Process Mining on Event Graphs: a Framework to Extensively Support Projects

Alessandro Berti

Process and Data Science group, Lehrstuhl für Informatik 9 52074 Aachen, RWTH Aachen University, Germany,
a.berti@pads.rwth-aachen.de

Abstract. Most business processes are supported nowadays by information systems, that record event data about the executions of the processes. Process mining plays an important role in linking the BPM field with data science, helping to identify the bottlenecks and the unwanted behavior, and to adopt strategies to improve the process, measuring the eventual benefit. While many algorithmic techniques have been developed for discovery, conformance checking and other process mining techniques, extracting data from today's information systems requires the specification of a complex query that extracts the required information and groups the events in cases. The research project described in this paper proposes a novel framework to support process mining analysis, that uses the advances in graph algorithms and in-memory data processing in order to reduce the costs of extraction and transformation of the event data contained in the information systems. At the end of the project, a set of pre-processing, discovery and conformance checking techniques, that do not require the specification of a case notion, will be made available in different environments, technologies and languages, e.g., PM4Py, Spark, Neo4J, Celonis. In comparison to related work, this research aims to obtain a complete and scalable framework that supports process mining from the Extraction, Transformation and Load phase (from relational and non-relational databases) to the effective analysis/usage of the data, and to get a class of process models fully capturing the lifecycle and the interactions between different classes. Since the framework is aimed at real-life, complex, information systems, a goal of the project is to attain significantly better scalability than existing approaches.

Key words: process mining, databases, event graphs, node encodings, process querying, process discovery, conformance checking

1 Motivation and Introduction

Process mining is a growing branch (including process discovery, conformance checking, model repair, process prediction, etc.) of business process management that aims to extract useful information and insights from event data contained in the information systems. Retrieving event data in a format that is useful for process mining analysis may be a challenge and requires knowledge of the particular information system and process mining expertise. To cite the process

mining manifesto [25] (G.P.2): "given the thousands of tables in the database of an ERP system like SAP, without concrete questions it is impossible to select the tables relevant for data extraction". The PM^2 process mining methodology [27] describes the different phases of a process mining project: (P1-2) planning and extraction, (P3) data processing (P4-P6) mining, analysis, evaluation and improvement. Phases (P1-2) and (P-3) are usually the most time-expensive[1]. In the extraction process, the specification of a *case notion* is particularly difficult. Indeed, most process mining techniques require that events belonging to the same execution of a business process are grouped. For some information systems (ERP, CRM), there may be several possible case notions: consider for example, a CRM system, where business opportunities and marketing campaigns are deeply intertwined. Moreover, the specification of a case notion may require a complex query to the data, taking into account events related to different database entities through a join.

This paper proposes a research project that aims to reduce the time spent in extracting and pre-processing data (both from relational and non-relational databases), and to avoid the necessity to specify a case notion, possibly covering the lifecycle of a process mining project. The resulting class of process schemas, similarly to the OCBC technique [12], should be able to capture the interactions between different perspectives. The research aims to abstract an information system as an *event graph* containing the following elements:

- *Events* happening in the considered information system, stored along with an ID and the timestamp.
- *Event classes*: features that describe the events.
- *Object perspectives*: objects contained in the information system, that may be created/changed/deleted by events happening in the system.
- *Class perspectives*: entities that group the objects of the information systems.
- *Process clusters*: potentially overlapping sets of class perspectives that are strongly intertwined and support a specific part of the information system.

This structure aims to infer the relationships between the different events and objects, that with the current techniques would require the presence of the schema,

[1] In some cases, 80% of the time is spent in (P1-2) and (P-3), see https://www.cloverdx.com/customers/case-study-processgold-improves-data-extract-preparation-accelerate-process-mining

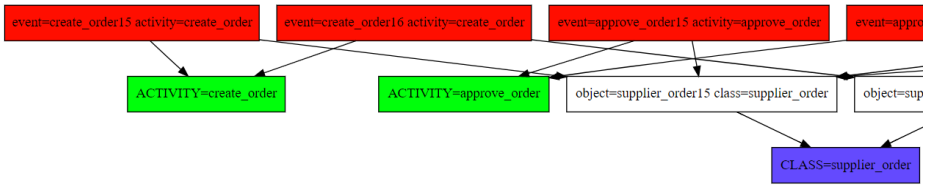

Fig. 1. Partial view of an example event graph. The red nodes are the events, the green nodes are the event classes, the white nodes are the object perspectives, the blue nodes are the class perspectives.

directly from the event graph, using different graph algorithms. A partial view of an example event graph is contained in Fig. 1. Pre-processing, discovery and conformance checking on event graphs are possible thanks to aggregations and advances in the graph algorithms (random walks [15], node encodings [11], graph simulations for pattern matching [13], etc.).

2 Related Work

Related work is summarized here, and could be split in:

1. Work covering the extraction and pre-processing phase ((P1-3) of the PM^2 methodology): translation of SPARQL queries into SQL queries on databases [4], OpenSLEX meta-model [6, 7] (by ingesting generic database logs, e.g., redo logs, into an easy-to-query meta-model instance).
2. Work on artifact-centric models ((P4-6) of the PM^2 methodology): GSM models [14], discovery of artifact-centric models [20, 16] and behavioral conformance [9]. The approach [16] partly address (P1-3) when timestamps are contained as columns of the schema, but not with generic database logs.
3. Work on the discovery of models where several case notions are considered ((P4-6) of the PM^2 methodology): Composite State Machine Miner [26], interacting processes with overlapping instances [10], Object-centric behav-

Table 1. Summary of the research questions, of the techniques and technologies that are involved in the project, and of the specific goals of the project for each research question.

Question	**Motivation**	**Specific Goals**
(P1-2) How to extract data from databases?	The Extraction, Transformation and Load procedure is among the most time-consuming parts of a process mining projects. Despite that, a scalable approach that works without problems with relational and non-relational databases is still missing.	Support the automatic transformation of the raw content of database logs into event graphs. For relational databases, some work has been done in [6, 7]. The goal is to add extraction support for non-relational databases. The event graph obtained from an information system should be stored into an efficient intermediate structure.
(P3) How to pre-process event data without specifying a case notion?	Pre-processing is an important step after extracting an event log from a database, since the cases might be incomplete or affected by noise. Not having a case notion is an impediment for pre-processing.	Support the filtering operations on event data, without a case notion, can be made possible on the event graph by using some graph algorithms to retrieve specific patterns and to cluster events. The goal is to obtain some filtering features that are similar to the ones provided by commercial software on logs having a single case notion.
(P4-6) How to discover a process model without specifying a case notion?	Despite a number of techniques is currently available, they suffer from scalability issues, rely on the information provided by the relational schema and/or cannot cope with noise in the log.	The goal is to obtain a class of succint descriptive diagrams, able to represent several class perspectives, along with their interactions.
(P4-6) How to check conformance without specifying a case notion?	With current techniques, it is not possible to support fine-grained conformance checking directly at the database level, without requiring the definition of an artifact structure on top of the database or the extraction of the events from the database.	The goal is to apply classic, well-known, conformance checking techniques, like token-based replay and alignments, on top of MVP models enriched by normative elements (see also [22]). Moreover, a way to check declarative constraints on the event graph shall be elaborated.

ioral constraint models (OCBC) [12]. In particular, OCBC models offer a theoretically powerful approach to discover models without requiring the specification of a case notion, and to perform conformance checking on top of such models, but do not cover the extraction and pre-processing phase and suffer from scalability problems.
4. Work on modeling of dynamics of multi-dimensional processes: BPMN with data annotations [30], relational process structure [31, 32].

3 Research Design and Methodology

In this section, an introduction to some aspects of design and methodology in the research is proposed.

- *Main goal/Artifact*: The main goal of the research project described in this paper is to provide a scalable framework for pre-processing event data, and a class of process models that is able to fully capture the lifecycle and the interactions of the different perspectives.
- *Problem relevance*: several techniques for the extraction of event logs and artifact-centric models from relational databases have been proposed [4, 7, 9, 10, 16, 12]. Non-relational databases, that are growing in importance, are still not covered by an extraction methodology aimed at obtaining artifact-centric models. This is because the information contained in the schema is important for the existing approaches. By representing the information as an event graph, the aim is to relax the need of a schema and infer the information needed directly from the relationships between the nodes in the graph.
- *Design evaluation*: several types of assessment will be performed on the proposed techniques: *automatic assessment* of the performance in extraction/processing, *semi-assisted assessment* of the quality of the event logs extracted from relational/non-relational databases (aided by clustering and concept drift detection techniques); *expert evaluation* of the quality of the provided models and of the conformance checking techniques. The expert evaluation will involve the production of cases studies with CRM/ERP companies.

Some other high-levels goals of the project are: scalability, usage of cloud/parallel computations, provision of connectors for relational and non-relational databases.

4 Current Outputs of the Research Project

The research project is still in an early phase. An initial short paper containing some ideas about discovery of multiple viewpoint models on top of event graphs is [3]. The code supporting the research is available in a fork of the PM4Py library, available at the address *https://github.com/javert899/pm4py-source*. At the moment, the ingestion of event graphs from some intermediate formats as XOC (format used by OCBC models) and OpenSLEX is available along with the discovery of multiple viewpoint models from event graphs. The documentation of these features is contained in a scientific paper that is under review.

References

1. Berti, A., van Zelst, S. J., van der Aalst, W.: Process Mining for Python (PM4Py): Bridging the Gap Between Process-and Data Science. International Conference on Process Mining demos (in print) (2019)
2. Berti, A., van der Aalst, W.: Reviving Token-based Replay: Increasing Speed While Improving Diagnostics ATAED (in print) (2019)
3. Berti, A., van der Aalst W.: StarStar Models: Using Events at Database Level for Process Analysis SIMPDA 2018 2270, 60-64 (2018)
4. Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M. et al.: Ontop: Answering SPARQL queries over relational databases. Semantic Web, 8(3), 471-487 (2017)
5. Cheng, L., Van Dongen, B., van der Aalst, W.: Scalable Discovery of Hybrid Process Models in a Cloud Computing Environment. IEEE Transactions on Services Computing (2019)
6. de Murillas, E. G. L., van der Aalst, W., Reijers, H. A.: Process mining on databases: Unearthing historical data from redo logs. In International Conference on Business Process Management (pp. 367-385). Springer, Cham (2016)
7. de Murillas, E. G. L., Reijers, H. A., van der Aalst, W.: Connecting databases with process mining: a meta model and toolset. Software & Systems Modeling, 1-39 (2018)
8. Di Ciccio, C., Maggi, F. M., Mendling, J.: Efficient discovery of target-branched declare constraints. Information Systems, 56, 258-283 (2016)
9. Fahland, D., De Leoni, M., Van Dongen, B., van der aalst, W. Behavioral conformance of artifact-centric process models. In International Conference on Business Information Systems (pp. 37-49). Springer, Berlin, Heidelberg (2011)
10. Fahland, D., De Leoni, M., Van Dongen, B., van der Aalst, W.: Conformance checking of interacting processes with overlapping instances. In International Conference on Business Process Management (pp. 345-361). Springer, Berlin, Heidelberg (2011)
11. Grover, A., Leskovec, J.: node2vec: Scalable Feature Learning for Networks (unpublished)
12. Li, G., de Carvalho, R. M., van der Aalst, W.: Automatic discovery of object-centric behavioral constraint models. In International Conference on Business Information Systems (pp. 43-58) Springer, Cham. (2017)
13. Henzinger, M. R., Henzinger, T. A., Kopke, P. W.: Computing simulations on finite and infinite graphs. In Proceedings of IEEE 36th Annual Foundations of Computer Science (pp. 453-462) (1995)
14. Hull, R., Damaggio, E., De Masellis, R., Fournier, F., Gupta, M., Heath III, F. T. et al.: Business artifacts with guard-stage-milestone lifecycles: managing artifact interactions with conditions and events. In Proceedings of the 5th ACM international conference on Distributed event-based system (pp. 51-62). ACM (2011)
15. Lovsz, L.: Random walks on graphs: A survey. Combinatorics, Paul erdos is eighty, 2(1), 1-46 (1993)
16. Lu, X., Nagelkerke, M., van de Wiel, D., Fahland, D.: Discovering interacting artifacts from ERP systems. IEEE Transactions on Services Computing, 8(6), 861-873 (2015)
17. Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., Vassilakis, T.: Dremel: interactive analysis of web-scale datasets. Proceedings of the VLDB Endowment, 3(1-2), 330-339 (2010)

18. Maggi, F. M., Bose, R. J. C., van der Aalst, W.: A knowledge-based integrated approach for discovering and repairing declare maps. In International Conference on Advanced Information Systems Engineering (pp. 433-448). Springer, Berlin, Heidelberg (2013)
19. Miller, J. J.: Graph database applications and concepts with Neo4j. In Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA (Vol. 2324, No. S 36) (2013)
20. Nooijen, E. H., Van Dongen, B., Fahland, D.: Automatic discovery of data-centric and artifact-centric processes. In International Conference on Business Process Management (pp. 316-327). Springer, Berlin, Heidelberg (2012)
21. Pesic, M., Schonenberg, H., van der Aalst, W.: Declare: Full support for loosely-structured processes. In 11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007) (pp. 287-287). IEEE (2007)
22. Sander, J.J.L., Erik, P., Moe, T.W.: Directly Follows-Based Process Mining: Exploration & a Case Study. International Conference on Process Mining (in print) (2019)
23. van der Aalst, W.: Discovering the Glue Connecting Activities. In It's All About Coordination (pp. 1-20). Springer, Cham (2018)
24. van der Aalst, W., De Masellis, R., Di Francescomarino, C., Ghidini, C.: Learning hybrid process models from events. In International Conference on Business Process Management (pp. 59-76). Springer, Cham (2017)
25. van der Aalst, W., Adriansyah, A., De Medeiros, A. K. A. et al.: Process mining manifesto. In International Conference on Business Process Management (pp. 169-194). Springer, Berlin, Heidelberg (2011)
26. van Eck, M. L., Sidorova, N., van der Aalst, W.: Discovering and exploring state-based models for multi-perspective processes. In International Conference on Business Process Management (pp. 142-157). Springer, Cham (2016)
27. van Eck, M. L., Lu, X., Leemans, S. J., van der Aalst, W.: PM^2: A Process Mining Project Methodology. In International Conference on Advanced Information Systems Engineering (pp. 297-313). Springer, Cham (2015)
28. van Zelst, S. J., Bolt, A., Hassani, M., van Dongen, B. F., van der Aalst, W.: Online conformance checking: relating event streams to process models using prefix-alignments. International Journal of Data Science and Analytics, 1-16 (2017)
29. Veit, F., Geyer-Klingeberg, J., Madrzak, J., Haug, M., Thomson, J.: The Proactive Insights Engine: Process Mining meets Machine Learning and Artificial Intelligence. In BPM (Demos) (2017)
30. Meyer, Andreas, et al.: Modeling and enacting complex data dependencies in business processes. Business process management (pp. 171-186) Springer, Berlin, Heidelberg (2013)
31. Steinau, Sebastian, Kevin Andrews, and Manfred Reichert: Modeling Process Interactions with Coordination Processes OTM Confederated International Conferences - On the Move to Meaningful Internet Systems. Springer, Cham (2018)
32. Steinau, Sebastian, Kevin Andrews, and Manfred Reichert: The relational process structure. International Conference on Advanced Information Systems Engineering. Springer, Cham (2018)