Stochastic Process Mining

Tian Li[1,2,*,†]

[1]The University of Melbourne, Australia
[2]RWTH Aachen University, Germany

Abstract

Process mining extracts event logs from information systems to derive insights into organizational business processes. Stochastic process mining specifically emphasizes techniques that integrate the frequency and probability of various process behaviors, allowing organizations to comprehend their business processes by differentiating between routine and exceptional occurrences. The ambition of my Ph.D. project is to improve state-of-the-art stochastic process discovery techniques and provide novel measures applicable to stochastic conformance checking.

Keywords

Stochastic Process Mining, Stochastic Process Discovery, Stochastic Conformance Checking

1. Introduction

Information systems in modern organizations continuously track the processes executed by employees, managers, and customers, generating large amounts of event data. Such data can be extracted as an event log, which is a collection of traces, where each trace is a sequence of activities recorded during the execution of a process. By analyzing the event logs, process mining provides systematic approaches to understanding, monitoring, and optimizing real-world processes.

Stochastic process mining specifically focuses on techniques that incorporate the frequency and probability of different process behavior. This perspective is crucial for organizations because it enables them to differentiate between routine operations and exceptional cases. Without this capability, business analysts risk misallocating resources by directing attention to rare behaviors that have a minimal impact on overall process performance [1].

The contributions of this project will be threefold. First, we investigate alternative formalisms to model the stochastic perspective of the process. Second, we propose novel discovery techniques to construct stochastic models from event data. Third, we propose measures for stochastic conformance checking. Section 2 introduces research questions and proposes solutions. Subsequently, Section 3 outlines the progress so far and future plans. Finally, Section 4 presents the outlook.

2. Research Background and Problems

This section serves to motivate and formalize the research questions for my Ph.D. project.

2.1. Research Question One

Two types of stochastic process modeling formalism have been proposed for stochastic process modeling: action graph-based models and Petri net-based models. The first type is the stochastic action graph introduced by Alkhammash et al. [2], which is similar to the directed graphs adopted by practitioners. The nodes and arcs in the model are annotated with numbers that reflect the frequencies of the actions

Doctoral Consortium co-located with 23rd International Conference on Business Process Management (BPM 2025), Seville, Spain, 31st Aug to 5th Sep 2025.

*Corresponding author.
tian.li.2@unimelb.edu.au (T. Li)
0000-0003-1288-3149 (T. Li)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4032/paper-16.pdf

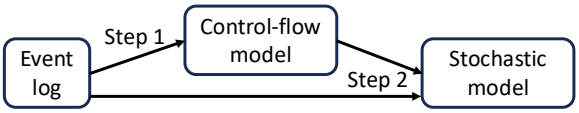

Figure 1: Two-stage approach.

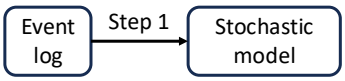

Figure 2: One-stage approach.

and "can follow" dependencies inferred from the event log. The second type consists of Petri nets with stochastic extensions. For instance, Leemans et al. [3] introduced the formalization of generalized stochastic labeled Petri nets, including silent transitions and annotated immediate transitions with weights. The probability of firing an immediate transition is determined by the relative weights of all enabled immediate transitions.

Although these formalisms have been studied thoroughly, other process modeling languages, such as Business Process Model and Notation (BPMN) and causal nets, can also be extended with the stochastic perspective. In practice, BPMN is a set of diagramming conventions used to describe business processes. Causal nets are a declarative process modeling formalism that relies on a small number of modeling constructs, but is expressive. Consequently, the first research question (RQ1) is formalized as: *What is the appropriate modeling formalism to model the stochastic perspective of the process?*

Proposed solution Business Process Model and Notation (BPMN) models are widely used by practitioners for decision making. Causal nets (C-nets) are used in multiple process discovery techniques. Many Markov models, such as Markov chains, Markov decision processes, or semi-Markov decision processes, have different characteristics. Similarly, transition systems have also been enhanced with probabilistic information and have been used for prediction. We propose a comparison study to justify the benefits of the selected stochastic process models compared to other stochastic process modeling languages.

2.2. Research Question Two

Although conventional process discovery methods excel in many areas, they have the limitation of ignoring the frequencies of the traces recorded in the event logs in the constructed models. *Stochastic process discovery* techniques mine models that pair traces with indications on how likely one can expect to see them in future executions of the process. Most techniques achieve this through an indirect method, which involves a conventional process discovery technique to construct the process flow and then add probability weights based on how often each trace appears, as illustrated in Fig. 1.

The technique in [4] is the first two-stage discovery framework to discover generalized stochastic Petri nets with timed transitions, which allow performance analysis. [5] introduced several weight estimators based on statistics computed on log and model. In 2024, two other algorithms were proposed to perform the discovery of stochastic processes with optimal stochastic quality guarantee [6, 7].

At the start of my Ph.D., the second research question (RQ2) was formalized as: *Given an event log and a process model that describes the control flow of the process observed in the event log, how do we construct a stochastic process model that maintains the same control flow while being capable of reproducing the probability of the observed process?*

Proposed solution We propose to define two-stage stochastic process discovery as finding a model with an optimal stochastic conformance checking measure over a given representation bias. Our strategy is to turn the given control flow model into a stochastic process model that assigns a weight parameter to every transition. Then, stochastic discovery is posed as an optimization problem, where values for the weights must be found so that a stochastic conformance measure is maximized.

2.3. Research Question Three

The one-stage techniques operate without relying on an initial control-flow model, but calculate control-flow and stochastic aspects simultaneously, as illustrated in Fig. 2. Toothpaste Miner [8] is the first

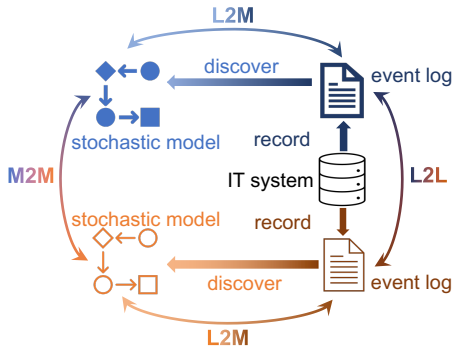

Figure 3: The scenario of L2L, L2M, and M2M stochastic conformance.

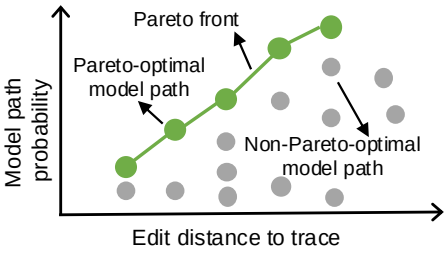

Figure 4: Pareto front for model paths.

single-stage technique, which applies a set of reduction and abstraction rules to generate a stochastic labeled Petri net (SLPN). Another technique *GASPD* is based on grammatical inference [2], which discovers a family of direct action graphs from an input event log.

However, these two one-stage approaches do not guarantee the stochastic quality of the constructed models. Thus, our third research question (RQ3) is: *Given an event log, how to directly construct a stochastic process model of manageable size while capable of reproducing the probability of the observed process in the event log?*

Proposed solution We plan to address inherently competing objectives: Reducing the complexity of the model while maintaining its stochastic quality. Thus, the one-stage discovery is addressed through a trade-off between model complexity and accuracy.

2.4. Research Question Four

Another core problem studied in process mining is conformance checking, which quantifies how much a process model agrees with an event log. The technique that ignores the stochastic perspective of processes can be misleading. For instance, non-stochastic-aware conformance checking cannot distinguish the discrepancy between event log $[\langle a, b\rangle^{50}, \langle b, a\rangle^{50}]$ and a stochastic process model that describes the stochastic language $[\langle a, b\rangle^{0.9}, \langle b, a\rangle^{0.1}]$. However, the model emphasizes that step b should occur after a more frequently than a after b, while the log suggests that both orders are equally likely.

Beyond regular Log-to-Model (L2M) scenarios, Model-to-Model (M2M) and Log-to-Log (L2L) conformance checking also benefit from a stochastic perspective. In many real-life scenarios, processes are influenced by internal or external requirements and change over time. Consequently, a stochastic process model designed to describe the behavior of the system can become outdated as time progresses. One can detect and quantify changes in stochastic behavior by comparing the latest discovered model with the original model using M2M stochastic conformance. Similarly, event logs that cover long periods or merge data from multiple organizations may contain different versions of the process behavior. Conclusions drawn from such logs may be misleading or biased when addressing specific regional or temporal issues. We illustrate these scenarios of stochastic conformance checking in Fig. 3.

In light of this, we establish our fourth research question (RQ4) as: *How to quantify stochastic conformance checking for log-to-log, log-to-model, and model-to-model conformance scenarios?*

Proposed solution In essence, an event log can be considered as a finite sample drawn from a probability distribution over traces, while stochastic process models represent probability distributions over traces. We propose adapting established statistical distances between probability distributions for stochastic conformance checking. Specifically, we review interesting features of statistical distances and discuss their use in the context of L2L, L2M, and M2M conformance checking. To expand the applicability of the distances, we propose the necessary adaptations.

2.5. Research Question Five

Although existing techniques [9, 10, 11, 12] quantify stochastic conformance by computing a numerical value between a stochastic model and an event log, they are not applicable if an aggregated event log is not available. Moreover, a single trace is not explicitly matched to the model.

Conformance checking techniques, such as alignments, identify a path allowed by the model with as few deviations as possible from an observed trace. However, when considering a stochastic perspective, if the selected path is unlikely according to the stochastic process model, it may not be the most likely explanation of the path through the model. As a consequence, further diagnostics are based on process behavior that is less relevant to be followed by design. For instance, a path with a probability of 10% according to the model and an edit distance of 3 to the trace may be a better match than a path with a probability of 0.1% and an edit distance of 2. The scenario highlights two possibly competing objectives when matching the trace to the stochastic model of the process: the probability of the selected path allowed by the model and its edit distance to the trace.

Therefore, we formalize our fifth research question (RQ5) as: *Given an observed trace, how to match it to a stochastic process model by identifying a likely model path with a low edit distance to the trace?*

Proposed solution The trade-off between the probability of the selected path allowed by the model and its edit distance to the trace, which we illustrate in Fig. 4. The Pareto front consists of model paths and indicates that any reduction in edit distance would necessarily decrease the model path probability, and any increase in model path probability would necessarily increase its edit distance to the trace. We propose a stochastic alignment technique that matches a single trace to a stochastic process model and produces an alignment that balances the importance of the normative behavior (the probability of the model path) and the alignment cost (the edit distance between the model path and trace). Business analysts can explicitly weigh the trade-off with a user-defined parameter.

3. Results and Road map

We first outline the progress to date in this Ph.D. project:

- Literature review to identify research gaps for stochastic process mining.
- Study 1: Define and implement the two-stage discovery problem with an optimality guarantee. The results of this study were published in CAiSE 2024 [13].
- Study 2: Identify statistical distances for the quantification of stochastic conformance. The results of this study are currently under review.
- Study 3: Introduce stochastic alignments that account for alignment cost and the probability of the model path. The results of the study are to be presented at the BPM 2025.

To complete the Ph.D. project and thesis, my future research plan is as follows:

- Study 4: Identify alternative process modeling formalisms other than SLPNs to model the stochastic perspective of the process.
- Study 5: Design a one-stage stochastic process discovery technique, and conduct an extensive evaluation.

Studies 4 and 5 are based on the existing work achieved in studies 1, 2, and 3. The quality of the discovered stochastic process models can be evaluated with the stochastic conformance checking measures discussed in the study.

However, two challenges may prevent the project from achieving our target. First, the stochastic discovery that guarantees stochastic optimality requires solving a nonconvex optimization problem. This poses a computational challenge as there are multiple locally optimal points, and the computed result may not be globally optimal. Second, the model-to-model stochastic conformance can be hard to measure because of potentially infinite process behavior. The conformance between two stochastic models can be computed by sampling, however, this does not guarantee an exact result.

4. Outlook

This Ph.D. project will contribute to the business process management (BPM) community by developing novel techniques for stochastic process mining. We treat the probability of process behavior as a first-class citizen due to its close link to simulation, prediction, and recommendation.

We aim to consider the appropriate representation bias for stochastic process modeling. Then, we plan to explore two types of stochastic discovery algorithms, i.e., a one-stage approach that directly constructs a stochastic model from the input event log and a two-stage approach that indirectly constructs a stochastic model using the event log. Furthermore, two types of stochastic conformance checking are investigated, one is the statistical distance-based techniques that measure numerically; the other is an alignment technique that returns an explicit artifact to match a trace with the given stochastic process model.

Simultaneously, to bridge the theory-practice divide, we will create user stories that tie all our research questions together and demonstrate the practical application of our findings in real-world scenarios. The goal is to illustrate how practitioners can leverage our research to address real-life challenges. For example, we will showcase how practitioners can perform stochastic alignments to explain deviations in an observed trace using the discovered stochastic process models.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. J. J. Leemans, A. Polyvyanyy, Stochastic-aware precision and recall measures for conformance checking in process mining, Inf. Syst. 115 (2023) 102197.

[2] H. Alkhammash, A. Polyvyanyy, A. Moffat, Stochastic directly-follows process discovery using grammatical inference, in: CAiSE, volume 14663 of *LNCS*, Springer, 2024, pp. 87–103.

[3] S. J. J. Leemans, A. F. Syring, W. M. P. van der Aalst, Earth movers' stochastic conformance checking, in: BPM Forum, volume 360 of *LNBIP*, Springer, 2019, pp. 127–143.

[4] A. Rogge-Solti, W. M. P. van der Aalst, M. Weske, Discovering stochastic petri nets with arbitrary delay distributions from event logs, in: BPM Workshops, volume 171 of *LNBIP*, Springer, 2013, pp. 15–27.

[5] A. Burke, S. J. J. Leemans, M. T. Wynn, Stochastic process discovery by weight estimation, in: ICPM Workshops, volume 406 of *LNBIP*, Springer, 2020, pp. 260–272.

[6] T. Brockhoff, M. S. Uysal, W. M. P. van der Aalst, Wasserstein weight estimation for stochastic petri nets, in: ICPM, IEEE, 2024, pp. 81–88.

[7] P. Cry, A. Horváth, P. Ballarini, P. L. Gall, A framework for optimisation based stochastic process discovery, in: QEST+FORMATS, volume 14996 of *LNCS*, Springer, 2024, pp. 34–51.

[8] A. Burke, S. J. J. Leemans, M. T. Wynn, Discovering stochastic process models by reduction and abstraction, in: Petri Nets, volume 12734 of *LNCS*, Springer, 2021, pp. 312–336.

[9] H. Alkhammash, A. Polyvyanyy, A. Moffat, L. García-Bañuelos, Entropic relevance: A mechanism for measuring stochastic process models discovered from event data, Inf. Syst. 107 (2022) 101922.

[10] S. J. J. Leemans, W. M. P. van der Aalst, T. Brockhoff, A. Polyvyanyy, Stochastic process mining: Earth movers' stochastic conformance, Inf. Syst. 102 (2021) 101724.

[11] T. Li, S. J. J. Leemans, A. Polyvyanyy, The jensen-shannon distance for stochastic conformance checking, in: ICPM Workshops, volume 533 of *LNBIP*, Springer, 2024, pp. 70–83.

[12] E. G. Rocha, S. J. J. Leemans, W. M. P. van der Aalst, Stochastic conformance checking based on expected subtrace frequency, in: ICPM, IEEE, 2024, pp. 73–80.

[13] S. J. J. Leemans, T. Li, M. Montali, A. Polyvyanyy, Stochastic process discovery: Can it be done optimally?, in: CAiSE, volume 14663 of *LNCS*, Springer, 2024, pp. 36–52.