Robust Discovery of Complex Process-Structures
(Extended Abstract)

Lisa Luise Mannel
Chair of Process and Data Science (PADS)
Department of Computer Science, RWTH Aachen University, Ahornstr. 55, 52074, Aachen, Germany
mannel@pads.rwth-aachen.de

***Abstract*—More and more companies gather and analyze the data generated by their processes. An often crucial step within this analysis is the discovery of the process based on the collected data. Unfortunately, there are still real-live scenarios where algorithms developed for this purpose fail to meet expectations. The research project presented in this paper develops the discovery algorithm eST-Miner, which introduces new concepts to address well-known challenges and aims to flexibly balance the strengths and weaknesses known from existing approaches.**

***Index Terms*—process discovery, Petri nets, implicit places**

I. Introduction

In process discovery, a process model is constructed aiming to reflect and summarize behavior defined by the example executions in a given event log. This task is challenging for a variety of reasons. Ideally, a discovery algorithm returns a model which is able to produce the important behavior represented within the event log (fitness) and at the same time does not allow for much unobserved behavior (precision), while remaining simple enough to be understood by a human interpreter. In real life logs, noise (errors, deviations) often further complicates this task. It is rarely possible to fulfill all requirements simultaneously. Based on the capabilities and focus of the used algorithm, the discovered models can vary greatly, and different trade-offs are possible.

II. Research Goal

The presented research project aims to develop a discovery algorithm that is able to reasonably balance the different quality aspects described above and additionally provides the flexibility to the users to skew the algorithm's behavior to fit their needs. In particular, it focuses on the discovery of models with high precision and guaranteed minimal fitness, which are not overfitting to include noise but still represent well-defined infrequent behavior. Additionally, the algorithm should maintain reasonable model simplicity and computational efficiency. Figure 1 shows an example of a model, which the developed discovery algorithm should return based on the given log, even if noise were added.

To achieve this overall research goal, several subgoals and desired properties have been defined as shown in Table I. In the following these goals will be explained and motivated. Section III summarizes the results achieved so far, with in-depth explanations deferred to the corresponding papers. Future research is described in more detail in Section IV. Finally, Section V positions the research project with respect to the state of the art and concludes the paper.

TABLE I
Overview of the desired properties of the developed algorithm

Representational Bias	**Frequencies & Noise**	**Guarantees**
Precise Petri nets: 1) long-term dependencies 2) self-loops 3) silent transitions (skips, loops, etc) 4) duplicate transitions (counting, process variations)	*Robustness:* 1) filter out noise (errors, deviations) 2) retain infrequent behaviour (rare but well-defined)	*Reliability:* 1) fitness 2) precision 3) rediscoverability

The research methodology is guided by the principles of Design Science Research [1]. For each subgoal, the developed solutions are formalized and implemented. Theoretical proofs of correctness and applicability are complemented by practical experiments based on adequate artificial examples as well as real-life data. Finally, the results are compared to existing algorithms to evaluate the contribution.

III. Completed Research

To allow for accurate process models, the chosen representational bias (Table I) includes *long-term dependencies*, as illustrated by the places p_9 and p_{10} in Figure 1. The basic idea of the eST-Miner presented in [2] is to evaluate all possible places in an efficient way using a noise parameter τ, and to return the places fitting a fraction of traces based on τ. Therefore, it can reliably find frequent dependencies, including long-term dependencies, *without being hampered by noise*. Heuristic approaches based on log characteristics can be utilized to immensely increase efficiency further, with very minor drawbacks to fitness and precision [3]. A variant of the algorithm discovers the novel class of *uniwired Petri nets* [3], further boosting efficiency and simplifying the returned models while still being able to discover long-term dependencies. Finally, to retain only meaningful places, the eST-Miner removes *implicit places*, i.e., places that do not further restrict the models behavior. The novel approach introduced in [4] uses the even log as an additional source of information.

I would like to thank my doctoral advisor Prof. Wil van der Aalst for his constant support. We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research.

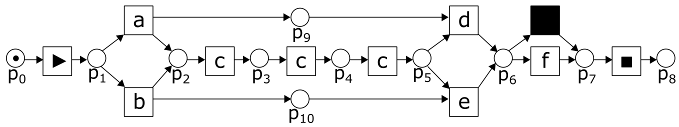

Fig. 1. The model above can be considered a desirable result for the event log $L = [\langle \blacktriangleright, a, c, c, c, d, \blacksquare \rangle^{104}, \langle \blacktriangleright\ b, c, c, c, e, \blacksquare \rangle^{90}, \langle \blacktriangleright\ a, c, c, c, d, f, \blacksquare \rangle^{17}, \langle \blacktriangleright b, c, c, c, e, f, \blacksquare \rangle^{21}]$, but is hard to discover for existing algorithms.

IV. Planned Research

Considering the research goals in Table I, several directions remain open for investigation. The present lack of *silent and duplicate transition labels* in the returned models severely hampers the attainability of high precision values. Certain common log structures, some of which are illustrated in Figure 1, cannot be expressed by Petri nets restricted to uniquely labeled transitions without resorting to flowerlike submodels. Examples are activities that can either happen once or be skipped (f), or activities that happen at least once but can be repeated, activities that happen a certain number of times (c), or logs that include data on very dissimilar subprocesses that would result in completely different submodels. The identification of such patterns and their inclusion in the model by the adequate use of non-unique transitions labels is bound to significantly increase the models precision.

Noise, and its differentiation from infrequent behavior, is often an issue in real life logs. The eST-Miner aims to achieve *robustness to noise* by requiring a place to replay only a certain fraction of traces to be considered fitting. This allows the algorithm to disregard patternless infrequent behavior while rare, yet well-defined, behavior is still discovered. An issue caused by this approach is that the traces replayable by the model are limited to the intersection of the traces replayable by all places. To increase the number of replayable traces without jeopardizing simplicity or precision, the targeted insertion of *duplicate or silent transitions* will be investigated.

The novel *implicit place removal strategy* introduced in [4] works well for places perfectly fitting the event log. Worth further investigation is its combination with the eST-Miner's noise handling approach or, in a more general setting, with Petri nets that do not perfectly fit the log. In this context its use for repairing a given model or adapting a model within an online discovery setting would be of interest as well.

The algorithm allows for a large variety of optional extensions and applications, some of which use *heuristics* to improve the quality of the returned model or to increase efficiency. Such approaches usually employ some kind of ranking on the activities or candidate places. Prior research has focused on simple rankings based on the directly and eventually follows relations for candidate places and the occurrence patterns of transitions [3], [5]. Besides refining these basic approaches, the use of alternative concepts can be investigated, which would be reusable for several extensions of the eST-Miner.

The approach can be adapted towards *new application areas*. One such option would be the online setting, where the model is build and adapted to comply with traces that become available over time. Another possible extension would be the evaluation of places with respect to a partial orders rather than single traces [6], which could be applied, for example, in the context of process discovery on uncertain event data [7].

V. Related Work and Summary

The combination of properties and capabilities listed in Table I and detailed in the previous sections set the eST-Miner apart from existing discovery approaches. Algorithms based on simple log abstractions, like the well-known Inductive Miner [8] and Alpha Miner [9], lack the ability to discover complex control-flow structures and are thus limited in their precision. Region-based approaches [10], [11] guarantee high precision only for models with unique transition labels, and have issues with overfitting as well as time and space requirements. Heuristic and genetic approaches do not provide guarantees and returned models are often hard to interpret.

The eST-Miner combines the ability to reliably discover long-term dependencies with unique noise handling capabilities and can provide guarantees with respect to fitness, precision and rediscoverability. Future projects will investigate the targeted addition of silent and duplicate transitions to improve fitness and precision, optimize the heuristic approaches and explore the adaption of the concept to other application areas. Results will be evaluated theoretically and experimentally with respect to the formulated goals as well as previous solutions. The existing extensions, variants and parameters allow to flexibly tailor the basic concept of the eST-Miner towards the unique needs of each applicant. The developed ideas and concepts can be valuable in other contexts as well, and thus contribute to the field of process mining as a whole.

References

[1] A. Hevner, S. March, J. Park, and S. Ram, "Design science in information systems research," *Management Information Systems Quarterly*, vol. 28, pp. 75–, 03 2004.

[2] L. L. Mannel and W. M. P. van der Aalst, "Finding complex process-structures by exploiting the token-game," in *Application and Theory of Petri Nets and Concurrency*. Springer Nature Switzerland AG, 2019.

[3] L. L. Mannel, Y. Epstein, and W. M. P. van der Aalst, *Improving the State-Space Traversal of the eST-Miner by Exploiting Underlying Log Structures*, 01 2020, pp. 334–347.

[4] L. L. Mannel, R. Bergenthum, and W. M. P. van der Aalst, "Removing implicit places using regions for process discovery," in *Proceedings of the International Workshop on Algorithms & Theories for the Analysis of Event Data (ATAED) 2020*, vol. 2625. CEUR-WS.org, pp. 20–32.

[5] L. L. Mannel and W. M. P. van der Aalst, "Finding uniwired Petri nets using eST-miner," in *Business Process Management Workshops*, 2019.

[6] R. Bergenthum, "Firing partial orders in a petri net," in *Application and Theory of Petri Nets and Concurrency*. Springer Int. Publishing, 2021.

[7] M. Pegoraro, M. S. Uysal, and W. M. P. van der Aalst, "Discovering Process Models from Uncertain Event Data," Sep 2019. [Online]. Available: https://publications.rwth-aachen.de/record/782564

[8] S. Leemans, D. Fahland, and W. van der Aalst, "Discovering block-structured process models from event logs - a constructive approach," *Application and Theory of Petri Nets and Concurrency*, 2013.

[9] E. Badouel, "On the α-reconstructibility of workflow nets," in *Application and Theory of Petri Nets*. Springer Berlin Heidelberg, 2012.

[10] E. Badouel, L. Bernardinello, and P. Darondeau, *Petri Net Synthesis*, ser. Text in Theoretical Computer Science, EATCS Series. Springer, 2015.

[11] J. M. van der Werf, B. van Dongen, C. Hurkens, and A. Serebrenik, "Process discovery using integer linear programming," in *Applications and Theory of Petri Nets*. Berlin, Heidelberg: Springer, 2008.