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 algo-
rithms 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 main-
tain 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.

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 repre-
sentational 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 $\tau$
and to return the places fitting a fraction of traces based
on $\tau$. 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 unwired Petri nets [3],
further boosting efficiency and simplifying the returned models
while still being able to discover long-term dependencies. Fi-
nally, 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.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0
International (CC BY 4.0).

---

**TABLE I**

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

---

Robustness:

- Filter out noise (errors, deviations)
- Retain infrequent behaviour (rare but well-defined)

Reliability:

- Fitness
- Precision
- Rediscoverability

**TABLE II**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Precise Petri nets</th>
<th>Robustness</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitness</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Precision</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Reliability</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

---

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0
International (CC BY 4.0).
Fig. 1. The model above can be considered a desirable result for the event log $L = \{a, c, c, c, d, \}^{104}, \{b, c, c, c, e, f\}^{90}, \{a, c, c, d, f, \}^{17}, \{b, c, c, e, f, \}^{21}\}$, 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 ($e$), 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


Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).