Leveraging frequencies in event data

a pledge for stochastic process mining

Sander J.J. Leemans[1]

[1] RWTH, Aachen, Germany

Abstract

Process mining aims to obtain insights from event logs. In this extended abstract, we will show that it is useful to take the frequency perspective (that is, stochastic behaviour) into account, and will discuss several stochastic process mining techniques.

Keywords

process mining, stochastic process mining, stochastic process discovery, stochastic conformance checking

1. Process mining

Organisations run on processes: processing an order, onboarding a new hire, getting travel approval; many work performed in organisations can be considered as processes. Process mining aims to optimise these processes through event logs: records of executions of processes, typically obtained from information systems that support the processes.

Figure 1 shows an overview the context and common tasks of process mining. A process is running in an organisation, and through information systems an event log is recorded. Using a process discovery technique, a process model can be discovered. Process discovery techniques need to trade-off several potentially competing model quality aspects, such as readability and filtering noise. Ideally, a process model would be compared to the actual process, however as that is assumed to be unknown, this relation can only be theoretically proven under certain assumptions, or estimated. Rather, in practice a model should be compared to (a separate test) event log, for instance on the quality dimensions of simplicity, fitness – the fraction of behaviour in the event log that is in the process model, and precision – the fraction of behaviour of the model that was observed in the event log.

A process model expresses a set of potential traces that the model supports, and it may be difficult to fully interpret insights gained from such a model by itself. Therefore, in process mining projects, the model is typically enhanced with frequency or performance information, after which the project may continue with repeated drill-down filters, hypotheses and verification [1]. In more advanced settings, process models can be

Algorithms and Theories for the Analysis of Event Data (ATAED'22)
EMAIL: s.leemans@bpm.rwth-aachen.de (S. J.J. Leemans)
URL: https://leemans.ch/ (S. J.J. Leemans)
ORCID: 0000-0002-5201-7125 (S. J.J. Leemans)

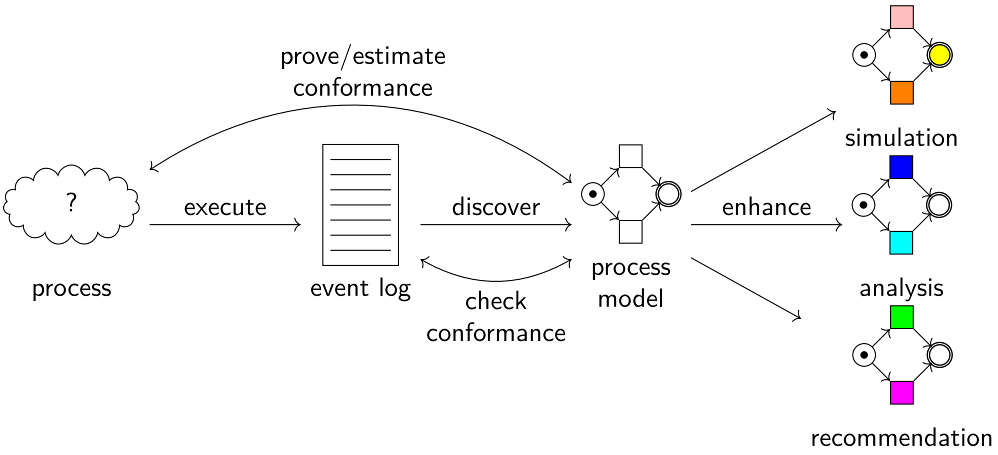

Figure 1: An overview of common tasks of process mining.

simulated. This provides a baseline for, after applying certain changes, comparing process redesigns in a what-if analysis. Finally, if process mining is integrated into daily operations, process models can be used to recommend interventions for traces that are still in the process [2].

2. Frequencies in process mining: the stochastic perspective

Let us consider two event logs:

$$
\begin{aligned}
L_1 = [&\langle \mathit{register}, \mathit{check}, \mathit{accept}\rangle^{10000}, \\
&\langle \mathit{register}, \mathit{check}, \mathit{reject}\rangle^{10000}, \\
&\langle \mathit{register}, \mathit{accept}\rangle^{1}]
\end{aligned}
\qquad
\begin{aligned}
L_2 = [&\langle \mathit{register}, \mathit{check}, \mathit{accept}\rangle^{9500}, \\
&\langle \mathit{register}, \mathit{check}, \mathit{reject}\rangle^{9500} \\
&\langle \mathit{register}, \mathit{accept}\rangle^{1001}]
\end{aligned}
$$

These logs have an equivalent control flow: the set of traces in both logs is the same. However, it is obvious that these logs are not from the same process: in L_1, $\langle \mathit{register}, \mathit{accept}\rangle$ occurs once, while in L_2 it occurs more than a thousand times. Any process mining techniques ignoring the stochastic perspective will consider these logs as to come from the Thus, these logs are different mostly because of their frequencies; in process models, we refer to this as the *stochastic perspective*.

The stochastic perspective is obviously present in a process: behaviour has a certain likelihood of appearing. Consequently, an event log derived from a process also has a stochastic perspective: behaviour has a certain likelihood of being recorded in the event log. Thus, the stochastic perspective *is there*.

On th right side of Figure 1, we need an idea of how often behaviour occurs in order to perform analysis: it matters whether behaviour is exceptional or common, and average performance measures are weighted by definition on the multiplicity of behaviour. For simulation, simulation software needs to know how likely each path or decision in the process is. Similarly, recommendation needs to be aware of how likely behaviour is in order to steer towards more likely favourable outcomes [3]. Thus, the most useful parts of process mining *need* the stochastic perspective. This leaves an obvious gap between the stochastic-having event logs and the stochastic-needing analysis, simulation and recommendation: process models with a stochastic perspective: *stochastic process models.* A stochastic process model not only expresses what behaviour can happen, but also how likely each trace is.

2.1. Analysis, simulation & recommendation

Without existing stochastic process models, existing analysis, simulation and recommendation techniques, which inherently use the stochastic perspective, must obtain this stochastic information in an ad-hoc fashion from the event log [3, 4]. Consequently, such techniques have no idea of the quality of the stochastic perspective they operate on and risk testing on their training logs, which is not good practice. Without explicit stochastic information, one cannot write it down, cannot reason about it, and adjust it in process redesign efforts.

2.2. Precision

Another area where considering stochastic process information is beneficial is in the evaluation of process models: we already discussed fitness, and most fitness measures take the stochastic information into account implicitly: the more likely behaviour in the log, the higher its influence on the fitness measure. Precision measures express the fraction of behaviour of the model that was seen in the event log:

$$\text{precision} = \frac{|\text{model} \cap \text{log}|}{|\text{model}|}$$

An inherent problem with this intuitive informal definition is that one needs a count of behaviour in the model. One cannot simply count traces, as models may express infinitely much behaviour through loops.

We illustrate this for one non-stochastic precision technique [5]. This approach considers the outgoing edges of the state space of a model: they divide the number of edges taken by the total number of edges to arrive at a number. However, these techniques do not consider at all what lies beyond edges that were seen in the model. Thus, unseen behaviour is only counted proportionally to the number of edges that go into that area, irrespective of the “size” of the unseen part [5].

For a stochastic process model, this is not a problem as we have a notion of size: in the state space, it is known exactly how likely each edge is, and that is exactly equal to the size – the probability mass – of the model that lies behind it. Thus, for stochastic process models more intuitive precision measures can be defined.

2.3. Reliability of conclusions

If we consider Figure 1 again, inaccuracies or imprecisions can be introduced at many steps of these common process mining tasks:

- When recording the event log from a process, the quality of the recording may vary, or extraction may be biased;
- When discovering a process model, a process discovery technique may need to make well-known trade-offs between potentially competing quality criteria;
- When estimating the quality of a process model with respect to the process, assumptions and bias may be tested based on a process model [6];
- When comparing a process model with an event log using a conformance checking technique, such a technique will try to squeeze a trace from the log onto the best-fitting path through the model [7]; There might be multiple such best-fitting paths, and there is no guarantee that a best-fitting path is the most likely explanation, yielding ambiguities and potentially inaccuracies;
- When enhancing a process model for analysis, simulation or recommendation, behaviour where log and model do not agree on (non-conforming parts) needs to be handled [4].

All of these steps may be sources of inaccuracies and imprecisions, which may propagate and aggregate over a process mining project. We conjecture that the use of stochastic process models makes it easier to quantify and study these inaccuracies and imprecisions, such that the reliability of conclusions can be quantified [8], and improved.

3. Existing stochastic process mining techniques

Next, we discuss some stochastic process mining techniques that mimic standard non-stochastic techniques: stochastic process discovery and stochastic conformance checking. Furthermore, we discuss completely new types of techniques that require considering stochastic process behaviour.

3.1. Stochastic process discovery

In stochastic process discovery, the aim is to automatically discover a stochastic process model – such as a stochastic labelled Petri net [14] – from an event log. Most stochastic process mining techniques take an existing non-stochastic process model and construct a stochastic process perspective on top if it. For instance, [9] constructs a stochastic perspective through time: it estimates the delay distribution of process steps, which in turn determines their likelihood. Another approach constructs a stochastic perspective for a process model using several estimators, ranging from simple counting to alignments [10].

A technique that does not start from an existing is [11], which starts from the behaviour in the event log and using reduction rules compacts, summarises and abstracts the stochastic behaviour until a suitable model remains.

A completely different approach is taken by [12], which constructs declarative constraints on the log, such that these constraints hold with a certain likelihood. As such, these models describe multiple options for stochastic behaviour, rather than a single stochastic language.

3.2. Stochastic conformance checking

A stochastic conformance checking technique compares with one another an event log and a stochastic process model, or two stochastic process models, or two event logs. Stochastic entropy [13] consists of two measures: recall is the entropy of the common behaviour of log and model – minimal number of bits required to describe the behaviour – divided by the entropy of the log. Precision is then the entropy of the common behaviour of the log and model divided by the entropy of the model.

Another stochastic conformance checking technique is the Earth Movers' Distance [14], which considers both log and model (or any other combination; loops need to be unfolded) as distributions over traces, and then applies the Wasserstein distance principle, which finds the least-cost way to transform one dstribution into the other. That is, both distributions are piles of earth, and the distance says how much earth – trace probability mass – need to be transported over what distance – trace difference – in order to transform one pile into the other. Besides a single conformance number, this measure can also provide detailed insights when projected on a model.

3.3. Goodies

Next to stochastic extensions of techniques, the concept of stochastic process behaviour has enabled some new types of analyses and techniques.

Some event logs have hundreds of activities, which make them challenging to analyse. A way to simplify models is to apply a trace-based filter, thereby focusing the analysis on, for instance, platinum customers, or orders with a value over a certain amount. Cohort analysis recommends filters based on the difference between traces that pass the filter and all the other traces: the filter that is associated with the largest difference in stochastic behaviour would simplify the model the most [15].

Most process mining insights are associational. However, as in associational insights there is no difference between cause and effect, for redesign or what-if analyses it is beneficial to perform causal reasoning. Recent studies have introduced causal reasoning: to discover causal rules from event logs [16], to discover causal relations between decisions in a process model [17], and to perform root-cause analysis [18]. All of these techniques are inherently enabled by the concept of stochastic process behaviour.

Further techniques targeting stochastic behaviour are the detection of differences in a stochastic process over time (concept drift) [19]; to discover anomalies in event logs without the use of process models [20]; and the combination of data-aware and stochastic process models [21].

4. Conclusion

Process mining is an exciting field of research. In this pledge for consideration of the stochastic perspective of process behaviour, we have shown several challenges of process mining concepts and techniques that may benefit from having a stochastic perspective. Several recent stochastic process mining techniques were discussed, both drop-in replacements for well-known process discovery and conformance checking techniques, as well as new techniques that leverage the stochastic perspective of behaviour of event logs to enable new types of analysis.

References

[1] M. L. van Eck, X. Lu, S. J. J. Leemans, W. M. P. van der Aalst, PM ^2 : A process mining project methodology, in: J. Zdravkovic, M. Kirikova, P. Johannesson (Eds.), Advanced Information Systems Engineering - 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings, volume 9097 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 297–313. URL: https://doi.org/10.1007/978-3-319-19069-3_19. doi:10.1007/978-3-319-19069-3_19.

[2] M. Shoush, M. Dumas, When to intervene? prescriptive process monitoring under uncertainty and resource constraints, CoRR abs/2206.07745 (2022). URL: https://doi.org/10.48550/arXiv.2206.07745. doi:10.48550/arXiv.2206.07745. arXiv:2206.07745.

[3] I. Verenich, M. Dumas, M. L. Rosa, H. Nguyen, Predicting process performance: A white-box approach based on process models, J. Softw. Evol. Process. 31 (2019). URL: https://doi.org/10.1002/smr.2170. doi:10.1002/smr.2170.

[4] S. J. J. Leemans, D. Fahland, W. M. P. van der Aalst, Exploring processes and deviations, in: BPM Workshops, volume 202 of *LNBIP*, 2014, pp. 304–316.

[5] A. Adriansyah, J. Munoz-Gama, J. Carmona, B. F. van Dongen, W. M. P. van der Aalst, Measuring precision of modeled behavior, Inf. Syst. E Bus. Manag. 13 (2015) 37–67. URL: https://doi.org/10.1007/s10257-014-0234-7. doi:10.1007/s10257-014-0234-7.

[6] G. Janssenswillen, B. Depaire, Towards confirmatory process discovery: Making assertions about the underlying system, Bus. Inf. Syst. Eng. 61 (2019) 713–728. URL: https://doi.org/10.1007/s12599-018-0567-8. doi:10.1007/s12599-018-0567-8.

[7] W. M. P. van der Aalst, A. Adriansyah, B. F. van Dongen, Replaying history on process models for conformance checking and performance analysis, WIREs Data Mining Knowl. Discov. 2 (2012) 182–192. URL: https://doi.org/10.1002/widm.1045. doi:10.1002/widm.1045.

[8] K. Goel, S. J. J. Leemans, N. Martin, M. T. Wynn, Quality-informed process mining: A case for standardised data quality annotations, ACM Trans. Knowl. Discov. Data 16 (2022) 97:1–97:47. URL: https://doi.org/10.1145/3511707. doi:10.1145/3511707.

[9] A. Rogge-Solti, W. M. P. van der Aalst, M. Weske, Discovering stochastic petri nets with arbitrary delay distributions from event logs, in: N. Lohmann, M. Song,

P. Wohed (Eds.), Business Process Management Workshops - BPM 2013 International Workshops, Beijing, China, August 26, 2013, Revised Papers, volume 171 of *Lecture Notes in Business Information Processing*, Springer, 2013, pp. 15–27. URL: https://doi.org/10.1007/978-3-319-06257-0_2. doi:10.1007/978-3-319-06257-0_2.

[10] A. Burke, S. J. J. Leemans, M. T. Wynn, Stochastic process discovery by weight estimation, in: S. J. J. Leemans, H. Leopold (Eds.), Process Mining Workshops - ICPM 2020 International Workshops, Padua, Italy, October 5-8, 2020, Revised Selected Papers, volume 406 of *Lecture Notes in Business Information Processing*, Springer, 2020, pp. 260–272. URL: https://doi.org/10.1007/978-3-030-72693-5_20. doi:10.1007/978-3-030-72693-5_20.

[11] A. Burke, S. J. J. Leemans, M. T. Wynn, Discovering stochastic process models by reduction and abstraction, in: D. Buchs, J. Carmona (Eds.), Application and Theory of Petri Nets and Concurrency - 42nd International Conference, PETRI NETS 2021, Virtual Event, June 23-25, 2021, Proceedings, volume 12734 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 312–336. URL: https://doi.org/10.1007/978-3-030-76983-3_16. doi:10.1007/978-3-030-76983-3_16.

[12] F. M. Maggi, M. Montali, R. Peñaloza, A. Alman, Extending temporal business constraints with uncertainty, in: D. Fahland, C. Ghidini, J. Becker, M. Dumas (Eds.), Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings, volume 12168 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 35–54. URL: https://doi.org/10.1007/978-3-030-58666-9_3. doi:10.1007/978-3-030-58666-9_3.

[13] S. J. J. Leemans, A. Polyvyanyy, Stochastic-aware conformance checking: An entropy-based approach, in: S. Dustdar, E. Yu, C. Salinesi, D. Rieu, V. Pant (Eds.), Advanced Information Systems Engineering - 32nd International Conference, CAiSE 2020, Grenoble, France, June 8-12, 2020, Proceedings, volume 12127 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 217–233. URL: https://doi.org/10.1007/978-3-030-49435-3_14. doi:10.1007/978-3-030-49435-3_14.

[14] 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. URL: https://doi.org/10.1016/j.is.2021.101724. doi:10.1016/j.is.2021.101724.

[15] S. J. J. Leemans, S. Shabaninejad, K. Goel, H. Khosravi, S. W. Sadiq, M. T. Wynn, Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining, in: G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, H. C. Mayr (Eds.), Conceptual Modeling - 39th International Conference, ER 2020, Vienna, Austria, November 3-6, 2020, Proceedings, volume 12400 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 92–102. URL: https://doi.org/10.1007/978-3-030-62522-1_7. doi:10.1007/978-3-030-62522-1_7.

[16] Z. D. Bozorgi, I. Teinemaa, M. Dumas, M. L. Rosa, A. Polyvyanyy, Process mining meets causal machine learning: Discovering causal rules from event logs, in: ICPM, IEEE, 2020, pp. 129–136.

[17] S. J. J. Leemans, N. Tax, Causal reasoning over control-flow decisions in process models, in: X. Franch, G. Poels, F. Gailly, M. Snoeck (Eds.), Advanced Information Systems Engineering - 34th International Conference, CAiSE 2022, Leuven, Belgium,

June 6-10, 2022, Proceedings, volume 13295 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 183–200. URL: https://doi.org/10.1007/978-3-031-07472-1_11. doi:10.1007/978-3-031-07472-1_11.

[18] M. S. Qafari, W. M. P. van der Aalst, Root cause analysis in process mining using structural equation models, in: BPM Workshops, volume 397 of *LNBIP*, 2020.

[19] T. Brockhoff, M. S. Uysal, W. M. P. van der Aalst, Time-aware concept drift detection using the earth mover's distance, in: B. F. van Dongen, M. Montali, M. T. Wynn (Eds.), 2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4-9, 2020, IEEE, 2020, pp. 33–40. URL: https://doi.org/10.1109/ICPM49681.2020.00016. doi:10.1109/ICPM49681.2020.00016.

[20] T. Nolle, Process learning for process autonomous anomaly correction (extended abstract), in: W. M. P. van der Aalst, R. M. Dijkman, A. Kumar, F. Leotta, F. M. Maggi, J. Mendling, B. T. Pentland, A. Senderovich, M. Sepúlveda, E. S. Asensio, M. Weske (Eds.), Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2021 co-located with 19th International Conference on Business Process Management (BPM 2021), Rome, Italy, September 6th - to - 10th, 2021, volume 2973 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 6–10. URL: http://ceur-ws.org/Vol-2973/paper_137.pdf.

[21] F. Stertz, J. Mangler, S. Rinderle-Ma, Data-driven improvement of online conformance checking, in: 24th IEEE International Enterprise Distributed Object Computing Conference, EDOC 2020, Eindhoven, The Netherlands, October 5-8, 2020, IEEE, 2020, pp. 187–196. URL: https://doi.org/10.1109/EDOC49727.2020.00031. doi:10.1109/EDOC49727.2020.00031.