# Stage-based Business Process Mining

Hoang Nguyen[1]

Supervisors: Marcello La Rosa[1], Marlon Dumas[2] and Arthur H.M. ter Hofstede[1]

[1] Queensland University of Technology, Australia
huanghuy.nguyen@hdr.qut.edu.au, {a.terhofstede,m.larosa}@qut.edu.au
[2] University of Tartu, Estonia
{marlon.dumas}@ut.ee

**Abstract.** Evidence-based BPM has gained significant momentum in recent years, thanks to the widespread adoption of enterprise systems that store detailed business process execution data in *event logs*. Techniques for analyzing business processes using event logs are termed "process mining" techniques. Their objective is to aid business analysts in improving business processes by learning knowledge from massive data. To date, techniques for process mining abound. For example, one can measure processing time and waiting time, diagnose process delays and quality issues, and replay an entire event log over a process model discovered from the log itself. However, these techniques often suffer from limited applicability, particularly when used on top of unpredictable processes such as patient treatment processes in healthcare as opposed to predictable processes such as a car manufacturing process. They failed to extract a highly fit process model, awkward in measuring process performance, and inaccurate in predictive monitoring. In addition, they are confused at how to divide the problem into sub-problems for better solutions. This research aims at designing a novel set of techniques based on a notion of business process stages which can improve over existing process mining techniques.

**Keywords:** Business process management, process mining, multistage, stage-based, decomposition

## 1 Research Motivation

Process Mining [1] was initiated from the field of Business Process Management that oversees and improves human work in organizations [2]. Therefore, Process Mining also concerns with common tasks in BPM such as process performance analysis, conformance checking and root cause analysis. However, differing from the social science branch of BPM concerning interviews, workshops and surveys for data collection, Process Mining focuses on analysing large and rich business process data (called *event logs*) available in enterprise IT systems in order to extract useful knowledge [3]. Process Mining thus is a bridge between BPM and data mining.

Like data mining, process mining techniques exploit data features (or variables) in event logs to learn useful knowledge for process improvement. These techniques fall into a number of categories. *Process discovery* [4] is to derive *process models* from event logs. *Conformance checking* [5] is to align an event log with a process model to verify whether the process execution complies with the process design. *Performance analysis* [6] is to measure process performance

X. Franch, J. Ralyté, R. Matulevičius, C. Salinesi, and R. Wieringa (Eds.):
CAiSE 2017 Forum and Doctoral Consortium Papers, pp. 161-169, 2017.
Copyright 2017 for this paper by its authors. Copying permitted for private and academic purposes.

metrics to identify bottlenecks. *Deviance mining* [7] is to derive business rules from event logs that can explain the root cause of positive or negative deviants. *Predictive monitoring* [8] aims at building predictive models that allow one to make forecasts of process performance. Finally, *comparative analysis* [9] is to contrast process variants and extract distinguishing behaviors. From another perspective, these techniques fall into two themes: *structural analysis* and *behavioral analysis*. In structural analysis, the purpose is to search for a structure from event logs that highly represents the process, e.g. process models, which can help to do performance analysis, conformance checking and serve as a basis for process re-engineering. In behavioral analysis, the purpose is to search for a set of behaviors (e.g. activity patterns) that are strongly correlated with a target variable, e.g. long case duration. Behavioral analysis is common in deviance mining and predictive monitoring based on trained classifiers such as decision trees [8], random forests [10], and neural networks [11]. In addition, some works also regard process models as a source of generalized behaviors for descriptive analysis [9].

Thus far, the main challenge to process mining is that many event logs exhibit a highly complex *feature space*. For example, real-life event logs can be found on the Business Process Intelligence web site from 2011 to 2017[1]. Notably, they are often *knowledge-intensive processes* [12] such as patient treatment, insurance claim handling, IT incident handling, and loan application assessment. Their feature space often includes, but not limited to, activities, humans, data payload, process context [13] and timestamps. Three main challenges of this feature space are the heterogeneity of case context, the decomposition into sub-processes, and the variability of data features. Different case contexts exist because process cases, e.g. customer orders or patients, are often prioritized based on different types, e.g. low-value and high-value cases, and processed differently. Mixing case contexts therefore can create greater variation in data features, thus makes it more difficult for process structural analysis. Sub-processes often exist and could be in sequence, in parallel or overlap. They are interrelated but fairly independent. Ignoring these sub-processes in one analysis might be the cause of inaccurate models. Moreover, the inherent variability of data features in business environment is a challenge to frequent feature mining for business processes. In many cases, it is the combination of these three challenges that creates a very heterogeneous feature space. Consequently, the current problems faced by process mining are *scalability* and *accuracy*. For example, process discovery techniques struggle with ill-structured processes [14]. Mining human-readable rules from event logs remains an issue [7]. The error rate of predictive monitoring remains remarkably high [15, 11].

Various process mining techniques have been proposed to deal with the above complex feature space. A common approach is based on *decomposition* of event logs into clusters, thus able to work with clusters (i.e. a higher abstraction level) instead of individual events. It is also known as *divide and conquer* approach which has been implemented for process discovery [16–20], conformance checking [19], performance analysis [21], deviance mining [22], and predictive monitoring [23]. Decomposition can be horizontal (i.e. by cases) or vertical (i.e. by activities). However, although scalability has been improved, the accuracy issue remains [18, 19, 23, 11]. Proposed techniques seem to be ad hoc while they only work with some specific datasets and struggle with others. There are several reasons learned from empirical results. For structural analysis, the proposed decompositions may underrepresent the real process structure [24]. Thus, when the

[1] www.win.tue.nl/bpi/doku.php?id=2017:challenge

models are tested against the logs, the result has low fitness and precision [6]. For behavioral analysis, despite the use of decomposition and strong classifiers, the accuracy could be affected due to the limited coverage of the process feature space, e.g. when a classification model contains only control-flow features but many process cases are driven by resources and context [11].

From the above background, this research proposes a novel process mining approach based on a notion of business process stages. Semantically, stages are a common way that humans use to divide their work into manageable parts. A stage thus is also a sub-process. For example, an outpatient treatment process involves stages such as reception, diagnosis, medication, and consultation. Stages have also been observed in BPM research and real datasets, including patient treatment[9], IT service delivery[25], government agency processes [26], bank loan application [27], and product development [28]. Traditionally, process stages have been studied in different disciplines. For example, in manufacturing it is known through the *state space model* for fault diagnosis [29, 30]. In patient flow research, it is called *compartment model* [31–33]. In product development, it is known as the *stage-gate model* [34]. Recently, stage-based analysis has been studied in process mining for inter-organizational comparative analysis but only on a manual basis [9, 35]. Continuing this stream, this research aims to develop stage-based techniques for knowledge-intensive processes taking advantage of event logs and foundational techniques of process mining.

The intuition here is that stages can help to improve process mining techniques. Intuitively, data features within the same stages tend to exhibit stronger relationship than those from different stages; thus, stage-based techniques could produce better result than those applied to the whole process. For example, stages could provide a vertical decomposition of event logs (i.e. by stages) in order to improve the quality of process models. The first question is how to discover stages from event logs that mimic the actual stage decomposition. Once stages have been correctly discovered, they can be used to discover process models by stages instead of one flat model for the whole log. Another application of stages is to measure *flow performance* [36]. This kind of performance is of particular interest in service organizations such as hospitals, product development and IT services because they are concerned with how smooth cases are pulled through the organizations. Since a stage decomposition consists of adjacent stages, each is a fairly independent queueing system, it is thus allowed to measure flow of cases (i.e. queuing items) based on queuing measures computed from event logs, e.g. arrival rate, departure rate, and length of queue. In addition, in predictive monitoring, it could be more accurate to build classifiers within a stage to provide prediction within that stage only, combined with inter-stage classifiers to provide a final prediction.

## 2 Research Problems & Research Questions

The previous section has discussed current research problems in detail. They are summarized as follows.

1. Current process discovery techniques suffer from low accuracy for ill-structured processes
2. Current process performance analysis techniques are limited in measuring the flow performance of business processes
3. Current predictive process monitoring techniques suffer from high error rate for ill-structured processes

Our research will be structured to address the following research questions:

1. How to discover business process stages from event logs?
2. How to mine business process performance from event logs based on stages
3. How to discover process models from event logs based on stages?
4. How to perform predictive process monitoring in stages?

## 3 Research Approach

This research project aims at developing stage-based techniques that can produce better result than existing techniques. We consider *Design Science* (DS) [37] as a relevant research method as its nature is to produce knowledge based on the development of artifacts (e.g. models, frameworks, and methods) to solve a problem [37]). In our research, the problems would be the research problems and the artifacts would be computer software that implements our proposed techniques.

Following the Design Science method, this project will primarily undergo five main steps to develop a technique [37]: (i) Define the problem; (ii) Suggest a solution; (iii) Develop artifacts; (iv) Evaluate the artifacts; (v) Conclude. Among these, the validity of DS-based research is mainly determined by the evaluation of the artifacts [38]. There are different validation approaches including observational, analytical, experimental, testing and descriptive [39]. This project will mainly take the experimental approach given the data-driven nature of the research.

A rigorous approach to experimental evaluation thus is vital to this project. The evaluations will generally consist of two parts: data-based and user-based. The former makes use of objective and quantitative measures while the latter involves humans, where needed, in qualitative assessment. Outline of research experiments are given below.

- Experiments will be carried out on event logs of varied characteristics
- Evaluation will be performed based on well-established criteria in Data Mining and Process Mining
- Controlled experimentation [40] will be conducted with stakeholders, where needed, to evaluate the subjective aspect of the research criteria
- The proposed technique will be benchmarked against baselines available in the literature

In regards to data collection and analysis, event logs are the main datasets used for experiments in this research. Access to data in different ways is planned as follows:

- Synthetic datasets will be created for the first validation using business process simulation software, e.g. BIMP[2] and CPN-Tools[3].
- Real-life datasets will be sourced from repositories of publicly available logs and industrial as well as academic partners. The publicly available logs are provided on academic public data repositories such as 3TU.Datacentrum[4] of Eindhoven University of Technology which have been used as benchmarking data for experiments in previous research in Process Mining.

[2] bimp.cs.ut.ee
[3] www.cpntools.org
[4] data.3tu.nl/repository/collection:event_logs

- The student may request for access to datasets of other research projects within the BPM Discipline. The request will be in compliance with the Ethics Clearance of the projects.
- The student will contact the pool of industrial partners of the BPM Discipline such as Commonwealth Bank of Australia, Suncorp and St Andrews War Memorial Hospital, to access further real-life logs, should this be needed.

## 4 Preliminary Results

The research thus far has carried out towards addressing the first two research questions: mining process stages from event logs and mining process performance based on stages. The result is reported in the following sections.

### 4.1 Mining Business Process Stages from Event Logs

Process mining techniques suffer from scalability issues when applied to large event logs, both in terms of computational requirements and in terms of interpretability of the produced outputs. For example, process models discovered from large event logs are often spaghetti-like and provide limited insights [14].

A common approach to tackle this limitation is to decompose the process into stages, such that each stage can be mined separately. This idea has been successfully applied in the context of automated process discovery [24] and performance mining [41]. The question is then how to identify a suitable set of stages and how to map the events in the log into stages. For simpler processes, the stage decomposition can be manually identified, but for complex processes, automated support for stage identification is required. Accordingly, several automated approaches to stage decomposition have been proposed [18, 19, 42]. However, these approaches have not been designed with the goal of approximating manual decompositions, and as we show in this work, the decompositions they produce turn out to be far apart from the corresponding manual decompositions.

This paper puts forward an automated technique to split an event log into stages, in a way that mimics manual stage decompositions. The proposed technique is designed based on two key observations: (i) that stages are intuitively fragments of the process in-between two milestone events; and (ii) that the stage decomposition is modular, meaning that there is a high number of direct dependencies inside each stage (high cohesion), and a low number of dependencies across stages (low coupling) – an observation that has also been applied in the context of process model decomposition [43] and more broadly in the fields of systems design and programming in general. For example, a loan origination process at a bank has multiple stages such as the application is assessed (accepted/rejected milestone), offered (offer letter sent milestone), negotiated (agreement signed milestone), and settled (agreement executed milestone). There may be many back-and-forth or jumps inside a stage, but relatively little across these stages.

The proposed technique starts by constructing a graph of direct control-flow dependencies from the event log. Candidate milestones are then identified by using techniques for computing graph cuts. A subset of these potential cut points is finally selected in a way that maximizes the modularity of the resulting stage decomposition according to a modularity measure borrowed from the field of social network analysis. The technique has been evaluated using real-life logs in terms of its ability to approximate manual decompositions using a well-accepted measure for the assessment of cluster quality.

### 4.2 Mining Process Performance Based on Staged Process Flows

*Process Performance Mining* (PPM) is a subset of process mining techniques concerned with the analysis of processes with respect to performance dimensions, chiefly *time* (how fast a process is executed); *cost* (how much a process execution costs); *quality* (how well the process meets customer requirements and expectations); and *flexibility* (how rapidly can a process adjust to changes in the environment) [2].

Along the time and flexibility dimensions, one recurrent analysis task is to understand how the temporal performance of a process evolves over a given period of time – also known as *flow performance* analysis in lean management [44]. For example, a bank manager may wish to know how the waiting times in a loan application process have evolved over the past month in order to adjust the resource allocation policies so as to minimize the effects of bottlenecks.

Existing PPM techniques are not designed to address such flow performance questions. Instead, these techniques focus on analyzing process performance in a "snapshot" manner, by taking as input an event log recorded during a period of time and extracting aggregate measures such as mean waiting time, processing time or cycle time of the process and its activities. For example, both the Performance Analysis plugins of ProM [45] and Disco [46] calculate aggregate performance measures (e.g. mean waiting time) over the entire period covered by an event log and display these measures by color-coding the elements of a process model. These tools can also produce animations of the flow of cases along a process model over time. However, extracting flow performance insights from these animations requires close and continuous attention from the analyst in order to detect visual cues of performance trends, bottleneck formation and dissolution, and phase transitions in the process performance. In other words, animation techniques allow analysts to get a broad picture of performance issues, but not to precisely quantify the evolution of process performance over time.

In this setting, this paper presents a PPM approach designed to provide a precise and quantifiable picture of flow performance. The approach relies on an abstraction of business processes called *Staged Process Flow* (SPF). An SPF breaks down a process into a series of queues corresponding to user-defined stages. Each stage is associated with a number of performance characteristics that are computed at each time point in an observation window. The evolution of these characteristics is then plotted via several visualization techniques that collectively allow flow performance to be analyzed from multiple perspectives in order to address the following questions:

Q1. How does the overall process performance evolve over time?
Q2. How does the formation and dissolution of bottlenecks affect the overall process performance?
Q3. How do changes in demand and capacity affect the overall process performance?

The paper demonstrates the advantages of the SPF approach over state-of-the-art process performance mining tools using real-life event logs of a Dutch bank and IT department of Volvo Belgium.

## 5 Conclusion and Future Work

This paper describes an overall approach of stage-based process mining based on observed gaps in current process mining techniques. So far, we have proposed

two stage-based techniques: one for discovering business process stages from event logs and one for mining process flow performance from event logs based on stages. The former work shows that our stage decomposition technique can provide results that are measurably much closer to the ground truth than the baselines. The latter work shows that it provides insights and addresses questions that cannot be answered by existing performance mining techniques. In the future, we will continue developing stage-based techniques for process discovery and predictive process monitoring.

## References


1. van der Aalst, W.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
2. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.: Fundamentals of Business Process Management. Springer (2013)
3. Van Der Aalst, W., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., Buijs, J.: Process mining manifesto. In: Business Process Management, Springer 169–194
4. Tiwari, A., Turner, C.J., Majeed, B.: A review of business process mining: state-oftheart and future trends. Business Process Management Journal **14**(1) (2008) 5–22
5. Rozinat, A., van der Aalst, W.M.: Conformance checking of processes based on monitoring real behavior. Information Systems **33**(1) (2008) 64–95
6. van der Aalst, W., Adriansyah, A., van Dongen, B.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(2) (2012) 182–192
7. Nguyen, H., Dumas, M., La Rosa, M., Maggi, F.M., Suriadi, S.: Mining business process deviance: a quest for accuracy. In: On the Move to Meaningful Internet Systems, Springer (2014) 436–445
8. Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C.: Predictive monitoring of business processes. In: International Conference on Advanced Information Systems Engineering, Springer (2014) 457–472
9. Suriadi, S., Mans, R.S., Wynn, M.T., Partington, A., Karnon, J.: Measuring patient flow variations: A cross-organisational process mining approach. In: Proc. of AP-BPM, Springer (2014) 43–58
10. Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., Maggi, F.M.: Complex symbolic sequence encodings for predictive monitoring of business processes. In: International Conference on Business Process Management, Springer (2015) 297–313
11. Tax, N., Verenich, I., La Rosa, M., Dumas, M.: Predictive business process monitoring with LSTM neural networks. arXiv preprint arXiv:1612.02130 (2016)
12. Di Ciccio, C., Marrella, A., Russo, A.: Knowledge-intensive processes: Characteristics, requirements and analysis of contemporary approaches. Journal on Data Semantics **4**(1) (2014) 29–57
13. Van der Aalst, W.M., Dustdar, S.: Process mining put into context. IEEE Internet Computing **16**(1) (2012) 82–86
14. van der Aalst, W.M.: Process mining: Discovering and improving spaghetti and lasagna processes. In: Proc. of CIDM, IEEE (2011)
15. van Dongen, B.F., Crooy, R.A., van der Aalst, W.M.: Cycle time prediction: When will this case finally be finished? In: OTM Confederated International Conferences, Springer (2008) 319–336
16. Weerdt, D.: Leveraging process discovery with trace clustering and text mining for intelligent analysis of incident management processes. In: IEEE Congress on Evolutionary Computation. (2012) 1–8
17. Rebuge, A., Ferreira, D.R.: Business process analysis in healthcare environments: A methodology based on process mining. Information Systems **37**(2) (2012) 99–116

18. Carmona, J., Cortadella, J., Kishinevsky, M.: Divide-and-conquer strategies for process mining. In: Proc. of BPM, Springer (2009) 327–343
19. Van Der Aalst, W.M.: A general divide and conquer approach for process mining. In: Proc. of FedCSIS, IEEE (2013) 1–10
20. Conforti, R., Dumas, M., García-Bañuelos, L., La Rosa, M.: BPMN miner: Automated discovery of BPMN process models with hierarchical structure. Information Systems **56** (2016) 284–303
21. van Dongen, B.F., Adriansyah, A.: Process mining: fuzzy clustering and performance visualization. In: Proc. of BPM Workshops, Springer (2010) 158–169
22. Ghattas, J., Peleg, M., Soffer, P., Denekamp, Y.: Learning the context of a clinical process. In: Business process management workshops, Springer (2010) 545–556
23. Di Francescomarino, C., Dumas, M., Maggi, F.M., Teinemaa, I.: Clustering-based predictive process monitoring. IEEE Transactions on Services Computing (2016)
24. Hompes, B., Verbeek, H., van der Aalst, W.M.: Finding suitable activity clusters for decomposed process discovery. In: Proc. of SIMPDA, Springer (2014) 32–57
25. Naldi, M., La Pinta, F., Lombardozzi, M., Picciano, M.: A phase model of the service delivery process for bundle services. In: Computer Modeling and Simulation (EMS), 2012 Sixth UKSim/AMSS European Symposium on, IEEE (2012) 263–268
26. van Dongen, B.: BPI Challenge 2015 Municipality 1 (2015)
27. van Dongen, B.F.: BPI Challenge 2012. Eindhoven University of Technology. Dataset (2012) http://dx.doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f.
28. Pietzsch, J.B., Shluzas, L.A., Pat-Cornell, M.E., Yock, P.G., Linehan, J.H.: Stage-gate process for the development of medical devices. Journal of Medical Devices **3**(2) (2009) 021004
29. Sulek, J.M., Marucheck, A., Lind, M.R.: Measuring performance in multi-stage service operations: An application of cause selecting control charts. Journal of Operations Management **24**(5) (2006) 711–727
30. Shi, J., Zhou, S.: Quality control and improvement for multistage systems: A survey. IIE Transactions **41**(9) (2009) 744–753
31. McClean, S.I., Millard, P.H.: A three compartment model of the patient flows in a geriatric department: a decision support approach. Health care management science **1**(2) (1998) 159–163
32. Mackay, M., Lee, M.: Choice of models for the analysis and forecasting of hospital beds. Health Care Management Science **8**(3) (2005) 221–230
33. Harrison, G.W., Escobar, G.J.: Length of stay and imminent discharge probability distributions from multistage models: variation by diagnosis, severity of illness, and hospital. Health care management science **13**(3) (2010) 268–279
34. Cooper, R.G.: Stage-gate systems: a new tool for managing new products. Business horizons **33**(3) (1990) 44–54
35. Partington, A., Wynn, M., Suriadi, S., Ouyang, C., Karnon, J.: Process mining for clinical processes: A comparative analysis of four australian hospitals. ACM Transactions on Management Information Systems **5**(4) (2015) 19
36. Anupindi, R., Chopra, S., Deshmukh, S.D., Mieghem, J.A.V., Zemel, E.: Managing Business Process Flows (3rd ed.). Prentice Hall (2012)
37. Dresch, A., Lacerda, D.P., Antunes Jr, J.A.V.: Design Science Research: A Method for Science and Technology Advancement. Springer (2014)
38. Pries-Heje, J., Baskerville, R.: The design theory nexus. MIS Quarterly (2008) 731–755
39. Von Alan, R.H., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly **28**(1) (2004) 75–105
40. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage Learning (2002)
41. Nguyen, H., Dumas, M., ter Hofstede, A.H., La Rosa, M., Maggi, F.M.: Business process performance mining with staged process flows. In: Proc. of CAiSE, Springer (2016)
42. Verbeek, H., van der Aalst, W.M., Munoz-Gama, J.: Divide and conquer. Technical report, BPM Center Report Series (2016)

43. Reijers, H.A., Mendling, J., Dijkman, R.M.: Human and automatic modularizations of process models to enhance their comprehension. Inf. Syst. **36**(5) (2011) 881–897
44. Modig, N., Ahlström, P.: This is lean: Resolving the efficiency paradox. Rheologica (2012)
45. Hornix, P.T.: Performance analysis of business processes through process mining. Master's thesis, Eindhoven University of Technology (2007)
46. Gunther, C.W., Rozinat, A.: Disco: Discover your processes. In: Proc. of BPM Demos. Volume 940 of CEUR Workshop Proceedings. (2012) 40–44