussed in this section using the notion of a *process cube*.

In a process cube, events are organized using different *dimensions* (e.g., case types, regions, subprocesses, departments, and time windows). The cells in such a process cube can be analyzed using process mining techniques by creating a sublog per cell. The results of different cells can be compared. Note that a process cube does not need to be limited to a single process: All events recorded in an organization can be organized in a single cube.

Process cubes are closely related to the cubes used in OLAP (Online Analytical Processing). Figure 12.14 shows an example of an OLAP cube. The example cube has three dimensions and the elements in each cell refer to sales transactions. Assume we are interested in the number of items sold. In this case, the OLAP tool can be used to show the number of products sold for each cell in the OLAP cube. Suppose that in the fourth quarter (Q4) very few iPhones were sold in region West. Then one can drill down into this cell. For instance, one can look at individual sales, view the sales per month (refinement of Q4 into October, November, and December), or view the sales per shop (refinement of the region dimension). When drilling down the information is refined. Pivoting the data, often referred to as “slicing and dicing”, helps to see particular patterns. By “slicing” the OLAP cube, the analyst can zoom into a selected slice of the overall data, e.g., only looking at sales of the

Fig. 12.15 Each cell in a process cube contains a collection of events that can be converted into an event log. Any process mining technique can be applied to the corresponding event log and subsequently results (e.g., a process model and social network) are associated to the cell. OLAP operations such as slice, dice, roll-up, and drill-down facilitate exploration and comparison of behavior



iPod nano. Slicing effectively removes a dimension from the cube. The term “dicing” refers to applying filters on (possibly) multiple dimensions of the cube. Dicing corresponds to selecting a subcube rather than removing a dimension by selecting a value for it. Views can also be changed by rotating the cube, e.g., swapping the rows and columns. The results can be viewed in tabular form or visualized using various charts. Many BI tools support the OLAP functionality described. These tools support a broad range of charts, e.g., pie charts, bar charts, radar plots, scatter plots, speedometers, Pareto charts, box plots, and scorecards. These can be used to view the data from different angles.

OLAP cubes and notions such as slicing and dicing are *not* process centric. BI products can analyze an OLAP cube with sales data as shown in Fig. 12.14, but do this without considering the underlying process. The sales events are immediately aggregated without trying to distill the underlying process.

Process cubes [22, 145] add process specific elements to OLAP cubes and can be viewed as a crossover between process mining and OLAP. Event attributes like activity, timestamp, resource, transaction type, etc. are handled in a specific manner and cells can be converted to logs as shown in Fig. 12.15. Note that results in OLAP cubes are numbers, e.g., the number of transactions in a shop or the average value of sales in October. Results associated to cells in a process cube may include a variety of models (e.g., process models, social networks) and are not limited to numbers (e.g., a sum or average). Earlier we compared process mining with spreadsheets (Sect. 1.3) and noted that spreadsheets are dealing with numbers rather than events and dynamic behavior. A similar distinction can be made between OLAP cubes and process cubes.

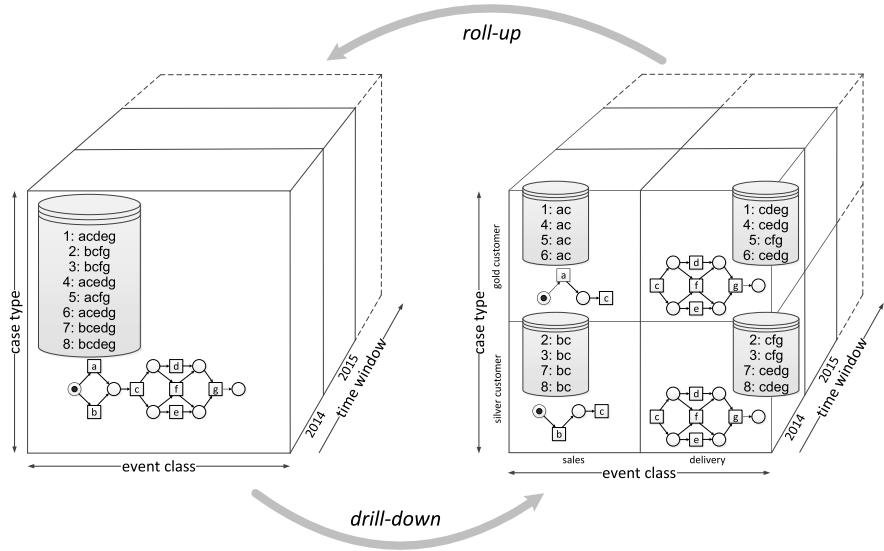


Fig. 12.16 Illustration of roll-up and drill-down. The drill-down operation uses case-based decomposition for the case type dimension and activity-based decomposition for the event class dimension [145]

Figure 12.16 (taken from [145]) is used to illustrate roll-up and drill-down operations. The cube has three dimensions: *case type*, *event class* and *time window*. In the left cube, there is only one case type and only one event class. The cube covers multiple time windows, but only one is shown (all cases completed in 2014). In this toy example, there are only eight cases (i.e., process instances) and seven distinct activities. The process may be split by identifying multiple case types and/or multiple event classes. The cube shown on the right-hand side of Fig. 12.16 has two case types (gold customer and silver customer) and two event classes (sales and delivery).

As shown in Fig. 12.16, cases 1, 4, 5, and 6 refer to gold customers. Hence, the cells in the “gold customer” row in Fig. 12.16(right) include events related to these four cases. The event class dimension is based on the event’s activity. The event class “sales” includes activities *a*, *b*, and *c*. The event class “delivery” refers to activities *c*, *d*, *e*, *f*, and *g*. The time window dimension uses the timestamps of events. A time window may refer to a particular day, week, month, or any other period.

Each dimension in a process cube may have an associated hierarchy, e.g., years composed of months and months composed of days. Using roll-up and drill-down operations, the granularity can be changed as shown in Fig. 12.16.

As mentioned, each cell in a process cube refers to a collection of events and possibly also process mining results (e.g., a discovered process model). Events may be included in *multiple* cells, e.g., sales and delivery cells share *c* events in Fig. 12.16. There are many situations where cells may be overlapping and share events. A person may be part of multiple departments or have multiple roles. The events gener-

ated by this person are therefore associated to the cells of the different departments and roles. Activities may also be part of multiple processes.

Figure 12.16 assumes a fixed case notion. In process cubes, we normally use a more relaxed case notion. This way we can look at the data from different angles as discussed in Sect. 5.5. When creating event logs from cells, we need to “flatten the event data” to have a clear process instance notion. When considering events related to customer orders, we may view the same event data from the viewpoint of orders, order lines, and deliveries (see Sect. 5.5). Therefore, a process cube needs to support multiple case notions. The same event may be part of multiple cells and multiple cases.

Given a process cube with suitably chosen dimensions, we can compare process mining results generated for an array of cells. We refer to this as *comparative process mining*. The goal is to highlight differences between cells. Next to cross-checking conformance (the log for cell i is replayed on the model for cell j), we can compare process models visually or overlay the models as is done in tools like myInvenio (see Fig. 11.15). In the context of ProM several implementations of the process cube concept exist [22, 145]. These have in common that arbitrary plug-ins can be applied to an array cells after which the results can be compared.

OLAP technology and the notion of process cubes can help to deal with large heterogeneous event collections. Also note the relation with event log decomposition (Sects. 12.2 and 12.3). The drill-down operation in Fig. 12.16 uses *case-based* decomposition for the case type dimension and *activity-based* decomposition for the event class dimension.

12.5 Streaming Process Mining

The process mining algorithms discussed before assume that all events are stored in a file or database. The algorithms can access all event data at any time. In some scenarios, such assumptions are unrealistic: Events arrive in *streams* and, if not processed immediately, the corresponding information is lost. Events may arrive so rapidly that it is not feasible to store them all in active storage where they can be processed. At best we may be able to archive the events, but there is no time to process these archived data. Moreover, at any point in time we may need to answer questions related to these streams of events (including the recent events). Hence, there is no time to access archived data. Questions may refer to recent data and need to be answered immediately. Handling streams of events can be viewed as “drinking from a firehose”: Trying to store and process all event data using conventional process mining algorithms is impossible. Therefore, dedicated algorithms are needed to handle streams of events.

Figure 12.17 shows a stream of events. Suppose that we would like to know the process at any point in time. The process may be changing over time. This phenomenon corresponds to the notion of *concept drift* [81] discussed in Sect. 10.6.3. This triggers the question: *Should the discovered model at time t describe the process over the last day, week, month or year?* New activities may appear and other

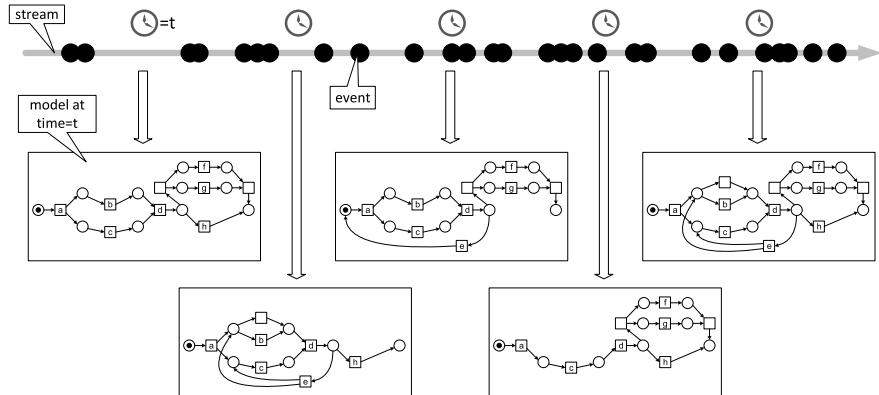


Fig. 12.17 Given a stream of events we would like to provide an up-to-date process model at any point in time

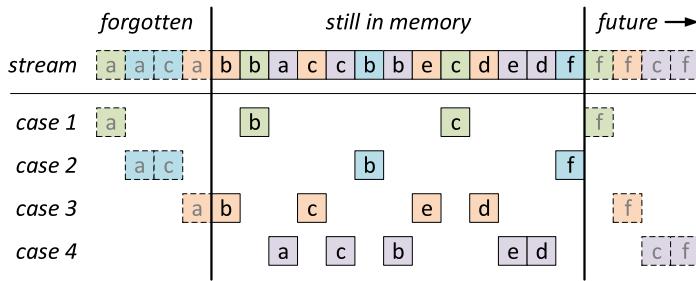


Fig. 12.18 Using a fixed window of events may lead to incorrect conclusions

activities may fade out. Also the ordering of activities may change. Concept drift refers to all perspectives, e.g., bottlenecks may emerge without changing the ordering of activities and the distribution of work over people may shift. However, for simplicity we focus on control-flow in this section.

In a streaming context, we cannot pass over the data multiple times. Processing time is limited, and questions need to be answered in (near) real-time. *Memory is limited and cannot be increased over time or when the arrival rate of events increases*. Given these constraints, we can often only aim for approximate results. There is a trade-off between handling larger volumes and ensuring accuracy. Often only summaries or samples can be stored. Yet, some streaming algorithms produce results of similar quality as traditional algorithms at a fraction of the computational cost. Hence, they can also be applied to large collections of non-streaming event data.

To understand the challenges, let us consider the example stream in Fig. 12.18. Suppose we are able to store a limited number of events and simply forget the oldest events when new events arrive. Many approaches for streaming data use a fixed-length “window” consisting of the last n elements for some (typically large) n . Ap-

plying this in the process mining context may lead to misleading results. As shown in Fig. 12.18, we may loose the prefixes of traces. The initial activities of a case may be forgotten, e.g., the prefix $\langle a, c \rangle$ is missing for case 2. This could lead to the incorrect assumption that cases can start with activity b . There is also the problem that we do not know whether cases are finished. The final f activity for case 1 may still occur in the future (or not).

Suppose we need to create a process model for the cases handled during the last month, but can only store 1 million events. Assume this corresponds to roughly 5% of events occurring per month. Hence, for every 20 events that happen on average only one can be stored.

- If we use the fixed-length window approach shown in Fig. 12.18, we have two problems. First of all, the traces may be incomplete (missing prefixes). We are only storing the last 1.5 day (5% of a month). Hence, this is a significant problem if cases are running for days rather than minutes. Second, the process model will be biased towards the most recent period and may not be representative for the whole month.
- We can also randomly sample events. Every event has a 5% probability of being stored in the fixed-length window. Each time a new event is stored, the oldest event is removed. This approach will also provide very misleading results. Suppose a case has 20 events. The probability that the whole case will be stored for a selected event is less than $0.05^{20-1} = 1.9 \times 10^{-25} \approx 0$. Many cases will be reduced to a few random events. Hence, the discovered process model on such data will not make any sense.

To address the problem, we can store *summarized data* or use *smarter forms of sampling*.

Instead of randomly sampling events, we can also sample cases, i.e., use a window where we only keep events of a selected 5% of all cases. We cannot *randomly sample cases* because we would need to keep a list of selected cases and a list of non-selected cases. We do not know the set of cases beforehand and, even if we would, we could not store this information. However, we can use hashing as a kind of controlled randomization. By using a suitable hash function, we can select cases without actually storing them [114]. This will allow us to retain all events of a selected subset of cases and thus create a representative process model.

Next to careful sampling, we can also *store summarized data rather than individual events* (see Fig. 12.19). This is based on the observation that many algorithms basically count frequencies of activities and local patterns like the directly-follows relation. The α -miner, the heuristic miner, and the directly-follows based inductive miners (i.e., IMD, IMFD, and IMCD) basically use the following sources of information: the frequencies of observed activities and the frequencies of the elements in the directly-follows relation. The number of unique activities is typically limited compared to the number of cases and events. Hence, these frequencies can be stored compactly. The memory that can be used for this has a fixed upper bound as shown in Fig. 12.19. In [27], a particular approach based on this idea is presented. The approach uses three queues $Q = (Q_{act}, Q_{df}, Q_{last})$.

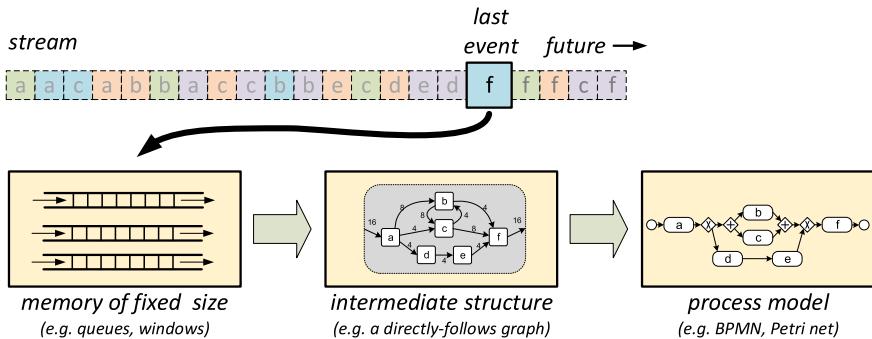


Fig. 12.19 To deal with streams, we can only use a predefined amount of memory. Instead of storing all events, we need to store summarized or sampled data. Often this limited memory is organized in such a way that some intermediate structure like a weighted directly-follows graph can be constructed from it (e.g., using queues $Q = (Q_{act}, Q_{df}, Q_{last})$). This intermediate structure can be used to create a process model at any point in time

- Queue Q_{act} is used to count the number of times an activity has occurred. It has elements of the form (a, n) where a is the activity and n the frequency.
- Queue Q_{df} is used to count the number of direct successions. It has elements of the form $((a, b), k)$ where k is the number of times that a was followed by b for the same case.
- Queue Q_{last} is used to keep track of the last activity executed for a particular case. It has elements of the form (c, a) where a is the last activity observed for c .

Queue Q_{last} is needed to construct Q_{df} . Each queue has maximum size and is updated each time a new event arrives. If a queue is full, then the addition of a new element implies the deletion of the least significant element. Queues can be sorted and operated using different heuristics. It is also possible to add *aging* to deal with concept drift. In this case the updates of the queues take into account an aging factor that gives more weight to recent events. Based on $Q = (Q_{act}, Q_{df}, Q_{last})$, we can apply existing process discovery techniques using information similar to the directly-follows graph (e.g., the α -miner, the heuristic miner, and the directly-follows based inductive miners). A detailed discussion of such approaches is beyond the scope of this book (see [27]).

Summarizing the above: We can adapt process mining to streams of event data by carefully sampling cases and by keeping only summarized data. We can use existing techniques for this [6, 114].

12.6 Beyond the Hype

The techniques described in this chapter enable the application of process mining when event logs are huge. As demonstrated, process mining techniques can exploit the MapReduce programming model, modern database technology (e.g., in-memory

databases and columnar databases), and large-scale distributed file systems (e.g., Hadoop). In this chapter, we focused on techniques to decompose event logs, but also discussed related notions such as process cubes and streaming process mining.

These topics fit perfectly with the current attention for “Big Data” in industry and society. Numerous books appeared in recent years. These discuss “Big Data” from different angles: distributed algorithms [114], analytics [17], societal impact [99, 100], and management [54]. However, the real challenges are often related to data acquisition, data preparation, and the interpretation of results. The majority of real-life applications of process mining would benefit more from a “data science mindset” rather than new Hadoop-like infrastructures.

Moreover, data sets that are “Big” today may be “small” tomorrow. The book “Concise Survey of Computer Methods” [107] by Peter Naur (1928–2016), published in 1974, used already the term “data science” and contains several chapters describing techniques for processing and managing “large datasets”. These were written at a time where hard disks had a capacity of just a few megabytes. Although the dimensions have changed dramatically, many of the core principles remained invariant. *To change data into real value, one needs to ask the right questions, use the right analysis techniques, and be able to interpret the results.* It does not suffice to just have a “Big Data” infrastructure.

Chapter 13

Analyzing “Lasagna Processes”

Lasagna processes are relatively structured and the cases flowing through such processes are handled in a controlled manner. Therefore, it is possible to apply all of the process mining techniques presented in the preceding chapters. This chapter characterizes Lasagna processes and discusses typical use cases for process mining. Moreover, the different stages of a process mining project for improving a Lasagna process are described. The resulting life-cycle model guides users of process mining tools like ProM. Moreover, different application scenarios are discussed.

13.1 Characterization of “Lasagna Processes”

Unlike Spaghetti processes, Lasagna processes have a clear structure and most cases are handled in a prearranged manner. There are relatively few exceptions and stakeholders have a reasonable understanding of the flow of work. It is impossible to define a formal requirement characterizing Lasagna processes. As a rule of thumb we use the following informal criterion: *a process is a Lasagna process if with limited efforts it is possible to create an agreed-upon process model that has a fitness of at least 0.8*, i.e., more than 80% of the events happen as planned and stakeholders confirm the validity of the model. This implies (assuming that a suitable event log can be extracted) that all of the process mining techniques presented in this book can be applied.

The spectrum ranging from Lasagna processes to Spaghetti processes is a *continuum*. Sometimes the terms “structured”, “semi-structured”, and “unstructured” are used to refer to the same continuum. In a *structured process* (i.e., Lasagna process) all activities are repeatable and have a well defined input and output. In highly structured processes most activities can, in principle, be automated. In *semistructured processes* the information requirements of activities are known and it is possible to sketch the procedures followed. However, some activities require human judgment and people can deviate depending on taste or the characteristics of the case being handled. In *unstructured processes* (i.e., Spaghetti process) it is difficult to define

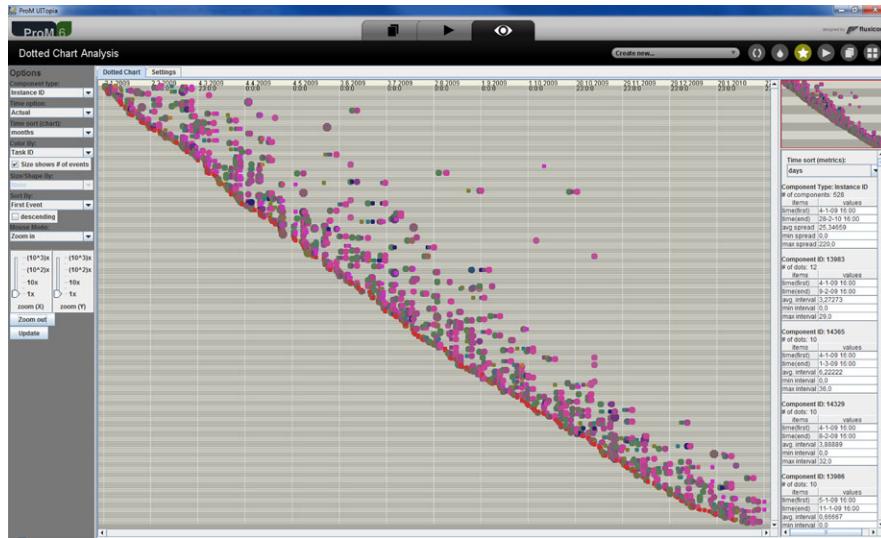


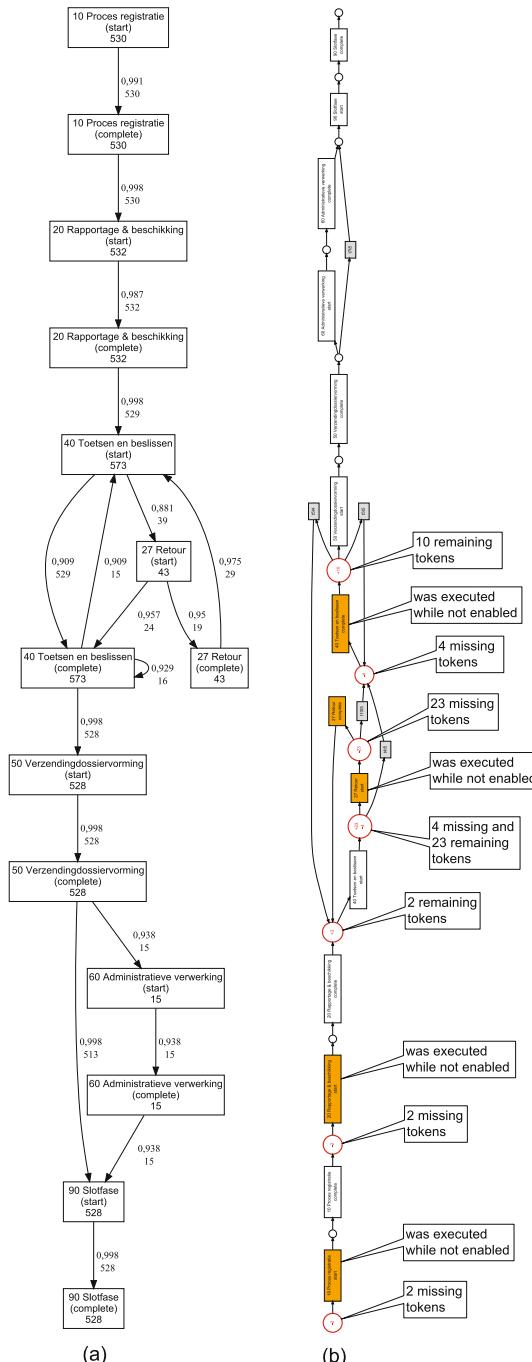
Fig. 13.1 Screenshot of ProM 6 showing a dotted chart for a WMO process of a Dutch municipality. Each line corresponds to one of the 528 requests that were handled in the period from 4-1-2009 until 28-2-2010. In total there are 5498 events represented as dots. The mean time needed to handle a case is approximately 25 days

pre- and post-conditions for activities. These processes are driven by experience, intuition, trial-and-error, rules-of-thumb, and vague qualitative information.

Let us consider an example of a Lasagna process. Figure 13.1 shows a dotted chart for one of the so-called WMO processes of a Dutch municipality. WMO (Wet Maatschappelijke Ondersteuning) refers to the social support act that came into force in The Netherlands on January 1st, 2007. The aim of this act is to assist people with disabilities and impairments. Under the act, local authorities are required to give support to those who need it, e.g., household help, providing wheelchairs and scootmobiles, and adaptations to homes. There are different processes for the different kinds of help. The dotted chart in Fig. 13.1 is based on the process for handling requests for household help. In a period of about one year, 528 requests for household WMO support were received. These 528 requests generated 5498 events each represented as a colored dot in Fig. 13.1. The color of the dot refers to the activity executed for the request, e.g., a red dot refers to activity “10 Process registratie” (register request) and a blue dot refers to activity “40 toetsen en beslissen” (evaluate and decide). The diagonal line of initial events shows that there is a steady flow of new requests. The dots also show that the time to completely handle requests is typically short (about one month).

Although no process model is shown in Fig. 13.1, the dotted chart already suggests that the process is a Lasagna process (regular arrival pattern, most cases are handled within one month, and clearly noticeable recurring patterns). Figure 13.2 demonstrates that this is indeed the case. The process model discovered by the heuristic miner shows that the process is highly structured and rather sequential.

Fig. 13.2 The C-net discovered using the heuristic miner (a) and the corresponding Petri net with missing and remaining tokens after replay (b). The numbers generated by the heuristic miner show the flow of tokens. The C-net was translated into an equivalent Petri net with silent transitions. The fitness was analyzed using ProM’s conformance checker (cf. Sect. 8.2). The fitness of the discovered process is 0.99521667. Of the 528 cases, 496 cases fit perfectly whereas for 32 cases there are missing or remaining tokens. The missing and remaining tokens show where the model and log deviate. For example, for two cases the activity “40 toetsen en beslissen” (evaluate and decide) was not started although it should have. Activity “20 Rapportage & beschikking” (report and intermediate decision) was started twice while this was not possible according to the model



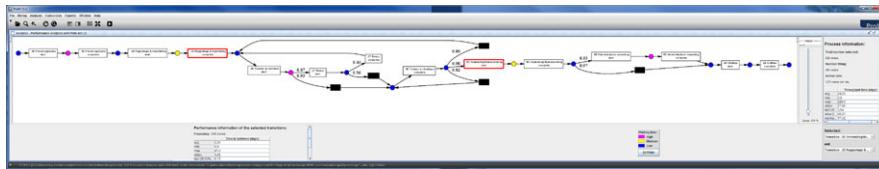


Fig. 13.3 Screenshot of ProM 5.2 while analyzing the bottlenecks in the process. The mean flow time of fitting cases is 24.66 days. Most time is spent on the activities “10 Process registratie”, “40 Toetsen en beslissen”, and “60 Administratieve verwerking”. The average time in-between the completion of activity “10 Rapportage & beschikking” and “50 Verzending/dossiervorming” is 2.24 days

The figure does not show the logic of splits and joins, e.g., one cannot see the difference between AND/OR/XOR-splits/joins.¹ ProM’s heuristic miner does not allow for the visualization of bindings used in Sect. 7.2. However, the logic of splits and joins is also discovered and can be shown if desired. When converting a C-net into a Petri net, EPC model, or BPMN model this information is taken into account. The discovered C-net in Fig. 13.2(a) is annotated with frequencies. The frequency of a node indicates how often the corresponding activity appeared in the event log. For instance, activity “20 Rapportage & beschikking” (report and intermediate decision) occurred 532 times. Arcs have a frequency indicating how often a token was passed along the arc when replaying the log. Figure 13.2(b) shows a WF-net obtained by using the corresponding conversion plug-in in ProM. The conformance checker of ProM shows that the fitness of model and log is 0.99521667. This shows that there are hardly any missing or remaining tokens when replaying all 528 cases. Figure 13.2(b) also shows some of the detailed diagnostics. The discovered process model and the high fitness value show that the WMO process is definitely a Lasagna process. This implies that, in principle, *all process mining techniques described in this book are applicable to this process* (assuming sufficient event data). Figure 13.3 shows one of many process mining techniques that can be applied. As explained in Sect. 9.4, delays can be analyzed by replaying the event log while taking timestamps into account. Figure 13.3 illustrates that it is possible to discover bottlenecks for a Lasagna process like the WMO process. Note that the plug-in used in Fig. 13.3 exploits the coupling between the event log and the discovered model (cf. Fig. 13.2).

In Sect. 13.4, we provide more examples of Lasagna processes. However, first we discuss typical use cases for process mining and present a life-cycle model for process mining projects.

¹In the remainder, we will never show the set of input and output bindings for C-nets discovered by the heuristic miner. The heuristic miner can visualize the logic of splits and joins, but this typically impairs the readability of the diagram.

13.2 Use Cases

The goal of process mining is to *improve* operational processes. In order to judge whether process mining efforts are successful, we need to define *Key Performance Indicators* (KPIs). In Sect. 3.3.2, we identified three classes of KPIs: KPIs related to *time* (e.g., lead time, service time, waiting time, and synchronization time), KPIs related to *costs*, and KPIs related to *quality*. Note that quality may refer to compliance, customer satisfaction, number of defects, etc. To evaluate suggested improvements, the effectiveness and efficiency of the *as-is* and *to-be* processes need to be quantified in terms of KPIs.

For Lasagna processes, process mining can result in one or more of the following *improvement actions*:

- *Redesign*. Insights obtained using process mining can trigger changes to the process, e.g., sequential activities no longer need to be executed in a fixed order, checks may be skipped for easy cases, decisions can be delegated if more than 50 cases are queueing, etc. Fraud detected using process mining may result in additional compliance regulations, e.g., introducing the 4-eyes principle for critical activities.
- *Adjust*. Similarly, process mining can result in (temporary) adjustments. For example, insights obtained using process mining can be used to temporarily allocate more resources to the process and to lower the threshold for delegation.
- *Intervene*. Process mining may also reveal problems related to particular cases or resources. This may trigger interventions such as aborting a case that has been queuing for more than 3 months or disciplinary measures for a worker that repeatedly violated compliance regulations.
- *Support*. Process mining can be used for operational support, e.g., based on historic information a process mining tool can predict the remaining flow time or recommend the action with the lowest expected costs.

Figure 2.4 in Chap. 2 illustrates the difference between a *redesign* (a permanent change requiring alterations to software or model) and an *adjustment* (a temporary change realized without modifying the underlying software or model).

As shown in Fig. 13.4, *use cases for process mining refer to a combination of KPIs and improvement actions*. Given a Lasagna process, some typical use cases for process mining are:

- Identification of bottlenecks to trigger a process redesign that reduces the overall flow time with 30%.
- Identification of compliance problems using conformance checking. Some of the compliance problems result in ad-hoc interventions whereas others lead to adjustments of the parameters used for work distribution.
- Harmonization of two processes after a merger based on a comparison of the actual processes. The goal of such a harmonization is to reduce costs.
- Predicting the remaining flow time of delayed cases to improve customer service.
- Providing recommendations for resource allocation aiming at a more balanced utilization of workers.

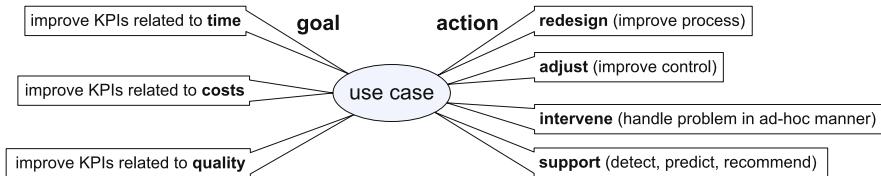


Fig. 13.4 Use cases for process mining combine goals (expressed in KPIs) and improvement actions, e.g., process mining can be used to shorten the flow time by providing insights that lead to a process redesign

- Identification of exceptional cases that generate too much additional work. By learning the profile of such cases, they can be handled separately to reduce the overall flow time.
- Visualization of the 10 most complicated or time consuming cases to identify potential risks.

These use cases illustrate the potential of process mining. It is easy to imagine the application of these use cases to the WMO process described earlier. For instance, results such as shown in Fig. 13.3 can be used to discover bottlenecks and to generate ideas for flow time reduction. The results of conformance analysis as depicted in Fig. 13.2(b) can be used to identify compliance problems, e.g., for the 32 cases having missing or remaining tokens one could analyze the social network of the people involved.

13.3 Approach

In Chap. 10, we described ten process mining related activities using the framework shown in Fig. 13.5. These ten activities are grouped into three categories: cartography (activities *discover*, *enhance*, and *diagnose*), auditing (activities *detect*, *check*, *compare*, and *promote*), and navigation (activities *explore*, *predict*, and *recommend*). Although the framework helps to understand the relations between the various process mining activities, it does not guide the user in conducting a process mining project. Therefore, this section introduces the *L^{*} life-cycle model for mining Lasagna processes*.

Several reference models describing the life-cycle of a typical data mining/BI project have been proposed by academics and consortia of vendors and users. For example, the *CRISP-DM* (CRoss-Industry Standard Process for Data Mining) methodology identifies a life-cycle consisting of six phases: (a) business understanding, (b) data understanding, (c) data preparation, (d) modeling, (e) evaluation, and (f) deployment [29]. CRISP-DM was developed in the late nineties by a consortium driven by SPSS. Around the same period SAS proposed the *SEMMA* methodology consisting of five phases: (a) sample, (b) explore, (c) modify, (d) model, and (e) assess. Both methodologies are very high-level and provide little support. Moreover, existing methodologies are not tailored towards process mining projects. Therefore, we propose the *L^{*}* life-cycle model shown in Fig. 13.6. This five-stage model

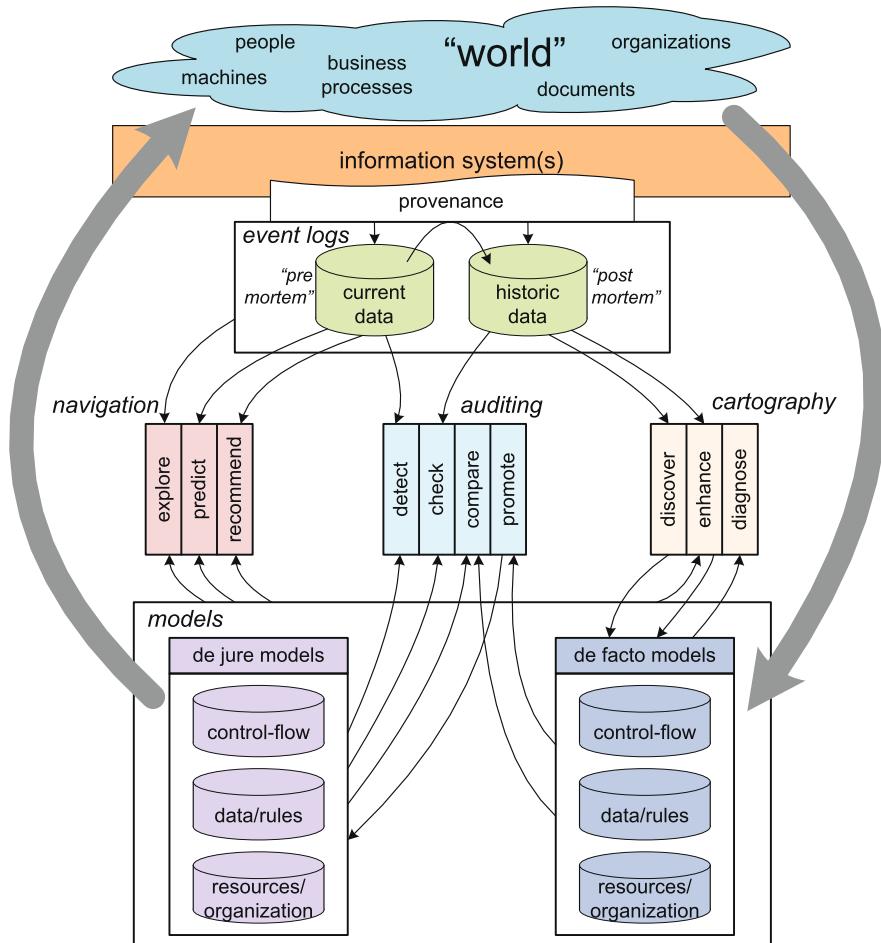


Fig. 13.5 The process mining framework introduced in Chap. 10. The framework identifies ten process mining activities (discover, check, enhance, etc.)

describes the life-cycle of a typical process mining project aiming to improve a Lasagna process.

In the remainder, we discuss each of the five stages. As shown in Fig. 13.6, the L^* life-cycle model refers to the ten process mining related activities (explore, discover, check, etc.) and the four improvement actions (redesign, adjust, intervene, and support) mentioned earlier.

13.3.1 Stage 0: Plan and Justify

Any process mining project starts with a planning and a justification of the planned activities. Before spending efforts on process mining activities, one should antic-

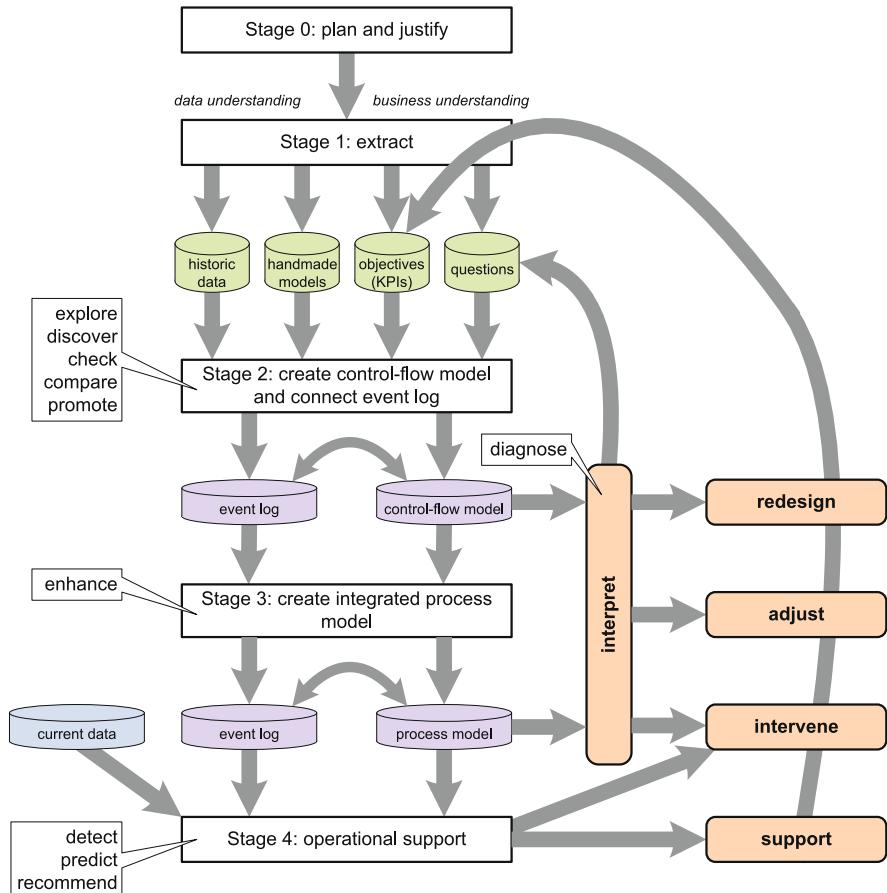


Fig. 13.6 The L^* life-cycle model describing a process mining project consisting of five stages: *plan and justify* (Stage 0), *extract* (Stage 1), *create control-flow model and connect event log* (Stage 2), *create integrated process model* (Stage 3), and *operational support* (Stage 4)

ipate benefits that may result from the project. There are basically three types of process mining projects:

- A *data-driven* (also referred to as “curiosity driven”) process mining project is powered by the availability of event data. There is no concrete question or goal, however, some of the stakeholders expect that valuable insights will emerge by analyzing event data. Such a project has an explorative character.
- A *question-driven* process mining project aims to answer specific questions, e.g., “Why do cases handled by team X take longer than cases handled by team Y?” or “Why are there more deviations in weekends?”.
- A *goal-driven* process mining project aspires to improve a process with respect to particular KPIs, e.g., cost reduction or improved response times.

For an organization without much process mining experience it is best to start with a question-driven project. Concrete questions help to scope the project and guide data extraction efforts.

Like any project, a process mining project needs to be planned carefully. For instance, activities need to be scheduled before starting the project, resources need to be allocated, milestones need to be defined, and progress needs to be monitored continuously.

13.3.2 Stage 1: Extract

After initiating the project, event data, models, objectives, and questions need to be extracted from systems, domain experts, and management.

In Chap. 5, we elaborated on data extraction. For example, Fig. 5.1 describes the process of getting from raw data to suitable event logs. Recall that event logs have two main requirements: (a) events need to be ordered in time and (b) events need to be correlated (i.e., each event needs to refer to a particular case).

As Fig. 13.6 shows, it is possible that there are already handmade (process) models. These models may be of low quality and have little to do with reality. Nevertheless, it is good to collect all models present and exploit existing knowledge as much as possible. For example, existing models can help in scoping the process and judging the completeness of event logs.

In a goal-driven process mining project, the objectives are also formulated in Stage 1 of the L^* life-cycle. These objectives are expressed in terms of KPIs. In a question-driven process mining project, questions need to be generated in Stage 1. Both questions and objectives are gathered through interviews with stakeholders (e.g., domain experts, end users, customers, and management).

13.3.3 Stage 2: Create Control-Flow Model and Connect Event Log

Control-flow forms the backbone of any process model. Therefore, Stage 2 of the L^* life-cycle aims to determine the de facto control-flow model of the process that is analyzed. The process model may be discovered using the process discovery techniques presented in Part III of this book (activity *discover* in Fig. 13.6). However, if there is a good process model present, it may be verified using conformance checking (activity *check*) or judged against the discovered model (activity *compare*). It is even possible to merge the handmade model and the discovered model (activity *promote*). After completing Stage 2 there is a control-flow model tightly connected to the event log, i.e., events in the event log refer to activities in the model. As discussed in Sect. 8.5.3, this connection is crucial for subsequent steps. If the fitness

of the model and log is low (say below 0.8), then it is difficult to move to Stage 3. However, by definition, this should not be a problem for a Lasagna process.

The output of Stage 2 may be used to answer questions, take actions, or to move to Stage 3. As Fig. 13.6 shows, the output (control-flow model connected to an event log) needs to be interpreted before it can be used to answer questions or trigger a redesign, an adjustment, or an intervention.

13.3.4 Stage 3: Create Integrated Process Model

In Stage 3, the model is enhanced by adding additional perspectives to the control-flow model (e.g., the organizational perspective, the case perspective, and the time perspective). Chapter 9 shows how these perspectives can be discovered and integrated, e.g., Fig. 9.16 describes the process of merging the different perspectives. The result is an integrated process model that can be used for various purposes. The model can be inspected directly to better understand the as-is process or to identify bottlenecks. Moreover, a complete process model can also be simulated as discussed in Sect. 9.6.

The output of Stage 3 can also be used to answer selected questions and take appropriate actions (redesign, adjust, or intervene). Moreover, the integrated process model is also input for Stage 4.

13.3.5 Stage 4: Operational Support

Stage 4 of the L^* life-cycle is concerned with the three operational support activities described in Chap. 10: *detect*, *predict*, and *recommend*. For instance, using short-term simulation (Sect. 9.6) or annotated transition systems (Sect. 10.4) it is possible to predict the remaining flow time for running cases. As shown in Fig. 13.6, Stage 4 requires current data (“pre mortem” data on running cases) as input. Moreover, the output does not need to be interpreted by the process mining analyst and can be directly offered to end users. For example, a deviation may result in an automatically generated e-mail sent to the responsible manager. Recommendations and predictions are presented to the persons working on the corresponding cases.

Note that operational support is the *most ambitious* form of process mining. This is only possible for Lasagna processes. Moreover, there needs to be an advanced IT infrastructure that provides high-quality event logs and allows for the embedding of an operational support system as described in Chap. 10.

The PM^2 process mining methodology presented in [175] can be viewed as a refinement of the L^* life-cycle. Using a case study conducted within IBM, the PM^2 methodology is explained. Moreover, selected ProM plug-ins are related to the different phases in [175].

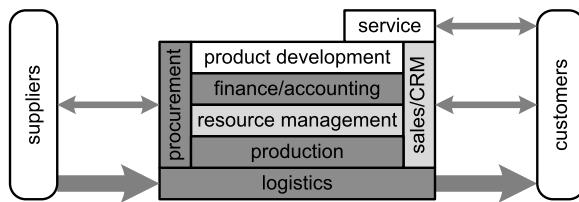


Fig. 13.7 Overview of the different functional areas in a typical organization. Lasagna processes are typically encountered in production, finance/accounting, procurement, logistics, resource management, and sales/CRM. Spaghetti processes are typically encountered in product development, service, resource management, and sales/CRM

13.4 Applications

In the last decade, we have applied process mining in over 150 organizations. Examples are municipalities (e.g., Alkmaar, Heusden, and Harderwijk), government agencies (e.g., Rijkswaterstaat, Centraal Justitieel Incasso Bureau, and Justice department), insurance related agencies (e.g., UWV), banks (e.g., ING Bank), hospitals (e.g., AMC hospital and Catharina hospital), multinationals (e.g., DSM and Deloitte), high-tech system manufacturers and their customers (e.g., Philips Health-care, ASML, Ricoh, and Thales), and media companies (e.g., Winkwaves). This illustrates the broad spectrum of situations in which process mining can be applied. In remainder of this section, we identify process mining opportunities in different functional areas and in different sectors and industries. Moreover, we briefly discuss two case studies involving Lasagna processes.

13.4.1 Process Mining Opportunities per Functional Area

Figure 13.7 shows the main *functional areas* that can be found in most organizations:

- *Product development* is concerned with all the preparations and engineering work needed to start producing a particular product. Products do not need to be physical objects (e.g., a car or copier); the product may also be a piece of information or a service (e.g., a new kind of insurance). Product development processes are typically Spaghetti-like because they have a lower frequency and depend on problem solving, expertise, and creativity rather than repetition, routine, and efficiency.
- *Production* is the functional area where the products are actually produced. Processes may range from classical manufacturing (assembling a car) to information creation (opening a bank account). Most production processes are Lasagna processes because they need to be reproducible and efficient.
- *Procurement* entails all activities to get the materials needed for production. Note that the input for the production process may also be information from other parties. The input materials need to be purchased, stocks need to be monitored,

deliveries need to be checked, etc. Processes in this functional area are typically Lasagna processes.

- The functional area *Sales/CRM* is concerned with all activities related to “lead-to-order” and “order-to-cash”. Besides the actual sales function, most organizations need to market their products and manage long-term relationships with their customers (CRM). Both Lasagna processes and Spaghetti processes can be found in this functional area. The handling of sales activities can be very structured whereas marketing-related activities may be rather unstructured.
- *Logistics* is concerned with the movements of products and materials, e.g., shipping the product to the customer and managing the storage space. Most processes in logistics are Lasagna processes.
- The functional area *Finance/accounting* deals with all financial aspects of an organization, e.g., billing customers, checking invoices, financial reporting, and auditing. Processes in this functional area are also typically Lasagna processes.
- *Resource management* is the functional area that makes sure there are sufficient resources to perform all other functions. HRM (Human Resource Management) is concerned with human resources and similar functions exist for machines, buildings, etc. Both Lasagna processes and Spaghetti processes can be found in this functional area, e.g., the handling of job applications may be very structured whereas the handling of a problematic employee may be rather ad-hoc.
- The functional area *Service* deals with all activities after the product has been shipped and paid for, e.g., activities related to product support, maintenance, repairing defective products, and help-desk operations. Service related processes are typically Spaghetti-like. Customers will use products in many different ways and repair processes are rather unpredictable for most products, e.g., no faults are found in the product returned by the customer or the wrong component is replaced and the product still malfunctions intermittently.

The characterization of the different functional areas in terms of Lasagna processes and Spaghetti processes is only intended as an indication. Both types of processes can be found in all of the functional areas. However, as shown in Fig. 13.7, it is possible to pinpoint typical functional areas for both types. For example, in most organizations product development processes are rather unstructured compared to production processes. This implies that most of the techniques presented in this book can be applied to production processes. However, for product development processes it is unlikely that all stages of the L^* life-cycle model (Fig. 13.6) can be executed. (Stages 3 and 4 are typically not possible for Spaghetti-like processes.)

13.4.2 Process Mining Opportunities per Sector

After contemplating on the presence of Lasagna and Spaghetti processes in the functional areas in one organization (Fig. 13.7), we now look at different sectors and industries.

The *primary sector* of the economy is concerned with transforming natural resources into primary products (e.g., agriculture, agribusiness, fishing, forestry and all mining and quarrying industries). Information technology tends to play a minor role in these industries. Hence, the application potential of process mining is limited. Of course there are exceptions. Consider for instance the tracking and tracing of food. In some countries meat and dairy products need to be tracked from source to sink. For example, meat products in supermarkets need to be linked to particular animals and farms. This requires the recording of events starting in the primary sector.

The *secondary sector* of the economy refers to the manufacturing of tangible products and includes the automotive industry, chemical industry, aerospace manufacturing, consumer electronics, etc. Organizations in the secondary sector typically have an organizational structure covering all functional areas depicted in Fig. 13.7. Hence, both Lasagna processes and Spaghetti processes can be encountered. An interesting observation across the different industries is that most manufacturers have become interested in monitoring their products after they have been sold. For example, Philips Healthcare is monitoring their medical equipment while being deployed in the field, e.g., their X-ray machines are connected to the Internet and the resulting logs are analyzed using ProM. The event logs of these X-ray machines provide vital information for marketing (What kind of features do customer use?), maintenance (When to service the machine?), development (Why do machines fail?), and testing (How to test machines under realistic circumstances?). In the future, more and more (consumer) products will be monitored remotely thus providing valuable information for the manufacturer.

The *tertiary sector* of the economy consists of all organizations that produce “intangible goods” such as services, regulations, and information. The term “services” should be interpreted in the broadest sense including transportation, insurance, wholesaling, retailing, entertainment, etc. Note that goods may be transformed in the process of providing the service (cf. preparing food in a restaurant). However, the focus is on serving the customer rather than transforming physical goods. In many industries in the tertiary sector, information plays a dominant role and many events are being recorded. This is the sector where the digital universe and the physical universe are aligned most. For example, an electronic bookstore can only sell a book if the information system indicates that the book is present. The bookstore would not be able to sell a particular book if the information system would indicate that it is out-of-stock; even if the book would be physically present in the warehouse.

Process mining can be used to improve a variety of Lasagna and Spaghetti processes encountered in the tertiary sector. Below we sketch some of the most interesting industries.

- The *healthcare* industry includes hospitals and other care organizations. Most events are being recorded (blood tests, MRI scans, appointments, etc.) and correlation is easy because each event refers to a particular patient. The closer processes get to the medical profession, the less structured they become. For instance, most diagnosis and treatment processes tend to be rather Spaghetti-like (see Fig. 14.1). Medical guidelines typically have little to do with the actual

processes. On the one hand, this suggests that these processes can be improved by structuring them. On the other hand, the variability of medical processes is caused by the different characteristics of patients, their problems, and unanticipated complications. Patients are saved by doctors deviating from standard procedures. However, some deviations also cost lives. Clearly, hospitals need to get a better understanding of care processes to be able to improve them. Process mining can help as event data is readily available [95].

- *Governments* range from small municipalities to large organizations operating at the national level, e.g., institutions managing processes related to unemployment, customs, taxes, and traffic offences. Both local and national government agencies can be seen as “administrative factories” as they execute regulations and the “products” are mainly informational or financial. Processes in larger government agencies are characterized by a high degree of automation. Consider, for example, tax departments that need to deal with millions of tax declarations. Processes in smaller government agencies (e.g., small municipalities) are typically not automated and managed by office workers rather than BPM systems. However, due to the legal requirements, all main events are recorded in a systematic manner. Consider, for example, the WMO process shown in Fig. 13.2; any municipality in The Netherlands is obliged to record the formal steps in such processes. Typical use cases for process mining in governments (local or non-local) are flow time reduction (e.g., shorten the time to get a building permit), improved efficiency, and compliance. Given the role of governments in society, compliance is of the utmost importance.
- *Banking* and *insurance* are two industries where BPM technology has been most effective. Processes are often automated and all events are recorded in a systematic and secure manner. Examples are the processing of loans, claims management, handling insurance applications, credit card payments, and mortgage payments. Most processes in banking and insurance are Lasagna processes, i.e., highly structured. Hence, all of the techniques presented in this book can be applied. Process discovery is less relevant for these organizations as most processes are known and documented. Typical uses cases in these industries involve conformance checking, performance analysis, and operational support.
- Organizations involved in *education* (e.g., high-schools and universities) are recording more and more information related to the study behavior of individuals. For instance, at TU/e we are applying process mining to analyze study behavior using a database containing detailed information about exam results of all students that ever studied computer science. Moreover, this database also contains information about high-school exam grades, etc. Some of these educational processes are structured, others are very unstructured. For example, it is very difficult to predict the remaining study time of students at a university because the curriculum often changes and students tend to have very different study patterns. Nevertheless, valuable insights can be obtained. By visualizing that few students follow the courses in the order intended, one can show that the design of a curriculum should not only focus on the “ideal student” (that passes all courses the first time), but also anticipate problems encountered by other students.

- The products manufactured by organizations in the secondary sector are distributed through various *retail* organizations. Here it is interesting to see that more and more information about products and customers is being recorded. Customers are tracked using loyalty cards or through online profiles. Products are tagged and the shop has real-time information about the number of items still available. A product that has an RFID tag has a unique identifier, i.e., two identical products can still be distinguished. This allows for the correlation of events and thus facilitates process mining.
- The *transportation* industry is also recording more and more information about the movement of people and products. Through tracking and tracing functionality the whereabouts of a particular parcel can be monitored by both sender and receiver. Although controversial, smartcards providing access to buildings and transportation systems can be used to monitor the movement of people. For example, the Dutch “ov-chipkaart” can be used to travel by train, subway, and bus. The traveler pays based on the distance between the entry point and exit point. The recorded information can be used to analyze traveling behavior. The booking of a flight via the Internet also generates lots of event data. In fact, the booking process involves only electronic activities. Note that the traveler interacts with one organization that contacts all kinds of other organizations in the background (airlines, insurance companies, car rental agencies, etc.). All of these events are being recorded, thus enabling process mining. The whole spectrum ranging from Lasagna processes to Spaghetti processes can be found in this industry.
- New technologies such as *cloud computing* and *Software-as-a-Service* (SaaS) have created a new industry that offers computing as a utility (like water and electricity). Google Apps. Salesforce.com, and Amazon EC2/S3 are examples of companies providing such utilities. The idea is not new: already in 1961 John McCarthy stated “If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility. The computer utility could become the basis of a new and important industry.” A well-known example of a SaaS provider that is using a cloud infrastructure is SalesForce.com. This company allows organizations to outsource the IT support of standard activities related to sales and CRM without worrying about scalability and maintenance. Users pay for using the software rather than owning it. Another example is the conference management system EasyChair that is currently probably the most commonly used system to host conferences and to manage the reviewing of scientific papers. To organize a conference, there is no need to install any software as everything is hosted and managed centrally. Organizations such as SalesForce.com and EasyChair have access to valuable event data. These data can be used to improve their software and to give advice to individual organizations. One of the challenges SaaS providers are facing is the need to deal with variability across organizations. Process mining can help analyzing differences between organizations using *cross-organizational process mining*, i.e., using process mining to compare similar processes within the same or in different organizations.

- The *capital goods* industry is also transforming from the situation in which customers purchase expensive machines to the situation in which customers only pay for the actual use of the machine. Note that this can be seen as a variant of the SaaS paradigm. The manufacturer of the machine remains being the owner and customers pay depending on usage and uptime of the machine. Clearly, such pricing models require the remote monitoring of capital goods. For instance, service provider and consumer need to agree on the actual use (e.g., hours of use or number of production cycles). Moreover, there may be Service Level Agreements (SLAs) specifying a fine if the machine is down for an extended period. Event data can be used as a basis for billing and checking SLAs. Moreover, the manufacturer gets insights into the way that machines are used, when they malfunction, and when they require maintenance.

These examples show that there are opportunities for process mining in all three economic sectors.

13.4.3 Two Lasagna Processes

To conclude this chapter, we briefly discuss two case studies analyzing Lasagna processes.

13.4.3.1 RWS Process

The Dutch national public works department, called “Rijkswaterstaat” (RWS), has 12 provincial offices. We analyzed the handling of invoices in one of these offices [160]. The office employs about 1,000 civil servants and is primarily responsible for the construction and maintenance of the road and water infrastructure in its province. To perform its functions, the RWS office subcontracts various parties such as road construction companies, cleaning companies, and environmental bureaus. Also, it purchases services and products to support its construction, maintenance, and administrative activities. The reason to employ process mining within RWS was twofold. First of all, RWS was involved in our longitudinal study into the effectiveness of WFM systems [116]. In the context of this study, RWS was interested to see the effects of WFM technology on flow times, response times, service levels, utilization, etc. Second, RWS was interested in better meeting deadlines with respect to the payment of invoices. Payment should take place within 31 days from the moment the invoice is received. After this period, the creditor is entitled (according to Dutch law) to receive interest over the outstanding sum. RWS would like to pay at least 90% of its invoices within 31 days. However, analysis of the event logs of RWS showed that initially only 70% of payments were paid in time.

Starting point for the analysis described in [160] was an event log containing information about 14,279 cases (i.e., invoices) generating 147,579 events. Figure 13.8 shows a C-net generated by the heuristic miner. This model shows that the RWS

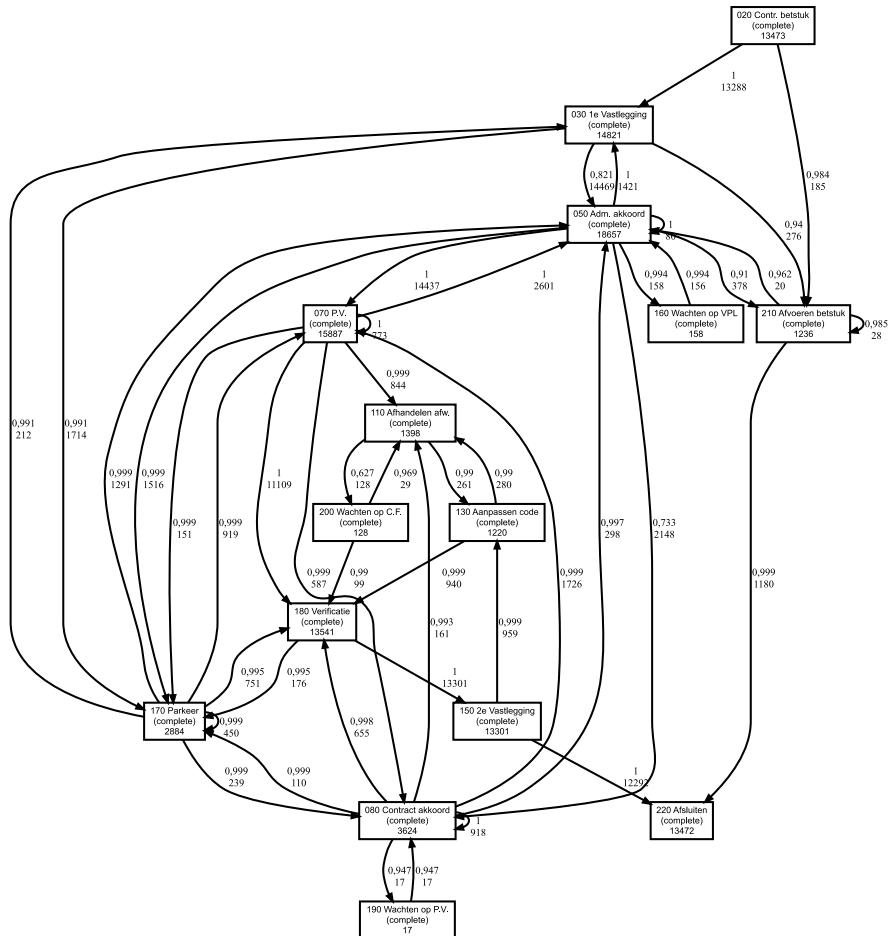


Fig. 13.8 Process model obtained using heuristic mining. The C-net describes the handling of invoices within one of the twelve provincial offices of RWS

process is fairly structured, but not as structured as the WMO process depicted in Fig. 13.2(a). After some efforts (filtering the log and tuning the parameters of the mining algorithm), it is possible to create a model with a fitness of more than 0.9. The log can be replayed on this model to highlight bottlenecks. Such analysis shows that several activities had to be re-done (as can be seen by the loops of length one or two in Fig. 13.8), i.e., work was sent “back-and-forth” between different activities and people thus causing delays.

The event log contains information about 271 resources, i.e., civil servants involved in the handling invoices. Figure 13.9 shows the social network based on the frequency of handovers (cf. Sect. 9.3.1). Figure 13.10 shows the same social network, but now only for the 13 resources that executed most activities. RWS could

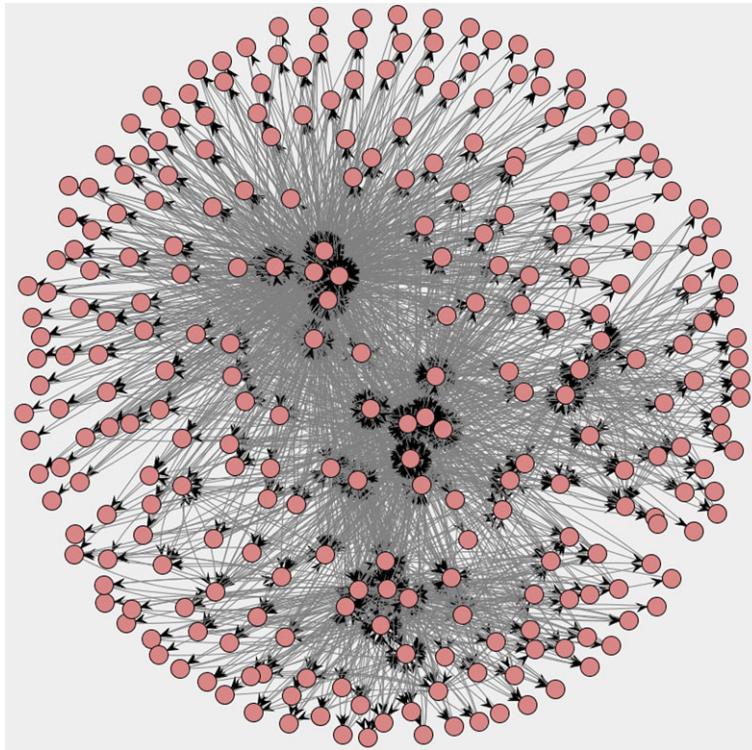


Fig. 13.9 Social network constructed based on handovers of work. Each of the 271 nodes corresponds to a civil servant. Two civil servants are connected if one executed an activity causally following an activity executed by the other civil servant

use these social networks to better understand how work is flowing through the organization. This analysis showed that some project leaders considered invoice approval to be of low priority, not realizing that because of their slow reaction time many invoices took more than 31 days. They were not aware of the impact of their actions and agreed to give the invoice approval a higher priority thus speeding up the process. See [160] for more information.

13.4.3.2 WOZ Process

In Sect. 13.1, we showed some analysis results for a WMO process of a municipality. To date, we have applied process mining in about a dozen municipalities. Moreover, we just started a new project (CoSeLoG) involving nine municipalities interested in cross-organizational process mining, i.e., analyzing differences between similar processes in different municipalities [35].

Processes in municipalities are typically Lasagna processes. To illustrate this we present another example. Figure 13.11 shows a so-called “WOZ process” discov-

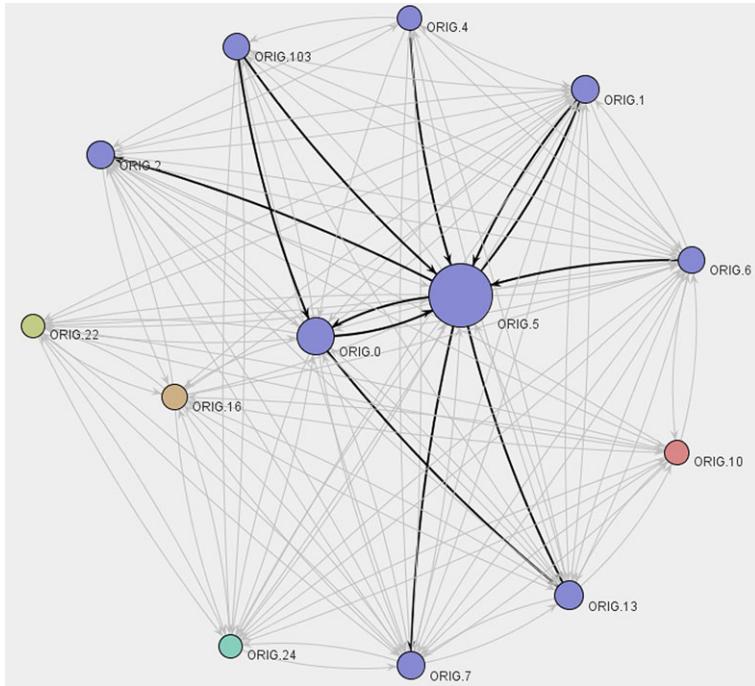


Fig. 13.10 Social network consisting of civil servants that executed more than 2000 activities in a 9 month period. The darker arcs indicate the strongest relationships in the social network. Nodes having the same color belong to the same clique. Names of resources have been anonymized for privacy reasons

ered for another municipality (i.e., different from the one for which we analyzed the WMO process). We applied the heuristic miner on an event log containing information about 745 objections against the so-called WOZ (“Waardering Onroerende Zaken”) valuation. Dutch municipalities need to estimate the value of houses and apartments. The WOZ value is used as a basis for determining the real-estate property tax. The higher the WOZ value, the more tax the owner needs to pay. Therefore, Dutch municipalities need to handle many objections (i.e., appeals) of citizens that assert that the WOZ value is too high. For this municipality we analyzed four processes related to objections and building permits. Here, we restrict ourselves to the WOZ process shown in Fig. 13.11.

The discovered WF-net has a good fitness: 628 of the 745 cases can be replayed without encountering any problems. The fitness of the model and log is 0.98876214 indicating that almost all recorded events are explained by the model. Hence, the WOZ process is clearly a Lasagna process. Nevertheless, it is interesting for the municipality to see the deviations highlighted in the model. Figure 13.12 shows a fragment of the diagnostics provided by the conformance checker (cf. Sect. 8.2).

The average flow time is approx. 178 days. Figure 13.13 shows some more performance-related diagnostics computed while replaying the event log contain-

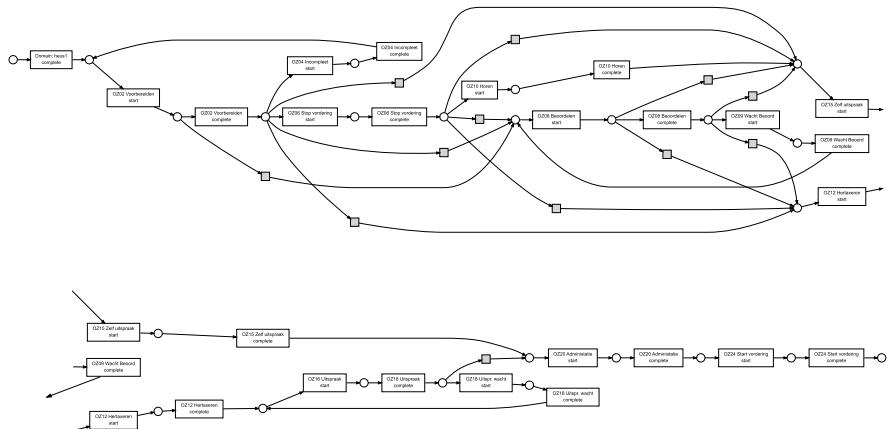


Fig. 13.11 WF-net discovered based on an event log of another municipality. The log contains events related to 745 objections against the so-called WOZ valuation. These 745 objections generated 9583 events. There are 13 activities. For 12 of these activities both start and complete events are recorded. Hence, the WF-net has 25 transitions

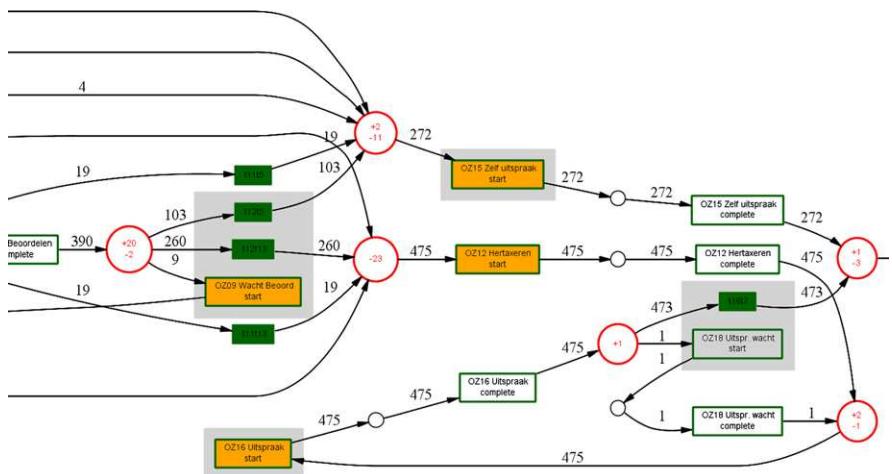


Fig. 13.12 Fragment of the WF-net annotated with diagnostics generated by ProM’s conformance checker. The WF-net and event log fit well (fitness is 0.98876214). Nevertheless, several low-frequent deviations are discovered. For example, activity “OZ12 Hertaxeren” (re-evaluation of WOZ value) is started 23 times without being enabled according to the model

ing timestamps. The standard deviation is approx. 53 days. ProM also visualizes the bottlenecks by coloring the places in the WF-net. Tokens tend to reside longest in the purple places. For example, the place in-between “OZ16 Uitspraak start” and “OZ16 Uitspraak complete” was visited 436 times. The average time spent in this place is 7.84 days. This indicates that activity “OZ16 Uitspraak” (final judgment)

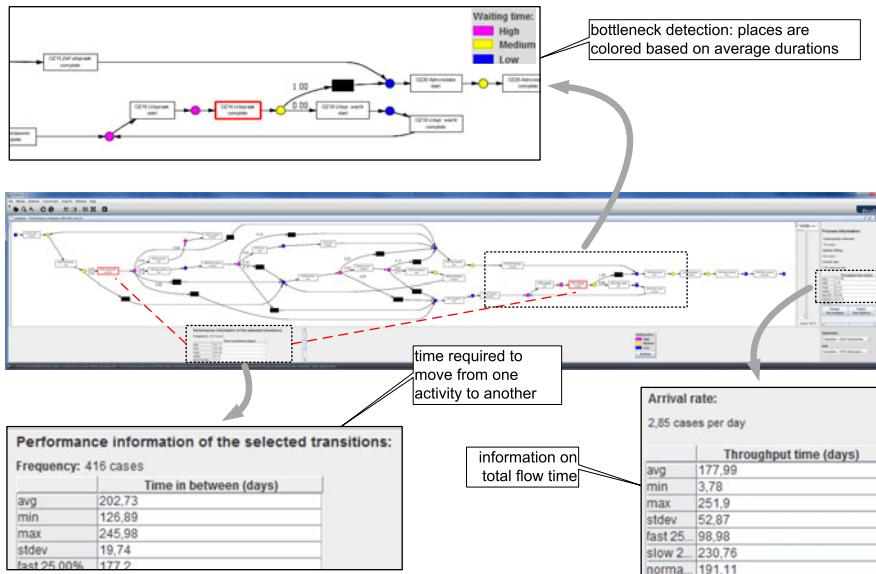


Fig. 13.13 Some diagnostics obtained by replaying the event log. These diagnostics explain why objections take on average approx. 178 days to be handled

takes about a week. The place before “OZ16 Uitspraak start” is also colored purple; on average it takes 138 days to start this activity after enabling. As shown in Fig. 13.13, it is also possible to simply select two activities and measure the time that passes in-between these activities. On average 202.73 days pass in-between the completion of activity “OZ02 Voorbereiden” (preparation) and the completion of “OZ16 Uitspraak” (final judgment). Note that this is longer than the average overall flow time. This is explained by the observation that only 416 of the objections (approx. 56%) follow this route; the other cases follow the branch “OZ15 Zelf uitspraak” which, on average, takes less time.

The event log also contains information about resources. The 9583 events are executed by 20 resources. Most activity instances have a start and complete event. These are typically done by the same person. However, in exceptional situations an activity is started by one person and completed by another. Table 13.1 shows the resource-activity matrix introduced in Sect. 9.3. The table shows that some people executed many activities (e.g., user 8 generated 2621 events) whereas others executed just a few activities (e.g., users 13 and 14 generated only one event). Figure 13.14 shows a social network based on the user profiles shown in Table 13.1. Persons that have similar profiles are connected and the strength of a connection depends on the degree of similarity (here we used the correlation coefficient). This information can be used to group people. Figure 13.14 shows four cliques discovered by ProM’s social network analyzer: *clique 1* consists of users 1, 2, 3, 8, 12, 13, 14, 16, and 17, *clique 2* consists of users 4, 5, 6, 9, 11, 18, and 19, *clique 3* consists of users 7 and 15, and *clique 4* consists of users 10 and 20. Consider, for example,

Table 13.1 Resource-activity matrix showing the number of times each user performed a particular activity: a_1 = “Domain: heus1”, a_2 = “OZ02 Voorbereiden”, a_3 = “OZ04 Incompleet”, a_4 = “OZ06 Stop vordering”, a_5 = “OZ08 Beoordelen”, a_6 = “OZ09 Wacht Beoord”, a_7 = “OZ10 Horen”, a_8 = “OZ12 Hertaxeren”, a_9 = “OZ15 Zelf uitspraak”, a_{10} = “OZ16 Uitspraak”, a_{11} = “OZ18 Uitspr. wacht”, a_{12} = “OZ20 Administatie”, a_{13} = “OZ24 Start vordering”. The names of users have been anonymized for privacy reasons

User	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
User 1	0	0	51	0	0	0	0	0	0	0	0	0	0
User 2	1	2	0	0	2	0	0	0	0	38	0	69	0
User 3	0	9	0	0	0	0	0	0	0	0	0	0	0
User 4	2	0	0	0	0	0	0	0	0	0	0	0	0
User 5	117	0	4	0	3	0	0	0	0	1	0	20	6
User 6	172	6	14	0	7	3	0	0	1	2	0	48	53
User 7	1	41	8	14	275	8	8	865	55	180	0	128	5
User 8	2	868	7	6	105	0	0	79	266	441	0	844	3
User 9	90	0	2	0	1	2	0	0	1	2	0	27	28
User 10	0	0	0	899	0	0	0	0	0	0	0	0	1019
User 11	336	1	3	1	4	2	0	0	0	1	0	18	23
User 12	1	645	13	21	419	3	0	3	217	281	1	334	9
User 13	0	1	0	0	0	0	0	0	0	0	0	0	0
User 14	0	0	0	0	0	0	0	0	0	1	0	0	0
User 15	0	0	0	0	0	0	0	2	2	0	0	2	0
User 16	1	3	3	2	1	0	0	1	2	3	1	0	0
User 17	0	4	0	0	0	0	0	0	0	0	0	0	0
User 18	9	0	0	0	0	0	0	0	0	0	0	0	0
User 19	13	1	0	0	1	0	0	0	0	0	0	4	0
User 20	0	0	0	21	0	0	0	0	0	0	0	0	258

clique 4. The two persons in this clique (users 10 and 20) only execute a_4 (“OZ06 Stop vordering”) and a_{13} (“OZ24 Start vordering”). Hence, it makes perfect sense that they are grouped together. For organizations it is interesting to see whether such clusters correspond to existing roles. Unexpected outcomes may trigger a redistribution of work.

The municipality for which we analyzed the WOZ process, provided us with several other event logs. For instance, event logs related to the handling of building permits. All of these processes can be classified as Lasagna processes and in principle all of the process mining techniques discussed in this book can be applied. The application of conformance checking on the processes of this municipality is discussed in more detail in [121]. For example, there it is shown that, despite the presence of a WFM system, processes still deviate from the normative models. The municipality was using eiStream WFM system (formerly known as Eastman Software and today named Global 360), therefore, we did not expect any deviations. However, as discussed in [121], process mining could reveal misconfigurations of the WFM

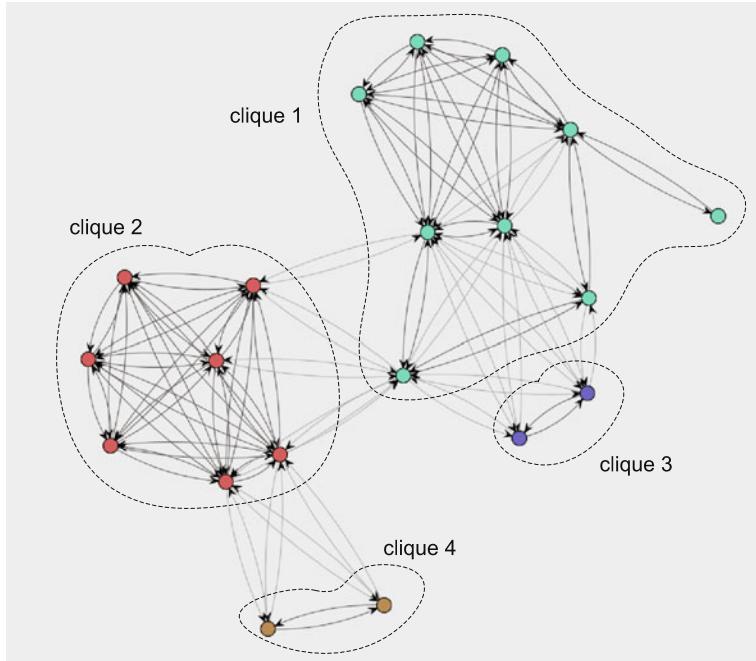


Fig. 13.14 Social network based on similarity of profiles. People that execute similar collections of activities are related and clustered in cliques

system. In [124], it is shown that, based on the event logs of this municipality, it is possible to discover simulation models covering all perspectives (control-flow, data dependencies, performance characteristics, and organizational characteristics). In Sect. 9.6, we showed how these perspectives can be merged into a single CPN model that can be simulated by CPN Tools. Although we did not conduct short-term simulations for this municipality, the validation of the models described in [124] shows that accurate simulations are possible for the selected process. Similarly, we showed in [167] that accurate time predictions are possible for the WOZ process of this municipality. In [167], various annotated transition systems are constructed using the approach described in Sect. 10.4. Each of these annotated transition systems is learned using one half of the event log, and evaluated using the other half. This illustrates that operational support is indeed possible for Lasagna processes.

Chapter 14

Analyzing “Spaghetti Processes”

Spaghetti processes are the counterpart of Lasagna processes. Because Spaghetti processes are less structured, only a subset of the process mining techniques described in this book are applicable. For instance, it makes no sense to aim at operational support activities if there is too much variability. Nevertheless, process mining can help to realize dramatic process improvements by uncovering key problems.

14.1 Characterization of “Spaghetti Processes”

As explained in the previous chapter, there is a continuum of processes ranging from highly structured processes (Lasagna processes) to unstructured processes (Spaghetti processes). In this chapter we focus on Spaghetti processes.

Figure 14.1 shows why unstructured processes are called Spaghetti processes. Only when zooming in one can see individual activities. Figure 14.2 shows a tiny fragment of the whole process. The fragment shows that activity “O_Bloedkweek 1” (a particular blood test) was scheduled 412 times and 230 times followed by “O_Bloedkweek 2” (another test). These activities are frequent. However, there are also several activities that are executed for only one of the 2765 patients.

The process model depicted in Fig. 14.1 was obtained using the heuristic miner with default settings. Hence, low frequent behavior has been filtered out. Nevertheless, the model is too difficult to comprehend. Note that this is not necessarily a problem of the discovery algorithm. Activities are only connected if they frequently followed one another in the event log (cf. Sect. 7.2). Hence, the complexity shown in Fig. 14.1 reflects reality and is not caused by the discovery algorithm!

Figure 14.1 is an extreme example used to explain the characteristics of a Spaghetti process. Given the data set it is not surprising that the process is unstructured; the 2765 patients did not form a homogeneous group and included individuals with very different medical problems. The process model can be simplified dramatically by selecting a group of patients with similar problems. However, also for more homogeneous groups of patients (e.g., people that had heart surgery), the resulting process model is often Spaghetti-like.

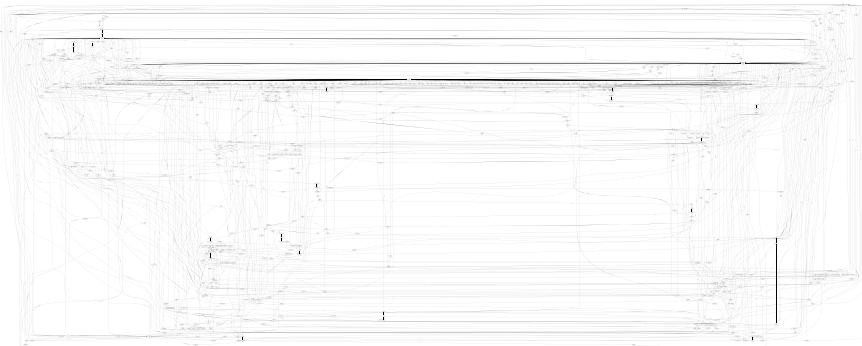


Fig. 14.1 Spaghetti process describing the diagnosis and treatment of 2765 patients in a Dutch hospital. The process model was constructed based on an event log containing 114,592 events. There are 619 different activities (taking event types into account) executed by 266 different individuals (doctors, nurses, etc.)

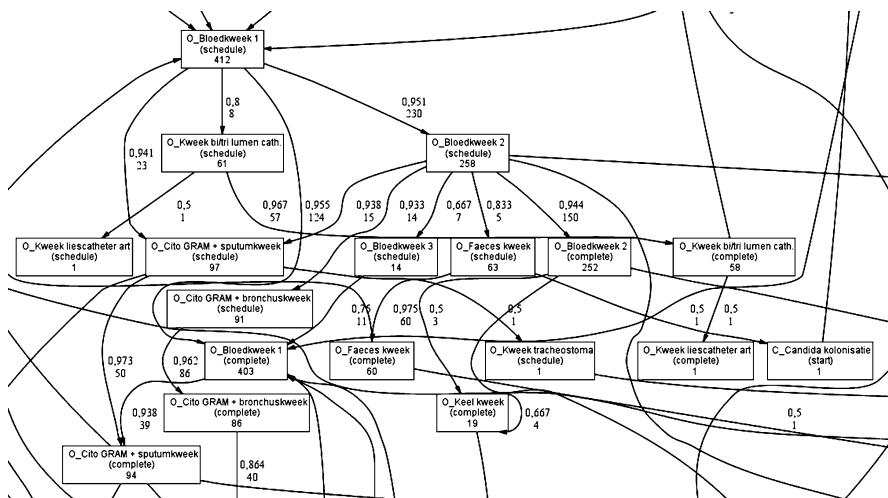


Fig. 14.2 Fragment of the Spaghetti process of Fig. 14.1 showing 18 activities of the 619 activities (2.9%)

Let us consider another, less extreme, example. Figure 14.3 shows the dotted chart for a process of one of the largest Dutch housing agencies (see also Figs. 9.3 and 9.4). Each case corresponds to a housing unit (accommodation such as a house or an apartment). The process starts when the tenant leasing the unit wants to stop renting it. The process ends when a new tenant moves into the unit after addressing all formalizaties. In-between, activities such as “registering the new address”, “first inspection”, “final inspection”, “finalize contract”, “return deposit”, “sign contract”, “repair”, and “update price” are executed. Figure 14.3 is based on an event log containing information about 208 units that changed tenant. There are 74 different

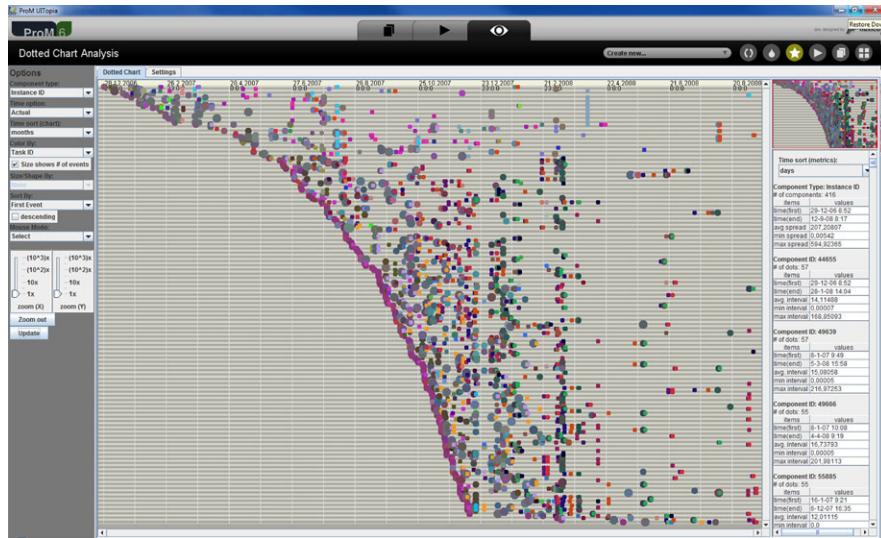


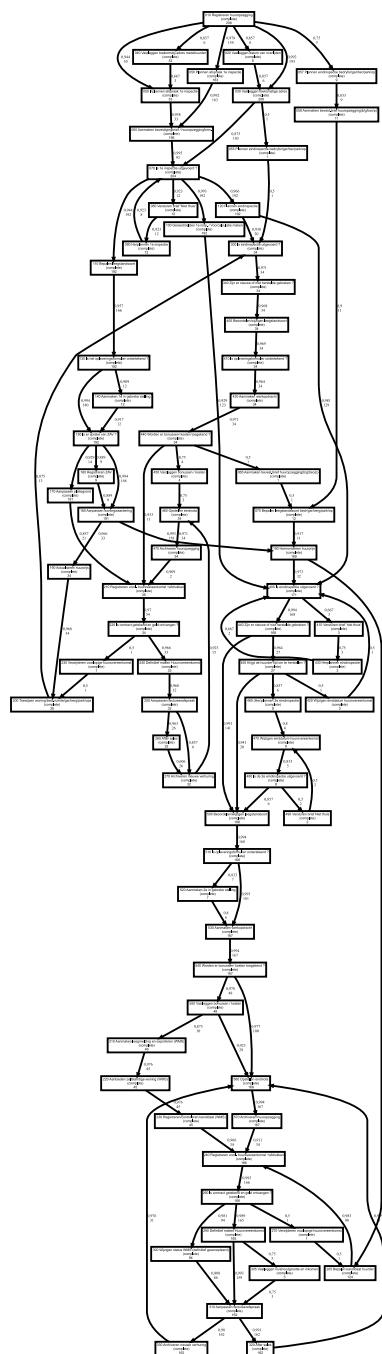
Fig. 14.3 Dotted chart created using an event log of a Dutch housing agency. Each line corresponds to a case (house or apartment). The event log contains 208 cases that generated 5987 events. There are 74 different activities

activities. In total 5987 activities were executed for the 208 units. As Fig. 14.3 shows there is a huge variance in flow time. For some units it takes a very long time to change ownership (sometimes more than a year) for others this is matter of days. The initial events of the 208 cases do not form a straight line; the curve shows that the arrival rate of new cases is increasing during the period covered by the event log.

Figure 14.4 shows a process model discovered using the heuristic miner. Although the model does not look as Spaghetti-like as Fig. 14.1, it is rather complicated considering the fact that it is based on only 208 cases. The 208 cases generate 203 unique traces, i.e., almost all cases follow a path that is not followed by any of the other cases. This observation, combined with the complexity of the model suggests that the log is far from complete thus complicating analysis.

The processes of the Dutch hospital and housing agency illustrate the challenges one is facing when dealing with Spaghetti processes. Nevertheless, such processes are very interesting from the viewpoint of process mining as they often allow for various improvements. A highly-structured well-organized process is often less interesting in this respect; it is easy to apply process mining techniques but there is also little improvement potential. Therefore, one should not shy away from Spaghetti processes as these are often appealing from a process management perspective. *Turning Spaghetti processes into Lasagna processes can be very beneficial for an organization.*

Fig. 14.4 C-net for the event log of the housing agency. The model was obtained using the heuristic miner (with default settings). The model was discovered based on an event log with 5987 events. All 208 cases start with activity “010 Registreren huuropzegging” (register request to end lease). Some of the activities are relatively infrequent, e.g., activity “020 Vastleggen datum van overlijden” occurred only 6 times (this activity is only executed if the tenant died)



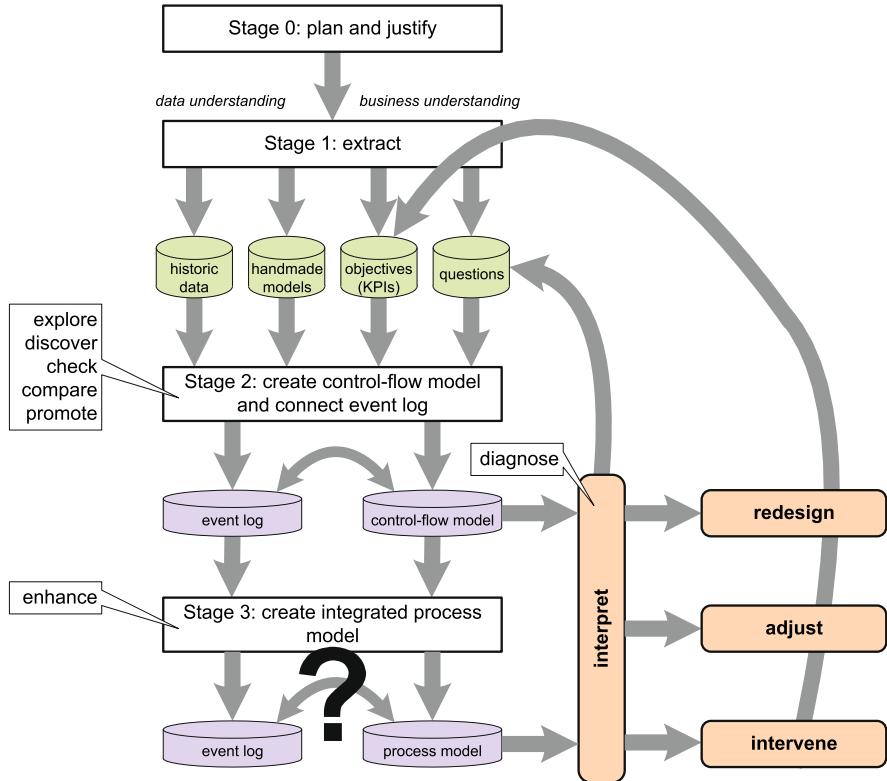


Fig. 14.5 The part of the L^* life-cycle model applicable to Spaghetti processes: stages 0, 1 and 2 are also possible for unstructured processes. However, creating an integrated process model covering all perspectives (Stage 3) is often not possible. Instead separate models are generated for the other perspectives, e.g., a social network

14.2 Approach

In Sect. 13.3 we introduced the L^* life-cycle model describing an idealized process mining project aiming at improving a Lasagna process. Only the initial stages are applicable for Spaghetti processes. Figure 14.5 shows the most relevant part of the L^* life-cycle model. Note that Stage 4 has been removed because operational support is impossible for the processes just described. To enable history-based predictions and recommendations it is essential to first make the “Spaghetti-like” process more “Lasagna-like”. In fact, Stage 3 will also be too ambitious for most Spaghetti processes. It is always possible to generate process models as shown in Figs. 14.1 and 14.4 (Stage 2). Moreover, it is also possible to view dotted charts, create social networks, etc. However, it is very unlikely that all of these can be folded into a meaningful comprehensive process model as the basis (the control-flow discovered) is too weak.

In Sect. 6.4 we discussed the challenges related to process mining. They are of particular relevance when dealing with Spaghetti processes. Event logs do not contain negative examples, i.e., only positive example behavior is given. The fact that something does not happen in an event log does not mean that it cannot happen. For example, Fig. 14.4 is based on an event log in which almost all cases follow a unique path (the 208 cases generate 203 different traces). Therefore, the discovery algorithm needs to generalize. For more complex processes, i.e., processes that are large and that allow for many behaviors, the event log is typically far from complete (cf. Sect. 6.4.2). To further complicate matters, there may be noisy behavior, i.e., infrequent behavior that the user is not interested in. Because of these complications, a discovery algorithm needs to carefully balance the four quality dimensions introduced earlier: *fitness*, *simplicity*, *precision*, and *generalization* (see Fig. 6.22). The process models shown in Figs. 14.1 and 14.4 illustrate the relevance of these considerations. For the characteristics of the different process discovery algorithms we refer to Part III of this book. Here, we only stress the importance of carefully *filtering* the event log before discovery.

Let us first consider the *filtering of activities* based on their characteristics, e.g., absolute or relative frequency. Figure 14.6(a) shows a filtering plug-in selecting all activities that occurred in at least 5% of all cases. This ProM 5.2 plug-in is applied to the event log used to construct Fig. 14.1, i.e., activities that do not appear frequently are removed from the event log. As a result, the process model will be simpler as fewer activities are included. Figure 14.6(b) shows a filtering plug-in in ProM 6 applied to the event log used to construct Fig. 14.4. In this case the top 80% of activities are included; all other activities are removed from the log. The effect of filtering is shown in Fig. 14.6(c). This C-net was obtained by selecting all activities that occur in at least 50% of all cases handled by the housing agency. A comparison of the process model obtained using the original event log (Fig. 14.4) with the process model obtained using the filtered event log (Fig. 14.6(c)), demonstrates the effect of filtering. The discovered model shows only 28 of the 74 activities appearing in the event log of the housing agency.

In principle, any model *can be made as simple as desired* by simply abstracting from infrequent activities. In the extreme case, the model contains only the most frequent activity. Such a model is not very useful. However, it shows that filtering can be used to seamlessly simplify models. Interestingly, it is sometimes useful to also abstract from very frequent activities that are interleaved with other activities (e.g., some system action executed after every update). These clutter the diagram while being less relevant. Note that there may be multiple criteria for selecting/removing activities (e.g., average costs, duration, and risks).

Besides the simple activity-based filtering illustrated by Fig. 14.6 there are more advanced types of filtering that transform low-level patterns into activities [77]. Moreover, the cases in the log can be partitioned in homogeneous groups as shown in [13, 62, 78]. The basic idea is that one *does not try to make one large and complex model for all cases, but simpler models for selected groups of cases*. Here, one can use the classical clustering techniques described in Sect. 4.3 and adapt them for process mining. To apply these techniques, feature extraction is needed to describe

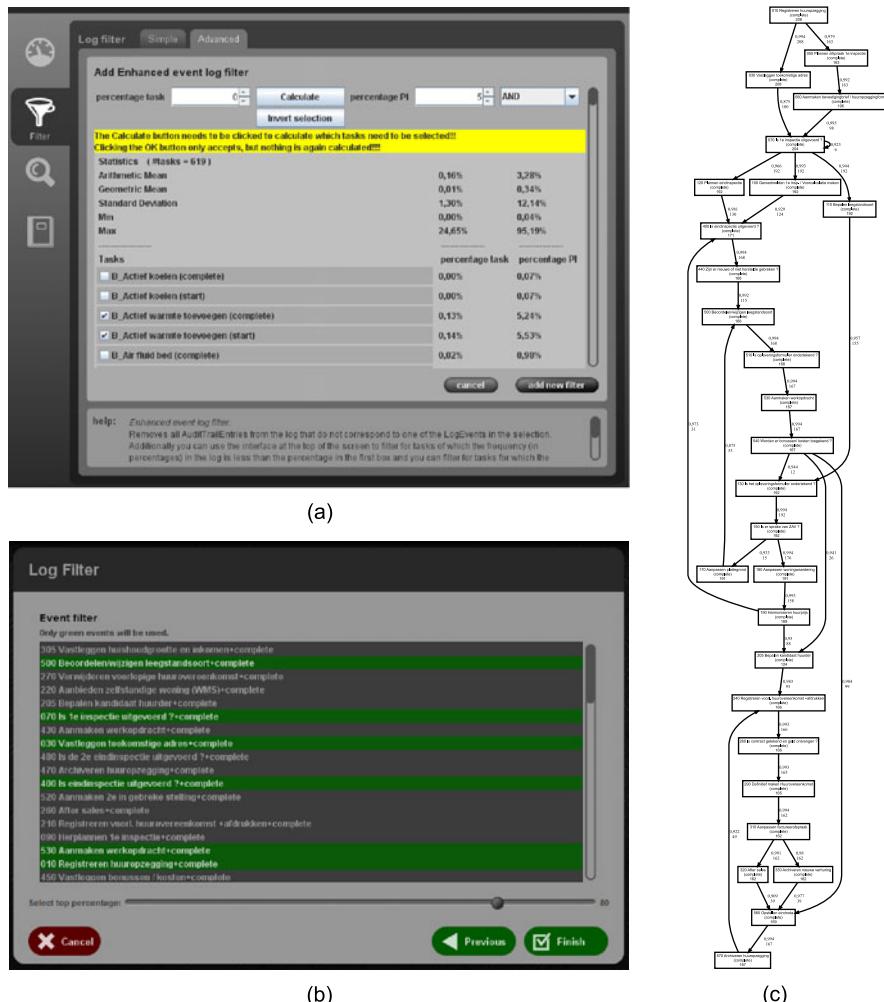


Fig. 14.6 Filtering the event log before process discovery: (a) selecting activities that occur for at least 5% of all 2765 patients, (b) selecting the top 80% of the 74 activities conducted by employees of the housing agency, (c) C-net discovered based on a filtered log (the event log of the housing agency after removing the activities occurring for less than 50% of the units)

cases in terms of a vector of variables (the features). By using a hierarchical clustering technique as shown in Fig. 14.7, one can view the same process at multiple levels. Cutting the dendrogram close to the root results in a few more complex models. Cutting the dendrogram closer to the leaves of the tree results in many simple models.

In the next chapter, we describe an alternative way to simplify process models. In contrast to filtering, simplification and abstraction techniques are directly applied to the process graph. This so-called *fuzzy mining* approach views process models

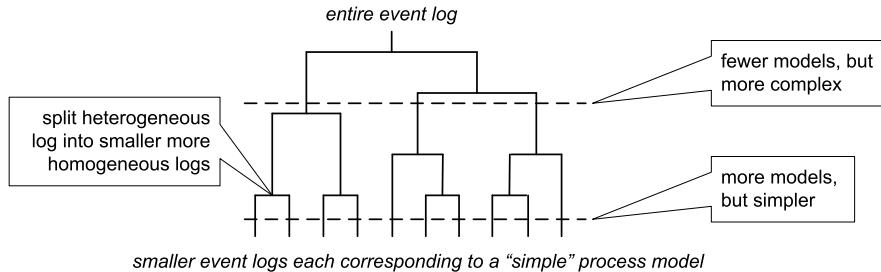


Fig. 14.7 Hierarchical clustering applied to heterogeneous event logs. The whole event log is partitioned into smaller, more homogeneous, event logs. This process is repeated until it is possible to create a “simple model” for each of the smaller logs. The resulting dendrogram can be cut closer to the root or closer to the leaves. This reflects the trade-off between the simplicity of models and the number of models

as if they are geographic maps (e.g., road maps or hiking maps). Depending on the map, insignificant roads and cities can be removed and streets and suburbs can be amalgamated into bigger structures. Figure 14.8 shows the effect this approach on the event log of the housing agency (i.e., the log used to construct the model in Fig. 14.4). Section 15.1.3 will elaborate further on the cartography metaphor used by the fuzzy mining approach.

14.3 Applications

In the previous chapter, we provided a systematic overview of the different sectors, industries, and functional areas where process mining can be used. In this section, we briefly revisit this overview for Spaghetti processes. Moreover, we give some pointers to case studies describing the analysis of highly unstructured processes.

14.3.1 Process Mining Opportunities for Spaghetti Processes

Many of the use cases presented in Sect. 13.2 also apply to Spaghetti processes. However, the “stakes are higher”; it will take more time to thoroughly analyze the process, but the potential gains are typically also more substantial.

Figure 14.9 highlights the functional areas where typically Spaghetti processes can be found.

Processes in *Product development* tend to be rather unstructured because they are low frequent (compared to production processes) and rely on creativity and problem-solving capabilities. For example, we have been mining event logs from Software Configuration Management (SCM) systems such as CVS and Subversion. In addition to managing the artifacts created by software engineers, these systems also collect and store information on the software development process to answer

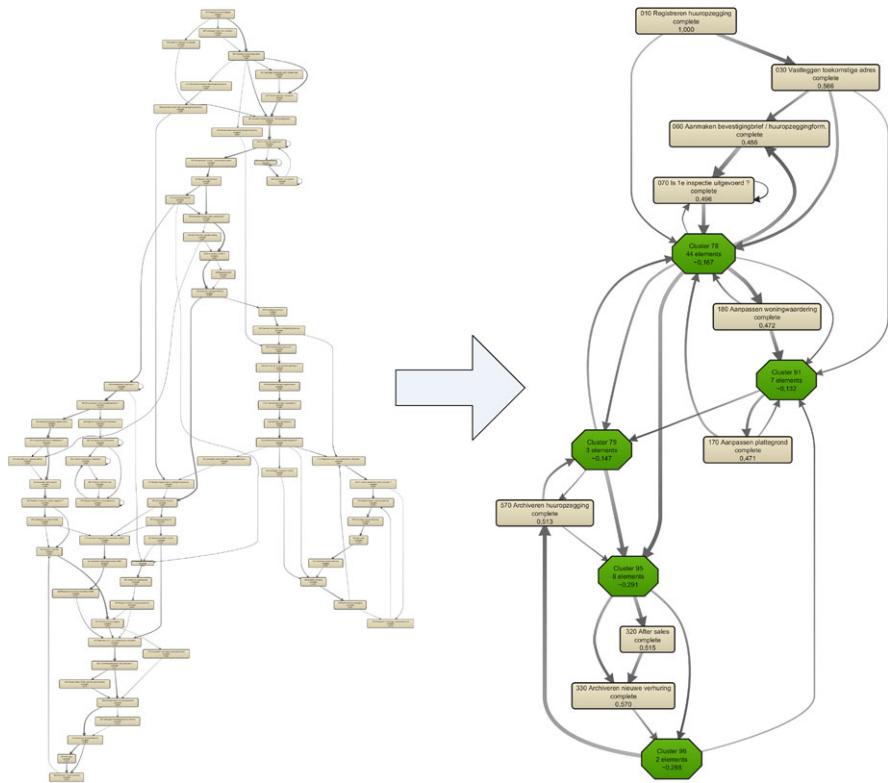
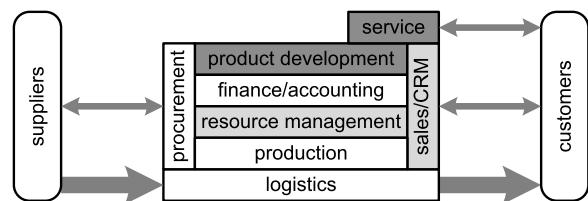


Fig. 14.8 Fuzzy mining applied to the event log of the housing agency. The cartography metaphor is used to support seamless abstraction and generalization. Both models provide a view on the same process. In the right model infrequent activities have been removed or amalgamated into cluster nodes. Moreover, infrequent arcs are removed based on the selected threshold

Fig. 14.9 Overview of the different functional areas in a typical organization. Spaghetti processes are typically encountered in product development, service, resource management, and sales/CRM



questions such as “Who created, accessed, or changed which documents?”, “When was a particular task completed?”, etc. Process discovery efforts using the event logs of SCM systems as input typically reveal Spaghetti-like processes as shown before.

Figure 14.9 indicates that one can also find Spaghetti processes in the functional area *Service*. An interesting development is that more and more products are monitored while being used in their natural habitat, e.g., modern high-end copiers, expensive medical devices, and critical production facilities collect event logs and can

be observed remotely. Later, we will show that ASML and Philips Healthcare already monitor the systems they manufacture. In the future, manufacturers will start monitoring also less expensive goods, e.g., cars, consumer electronics, and heating systems will be connected to the Internet for a variety of reasons. Manufacturers would like to know how their products are used, when they malfunction, and how to repair them.

Resource management and *Sales/CRM* are two functional areas where a mixture of Spaghetti and Lasagna processes can be encountered (cf. Sect. 13.4.1).

One can come across Spaghetti processes in all sectors and industries mentioned in Sect. 13.4.2. However, processes in the tertiary sector tend to be less structured than processes in the other two sectors. For instance, as is illustrated by Fig. 14.1, the healthcare industry is notorious in this respect. In general one can say that processes driven by humans that can operate in an autonomous manner are less structured. Situations, in which expertise, intuition, and creativity are important, stimulate self-government. Doctors in hospitals and engineers in large construction projects often need to deal with one-of-a-kind problems. Consumers that are using products also operate in an autonomous manner. Consider, for example, a television that can be monitored remotely to learn how it is used and when it malfunctions. Some users will watch television the whole day and constantly switch channels whereas other users only watch the news at 8 pm and then switch off the television. Self-directed behavior of consumers and professionals typically results in Spaghetti-like processes.

As mentioned earlier, *Spaghetti processes are interesting from the viewpoint of process mining*. First of all, it is interesting to learn from the amazing capabilities of humans to deal with complex unstructured problems. When automating parts of the process it is important to understand why processes are unstructured to avoid building counter-productive and inflexible information systems. Second, Spaghetti processes have the largest improvement potential. They are more difficult to analyze, but the prospective rewards are also higher.

14.3.2 Examples of Spaghetti Processes

We have encountered Spaghetti processes in a variety of organizations. In Chap. 13, we already mentioned several organizations where we applied process mining. In this section, we give three additional examples: ASML, Philips Healthcare, and AMC. The goal is not to describe the processes of these organizations in detail, but to provide pointers to applications of process mining in Spaghetti-like environments.

14.3.2.1 ASML

ASML is the world’s leading manufacturer of chip-making equipment and a key supplier to the chip industry. ASML designs, develops, integrates and services ad-

vanced systems to produce semiconductors. Process mining has been used to analyze the test process of wafer scanners in ASML [123].

Wafer scanners are complex machines consisting of many building blocks. They use a photographic process to image nanometric circuit patterns onto a silicon wafer. Because of competition and fast innovation, the time-to-market is very important and every new generation of wafer scanners is balancing on the border of what is technologically possible. As a result, the testing of manufactured wafer scanners is an important, but also time-consuming, process. Every wafer scanner is tested in the factory of ASML. When it passes all tests, the wafer scanner is disassembled and shipped to the customer where the system is re-assembled. At the customer's site, the wafer scanner is tested again. Testing is time-consuming and takes several weeks on both sites. Since time-to-market is very important, ASML is constantly looking for ways to reduce the time needed to test wafer scanners.

Figure 14.10 shows that the testing of wafer scanners is indeed a Spaghetti process [123]. The model was discovered based on an event log containing 154,966 events. The event log contained information about 24 carefully chosen wafer scanners (same type, same circumstances, and having complete logs). The number of events per case (i.e., the length of the executed test sequence) in this event log ranges from 2820 to 16250 events. There are 360 different activities, all identified by four-letter test codes. Each instance of these 360 activities has a start event and complete event. Figure 14.10 is based on just the complete events.

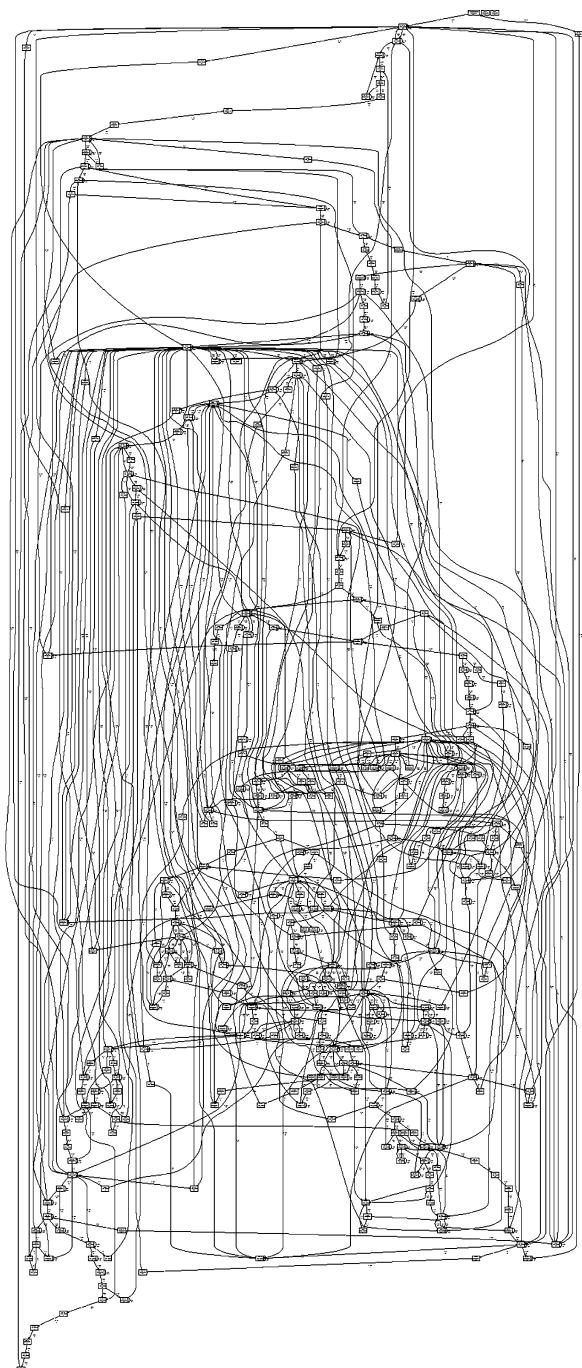
ASML also had a so-called reference model describing the way that machines should be tested. This reference model is at the level of job steps rather than test codes. However, ASML maintains a mapping from the lower level codes to these higher level activities. Comparing the reference model and our discovered model (both at the job step and test code level) revealed interesting differences. Moreover, using ProM's conformance checker we could show that the average fitness was only $\text{fitness}(L, N) = 0.375$, i.e., less than half of the events can be explained by the model (Sect. 8.2). When replaying, we discovered many activities that had occurred but that should not have happened according to the reference model and activities that should have happened but did not.

Both the discovered process models and the results of conformance checking showed that process mining can provide new insights that can be used to improve the management of complex Spaghetti-like processes. We refer to [123] for more details.

14.3.2.2 Philips Healthcare

Philips Healthcare is one of the leading manufacturers of medical devices, offering diagnosing imaging systems, healthcare information technology solutions, patient monitoring systems, and cardiac devices. Like ASML, Philips Healthcare is developing complex high-tech machines that record massive amounts of events. Since 2007 there has been an ongoing effort to analyze the event logs of these machines using process mining.

Fig. 14.10 Process model discovered for ASML’s test process



Philips Remote Services (PRS) is a system for the active monitoring of systems via the Internet. PRS has been established to deliver remote technical support, monitoring, diagnostics, application assistance, and other added value services. Low level events (e.g., pushing a button, changing the dosage) are recorded by the machine and subsequently sent to Philips via PRS. Using the Remote Analysis, Diagnostics And Reporting (RADAR) system, event logs are converted into an XML format and stored in the internal database of RADAR. Subsequently the collected event data are translated into MXML files to enable process mining.

Process mining has been applied extensively to the event logs generated by Allura Xper systems. These are X-ray systems designed to diagnose and possibly assist in the treatment of all kinds of diseases, like heart or lung diseases, by generating images of the internal body. These systems record three types of events:

- *User messages*. When a message is shown to the user (e.g., “Geometry restarting”) this is recorded in the event log.
- *Commands*. Both users and system components can invoke commands. These are all recorded. Commands typically have various parameters (e.g., voltage values).
- *Warnings and errors*. Whenever a problem occurs (or is anticipated) an event is recorded.

Each event has a timestamp and contains information about the component that generated the event.

It is possible to analyze the processes in Allura Xper systems from various angles. The concept of a “case” (i.e., process instance) may refer to a machine, a machine day, the execution of a particular procedure, the repair of a machine, etc. Figure 14.11 shows an example taken from [67]. Processes discovered for these systems tend to be Spaghetti-like. To simplify diagnosis, the log is often preprocessed as discussed in [77–79]. Moreover, fuzzy mining, as illustrated by Fig. 14.8, is used to further simplify the model [67].

Mining processes from the event logs generated by Allura Xper systems is very challenging. The machines consist of many components and can be used in many different ways. Moreover, logging is rather low-level and changes with every new version. Nevertheless, there are various opportunities for process and system improvements using process mining. These are listed below. Note that opportunities also apply to other types of systems that are monitored remotely.

- Process mining provides *insight* into how systems are actually used. This is interesting from a *marketing* point of view. For example, if a feature is rarely used, then this may trigger additional after sales activities. It is also possible that, based on process mining results, the feature is removed or adapted in future systems.
- *Testing* can be improved by constructing test scenarios based on the actual use of the machines. For instance, for medical equipment it is essential to prove that the system was tested under realistic circumstances.
- Process mining can be used to improve the *reliability* of next generations of systems. Better systems can be designed by understanding why and when systems malfunction.

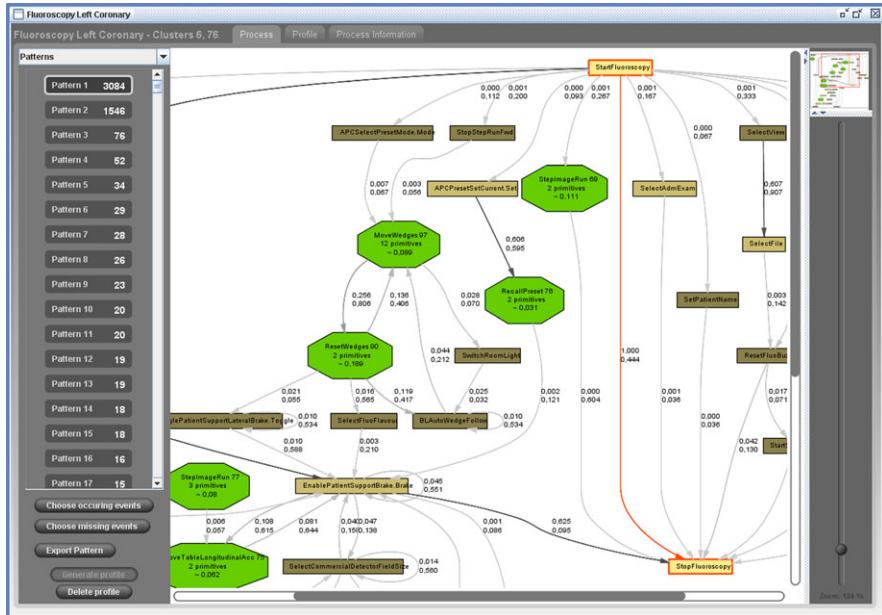


Fig. 14.11 Screenshot of a discovered process model for fluoroscopy runs in the context of the so-called “left coronary procedure” inside Allura Xper systems distributed all over the globe

- Process mining can also be used for *fault diagnosis*. By learning from earlier problems, it is possible to find the root cause for new problems that emerge. For example, we have analyzed under which circumstances particular components are replaced. This resulted in a set of *signatures*. When a malfunctioning X-ray machine exhibits a particular “signature” behavior, the service engineer knows what component to replace.
- Historic information can also be used to *predict* future problems. For instance, it is possible to anticipate that an X-ray tube is about to fail. Hence, the tube can be replaced before the machine starts to malfunction.

These examples show the potential of remote diagnostics based on process mining.

14.3.2.3 AMC Hospital

Hospitals are particularly interesting from a process mining point of view. By law, hospitals need to record more and more data in a systematic manner and all event data are connected to patients. Therefore, it is relatively straightforward to correlate events. For example, by Dutch law all hospitals need to record the diagnostic and treatment steps at the level of individual patients in order to receive payments. This so-called “Diagnose Behandeling Combinatie” (DBC) forces Dutch hospitals to record all kinds of events. There is also consensus that processes in hospitals

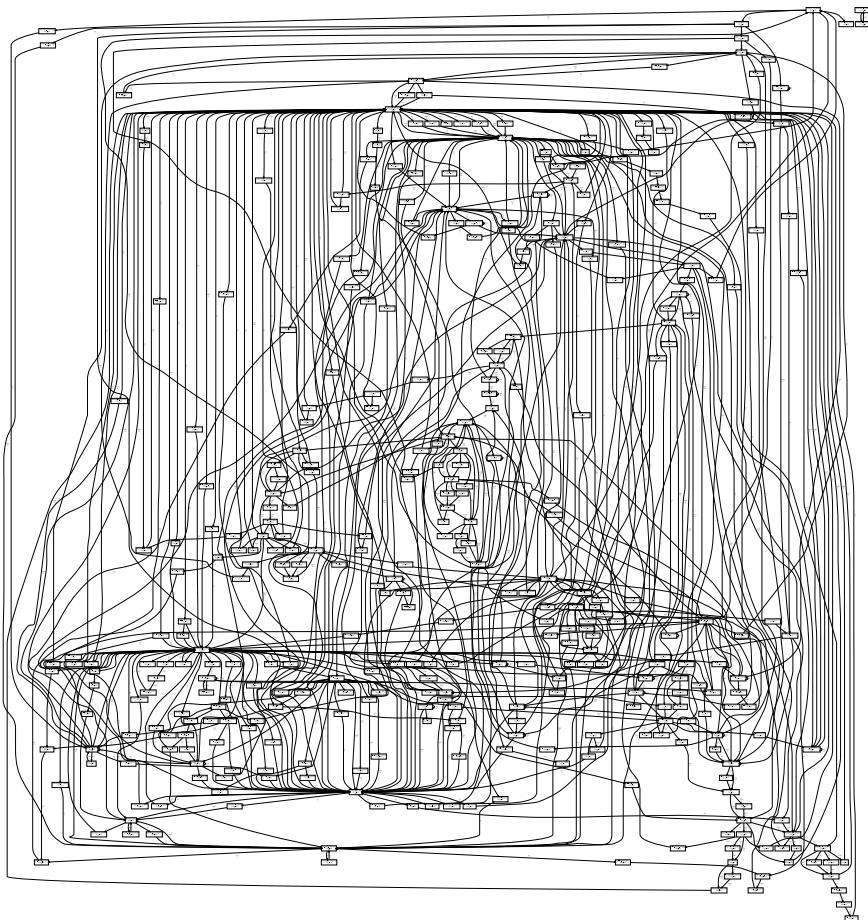


Fig. 14.12 Another Spaghetti process. The model is based on a group of 627 gynecological oncology patients. The event log contains 24331 events referring to 376 different activities

can be improved. Unlike most other domains, operational care processes are not tightly controlled by management. This, combined with the intrinsic variability of care processes, results in Spaghetti.

Some think that care processes in hospitals can be improved by simple principles from operations management or by introducing workflow technology. Process models such as the one shown in Fig. 14.1 demonstrate that this is not case. One needs to better understand the variability, before suggesting solutions.

We conducted several process mining experiments based on event data of the AMC hospital in Amsterdam [95]. Together with people of the AMC we have been investigating the introduction of workflow technology in this large academic hospital. This revealed many limitations of existing WFM/BPM systems when it comes to care processes. The variability in these processes is larger than in most other do-

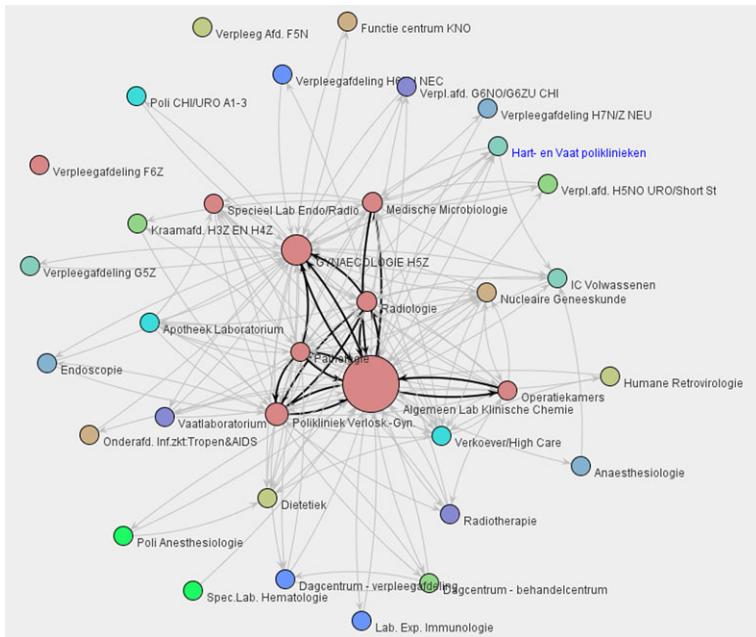


Fig. 14.13 Social network showing handovers between different organizational units of the AMC hospital

mains. This imposes unique requirements with respect to flexibility. Moreover, care processes combine flow oriented tasks with scheduled tasks [96]. As a result, conventional workflow technology is not applicable and a better understanding of the processes is needed.

Figure 14.12 shows an example of a process model constructed for the AMC hospital. The model was discovered based on event data of a group of 627 gynecological oncology patients treated in 2005 and 2006. All diagnostic and treatment activities have been recorded for these patients. Clearly, this is a Spaghetti process. However, as shown in [95] it is possible to create simple models for homogeneous groups of patients using the hierarchical clustering technique illustrated by Fig. 14.7. The same event log also contained information about resources. For instance, Fig. 14.13 shows a social network based on this log. As in earlier examples, the social network is based on handovers of work. However, now we do not look at individuals but at the level of organizational units. Figure 14.13 can be used to analyze the flow of work between different departments of the AMC hospital. For example, the social network reveals that most handovers take place between the gynecology department and the general clinical lab.

Experiences with process mining in several hospitals revealed important challenges when applying this new technology. The databases of hospitals contain lots of event data. Since any event can be linked to a patient, correlation is easy. However, for many events only the date (“31-12-2010”) is known and not the exact timestamp

(“31-12-2010:11.52”). Therefore, it may be impossible to deduce the order in which events took place. Another problem is related to the trade-off illustrated by the dendrogram in Fig. 14.7. The process model for a large group of patients is typically Spaghetti-like as illustrated by Fig. 14.12. It is possible to create simpler models by looking at smaller homogeneous groups of patients. However, the drawback is that often the number of cases per group gets rather small. If there are only few cases in such a homogeneous group, the result is not very reliable. Only for homogeneous groups with more cases, the result is more trustworthy.

Despite these challenges, process mining provides a “mirror” for managers, doctors, and IT specialists in hospitals. To improve care-flows and to provide better IT support, it is essential to face the inherent complexity of these Spaghetti processes.

Part VI

Reflection

Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining:
The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling
and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery:
An Introduction

Chapter 7
Advanced Process
Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance
Checking

Chapter 9
Mining Additional
Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining
Software

Chapter 12
Process Mining in the
Large

Chapter 13
Analyzing “Lasagna
Processes”

Chapter 14
Analyzing “Spaghetti
Processes”

Part VI: Reflection

Chapter 15
Cartography and
Navigation

Chapter 16
Epilogue

The final part of this book reflects on the relevance and positioning of process mining. Chapter 15 relates process mining to cartography and navigation. Limitations of traditional process models are revealed by studying the features of geographic maps. Many of these limitations can (and should) be addressed by process mining techniques. Moreover, navigation systems and mashups based on Google Maps illustrate how process maps can be actively used at run-time. Chapter 16 concludes this book by summarizing the benefits of process mining and listing challenges that need to be addressed to make process mining even more applicable.

Chapter 15

Cartography and Navigation

Process models can be seen as the “maps” describing the operational processes of organizations. Similarly, information systems can be looked at as “navigation systems” guiding the flow of work in organizations. Unfortunately, many organizations fail in creating and maintaining accurate business process maps. Often process models are outdated and have little to do with reality. Moreover, most information systems fail to provide the functionality offered by today’s navigation systems. For instance, workers are not guided by the information system and need to work behind the system’s back to get things done. Moreover, useful information such as the “estimated arrival time” of a running case is not provided. Process mining can help to overcome some of these problems.

15.1 Business Process Maps

The first geographical maps date back to the 7th Millennium BC. Since then cartographers have improved their skills and techniques to create maps thereby addressing problems such as clearly representing desired traits, eliminating irrelevant details, reducing complexity, and improving understandability. Today, most geographic maps are digital and of high quality. This has fueled innovative applications of cartography as is illustrated by modern car navigation systems (e.g., TomTom, Garmin, and Navigon), Google Maps, mashups using geo-tagging, etc. There are thousands of mashups using Google Maps, e.g., applications projecting information about traffic conditions, real estate, fastfood restaurants, or movie showtimes onto a selected map. People can seamlessly zoom in and out using such maps and interact with it, e.g., traffic jams are projected onto the map and the user can select a particular problem to see details.

Process models can be seen as the “*business process maps*” describing the operational processes of organizations [138]. Unfortunately, accurate business process maps are typically missing. Process models tend to be outdated and not aligned with reality. Moreover, unlike geographic maps, process models are typically not well understood by end users.

As indicated in Sect. 10.1.1, we suggest *adopting ideas from cartography*. In the remainder of this section, we discuss ways of improving process models inspired by cartographic techniques. Some of these ideas are already supported by existing process mining techniques, others point to further innovations.

15.1.1 Map Quality

Geographical maps are typically of high quality compared to business process maps. For example, the maps used by navigation systems are very accurate, e.g., when driving from Amsterdam to Rome relatively few discrepancies between reality and the map will be encountered.

Process models tend to provide an idealized view on the business process that is modeled. Imagine that road maps would view the real highway system through similar rose-tinted glasses, e.g., showing a road that is not there but that should have been there. This would be unacceptable. However, these are the kind of business process maps used in many organizations. Such a “PowerPoint reality” limits the use and trustworthiness of process models.

In Chap. 8 we showed various conformance checking techniques that can be used as a “reality check” for business process maps. For instance, using replay the fitness of a process model and an event log can be determined. We encountered many real-life processes in which the fitness of the model and the log is less than 0.4. This implies that less than 40% of the behavior seen in reality fits into the model.

Some will argue that road maps are easier to maintain than process models, because a road system evolves at a much slower pace than a typical business process. This is indeed the case. However, this makes it even more important to have accurate up-to-date business process maps!

Besides differences in quality, there are also huge differences in understandability. Most people will intuitively understand geographical maps while having problems understanding process models. The dynamic nature of processes makes things more complicated (cf. workflow patterns [155, 191]). Therefore, the perceived complexity is partly unavoidable. Nevertheless, ideas from cartography can help to improve the understandability of process models.

15.1.2 Aggregation and Abstraction

Figure 15.1 shows a map. The map *abstracts* from less significant roads and cities. Roads that are less important are not shown. A cut-off criterion could be based on the average number of cars using the road per day. Similarly, the number of citizens could be used as a cut-off criterion for cities. For example, in Fig. 15.1 cities of less than 50,000 inhabitants are abstracted from. Maps also *aggregate* local roads and local districts (neighborhoods, suburbs, centers, etc.) into bigger entities.



Fig. 15.1 Road map of The Netherlands. The map abstracts from smaller cities and less significant roads; only the bigger cities, highways, and other important roads are shown. Moreover, cities aggregate local roads and local districts

Figure 15.1, for instance, shows Eindhoven as a single dot while it consists of many roads, various districts (Strijp, Gestel, Woensel, Gestel, etc.), and neighboring cities (e.g., Veldhoven). People interested in Eindhoven can look at a city map to see more details.

Process models also need to abstract from less significant things. Activities can be removed if they are less frequent, e.g., activities that occur in less than 20% of completed cases are abstracted from. Also time and costs can be taken into account, e.g., activities that account for less than 8% of the total service time are removed unless the associated costs are more than € 50,000.

Aggregation is important for process mining because many event logs contain low-level events that need to be aggregated into more meaningful activities. In [77] it is shown how frequent low-level patterns can be identified and aggregated. Suppose that $x = \{\langle a, b, c \rangle, \langle a, b, b, c \rangle\}$, $y = \{\langle a, d, e, c \rangle, \langle a, e, d, c \rangle\}$, and $z = \{\langle d, d, d, a \rangle\}$

d	d	d	a	a	b	b	c	a	d	e	c	a	b	c	
z				x				y				x			

Fig. 15.2 A low-level trace is mapped onto a trace at a higher level of abstraction, e.g., the subsequence $\langle d, d, d, a \rangle$ is mapped onto z

are frequent low-level patterns that represent meaningful activities, e.g., the low-level subsequences a, b, c and a, b, b, c are possible manifestations of activity x . Now consider the low-level trace $\sigma = \langle d, d, d, a, a, b, b, c, a, d, e, c, a, b, c \rangle$. This trace can be rewritten into $\sigma' = \langle z, x, y, x \rangle$ showing the aggregated behavior (see Fig. 15.2). By preprocessing the event log in this way, it is possible to discover a simpler process model. Filtering, as described in Sect. 14.2, can be seen as another form of preprocessing. It is also possible to apply aggregation directly to the graph structure (see fuzzy mining [66] and Sects. 14.2 and 15.1.3).

Aggregation introduces multiple levels. For each aggregate node a kind of “city map” can be constructed showing the detailed low-level behavior. In principle there can be any number of levels, cf. country maps, state maps, city maps, district maps, etc.

15.1.3 Seamless Zoom

There may be different geographic maps of the same area using different scales. Moreover, using electronic maps it is possible to seamlessly zoom in and out. Note that, while zooming out, insignificant things are either left out or dynamically clustered into aggregate shapes (e.g., streets and suburbs amalgamate into cities). Navigation systems and applications such as Google Maps provide such a seamless zoom. Traditionally, process models are static, e.g., it is impossible to seamlessly zoom in to see part of the process in more detail. To deal with larger processes, typically a *static hierarchical decomposition* is used. In such a hierarchy, a process is composed of subprocesses, and in turn these subprocesses may be composed of smaller subprocesses.

Consider, for example, the WF-net shown in Fig. 15.3. The WF-net consists of atomic activities (a, b, \dots, l) partitioned over three subprocesses x, y , and z . The overall process is composed of these three subprocesses. Figure 15.4 shows the top-level view of this composition. The semantics of such a hierarchical decomposition is the “flattened” model, i.e., subprocesses at the higher level are recursively replaced by their inside structure until one large flat process model remains (in our example there are only two levels).

Figures 15.3 and 15.4 show the limitations of hierarchical decomposition. At the highest level one needs to be aware of all interactions at the lower levels. The reason is that higher levels in the decomposition need to be consistent with the lower levels, e.g., because there is a connection between activity l and activity b at the lower level, there also needs to be a connection between z and x at the higher level.

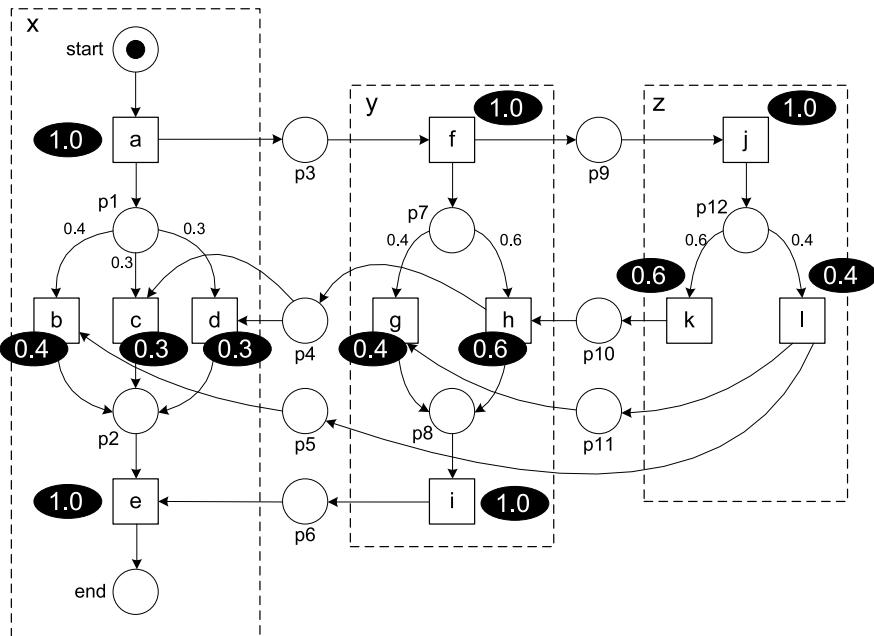
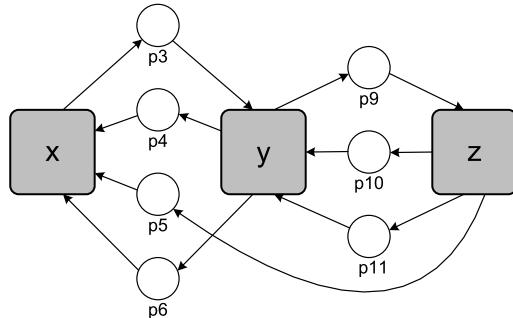


Fig. 15.3 A WF-net consisting of 12 atomic activities partitioned over three subprocesses x , y , and z . The average frequency of each activity is shown. For instance, activity h is executed for 60% of the cases

Fig. 15.4 Top-level view on the hierarchical WF-net shown in Fig. 15.3



This is not only the case for WF-nets, but holds for the hierarchy constructs in other languages such as BPMN, YAWL and EPCs. From a design point of view, hierarchical decomposition makes perfect sense. When designing a system it is important to ensure consistency between different levels and the possibility to “flatten” models provides clear execution semantics.

However, when viewing a process model it is important to be able to zoom out to see fewer details and zoom in to see more details. This implies that the view is not static, i.e., activities should not be statically bound to a particular level chosen at design time. Moreover, when abstracting from infrequent low-level behavior the

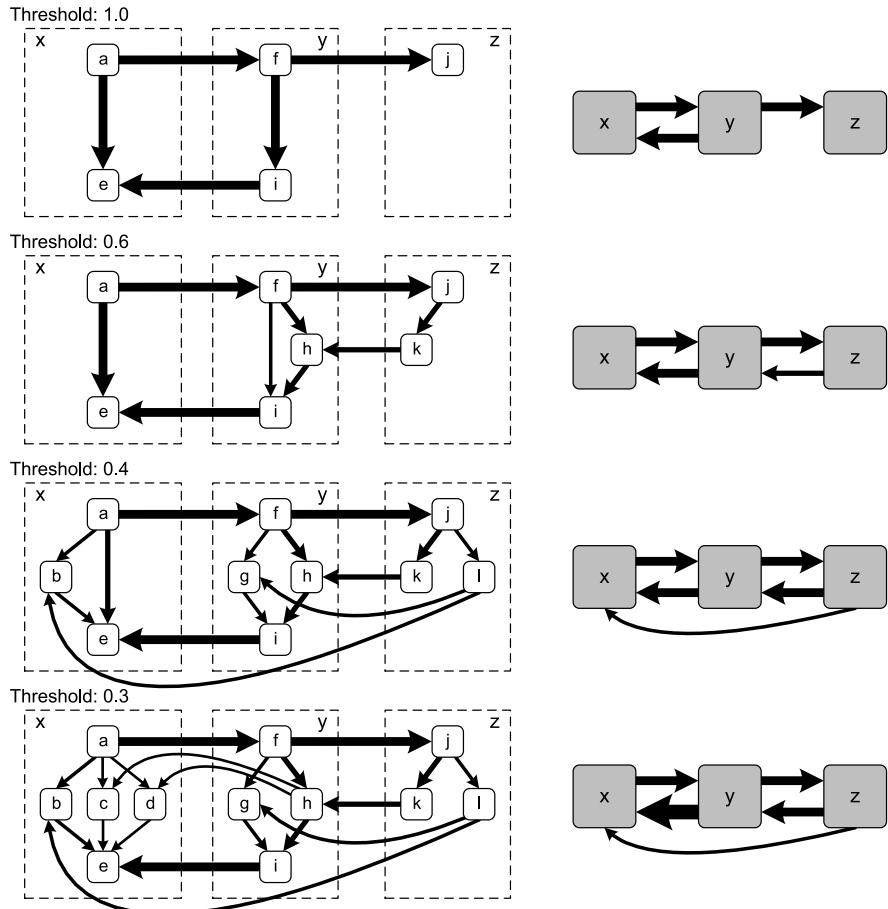


Fig. 15.5 The process specified by the WF-net of Fig. 15.3 viewed at different levels of abstraction. An activity and its corresponding connections are removed if the transition is less frequent than the threshold. Both the atomic view (*left*) and the aggregate view (*right*) are shown for four threshold values (0.3, 0.4, 0.6, and 1.0)

corresponding connections at higher levels should also be removed. For instance, if activity *l* is very infrequent, it is not sufficient to hide it at a lower level: the connection between *z* and *x* (i.e., place *p5*) should also be removed.

Figure 15.5 illustrates how processes can be viewed while taking into account the frequencies of activities. As shown in Fig. 15.3, activities *a*, *e*, *f*, *i*, and *j* have a frequency of 1, i.e., they are executed once for each case. Activities *h* and *k* are executed for 60% of the cases, and activities *b*, *g*, and *l* are executed for 40% of the cases. Activities *c* and *d* are least frequent and are executed for only 30% of the cases. Assume that we would like to seamlessly simplify the model by progressively leaving out more activities based on their frequencies. Figure 15.5 shows four different levels. Here, we abstract from the detailed process logic and only show ac-

tivities and their connections. Moreover, we show the intensity of connections by proportionally varying the width of the arcs. If the threshold is set to 0.3, then all activities are included. When the threshold is increased to 0.4, then activities c and d and their connections disappear. When the threshold is increased to 0.6, also activities b , g , and l and their connections disappear. If the threshold is set to 1, then only the most frequent activities are included. The left-hand side of Fig. 15.5 shows atomic activities and their relations. The right-hand side of the figure shows the connections if we assume that the activities are aggregated as shown in the original WF-net (cf. Fig. 15.3). It is important to note that the connection between z and x disappears when the threshold is higher than 0.4. If we abstract from the infrequent activities b and l , then we should also remove this connection. For the same reason the connection between z and y is not shown when the threshold is set to 1.

Figure 15.5 shows how one can seamlessly zoom in and zoom out to show more or less detail. This is very different from providing a static hierarchical decomposition and showing a particular level in the hierarchy as is done by the graphical editors of BPM systems, WFM systems, simulation tools, business process modeling tools, etc.

Thus far we assumed a static partitioning of atomic activities over three subprocesses. Depending on the desired view this partitioning may change. To illustrate this, we use an example event log consisting of 100 cases and 3730 events. This event log contains events related to the reviewing process of journal papers. Each paper is sent to three different reviewers. The reviewers are invited to write a report. However, reviewers often do not respond. As a result it is not always possible to make a decision after a first round of reviewing. If there are not enough reports, then additional reviewers are invited. This process is repeated until a final decision can be made (accept or reject). Figure 15.6 shows the process model discovered by the α -algorithm.

The α -algorithm does not allow for seamlessly zooming in and out. One would need to filter out infrequent activities from the log and subsequently apply the α -algorithm to different event logs. The *Fuzzy Miner* of ProM allows for seamlessly zooming in and out as is shown in Fig. 15.7 [65, 66]. The three fuzzy models shown in Fig. 15.7 are all based on the event log also used by the α -algorithm. Figure 15.7(a) shows the most detailed view. All activities are included. The color and width of the connections indicate their significance (like in Fig. 15.5). Figure 15.7(b) shows the most abstract view. The decision activity is typically executed multiple times per paper. Therefore, it is most frequent. The other 18 activities are partitioned over 4 so-called cluster nodes. Each cluster node aggregates multiple atomic activities. Using a threshold similar to the one used in Fig. 15.5, the Fuzzy Miner can seamlessly show more or less details. Figure 15.7(c) shows a model obtained using an intermediate threshold value. The top-level model shows the six most frequent activities. The other activities can be found in the three cluster nodes. Figure 15.7(d) shows the inner structure of an aggregate node consisting of 10 atomic activities. Note that the inner structure of an aggregate node shows the connections to nodes at the higher level.

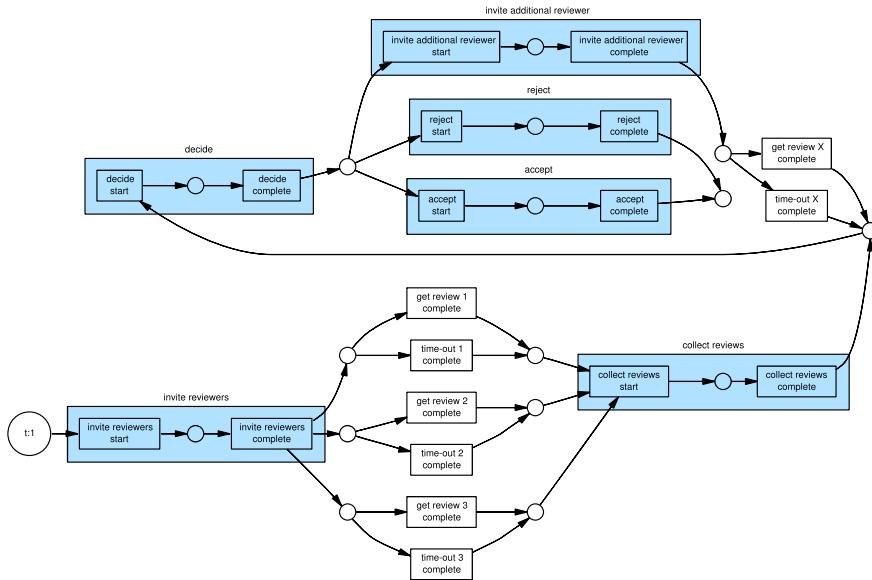


Fig. 15.6 WF-net discovered using the α -algorithm plug-in of ProM. An event log consisting of 100 cases and 3730 events was used to create the model. All activities are shown in the discovered model

When zooming out using Google maps, less significant elements are either left out or dynamically clustered into aggregate shapes. For example, streets and suburbs amalgamate into cities. This is similar to the zoom functionality provided by ProM's Fuzzy Miner as was demonstrated using Fig. 15.7. Note that in this particular example activities are aggregated and not removed. The Fuzzy Miner has many parameters that allow the user to influence the resulting model. Using different settings of these parameters it is also possible to abstract from activities (i.e., remove them) rather than aggregating them. Activities can also be removed by filtering the event log before applying a discovery algorithm (see Sect. 14.2).

15.1.4 Size, Color, and Layout

Cartographers not only eliminate irrelevant details, but also use colors to highlight important features. For instance, the map shown in Fig. 15.1 emphasizes the importance of highways using the color red. Moreover, graphical elements have a particular size to indicate their significance, e.g., the sizes of lines and dots may vary. For instance, in Fig. 15.1 the size of a city name is proportional to the number of citizens, e.g., Zaandstad is clearly smaller than Amsterdam. Geographical maps also have a clear interpretation of the x -axis and y -axis, i.e., the layout of a map is not arbitrary as the coordinates of elements have a meaning.

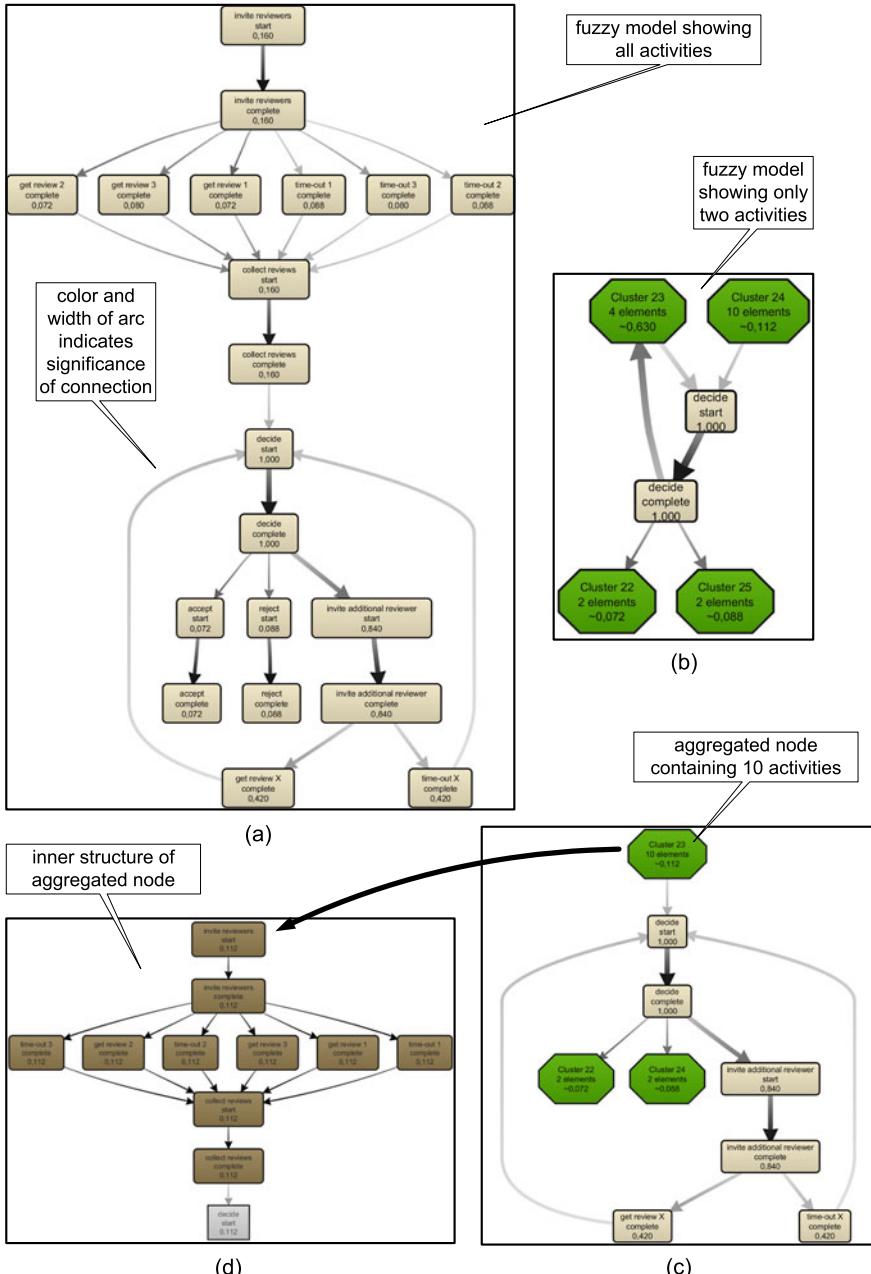


Fig. 15.7 Three business process maps obtained using ProM's Fuzzy Miner. The most detailed fuzzy model (a) shows all activities. The least detailed fuzzy model (b) shows only two activities; all other activities are aggregated into so-called “cluster nodes”. The third fuzzy model (c) shows six activities. For one of the aggregate nodes, the inner structure is shown (d)

All of this is in stark contrast with mainstream process models. The x -axis and y -axis of a process model have no meaning, e.g., the layout of the WF-net shown in Fig. 15.6 was generated automatically without assigning any semantics to the positions of activities. Although modeling tools allow for using colors, the color typically has no semantics. The different types of model elements (e.g., activities, gateways, events, connectors, and places) typically have a default color. Moreover, the size of a model element also has no semantics. Typically all elements of a particular type have the same size.

Because size, color, and layout are not employed when creating business process maps, the result is less intuitive and less informative. However, ideas from cartography can easily be incorporated in the construction of business process maps. Some examples:

- The *size of an activity* can reflect its frequency or some other property indicating its significance (e.g., costs or resource use).
- The *color of an activity* can reflect the mean service time of the activity. For example, activities that take longer than average are colored red whereas short running activities are colored green.
- The *width of an arc* can reflect the importance of the corresponding causal dependency.
- The *coloring of arcs* can be used to highlight bottlenecks.
- The *positioning of activities* can have a well-defined meaning. Similar to swim-lanes the y -axis could reflect the role associated to an activity. Similar to a Gantt chart, the x -axis could reflect some temporal aspect.

It is important to use these conventions in a consistent manner across different maps.

15.1.5 Customization

The same geographic area is typically covered by many different maps. There are different maps depending on the type of activity they intend to support, e.g., bicycle maps, hiking maps, and road maps. Obviously, these maps use different scales. However, there are more differences. For instance, a bicycle map shows bicycle paths that are not shown on motorists' map.

Figure 15.7 illustrates that multiple views can be created for the same reality captured in an event log. In earlier chapters we already showed that there is no such thing as *the* process model describing a process. Depending on the questions one seeks to answer, a customized process model needs to be created. In Sect. 6.4.4, we referred to this as taking a “2-D slice of a 3-D reality”. The same process can be viewed from different angles and at different levels of granularity. For noisy event logs one may prefer to focus on just the main behavior or also include less frequent behavior. For example, from an auditing point of view the low frequent behavior may be most interesting.

15.2 Process Mining: TomTom for Business Processes?

After comparing geographic maps with business process maps, we now explore the analogy between navigation systems and information systems. Section 10.1.3 already mentioned navigation activities in the context of the refined process mining framework (cf. Fig. 10.1) By establishing a close connection between business process maps and the actual behavior recorded in event logs, it is possible to realize TomTom-like functionality. Analogous to TomTom’s navigation devices, process mining tools can help end users (a) by navigating through processes, (b) by projecting dynamic information on process maps (e.g., showing “traffic jams” in business processes), and (c) by providing predictions regarding running cases (e.g., estimating the “arrival time” of a case that is delayed) [138].

15.2.1 Projecting Dynamic Information on Business Process Maps

The navigation systems of TomTom can be equipped with so-called “LIVE services” (cf. www.tomtom.com) showing traffic jams, mobile speed cameras, weather conditions, etc. This information is projected onto the map using current data.

In Chap. 9, we showed that a tight coupling between an event log and a process model can be used to extend process models with additional perspectives, e.g., highlighting bottlenecks, showing decision rules, and relating the process model to organizational entities. The same coupling can also be used to visualize “pre mortem” event data. Information about the current state of running cases can be projected onto the process model.

The idea is analogous to mashups using geo-tagging (e.g., Panoramio, HousingMaps, Funda, and Flickr). Many of these mashups use Google Maps. Consider, for example, the map shown in Fig. 15.8. Prospective customers can visit the site of Funda to look for a house that meets particular criteria. Information about houses that are for sale are projected onto a map. Figure 15.8 shows the houses that are for sale in Hapert. Figure 15.9 shows another example. Now the map shows traffic jams. Both maps are dynamic, i.e., the information projected onto these maps changes continuously.

Both historic and current event data can be used to “breathe life” into otherwise static business process maps. Similar to the visualization of traffic jams in Fig. 15.9, “traffic” in business processes can be visualized. Besides process maps, one can also think of other maps to project information on. Consider, for example, the social networks shown in Figs. 9.6 and 9.7. Work items waiting to be handled can be projected onto these models, e.g., cases that are waiting for a decision by a manager are projected onto the manager role. Some work items also have a geographic component, e.g., a field service engineer could be provided with a map like the one in Fig. 15.8 showing the devices that need maintenance. It is also possible to project work items onto maps with a temporal dimension (Gantt charts, agendas, etc.). For instance, a surgeon could view scheduled operations in his agenda. Hence, a variety of maps covering different perspectives can be used to visualize event related data.

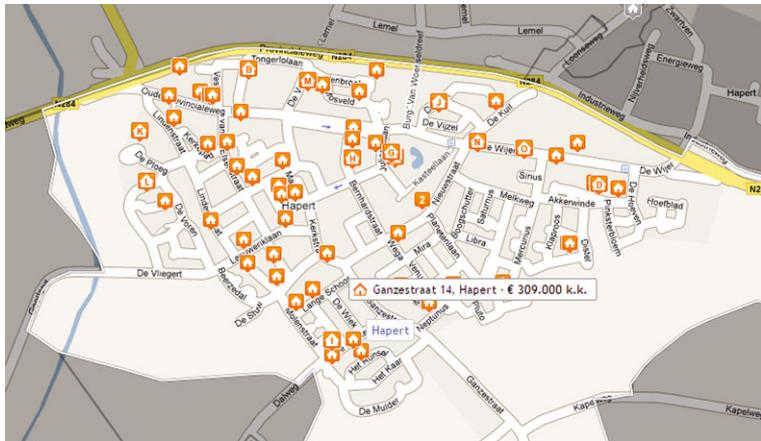


Fig. 15.8 Funda allows users to view maps with houses for sale that meet particular criteria, e.g., constraints related to size, volume, and pricing. The map shows the 53 houses for sale in the Dutch town Hapert



Fig. 15.9 Road map showing traffic jams: the car icons indicate problem spots and congested roads are highlighted. Modern navigation systems show such maps and, based on real-time traffic information, alternative routes are suggested to avoid bottlenecks

The YAWL system [41, 150] provides a visualization framework able to map pending work items and resources onto various maps, e.g., geographic maps, process maps, and organizational maps. YAWL also defines various distance notions

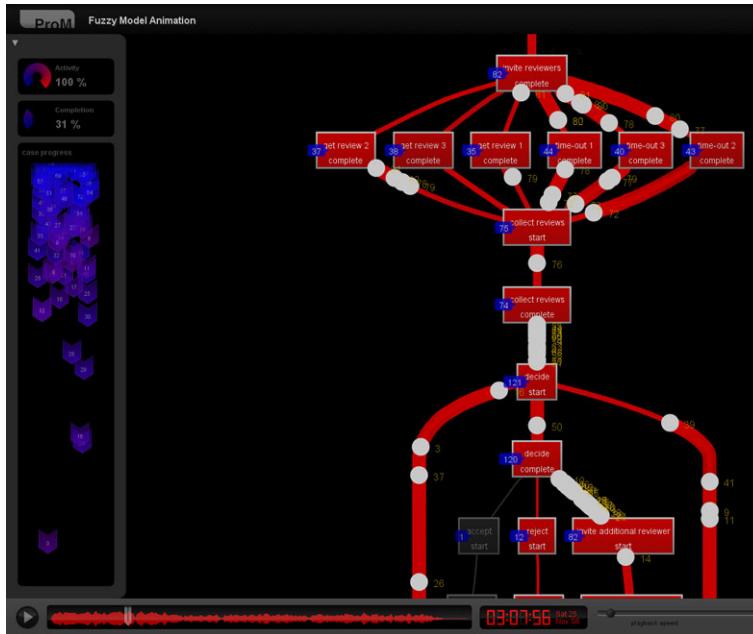


Fig. 15.10 The fuzzy model discovered earlier (cf. Fig. 15.7(a)) is used to replay the event log. The animation reveals the problem that many reviewers do not provide a report in time. As a result the editor of the journal cannot make a final decision and needs to invite additional reviewers. There is long queue of work items waiting for a decision and many pending invitations

based on these maps, for instance, a field service engineer can see the work items closest or most urgent.

From business process maps to business process movies

Once events in the log can be related to activities in the process model, it is possible to replay history on a case-by-case basis. This was used for conformance checking and model extension. Now we go one step further; we do not consider an individual case but all relevant cases at the same time. Assuming that events have a timestamp, all events in the log can be globally ordered, i.e., also events belonging to different cases can be sorted. After each event, the process is in a particular global state. One can think of this state as a *photograph* of the process. The state can be projected onto a business process map, a geographic map, or an organizational map. Since such a photograph is available after each event, it is also possible to create a *movie* by simply showing one photograph after another. Hence, it is possible to use event logs to create a “business process movie”. Fig. 15.10 shows an example using ProM’s Fuzzy Miner [65, 66]. The event log and the fuzzy model are converted into

an animation. The dots visible in Fig. 15.10 are moving along the arcs and refer to real cases. Such a business process movie provides a very convincing means to show problems in the as-is process. Unlike simulation, the animation shows reality and people cannot dismiss the outcomes by questioning the model. Therefore, business process movies help to expose the real problems in an organization.

15.2.2 Arrival Time Prediction

Whereas a TomTom device is continuously showing the *expected arrival time*, users of today's information systems are often left clueless about likely outcomes of the cases they are working on. This is surprising as many information systems gather a lot of historic information, thus providing an excellent basis for all kinds of predictions (expected completion time, likelihood of some undesirable outcome, estimated costs, etc.). Fortunately, as shown in Sect. 10.4, event logs can be used build predictive models.

The annotated transition system [164, 167] described in Sect. 10.4 can be used to predict the remaining flow time of a running case. The transition system is constructed using an event log L and a state representation function $l^{state}()$ or obtained by computing the state-space of a (discovered) process model. By systematically replaying the event log, the states are annotated with historic measurements. The mean or median of these historic measurements can be used to make predictions for running cases in a particular state. Each time the state of a case changes, a new prediction is made for the remaining flow time. Clearly, this functionality is similar to the prediction capabilities of a navigation device. Moreover, using different annotations, other kinds of predictions can be made. For instance, the transition system can be annotated with cost information to predict the total or remaining costs. Similarly, the outcome of a process or occurrence of an activity can be predicted.

Alternative approaches based on regression analysis, short-term simulation, or decision tree learning can be used to predict properties such as the remaining flow time of a running case. This illustrates that process mining can be used to extend information systems with predictive analytics.

15.2.3 Guidance Rather than Control

Car navigation systems provide *directions* and *guidance without controlling* the driver. The driver is still in control, but, given a goal (e.g. to get from A to B as fast as possible), the navigation system recommends the next action to be taken. In Sect. 10.5 we showed that predictions can be turned into recommendations. Recommendations are given with respect to a goal, e.g., to minimize costs, to minimize

the remaining flow time, or to maximize the likelihood of success. Such a goal is operationalized by defining a performance indicator that needs to be minimized or maximized. For every possible next action, the value of the performance indicator is predicted. This information is used to rank the possible actions and thus recommend the next step to be taken (cf. Fig. 10.12).

Recommendations based on process mining allow for systems that are flexible but also supporting operational decision making. Today's information systems typically do not provide a good balance between flexibility and support. The system is either restricting people in their actions or not providing any guidance. BPM systems offering more flexibility (e.g., case handling systems like BPM|one or declarative workflow systems like Declare), can be extended with a recommendation service based on process mining techniques [164].

The TomTom metaphor illustrates that many information systems lack functionality present in today's navigation devices [138]. However, high-quality process models tightly coupled to event logs enable TomTom-like functionalities such as predicting the "arrival time" of a process instance, recommending the next activity to be executed, and visualizing "traffic jams" in business processes.

Chapter 16

Epilogue

To conclude this book we summarize the main reasons for using process mining. Process mining can be seen as the “missing link” between data mining and traditional model-driven BPM. Although mature process mining techniques and tools are available, several challenges remain to further improve the applicability of the techniques presented in the preceding chapters. Therefore, we list the most important challenges. Finally, we encourage the reader to start using process mining today. For organizations that store event data in some form, the threshold to get started is really low.

16.1 Process Mining as a Bridge Between Data Mining and Business Process Management

Process mining is an important tool for modern organizations that need to manage non-trivial operational processes. On the one hand, there is an incredible growth of event data. On the other hand, processes and information need to be aligned perfectly in order to meet requirements related to compliance, efficiency, and customer service. The digital universe and the physical universe are amalgamating into one universe where events are recorded as they happen and processes are guided and controlled based on event data.

Part I of the book positioned process mining in the context of data science and introduced process mining as a new technology complementing existing approaches. The data scientist of the future needs to be able to analyze processes. This necessitates the process-centric “line of attack” presented in this book.

In *Part II*, we presented the two main disciplines that process mining is building on: Business Process Management (BPM) and data mining. Chapter 3 introduced several process modeling techniques and discussed the role of process models in the context of BPM. In Chap. 4, we introduced some of the basic data mining techniques.

Classical BPM approaches use process models as static descriptions or to drive a BPM/WFM system. If process models are just descriptive, they tend to be informal

and of low quality (i.e., not describing reality well). If models are used to configure a BPM/WFM system, they tend to force people to work in a particular manner. Data mining techniques aim to describe and understand reality based on historic data. However, most data mining techniques are *not* process-centric. Fortunately, process mining provides a link between both disciplines. Like other BPM approaches, process mining is process-centric. However, unlike most BPM approaches, it is driven by factual event data rather than hand-made models. Hence, process mining can be seen as a bridge between the preliminaries presented in Chaps. 3 and 4.

In *Part III*, we focused on the most challenging process mining task: *process discovery*. First, we discussed the input needed for process mining (Chap. 5). Then, we presented a very basic algorithm (Chap. 6) followed by an overview of more powerful process discovery techniques (Chap. 7). Unlike basic data mining techniques such as decision tree and association rule learning, process discovery problems are characterized by a complex search space as is illustrated by the many workflow patterns. Whereas the aim of many data mining techniques is to be able to deal with many records or many variables, the main challenge of process discovery is to adequately capture behavioral aspects.

Process mining is not limited to process discovery. In fact, process discovery is just one of many process mining tasks. Therefore, *Part IV* expanded the scope of process mining into several directions. These expansions have in common that the event log and the process model are tightly coupled, thus allowing for new forms of analysis and support. Chapter 8 presented various conformance checking techniques. As shown in Chap. 9, the organizational perspective, the case perspective, and the time perspective can be added to discovered process models or used to create complementary models. Recommendations and predictions (based on a combination of historic event data and partial traces of running cases) are examples of the operational support functionalities described in Chap. 10. Chapters 8, 9, and 10 illustrate the breadth of the process mining spectrum.

In *Part V*, we shifted the focus to software, scalability, and real-life applications. Chapter 11 elaborated on tool support for process mining. Next to ProM, it discusses 11 commercial process mining products. Chapter 12 discussed a range of approaches to deal with extremely large data sets. In Chaps. 13 and 14, we elaborated on two characteristic types of processes (“Lasagna processes” and “Spaghetti processes”) and showed how process mining can add value.

In this last part (*Part VI*), we started by taking a step back and reflected on the material presented in the preceding parts. In Chap. 15, we compared business process models, business process analysis, and business process support with geographic maps and navigation systems. This comparison revealed limitations of current BPM practices and confirmed the potential of process mining to “breathe life” into process models. Process mining provides not only a bridge between data mining and BPM; it also helps to address the classical divide between “business” and “IT”. IT people tend to have a technology-oriented focus with little consideration for the actual business processes that need to be supported. People focusing on the “business-side” of BPM are typically not interested in technological advances and the precise functionality of information systems. The empirical nature of process mining can bring both

groups of people together. Evidence-based BPM based on process mining helps to create a common ground for business process improvement and information systems development.

Learning more about process mining

The interested reader can take the online course *Process Mining: Data science in Action* based on this book and offered via Coursera. This Massive Open Online Course (MOOC) explains the key analysis techniques in process mining. Participants can learn various process discovery algorithms. These are used to automatically learn process models from raw event data. Various other process analysis techniques that use event data are presented. Moreover, the MOOC provides easy-to-use software, real-life data sets, and practical skills to directly apply the theory in a variety of application domains. Visit <https://www.coursera.org/course/procmin> for more information on the course and www.processmining.org for additional information.

16.2 Challenges

Existing process mining techniques and tools such as ProM are mature and can be applied to both Lasagna and Spaghetti processes. We have applied ProM in more than 150 organizations ranging from municipalities and hospitals to financial institutions and manufacturers of high-tech systems. Despite the applicability of process mining there are many interesting challenges; these illustrate that process mining is a young discipline.

Process discovery is probably the most important and most visible intellectual challenge related to process mining. As shown, it is far from trivial to construct a process model based on *event logs that are incomplete and noisy*. Unfortunately, there are still researchers and tool vendors that assume logs to be complete and free of noise. Although heuristic mining, genetic mining, and fuzzy mining (cf. Chap. 7) provide case-hardened process discovery techniques, many improvements are possible to construct more intuitive *80/20 models*, i.e., simple models that are able to explain the most likely/common behavior. Recently developed inductive mining approaches seem to provide a good basis for the next generation of process discovery techniques.

New process mining approaches should reconsider the *representational bias* to be used. Many approaches use a graph-based notation allowing for models that do not make much sense (deadlocks, disconnected parts, etc.). WF-nets, BPMN models, EPCs, etc. can represent processes that are not sound, e.g., a process having a deadlock or an activity that can never be activated. The search space of a technique using such a representational bias is too large. For instance, the α -algorithm can discover WF-nets that are not sound and the heuristic miner and the genetic miner can discover C-nets that deadlock. Approaches using process trees, in particular the

inductive mining approaches described in Chap. 7, do not suffer from this problem. However, process trees have difficulties expressing certain process constructs and existing techniques fail to duplicate activities when needed. Hence, there is still room for improvement.

Another challenge is the notion of *concept drift*, i.e., processes change while being observed. Existing process discovery approaches do not take such changes into account. It is interesting to detect when processes change and to visualize such changes.

Alignments are a powerful tool to relate modeled and observed behavior (see Sect. 8.3). However, conformance checking is not well supported by today's commercial process mining tools. Moreover, computing alignments is very time consuming compared to most forms of process discovery. Hence, there is a need for better performing conformance checking techniques.

Process mining heavily depends to the ability to extract suitable event logs. The scope and granularity of an event log should match the questions one would like to answer. Unfortunately, in some information systems event data are just a byproduct for debugging or scattered over many tables. Some systems also “forget” events, e.g., when a record is updated, the old values are simply overwritten. Earlier we used the term *business process provenance* to stress the importance of recording events in such a way that history is recorded correctly and cannot be tampered with. Event logs should be “first-class citizens” rather than some byproduct. Data elements in event logs should have clear *semantics*. Therefore, developers should not simply insert write statements without a reference to a commonly agreed-upon *ontology*. We encountered systems where parts of the logging depend on the language setting. For example, depending on the language setting of the system, an event attribute may have value “Off” in English, “Uit” in Dutch, or “Aus” in German. Semantically, these are all the same. However, such ad-hoc logging is making analysis more complex. Attributes of events and cases should refer to one or more ontologies that clearly define concepts and possible attribute values. Logging formats such as XES and SA-MXML (cf. Chap. 5) can relate event data to ontologies. However, the challenge is to make sure that organizations actually start producing semantically annotated event data.

Another challenge is produce process models that have a quality and understandability comparable to geographic maps. As shown in Chap. 15, we can learn many lessons from cartography.

Process mining can be used off-line and online. For off-line process mining, only historic (“post mortem”) data is needed and no tight coupling between the process mining software and existing enterprise information systems is needed. For online process mining (e.g., providing predictions and recommendations), operational support capabilities need to be embedded in enterprise information systems. From a technological point of view this may be challenging. It is difficult to embed such advanced functionality in legacy systems. Moreover, online process mining typically requires additional computing power. It is important to overcome these challenges as the value of operational support based on process mining is evident (cf. Chap. 10). For example, a process model showing the current status of running cases is much more interesting than a static process model not showing any “live data”.

Responsible process mining

Big Data is changing the way we do business, socialize, conduct research, and govern society. In today's society, event data are collected about anything, at any time, and at any place. Today's process mining tools are able to analyze such data and can handle event logs with billions of events by exploiting modern IT infrastructures. These amazing capabilities also imply a great responsibility. *Fairness, confidentiality, accuracy and transparency should be key concerns for any process miner.* There are foundational questions related to these concerns:

- *How to avoid unfair conclusions even if they are true?* Process mining should not be misused by management. People that are deviating or that delay the process may do this for good reasons. Management should show an interest in “positive deviants” and not blame individuals handling the difficult cases.
- *How to answer questions without revealing secrets?* Avoid the (un)intended leakage of information, for example, by using randomization or hashing. Most questions can be answered without revealing sensitive information.
- *How to answer questions with a guaranteed level of accuracy?* The curse of dimensionality and overfitting may lead to bogus results. The process mining tool will always return results (e.g., a model). Hence, the analyst always needs to assess and communicate the confidence level. Conformance checking and cross-validation are key.
- *How to clarify answers such that they become indisputable?* Results need to be explainable and traceable. Process mining should not be a black box, but provide insights understandable by humans. Analysis workflows should be reproducible by others.

It is important that process mining is done in a responsible manner. Moreover, one should look for *positive* ways to ensure fairness, confidentiality, accuracy and transparency. Simply blocking the use of event data will prevent any form of process improvement from happening.

16.3 Start Today!

As demonstrated in this book, process mining can be brought into play for many different purposes. Process mining can be used to diagnose the actual processes. This is valuable because in many organizations most stakeholders lack a correct, objective, and accurate view on important operational processes. Process mining can subsequently be used to improve such processes. Conformance checking can be used for auditing and compliance. By replaying the event log on a process model it is possible to quantify and visualize deviations. Similar techniques can be used to detect bottlenecks and build predictive models. Given the applicability of process

mining, we hope that this book encourages the reader to start using process mining *today*.

Getting example data

Lots of data sets are available via the websites such as www.processmining.org, <http://www.win.tue.nl/ieeetfpm>, and http://data.3tu.nl/repository/collection:event_logs. The last link refers to the repository of the 3TU data center. This site provides persistent data sets. Each event log has so called Digital Object identifier (DOI) reference. Clicking on a DOI reference in an article, report or website will seamlessly provide access to the event log. The collection of event logs includes several real-life logs, including the event logs used for the annual *BPI challenge*.

The threshold to start an off-line process mining project is really low. Most organizations have event data hidden in their systems. Once the data is located, conversion is typically easy. Most tools support XES, but can also read CSV files or access databases via JDBC. The freely available open-source process mining tool ProM can be downloaded from www.processmining.org. ProM can be applied to any MXML or XES file and supports all of the process mining techniques mentioned in the preceding chapters. After reading this book, installing the software, and extracting the event data, the reader is able experience the “magic” of process mining, i.e., discovering and improving processes based on facts rather than fiction.

References

1. ACSI. Artifact-Centric Service Interoperation (ACSI) Project Home Page. www.acsi-project.eu.
2. A. Adriansyah. *Aligning Observed and Modeled Behavior*. Phd thesis, Eindhoven University of Technology, April 2014.
3. A. Adriansyah, B. van Dongen, and W.M.P. van der Aalst. Conformance Checking using Cost-Based Fitness Analysis. In C.H. Chi and P. Johnson, editors, *IEEE International Enterprise Computing Conference (EDOC 2011)*, pages 55–64. IEEE Computer Society, 2011.
4. A. Adriansyah, B.F. van Dongen, and W.M.P. van der Aalst. Towards Robust Conformance Checking. In M. zur Muehlen and J. Su, editors, *BPM 2010 Workshops, Proceedings of the Sixth Workshop on Business Process Intelligence (BPI2010)*, volume 66 of *Lecture Notes in Business Information Processing*, pages 122–133. Springer, Berlin, 2011.
5. A. Adriansyah, J. Munoz-Gama, J. Carmona, B.F. van Dongen, and W.M.P. van der Aalst. Measuring Precision of Modeled Behavior. *Information Systems and e-Business Management*, **13**(1):37–67, 2015.
6. C. Aggarwal. *Data Streams: Models and Algorithms*, volume 31 of *Advances in Database Systems*. Springer, Berlin, 2007.
7. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pages 487–499, Santiago de Chile, Chile, 1994. Morgan Kaufmann Publishers Inc.
8. R. Agrawal, D. Gunopulos, and F. Leymann. Mining Process Models from Workflow Logs. In *Sixth International Conference on Extending Database Technology*, volume 1377 of *Lecture Notes in Computer Science*, pages 469–483. Springer, Berlin, 1998.
9. E. Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, 2010.
10. A.K. Alves de Medeiros. *Genetic Process Mining*. Phd thesis, Eindhoven University of Technology, 2006.
11. A.K. Alves de Medeiros, W.M.P. van der Aalst, and A.J.M.M. Weijters. Workflow Mining: Current Status and Future Directions. In R. Meersman, Z. Tari, and D.C. Schmidt, editors, *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, pages 389–406. Springer, Berlin, 2003.
12. A.K. Alves de Medeiros, A.J.M.M. Weijters, and W.M.P. van der Aalst. Genetic Process Mining: An Experimental Evaluation. *Data Mining and Knowledge Discovery*, **14**(2):245–304, 2007.
13. A.K. Alves de Medeiros, A. Guzzo, G. Greco, W.M.P. van der Aalst, A.J.M.M. Weijters, B. van Dongen, and D. Sacca. Process Mining Based on Clustering: A Quest for Precision. In A. ter Hofstede, B. Benatallah, and H.Y. Paik, editors, *BPM 2007 International Workshops (BPI, BPD, CBP, ProHealth, RefMod, Semantics4ws)*, volume 4928 of *Lecture Notes in Computer Science*, pages 17–29. Springer, Berlin, 2008.

14. A.K. Alves de Medeiros, W.M.P. van der Aalst, and A.J.M.M. Weijters. Quantifying Process Equivalence Based on Observed Behavior. *Data and Knowledge Engineering*, **64**(1):55–74, 2008.
15. D. Angluin and C.H. Smith. Inductive Inference: Theory and Methods. *Computing Surveys*, **15**(3):237–269, 1983.
16. E. Badouel and P. Darondeau. Theory of Regions. In W. Reisig and G. Rozenberg, editors, *Lectures on Petri Nets I: Basic Models*, volume 1491 of *Lecture Notes in Computer Science*, pages 529–586. Springer, Berlin, 1998.
17. B. Baessens. *Analytics in a Big Data World*. Wiley, Hoboken, New Jersey, 2014.
18. J. Becker, M. Kugeler, and M. Rosemann, editors. *Process Management: A Guide for the Design of Business Processes*, International Handbooks on Information Systems. Springer, Berlin, 2011.
19. R. Bergenthal, J. Desel, R. Lorenz, and S. Mauser. Process Mining Based on Regions of Languages. In G. Alonso, P. Dadam, and M. Rosemann, editors, *International Conference on Business Process Management (BPM 2007)*, volume 4714 of *Lecture Notes in Computer Science*, pages 375–383. Springer, Berlin, 2007.
20. A.W. Biermann. On the Inference of Turing Machines from Sample Computations. *Artificial Intelligence*, **3**:181–198, 1972.
21. A.W. Biermann and J.A. Feldman. On the Synthesis of Finite-State Machines from Samples of their Behavior. *IEEE Transaction on Computers*, **21**:592–597, 1972.
22. A. Bolt and W.M.P. van der Aalst. Multidimensional Process Mining Using Process Cubes. In K. Gaaloul, R. Schmidt, S. Nurcan, S. Guerreiro, and Q. Ma, editors, *Enterprise, Business-Process and Information Systems Modeling (BPMDS 2015)*, volume 214 of *Lecture Notes in Business Information Processing*, pages 102–116. Springer, Berlin, 2015.
23. A. Bolt, M. de Leoni, and W.M.P. van der Aalst. Scientific Workflows for Process Mining: Building Blocks, Scenarios, and Implementation. *International Journal on Software Tools for Technology Transfer*, 2016.
24. M. Bramer. *Principles of Data Mining*. Springer, Berlin, 2007.
25. L. Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3):199–231, 2001.
26. J.C.A.M. Buijs, B.F. van Dongen, and W.M.P. van der Aalst. Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity. *International Journal of Cooperative Information Systems*, **23**(1):1–39, 2014.
27. A. Burattin, A. Sperduti, and W.M.P. van der Aalst. Control-Flow Discovery from Event Streams. In *IEEE Congress on Evolutionary Computation (CEC 2014)*, pages 2420–2427. IEEE Computer Society, 2014.
28. J. Carmona and J. Cortadella. Process Mining Meets Abstract Interpretation. In J.L. Balcazar, editor, *ECML/PKDD 210*, volume 6321 of *Lecture Notes in Artificial Intelligence*, pages 184–199. Springer, Berlin, 2010.
29. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. CRISP-DM 1.0: Step-by-step data mining guide. www.crisp-dm.org, 2000.
30. E.M. Clarke, O. Grumberg, and D.A. Peled. *Model Checking*. The MIT Press, Cambridge, Massachusetts and London, UK, 1999.
31. B.D. Clinton and A. van der Merwe. Management Accounting: Approaches, Techniques, and Management Processes. *Cost Management*, **20**(3):14–22, 2006.
32. E.F. Codd. A Relational Model for Large Shared Data Banks. *Communications of the ACM*, **13**(6):377–387, 1970.
33. J.E. Cook and A.L. Wolf. Discovering Models of Software Processes from Event-Based Data. *ACM Transactions on Software Engineering and Methodology*, **7**(3):215–249, 1998.
34. J. Cortadella, M. Kishinevsky, L. Lavagno, and A. Yakovlev. Deriving Petri Nets from Finite Transition Systems. *IEEE Transactions on Computers*, **47**(8):859–882, 1998.
35. CoSeLoG. Configurable Services for Local Governments (CoSeLoG) Project Home Page. www.win.tue.nl/coselog.

36. D. Court, D. Elzinga, S. Mulder, and O.J. Vetzik. The Consumer Decision Journey. *McKinsey Quarterly*, **3**:1–9, 2009.
37. T. Curran and G. Keller. *SAP R/3 Business Blueprint: Understanding the Business Process Reference Model*. Upper Saddle River, 1997.
38. A. Datta. Automating the Discovery of As-Is Business Process Models: Probabilistic and Algorithmic Approaches. *Information Systems Research*, **9**(3):275–301, 1998.
39. S. Davidson, S. Cohen-Boulakia, A. Eyal, B. Ludaescher, T. McPhillips, S. Bowers, M. Anand, and J. Freire. Provenance in Scientific Workflow Systems. *Data Engineering Bulletin*, **30**(4):44–50, 2007.
40. M. De Leoni and W.M.P. van der Aalst. Data-Aware Process Mining: Discovering Decisions in Processes Using Alignments. In S.Y. Shin and J.C. Maldonado, editors, *ACM Symposium on Applied Computing (SAC 2013)*, pages 1454–1461. ACM Press, 2013.
41. M. de Leoni, W.M.P. van der Aalst, and A.H.M. ter Hofstede. Visual Support for Work Assignment in Process-Aware Information Systems: Framework Formalisation and Implementation. *Decision Support Systems*, **54**(1):345–361, 2012.
42. M. de Leoni, W.M.P. van der Aalst, and M. Dees. A General Process Mining Framework for Correlating, Predicting and Clustering Dynamic Behavior Based on Event Logs. *Information Systems*, **56**:235–257, 2016.
43. J. De Weerdt, M. De Backer, J. Vanthienen, and B. Baesens. A Multi-Dimensional Quality Assessment of State-of-the-Art Process Discovery Algorithms Using Real-Life Event Logs. *Information Systems*, **37**(7):654–676, 2012.
44. J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, **51**(1):107–113, 2008.
45. J. Desel and J. Esparza. *Free Choice Petri Nets*, volume 40 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, Cambridge, UK, 1995.
46. J. Desel, W. Reisig, and G. Rozenberg, editors. *Lectures on Concurrency and Petri Nets*, volume 3098 of *Lecture Notes in Computer Science*. Springer, Berlin, 2004.
47. P.C. Diniz and D.R. Ferreira. Automatic Extraction of Process Control Flow from I/O Operations. In M. Dumas, M. Reichert, and M.C. Shan, editors, *Business Process Management (BPM 2008)*, volume 5240 of *Lecture Notes in Computer Science*, pages 342–357. Springer, Berlin, 2008.
48. D. Donoho. 50 years of Data Science. Technical report, Stanford University, September 2015. Based on a presentation at the Tukey Centennial workshop, Princeton NJ, Sept. 18 2015.
49. M. Dumas, W.M.P. van der Aalst, and A.H.M. ter Hofstede. *Process-Aware Information Systems: Bridging People and Software through Process Technology*. Wiley & Sons, 2005.
50. M. Dumas, M. La Rosa, J. Mendling, and H. Reijers. *Fundamentals of Business Process Management*. Springer, Berlin, 2013.
51. A. Ehrenfeucht and G. Rozenberg. Partial (Set) 2-Structures – Part 1 and Part 2. *Acta Informatica*, **27**(4):315–368, 1989.
52. D. Fahland and W.M.P. van der Aalst. Model Repair: Aligning Process Models to Reality. *Information Systems*, **47**:220–243, 2015.
53. D.R. Ferreira and D. Gillblad. Discovering Process Models from Unlabelled Event Logs. In U. Dayal, J. Eder, J. Koehler, and H. Reijers, editors, *Business Process Management (BPM 2009)*, volume 5701 of *Lecture Notes in Computer Science*, pages 143–158. Springer, Berlin, 2009.
54. S. Finlay. *Predictive Analytics, Data Mining, and Big Data: Myths, Misconceptions and Methods*. Palgrave Macmillan, Hampshire, UK, 2014.
55. Forrester. The Forrester Wave: Enterprise Business Intelligence Platforms (Q4 2010). www.forrester.com, 2010.
56. Gartner. Magic Quadrant for Business Intelligence and Analytics Platforms. www.gartner.com, 2015.
57. G.M. Gavetti, R. Henderson, and S. Giorgi. Kodak and The Digital Revolution. Harvard Business School Case 705-448 (Revised November 2005), 2004.

58. S. Ghemawat and S.T. Leung H. Gobioff. The Google File System. *ACM SIGOPS Operating Systems Review*, **37**(5):29–43, 2003.
59. S. Goedertier, D. Martens, B. Baesens, R. Haesen, and J. Vanthienen. Process Mining as First-Order Classification Learning on Logs with Negative Events. In A. ter Hofstede, B. Benatallah, and H.Y. Paik, editors, *BPM 2007 International Workshops (BPI, BPD, CBP, ProHealth, ReMod, Semantics4ws)*, volume 4928 of *Lecture Notes in Computer Science*, pages 42–53. Springer, Berlin, 2008.
60. S. Goedertier, D. Martens, J. Vanthienen, and B. Baesens. Robust Process Discovery with Artificial Negative Events. *Journal of Machine Learning Research*, **10**:1305–1340, 2009.
61. E.M. Gold. Language Identification in the Limit. *Information and Control*, **10**(5):447–474, 1967.
62. G. Greco, A. Guzzo, L. Pontieri, and D. Saccà. Discovering Expressive Process Models by Clustering Log Traces. *IEEE Transaction on Knowledge and Data Engineering*, **18**(8):1010–1027, 2006.
63. P.D. Grünwald. *Minimum Description Length Principle*. MIT press, Cambridge, MA, 2007.
64. C.W. Günther. XES Standard Definition. www.xes-standard.org, 2009.
65. C.W. Günther. *Process Mining in Flexible Environments*. Phd thesis, Eindhoven University of Technology, September 2009.
66. C.W. Günther and W.M.P. van der Aalst. Fuzzy Mining: Adaptive Process Simplification Based on Multi-perspective Metrics. In G. Alonso, P. Dadam, and M. Rosemann, editors, *International Conference on Business Process Management (BPM 2007)*, volume 4714 of *Lecture Notes in Computer Science*, pages 328–343. Springer, Berlin, 2007.
67. C.W. Günther, A. Rozinat, W.M.P. van der Aalst, and K. van Uden. Monitoring Deployed Application Usage with Process Mining. BPM Center Report BPM-08-11, BPMcenter.org, 2008.
68. M. Hammer and J. Champy. *Reengineering the corporation*. Nicolas Brealey Publishing, London, 1993.
69. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT press, Cambridge, MA, 2001.
70. D. Harel and R. Marelly. *Come, Let's Play: Scenario-Based Programming Using LSCs and the Play-Engine*. Springer, Berlin, 2003.
71. J. Herbst. A Machine Learning Approach to Workflow Management. In *Proceedings 11th European Conference on Machine Learning*, volume 1810 of *Lecture Notes in Computer Science*, pages 183–194. Springer, Berlin, 2000.
72. J. Herbst. *Ein induktiver Ansatz zur Akquisition und Adaption von Workflow-Modellen*. PhD thesis, Universität Ulm, November 2001.
73. M. Hilbert and P. Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, **332**(6025):60–65, 2011.
74. B. Hompes, E. Verbeek, and W.M.P. van der Aalst. Finding Suitable Activity Clusters for Decomposed Process Discovery. In P. Ceravolo, B. Russo, and R. Accorsi, editors, *IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2014)*, volume 237 of *Lecture Notes in Business Information Processing*, pages 32–57. Springer, Berlin, 2015.
75. IEEE Task Force on Process Mining. Process Mining Manifesto. In F. Daniel, K. Barkaoui, and S. Dustdar, editors, *Business Process Management Workshops*, volume 99 of *Lecture Notes in Business Information Processing*, pages 169–194. Springer, Berlin, 2012.
76. S. Jablonski and C. Bussler. *Workflow Management: Modeling Concepts, Architecture, and Implementation*. International Thomson Computer Press, London, UK, 1996.
77. R.P. Jagadeesh Chandra Bose and W.M.P. van der Aalst. Abstractions in Process Mining: A Taxonomy of Patterns. In U. Dayal, J. Eder, J. Koehler, and H. Reijers, editors, *Business Process Management (BPM 2009)*, volume 5701 of *Lecture Notes in Computer Science*, pages 159–175. Springer, Berlin, 2009.
78. R.P. Jagadeesh Chandra Bose and W.M.P. van der Aalst. Context Aware Trace Clustering: Towards Improving Process Mining Results. In H. Liu and Z. Obradovic, editors, *Proceed-*

- ings of the SIAM International Conference on Data Mining (SDM 2009), pages 401–412. Society for Industrial and Applied Mathematics, 2009.
- 79. R.P. Jagadeesh Chandra Bose and W.M.P. van der Aalst. Trace Alignment in Process Mining: Opportunities for Process Diagnostics. In R. Hull, J. Mendling, and S. Tai, editors, *Business Process Management (BPM 2010)*, volume 6336 of *Lecture Notes in Computer Science*, pages 227–242. Springer, Berlin, 2010.
 - 80. R.P. Jagadeesh Chandra Bose, R. Mans, and W.M.P. van der Aalst. Wanna Improve Process Mining Results? It's High Time We Consider Data Quality Issues Seriously. In B. Hammer, Z.H. Zhou, L. Wang, and N. Chawla, editors, *IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2013)*, pages 127–134, Singapore, 2013. IEEE.
 - 81. R.P. Jagadeesh Chandra Bose, W.M.P. van der Aalst, I. Zliobaite, and M. Pechenizkiy. Dealing with Concept Drifts in Process Mining. *IEEE Transactions on Neural Networks and Learning Systems*, **25**(1):154–171, 2014.
 - 82. K. Jensen and L.M. Kristensen. *Coloured Petri Nets*. Springer, Berlin, 2009.
 - 83. D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age: Solving Problems with Visual Analytics*. VisMaster, <http://www.vismaster.eu/book/>, 2010.
 - 84. M. Kerremans. Automated Business Process Discovery Improves BPM Outcomes, Research Note G00164422. www.gartner.com, 2008.
 - 85. S.C. Kleene. Representation of Events in Nerve Nets and Finite Automata. In C.E. Shannon and J. McCarthy, editors, *Automata Studies*, pages 3–41. Princeton University Press, 1956.
 - 86. E. Lamma, P. Mello, M. Montali, F. Riguzzi, and S. Storari. Inducing Declarative Logic-Based Models from Labeled Traces. In G. Alonso, P. Dadam, and M. Rosemann, editors, *International Conference on Business Process Management (BPM 2007)*, volume 4714 of *Lecture Notes in Computer Science*, pages 344–359. Springer, Berlin, 2007.
 - 87. D. Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety (Research Note 949). Technical report, META Group, February 2001.
 - 88. S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Discovering Block-structured Process Models from Event Logs: A Constructive Approach. In J.M. Colom and J. Desel, editors, *Applications and Theory of Petri Nets 2013*, volume 7927 of *Lecture Notes in Computer Science*, pages 311–329. Springer, Berlin, 2013.
 - 89. S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In N. Lohmann, M. Song, and P. Wohed, editors, *Business Process Management Workshops, International Workshop on Business Process Intelligence (BPI 2013)*, volume 171 of *Lecture Notes in Business Information Processing*, pages 66–78. Springer, Berlin, 2014.
 - 90. S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Discovering Block-structured Process Models from Incomplete Event Logs. In G. Ciardo and E. Kindler, editors, *Applications and Theory of Petri Nets 2014*, volume 8489 of *Lecture Notes in Computer Science*, pages 91–110. Springer, Berlin, 2014.
 - 91. S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Scalable Process Discovery with Guarantees. In K. Gaaloul, R. Schmidt, S. Nurcan, S. Guerreiro, and Q. Ma, editors, *Enterprise, Business-Process and Information Systems Modeling (BPMDS 2015)*, volume 214 of *Lecture Notes in Business Information Processing*, pages 85–101. Springer, Berlin, 2015.
 - 92. F. Leymann and D. Roller. *Production Workflow: Concepts and Techniques*. Prentice-Hall PTR, Upper Saddle River, New Jersey, USA, 1999.
 - 93. Z. Manna and A. Pnueli. *The Temporal Logic of Reactive and Concurrent Systems: Specification*. Springer, New York, 1991.
 - 94. H. Mannila, H. Toivonen, and A.I. Verkamo. Discovery of Frequent Episodes in Event Sequences. *Data Mining and Knowledge Discovery*, **1**(3):259–289, 1997.
 - 95. R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker. Application of Process Mining in Healthcare: A Case Study in a Dutch Hospital. In *Biomedical Engineering Systems and Technologies*, volume 25 of *Communications in Computer and Information Science*, pages 425–438. Springer, Berlin, 2009.

96. R.S. Mans, N.C. Russell, W.M.P. van der Aalst, A.J. Moleman, and P.J.M. Bakker. Schedule-Aware Workflow Management Systems. In K. Jensen, S. Donatelli, and M. Koutny, editors, *Transactions on Petri Nets and Other Models of Concurrency IV*, volume 6550 of *Lecture Notes in Computer Science*, pages 121–143. Springer, Berlin, 2010.
97. R. Mans, W.M.P. van der Aalst, and E. Verbeek. Supporting Process Mining Workflows with RapidProM. In L. Limonad and B. Weber, editors, *Business Process Management Demo Sessions (BPMD 2014)*, volume 1295 of *CEUR Workshop Proceedings*, pages 56–60. [CEUR-WS.org](http://ceur-ws.org), 2014.
98. R. Mans, W.M.P. van der Aalst, and R. Vanwersch. *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*. Springer Briefs in Business Process Management. Springer, Berlin, 2015.
99. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2011.
100. V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Like, Work, and Think*. Houghton Mifflin Harcourt, Boston, 2013.
101. J. Mendling, G. Neumann, and W.M.P. van der Aalst. Understanding the Occurrence of Errors in Process Models Based on Metrics. In F. Curbera, F. Leymann, and M. Weske, editors, *Proceedings of the OTM Conference on Cooperative information Systems (CoopIS 2007)*, volume 4803 of *Lecture Notes in Computer Science*, pages 113–130. Springer, Berlin, 2007.
102. T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
103. M. Montali, M. Pesic, W.M.P. van der Aalst, F. Chesani, P. Mello, and S. Storari. Declarative Specification and Verification of Service Choreographies. *ACM Transactions on the Web*, **4**(1):1–62, 2010.
104. G.E. Moore. Cramming More Components Onto Integrated Circuits. *Electronics*, pages 114–117, April 1965.
105. J. Munoz-Gama and J. Carmona. Enhancing Precision in Process Conformance: Stability, Confidence and Severity. In N. Chawla, I. King, and A. Sperduti, editors, *IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011)*, pages 184–191, Paris, France, April 2011. IEEE.
106. J. Munoz-Gama, J. Carmona, and W.M.P. van der Aalst. Single-Entry Single-Exit Decomposed Conformance Checking. *Information Systems*, **46**:102–122, 2014.
107. P. Naur. *Concise Survey of Computer Methods*. Studentlitteratur Lund, Akademisk Forlag, Copenhagen, 1974.
108. A. Nerode. Linear Automaton Transformations. *Proceedings of the American Mathematical Society*, **9**(4):541–544, 1958.
109. T. Ohno. *Toyota Production System: Beyond Large-Scale Production*. Productivity Press, 1988.
110. OMG. Business Process Model and Notation (BPMN). Object Management Group, dtc/2010-06-05, 2010.
111. C.A. Petri. *Kommunikation mit Automaten*. PhD thesis, Institut für instrumentelle Mathematik, Bonn, 1962.
112. G. Press. A Very Short History of Data Science. Forbes Technology, <http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/>, 2013.
113. T. Pyzdek. *The Six Sigma Handbook: A Complete Guide for Green Belts, Black Belts, and Managers at All Levels*. McGraw Hill, New York, 2003.
114. A. Rajaraman and J.D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2011.
115. M. Reichert and B. Weber. *Enabling Flexibility in Process-Aware Information Systems: Challenges, Methods, Technologies*. Springer, Berlin, 2012.
116. H.A. Reijers and W.M.P. van der Aalst. The Effectiveness of Workflow Management Systems: Predictions and Lessons Learned. *International Journal of Information Management*, **25**(5):458–472, 2005.

117. W. Reisig and G. Rozenberg, editors. *Lectures on Petri Nets I: Basic Models*, volume 1491 of *Lecture Notes in Computer Science*. Springer, Berlin, 1998.
118. M. Rosemann, J. Recker, and C. Flender. Contextualisation of Business Processes. *International Journal of Business Process Integration and Management*, **3**(1):47–60, 2008.
119. A. Rozinat. *Process Mining: Conformance and Extension*. Phd thesis, Eindhoven University of Technology, November 2010.
120. A. Rozinat and W.M.P. van der Aalst. Decision Mining in ProM. In S. Dustdar, J.L. Fiadeiro, and A. Sheth, editors, *International Conference on Business Process Management (BPM 2006)*, volume 4102 of *Lecture Notes in Computer Science*, pages 420–425. Springer, Berlin, 2006.
121. A. Rozinat and W.M.P. van der Aalst. Conformance Checking of Processes Based on Monitoring Real Behavior. *Information Systems*, **33**(1):64–95, 2008.
122. A. Rozinat, A.K. Alves de Medeiros, C.W. Günther, A.J.M.M. Weijters, and W.M.P. van der Aalst. The Need for a Process Mining Evaluation Framework in Research and Practice. In A. ter Hofstede, B. Benatallah, and H.Y. Paik, editors, *BPM 2007 International Workshops (BPI, BPD, CBP, ProHealth, RefMod, Semantics4ws)*, volume 4928 of *Lecture Notes in Computer Science*, pages 84–89. Springer, Berlin, 2008.
123. A. Rozinat, I.S.M. de Jong, C.W. Günther, and W.M.P. van der Aalst. Process Mining Applied to the Test Process of Wafer Scanners in ASML. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, **39**(4):474–479, 2009.
124. A. Rozinat, R.S. Mans, M. Song, and W.M.P. van der Aalst. Discovering Simulation Models. *Information Systems*, **34**(3):305–327, 2009.
125. A. Rozinat, M. Wynn, W.M.P. van der Aalst, A.H.M. ter Hofstede, and C. Fidge. Workflow Simulation for Operational Decision Support. *Data and Knowledge Engineering*, **68**(9):834–850, 2009.
126. A.W. Scheer. *Business Process Engineering: Reference Models for Industrial Enterprises*. Springer, Berlin, 1994.
127. H. Smith and P. Fingar. *Business Process Management: The Third Wave*. Meghan Kiffer Press, 2006.
128. M. Sole and J. Carmona. Process Mining from a Basis of Regions. In J. Lilius and W. Penczek, editors, *Applications and Theory of Petri Nets 2010*, volume 6128 of *Lecture Notes in Computer Science*, pages 226–245. Springer, Berlin, 2010.
129. M. Song and W.M.P. van der Aalst. Supporting Process Mining by Showing Events at a Glance. In K. Chari and A. Kumar, editors, *Proceedings of 17th Annual Workshop on Information Technologies and Systems (WITS 2007)*, pages 139–145, Montreal, Canada, 2007.
130. M. Song and W.M.P. van der Aalst. Towards Comprehensive Support for Organizational Mining. *Decision Support Systems*, **46**(1):300–317, 2008.
131. R. Srikant and R. Agrawal. Mining Sequential Patterns: Generalization and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology (EDBT '96)*, pages 3–17, 1996.
132. A.H.M. ter Hofstede, W.M.P. van der Aalst, M. Adams, and N. Russell. *Modern Business Process Automation: YAWL and its Support Environment*. Springer, Berlin, 2010.
133. J.W. Tukey. The Future of Data Analysis. *Annals of Mathematical Statistics*, **33**(1):1–67, 1962.
134. V. Turner, J.F. Gantz, D. Reinsel, and S. Minton. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. International Data Corporation, Framingham, MA, USA, 2014. <http://www.emc.com/leadership/digital-universe/>.
135. A. Valmari. The State Explosion Problem. In W. Reisig and G. Rozenberg, editors, *Lectures on Petri Nets I: Basic Models*, volume 1491 of *Lecture Notes in Computer Science*, pages 429–528. Springer, Berlin, 1998.
136. W.M.P. van der Aalst. The Application of Petri Nets to Workflow Management. *The Journal of Circuits, Systems and Computers*, **8**(1):21–66, 1998.

137. W.M.P. van der Aalst. Business Process Management Demystified: A Tutorial on Models, Systems and Standards for Workflow Management. In J. Desel, W. Reisig, and G. Rozenberg, editors, *Lectures on Concurrency and Petri Nets*, volume 3098 of *Lecture Notes in Computer Science*, pages 1–65. Springer, Berlin, 2004.
138. W.M.P. van der Aalst. Using Process Mining to Generate Accurate and Interactive Business Process Maps. In A. Abramowicz and D. Flejter, editors, *Business Information Systems (BIS 2009) Workshops*, volume 37 of *Lecture Notes in Business Information Processing*, pages 1–14. Springer, Berlin, 2009.
139. W.M.P. van der Aalst. Business Process Simulation Revisited. In J. Barjis, editor, *Enterprise and Organizational Modeling and Simulation*, volume 63 of *Lecture Notes in Business Information Processing*, pages 1–14. Springer, Berlin, 2010.
140. W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, Berlin, 2011.
141. W.M.P. van der Aalst. Distributed Process Discovery and Conformance Checking. In J. de Lara and A. Zisman, editors, *International Conference on Fundamental Approaches to Software Engineering (FASE 2012)*, volume 7212 of *Lecture Notes in Computer Science*, pages 1–25. Springer, Berlin, 2012.
142. W.M.P. van der Aalst. A General Divide and Conquer Approach for Process Mining. In M. Ganzha, L. Maciaszek, and M. Paprzycki, editors, *Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*, pages 1–10. IEEE Computer Society, 2013.
143. W.M.P. van der Aalst. Business Process Management: A Comprehensive Survey. *ISRN Software Engineering*, **2013**:1–37, 2013. doi:[10.1155/2013/507984](https://doi.org/10.1155/2013/507984).
144. W.M.P. van der Aalst. Decomposing Petri Nets for Process Mining: A Generic Approach. *Distributed and Parallel Databases*, **31**(4):471–507, 2013.
145. W.M.P. van der Aalst. Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining. In M. Song, M. Wynn, and J. Liu, editors, *Asia Pacific Conference on Business Process Management (AP-BPM 2013)*, volume 159 of *Lecture Notes in Business Information Processing*, pages 1–22. Springer, Berlin, 2013.
146. W.M.P. van der Aalst. Data Scientist: The Engineer of the Future. In K. Mertins, F. Benaben, R. Poler, and J. Bourrieres, editors, *Proceedings of the I-ESA Conference*, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer, Berlin, 2014.
147. W.M.P. van der Aalst. Extracting Event Data from Databases to Unleash Process Mining. In J. vom Brocke and T. Schmiedel, editors, *BPM: Driving Innovation in a Digital World*, pages 105–128. Springer, Berlin, 2015.
148. W.M.P. van der Aalst and S. Dustdar. Process Mining Put into Context. *IEEE Internet Computing*, **16**(1):82–86, 2012.
149. W.M.P. van der Aalst and C. Stahl. *Modeling Business Processes: A Petri Net Oriented Approach*. MIT press, Cambridge, MA, 2011.
150. W.M.P. van der Aalst and A.H.M. ter Hofstede. YAWL: Yet Another Workflow Language. *Information Systems*, **30**(4):245–275, 2005.
151. W.M.P. van der Aalst and K.M. van Hee. *Workflow Management: Models, Methods, and Systems*. MIT press, Cambridge, MA, 2004.
152. W.M.P. van der Aalst, J. Desel, and A. Oberweis, editors. *Business Process Management: Models, Techniques, and Empirical Studies*, volume 1806 of *Lecture Notes in Computer Science*. Springer, Berlin, 2000.
153. W.M.P. van der Aalst, P. Barthelmess, C.A. Ellis, and J. Wainer. Proclcts: A Framework for Lightweight Interacting Workflow Processes. *International Journal of Cooperative Information Systems*, **10**(4):443–482, 2001.
154. W.M.P. van der Aalst, J. Desel, and E. Kindler. On the Semantics of EPCs: A Vicious Circle. In M. Nüttgens and F.J. Rump, editors, *Proceedings of the EPK 2002: Business Process Management using EPCs*, pages 71–80, Trier, Germany, November 2002. Gesellschaft für Informatik, Bonn.
155. W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski, and A.P. Barros. Workflow Patterns. *Distributed and Parallel Databases*, **14**(1):5–51, 2003.

156. W.M.P. van der Aalst, B.F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A.J.M.M. Weijters. Workflow Mining: A Survey of Issues and Approaches. *Data and Knowledge Engineering*, **47**(2):237–267, 2003.
157. W.M.P. van der Aalst, A.J.M.M. Weijters, and L. Maruster. Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, **16**(9):1128–1142, 2004.
158. W.M.P. van der Aalst, H.T. de Beer, and B.F. van Dongen. Process Mining and Verification of Properties: An Approach based on Temporal Logic. In R. Meersman and Z. Tari et al., editors, *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005*, volume 3760 of *Lecture Notes in Computer Science*, pages 130–147. Springer, Berlin, 2005.
159. W.M.P. van der Aalst, H.A. Reijers, and M. Song. Discovering Social Networks from Event Logs. *Computer Supported Cooperative work*, **14**(6):549–593, 2005.
160. W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek. Business Process Mining: An Industrial Application. *Information Systems*, **32**(5):713–732, 2007.
161. W.M.P. van der Aalst, M. Dumas, C. Ouyang, A. Rozinat, and H.M.W. Verbeek. Conformance Checking of Service Behavior. *ACM Transactions on Internet Technology*, **8**(3):29–59, 2008.
162. W.M.P. van der Aalst, M. Pesic, and H. Schonenberg. Declarative Workflows: Balancing Between Flexibility and Support. *Computer Science – Research and Development*, **23**(2):99–113, 2009.
163. W.M.P. van der Aalst, J. Nakatumba, A. Rozinat, and N. Russell. Business Process Simulation: How to Get it Right? In J. vom Brocke and M. Rosemann, editors, *Handbook on Business Process Management*, International Handbooks on Information Systems, pages 313–338. Springer, Berlin, 2010.
164. W.M.P. van der Aalst, M. Pesic, and M. Song. Beyond Process Mining: From the Past to Present and Future. In B. Pernici, editor, *Advanced Information Systems Engineering, Proceedings of the 22nd International Conference on Advanced Information Systems Engineering (CAiSE'10)*, volume 6051 of *Lecture Notes in Computer Science*, pages 38–52. Springer, Berlin, 2010.
165. W.M.P. van der Aalst, V. Rubin, H.M.W. Verbeek, B.F. van Dongen, E. Kindler, and C.W. Günther. Process Mining: A Two-Step Approach to Balance Between Underfitting and Overfitting. *Software and Systems Modeling*, **9**(1):87–111, 2010.
166. W.M.P. van der Aalst, K.M. van Hee, J.M. van der Werf, and M. Verdonk. Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor. *IEEE Computer*, **43**(3):90–93, 2010.
167. W.M.P. van der Aalst, M.H. Schonenberg, and M. Song. Time Prediction Based on Process Mining. *Information Systems*, **36**(2):450–475, 2011.
168. W.M.P. van der Aalst, K.M. van Hee, A.H.M. ter Hofstede, N. Sidorova, H.M.W. Verbeek, M. Voorhoeve, and M.T. Wynn. Soundness of Workflow Nets: Classification, Decidability, and Analysis. *Formal Aspects of Computing*, **23**(3):333–363, 2011.
169. W.M.P. van der Aalst, A. Adriansyah, and B. van Dongen. Replaying History on Process Models for Conformance Checking and Performance Analysis. *WIREs Data Mining and Knowledge Discovery*, **2**(2):182–192, 2012.
170. J.M.E.M. van der Werf, B.F. van Dongen, C.A.J. Hurkens, and A. Serebrenik. Process Discovery using Integer Linear Programming. *Fundamenta Informaticae*, **94**:387–412, 2010.
171. B.F. van Dongen. *Process Mining and Verification*. Phd thesis, Eindhoven University of Technology, 2007.
172. B.F. van Dongen and W.M.P. van der Aalst. Multi-Phase Process Mining: Building Instance Graphs. In P. Atzeni, W. Chu, H. Lu, S. Zhou, and T.W. Ling, editors, *International Conference on Conceptual Modeling (ER 2004)*, volume 3288 of *Lecture Notes in Computer Science*, pages 362–376. Springer, Berlin, 2004.
173. B.F. van Dongen, N. Busi, G.M. Pinna, and W.M.P. van der Aalst. An Iterative Algorithm for Applying the Theory of Regions in Process Mining. In W. Reisig, K. van Hee, and K. Wolf,

- editors, *Proceedings of the Workshop on Formal Approaches to Business Processes and Web Services (FABPWS'07)*, pages 36–55. Publishing House of University of Podlasie, Siedlce, Poland, 2007.
174. B.F. van Dongen, A.K. Alves de Medeiros, and L. Wenn. Process Mining: Overview and Outlook of Petri Net Discovery Algorithms. In K. Jensen and W.M.P. van der Aalst, editors, *Transactions on Petri Nets and Other Models of Concurrency II*, volume 5460 of *Lecture Notes in Computer Science*, pages 225–242. Springer, Berlin, 2009.
 175. M.L. van Eck, X. Lu, S.J.J. Leemans, and W.M.P. van der Aalst. PM2: A Process Mining Project Methodology. In J. Zdravkovic, M. Kirikova, and P. Johannesson, editors, *International Conference on Advanced Information Systems Engineering (CAiSE 2015)*, volume 9097 of *Lecture Notes in Computer Science*, pages 297–313. Springer, Berlin, 2015.
 176. R.J. van Glabbeek and W.P. Weijland. Branching Time and Abstraction in Bisimulation Semantics. *Journal of the ACM*, **43**(3):555–600, 1996.
 177. M. van Leeuwen and A. Siebes. StreamKrimp: Detecting Change in Data Streams. In *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 672–687. Springer, Berlin, 2008.
 178. J.J. van Wijk. The Value of Visualization. In *Visualization 2005*, pages 79–86. IEEE CS Press, 2005.
 179. H.M.W. Verbeek, T. Basten, and W.M.P. van der Aalst. Diagnosing Workflow Processes using Woflan. *The Computer Journal*, **44**(4):246–279, 2001.
 180. J. vom Brocke and M. Rosemann, editors. *Handbook on Business Process Management*, International Handbooks on Information Systems. Springer, Berlin, 2010.
 181. J. vom Brocke and M. Rosemann, editors. *Handbook on Business Process Management 1: Introduction, Methods, and Information Systems*, International Handbooks on Information Systems. Springer, Berlin, 2014.
 182. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.
 183. A.J.M.M. Weijters and J.T.S. Ribeiro. Flexible Heuristics Miner (FHM). BETA Working Paper Series, WP 334, Eindhoven University of Technology, Eindhoven, 2010.
 184. A.J.M.M. Weijters and W.M.P. van der Aalst. Rediscovering Workflow Models from Event-Based Data using Little Thumb. *Integrated Computer-Aided Engineering*, **10**(2):151–162, 2003.
 185. L. Wen, W.M.P. van der Aalst, J. Wang, and J. Sun. Mining Process Models with Non-Free-Choice Constructs. *Data Mining and Knowledge Discovery*, **15**(2):145–180, 2007.
 186. L. Wen, J. Wang, W.M.P. van der Aalst, B. Huang, and J. Sun. A Novel Approach for Process Mining Based on Event Types. *Journal of Intelligent Information Systems*, **32**(2):163–190, 2009.
 187. M. Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer, Berlin, 2007.
 188. G. Widmer and M. Kubat. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, **23**:69–101, 1996.
 189. Wikipedia. Observable Universe. http://en.wikipedia.org/wiki/Observable_universe, 2011.
 190. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.
 191. Workflow Patterns Home Page. <http://www.workflowpatterns.com>.
 192. M. zur Muehlen. *Workflow-based Process Controlling: Foundation, Design and Application of workflow-driven Process Information Systems*. Logos, Berlin, 2004.
 193. M. zur Muehlen and J. Recker. How Much Language Is Enough? Theoretical and Practical Use of the Business Process Modeling Notation. In Z. Bellahsene and M. Léonard, editors, *Proceedings of the 20th International Conference on Advanced Information Systems Engineering (CAiSE'08)*, volume 5074 of *Lecture Notes in Computer Science*, pages 465–479. Springer, Berlin, 2008.

Index

A

α -algorithm, 167, 171–174
 limitations, 174
ABPD, *see* Automated Business Process Discovery
Activity, 32
Activity instance, 131, 140, 177
Activity-based decomposition, 368, 373
Agglomerative hierarchical clustering, 94, 102, 286, 288
AHC, *see* Agglomerative hierarchical clustering
Algorithms, 12
Alignment, 256, 259
AMC Hospital, 424
Anonymize, 289
Apache Hadoop, 362, 364
Apriori algorithm, 106
Artificial negative events, 188, 239
ASML, 420
Association rule mining, 94, 104
 confidence, 104
 item-set, 105
 lift, 104
 support, 104
Auditing, 243, 245, 304
Automated Business Process Discovery, 16, 51

B

Bag, *see* Multi-set
BAM, *see* Business Activity Monitoring
Basel Accords, 50
Basel II Accord, 244
BI, *see* Business Intelligence
Big Data, 3, 9, 354
Bisimilarity, 180
Block-structured process models, 185

Bonferroni's correction, 355
Bonferroni's principle, 355, 356
BPI, *see* Business Process Intelligence
BPM, *see* Business Process Management, 27
BPM life-cycle, 30
BPM use cases, 45
BPMN, *see* Business Process Modeling Notation
BPR, *see* Business Process Reengineering
Branching bisimilarity, 59, 181
Business Activity Monitoring, 49
Business alignment, 243, 245
Business Intelligence, 49, 326
 products, 326
Business process improvement, 17
Business Process Intelligence, 16, 51
Business Process Management, 16, 27, 44, 45
Business process maps, 431
Business Process Modeling Notation, 28, 68, 70
Business process movie, 443
Business process provenance, 144, 302
Business Process Reengineering, 49

C

C-net, 72, 73, 205
Cartography, 303, 431
Case, 32, 134
Case perspective, 34, 275, 294
Case-based decomposition, 368, 369
Categorical variable, 91
 nominal, 92
 ordinal, 92
Causal net, *see* C-net
Celonis Process Mining (Celonis GmbH), 142, 340, 342, 343
CEP, *see* Complex Event Processing

- Classification, 93
 accuracy, 114
 error, 114
 F1 score, 114
 precision, 114
 recall, 114, 189
- Classifier, 133, 278
- Clustering, 94, 100
 centroid, 101
 cluster, 101
- CoBeFra, 337
- Colored Petri net, 64, 299
- Columnar databases, 360
- Comparative process mining, 381
- Complex Event Processing, 50
- Compliance, 50, 244
- Concept drift, 265, 320, 321, 381, 450
- Conformance, 33, 276, 301
- Conformance checking, 243, 244, 246, 264, 304
 cross-validation, 265
 footprint matrix, 263
 token-based replay, 246
- Conformance-related questions, 25
- Confusion matrix, 113
- Context, 318, 319
 case context, 319
 external context, 320
 process context, 319
 social context, 320
- Continuous Process Improvement, 48
- Control-flow perspective, 34, 275
- Corporate governance, 50, 244
- Corporate Performance Management, 49
- Correlation problem, 132, 143
- Coverability graph, 64
- CPI, *see* Continuous Process Improvement
- CPM, *see* Corporate Performance Management
- CPN, *see* Colored Petri net
- CPN Tools, 299
- CRM, *see* Customer Relationship Management
- Cross-organizational process mining, 401
- Cross-validation, 115, 116, 187, 271
 conformance, 265
 jack-knifing, 118
 k-fold cross-validation, 117, 188
 leave-one-out cross-validation, 118
 process discovery, 187, 271
 test set, 116
 training set, 116
- Curse of dimensionality, 119, 297
- Customer journey, 7
 stages, 8
 touchpoint, 7
- Customer Relationship Management, 27
- D**
- Data explosion, 4
- Data mining, 12, 46, 49, 89
 data set, 90, 91
 definition, 89
 descriptive, 94
 instance, 91
 predictive, 94
 variable, 91
- Data quality, 144, 148–151
- Data science, 3, 9, 10, 12, 15
- Data scientist, 10, 15
- Data source, 125
- Data warehouse, 127
- Databases, 13
- Decision mining, 294, 296
- Decision point, 294, 296
- Decision tree, 93–95, 97, 98
- Decision tree learning, 94, 95, 97, 98
- Declare, 201, 266, 309, 310
- Default classifier, 133
- Dendrogram, 103, 288, 417
- Dependency graph, 202, 204
- Desire lines, 43
- Digital universe, 3
- Digitalization, 7
- Digitization, 7
- Disco, 142
- Disco (Fluxicon), 142, 340–342
- Discovery, 33, 276, 301
- Distributed systems, 13
- DMAIC, 48
- Dotted chart, 278, 279, 281, 306
- E**
- 80/20 model, 186, 196, 449
- Enhancement, 33, 276, 301, 304
- Enterprise Discovery Suite (StereoLOGIC Ltd), 340
- Entropy, 97
- Enumerating model, 190
- EPC, *see* Event-driven process chains
- Episode mining, 107, 109
- Equivalence, 179
- Ethics, 14
- ETL, *see* Extract, Transform, and Load
- Event, 5
- Event correlation, 9
- Event data, 5

- Event log, 32, 33, 35, 128, 134
Event log metrics, 364, 365, 367
Event-driven process chains, 70
EXtensible Event Stream, *see* XES
Extension, 33, 276, 304
Extract, Transform, and Load, 127, 326
- F**
4-eyes principle, 266
False negatives, 114
False positives, 114
Feature extraction, 92, 312, 315, 416
Filtering, 128, 416, 438
Fitness, 166, 167, 188, 246, 247, 250, 253
Flat process model, 153
Flower model, 189, 218, 227
Footprint, *see* Footprint matrix
Footprint matrix, 168, 188, 263
Free-choice Petri net, 184
Functional areas, 397, 418
 finance/accounting, 398
 logistics, 398
 procurement, 397
 product development, 397
 production, 397
 resource management, 398
 sales/CRM, 398
 service, 398
Fuzzy mining, 207, 417
- G**
Generalization, 166, 190, 271, 272
Genetic process mining, 200, 207, 208, 210
Gini index of diversity, 100
Google File System (GFS), 361
GRC, 50
- H**
Hadoop, *see* Apache Hadoop
Hadoop Distributed File System (HDFS), 362
Heuristic miner, 201, 332, 388
Heuristic mining, 201, 202, 205
Hidden Markov model, 111
- I**
IM, *see* Inductive Miner
In-memory analytics, 360
In-memory database management system, 359
Incompleteness, 38, 186, 240
Inductive bias, 118
Inductive Miner, 222, 227, 233
 IMD framework, 235
 directly-follows based (IMD), 234, 235
 incompleteness (IMC), 234, 235
- incompleteness—directly-follows based (IMCD), 234
infrequent (IMF), 234, 338
infrequent—directly-follows based (IMFD), 234
- Inductive mining, 79, 222
 directly-follows graph, 223
 exclusive-choice cut, 225
 fall-through, 227, 235
 parallel cut, 225
 redo-loop cut, 226
 scalability, 236
 sequence cut, 225
- Inference, 41
Information gain, 99
Internet of content, 5
Internet of events, 5, 6
Internet of locations, 5
Internet of people, 5
Internet of things, 5
Interstage Business Process Manager
 Analytics (Fujitsu Ltd), 340
- ISO 9001:2008, 245
- K**
k-means clustering, 94, 100, 286
Key Performance Indicator, 85, 391
KNIME, 327
KPI, *see* Key Performance Indicator
- L**
L* life-cycle model, 392, 415
Language-rediscoverable, 230
Lasagna process, 22, 387
Lean Six Sigma, 46
Learning bias, 118
Linear regression, 93
Log-based ordering relations, 168
- M**
Machine learning, 13
MapReduce, 362, 364, 371, 372
 compute directly-follows graph, 372
 map task, 363
 reduce task, 363
Market basket analysis, 91, 104, 105
MDL, *see* Minimal Description Length
Minimal Description Length, 120, 189
Mining eXtensible Markup Language, *see* MXML
Minit (Gradient ECM), 142, 340, 343
Model-based process analysis, 83, 304
Moore's law, 3, 356, 357
Multi-set, 60

- Municipalities, 404
 WMO (Wet Maatschappelijke Ondersteuning) process, 388
 WOZ (Waardering Onroerende Zaken) process, 405
 MXML, 127, 138
 MyInvenio (Cognitive Technology), 340, 344
- N**
 Navigation, 305, 441
 Neural networks, 111
 Noise, 39, 185, 240
 Non-fitting, 196
 NoSQL database management systems, 361
 Numerical variable, 91
- O**
 Occam's razor, 120, 167
 Off-line process mining, 303
 OLAP, *see* Online Analytical Processing
 Online Analytical Processing, 49, 127, 160, 378
 Online process mining, 303, 305
 OpenXES, 142
 Operational support, 34, 50, 305, 306
 detect, 306, 307
 predict, 306, 311
 recommend, 306, 316
 Operations management, 16, 56
 Operations research, 16
 Optimization, 16
 Organizational perspective, 34, 275, 281
 Overfitting, 38, 100, 119, 120, 167, 190, 197
- P**
 PAIS, *see* Process-Aware Information System
 Pattern discovery, 94
 Perceptive Process Mining (Lexmark), 340, 344
 Performance analysis, 83
 Performance-related questions, 25
 Perspectives, 34, 275
 Petri net, 25, 26, 59
 bounded, 64
 deadlock free, 64
 enabled, 61
 firing rule, 61, 62
 firing sequence, 62
 labeled, 62
 live, 64
 marking, 26, 60
 place, 26, 60
 safe, 64
 state, 60
 token, 60
 transition, 26, 60
 Philips Healthcare, 421
 Photography trends, 6
 Play-In, 41, 243
 Play-Out, 41, 243
 PMLAB, 337
 Post mortem event data, 302
 Pre mortem event data, 302
 Precision, 166, 190, 269, 270, 272
 Predictive analytics, 13, 50
 Preprocessing the event log, 144, 281
 Primary sector, 399
 Privacy, 14
 Privacy issues, 289
 Process automation, 17
 Process cubes, 378–380
 Process discovery, 163, 164, 167, 197, 202, 207, 216, 218, 238, 304
 Process instance, *see* Case
 Process mining, 5, 13, 17, 25, 30, 33, 301
 Process Mining Framework, 301
 Process mining spectrum, 34, 321
 Process model, 26
 Process science, 15, 17, 55
 Process tree, 79, 80, 222
 conversion to Petri nets, 81
 semantics, 82
 Process-Aware Information System, 27
 Proclcts, 161
 Projection, 134, 374
 ProM, 52, 142, 331–334
 historical context, 331
 plug-ins, 332, 333
 ProM 5.2, 332
 ProM 6, 333
 ProMimport, 138
- Q**
 QPR ProcessAnalyzer (QPR), 340
 Quality dimensions, 188, 269
- R**
 R (analysis tool), 327
 Radio Frequency Identification, 4
 RapidMiner, 327, 338
 RapidProM, 46, 327, 338
 Reachability graph, 62
 Rediscovering process models, 178, 230
 Regression, 93
 Repair, 33, 268, 304
 Replay, 42, 243, 246, 248, 250, 254
 Representational bias, 118, 183, 449
 Resource, 131, 278, 281

- Resource-activity matrix, 281, 407
Responsible process mining, 451
RFID, *see* Radio Frequency Identification
Rialto Process (Exeura), 142, 340
Rijkswaterstaat, 402
Risk, 50, 244
RWS, *see* Rijkswaterstaat
- S**
SA-MXML, 138, 142
SAP HANA, 361
Sarbanes–Oxley Act, 50, 244
Scientific workflows, 338
Scoping the event log, 144
Secondary sector, 399
Semantically Annotated Mining eXtensible Markup Language, *see* SA-MXML
Sequence, 75, 133
Sequence mining, 107, 108
Short-term simulation, 88, 299, 315
Simple event log, 136
Simplicity, 166, 189, 269
Simulation, 87, 299
Six Sigma, 46–48
Slicing and dicing, 378
Smartphone, 6
SNP Business Process Analysis (SNP Schneider-Neureither & Partner AG), 142, 340
Social network, 135, 282
Social network analysis, 282
Sociometry, 282
Soundness, 77, 165
SOX, *see* Sarbanes–Oxley Act
Spaghetti process, 22, 411, 416
Statistics, 10, 12
Stochastics, 15
Streaming process mining, 381
Supervised learning, 92
 predictor variable, 92
 response variable, 92
- T**
Temporal logic, 84
Tertiary sector, 399, 420
Time perspective, 34, 275, 290
Timestamp, 128, 131, 143, 278, 290
TomTom metaphor, 441
Total Quality Management, 48
Toyota Production System, 46
TQM, *see* Total Quality Management
Trace, 32, 35, 134
Trace equivalence, 59, 180
- Transaction type, 131, 277, 278
Transactional life-cycle model, 131, 142, 177, 293
Transition system, 58, 212
 learning, 212
 state, 58
 transition, 58
True negatives, 114
True positives, 114
- U**
Underfitting, 38, 119, 120, 167, 190, 197
Unsupervised learning, 92, 94
- V**
Verification, 83
Vicious circle, 71
Visual analytics, 13, 14, 50, 281, 306
- W**
WebMethods Process Performance Manager (Software AG), 340
WEKA, 327
WF-net, 165
WF-nets, *see* Workflow net
WFM, *see* Workflow Management, 27
Workflow Management, 17, 27
Workflow mining, 51
Workflow net, 65
 case, 65
Workflow Patterns Initiative, 29, 57, 66
- X**
XES, 127, 138, 140
 classifier, 140
 concept extension, 140
 extensions, 138
 global attributes, 140
 life-cycle extension, 142
 meta model, 138
 organizational extension, 142
 semantic extension, 142
 serialization, 140
 standard extensions, 140
 time extension, 142
XESame, 142
- Y**
YAWL, 66, 67
Yerkes–Dodson law, 56, 289, 318
Yet Another Workflow Language, *see* YAWL
Yin and Yang in process mining, 25