Unearthing the Real Process Behind the Event Data: The Case for Increased Process Realism (Extended Abstract)

Gert Janssenswillen[1]

UHasselt - Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium
gert.janssenswillen@uhasselt.be

Abstract. Companies in the 21st century possess a large amount of data about their products, customers and transactions. The increase in available event data gave rise to process mining, a discipline that focuses on extracting insights about processes from event logs. However, correctly displaying business processes is not a trivial task. The concept of process realism is introduced in this dissertation — stressing the need for reliable process analysis results for evidence-based decision making — which is approached from two angles. Firstly, quality dimensions and measures for process discovery are analysed on a large scale and compared with each other on the basis of empirical experiments. Secondly, by developing a transparent and extensible tool-set, a framework is offered to analyse process data from different perspectives. Exploratory and descriptive analysis of process data and testing of hypotheses again leads to increased process realism. Based on both approaches, recommendations are made for future research, and a call is made to give the *process realism* mindset a central place within process mining analyses.

Keywords: Process mining · Event data · Conformance Checking

1 Introduction

In current times, organisations possess a tremendous amount of data concerning their customers, products and processes. Many activities which are taking place in their operational processes are being recorded in event logs [2]. Techniques from the process mining field, which has grown steadily over the last decades, can be applied to gain insights into these event data [1]. Over the past decade, a lot of attention has been given to the discovery of process models from logs [4,5], and the quality measurement of these models [3,6,7].

The results of process mining analyses, if acted upon, can have important ramifications for business operations in two ways. Firstly, improvements to the performance of processes. Performance — or a lack thereof — can be expressed in many different manners, such as the time spent on the process or the incurred operational costs. Secondly, improvements in compliance with rules and regulations, whether imposed internally in organisations or by (inter)national laws, are important to prevent fraud and other types of risk. Both of these aspects strongly rely on the ability to accurately delineate the process and all its relevant characteristics based on the process data that has been extracted from the organisation's information systems.

This dissertation aims to contribute from several angles related to this accurate representation of processes, introducing the notion of the concept *process realism*. Process realism can be defined as *the interest or concern for the actual or real process, as distinguished from the abstract, speculative, etc.,* or, *the tendency to view or represent processes as they really are.* In order to optimise processes, evidence-based decision making is needed. Consequently, it is essential to map these processes in a realistic way. Blindly relying on both partial and/or inconsistent data and on algorithms can lead to wrong actions being taken.

Process realism is approached from two perspectives. First, quality dimensions and measures for process discovery results are analysed on a large scale and compared with each other on the basis of empirical experiments (Part II of the thesis). Which measures are best suited to assess the quality of a discovered process model? What are their weaknesses and strengths? Which challenges still need to be overcome in order to evolve towards reliable quality measurement? The results of these experiments are discussed in Section 2.

In addition to the focus on process models, process realism is also approached from a data point of view. By developing a transparent and extensible tool-set, a framework is offered to analyse process data from different perspectives (Part III of the thesis). Exploratory and descriptive analysis of process data and testing of hypotheses again leads to increased process realism. This led to the creation of `bupaR`, an open-source software suite for process analysis in R. This part of the dissertation is further discussed in Section 3.

2 Process Model Quality

When looking at the results of a process discovery algorithm, the outcome is often too easily (mis)taken for absolute truth about the underlying process. However, the fact that it was discovered from a sample of event data, which probably also contains measurement errors, tells us that this is not necessarily the case. Simultaneously, it is not a reliable representation of the original event data either, because of the filters and other choices and assumptions imposed by the discovery algorithm used. As such, awareness about whether you are describing the event data, making assertions about the underlying process, or an ambiguous mix of both is currently missing.

Being able to accurately quantify the quality of discovered process models, which is an important component of conformance checking, is critical for process discovery. Only through accurate quality measurement can the trustworthiness of discovered process models be assessed, to see whether the insights they deliver are reliable. It is crucial to know whether a discovered process model is a precise and fitting representation of the event data or the underlying process. Many quality measures — fitness, precision and generalization — have been developed over the past years, but they have so far only been evaluated narrowly on how they compare to each other. Moreover, it is not clear how to interpret or combine different dimensions such as precision and generalization. The research objective related to process model quality is therefore twofold:

1. examine the measures in terms of validity, sensitivity and feasibility, and
2. analyse their ability to quantify the quality of the model as a representation of the underlying process, i.e. the system.

A summary of the results of Part II is shown in Table 1. The *unbiased estimator* column for fitness and precision measures refers to the ability of measures to act as an unbiased estimator of the correspondence between the process model and the underlying process (i.e, the system), when applied between a log and a model. We refer to this as system-fitness (system-precision) and log-fitness (log-precision), respectively. For generalization measures, *unbiased estimator* refers to the ability of generalization measures to unbiasedly estimate system-fitness, which is how generalization is most often defined.[1]

Table 1. Summary of results. F = Fitness, P = Precision, G = Generalization, ab = Alignment-Based, ne = Behavioural, tb = Token-Based, ba = Best Align, oa = One Align.

Measure	**Feasibility**	**Validity**	**Sensitivity**	**Unbiased estimator**
F ab	✗	✓	✓	⚠
ne	✓	✓	✗	⚠
tb	⚠	✓	✓	⚠
P ab	⚠	⚠	⚠	⚠
ne	✓	✓	✓	✗
ba	✗	✓	✓	✗
oa	⚠	✓	⚠	✗
G ab	✓	✗	✗	✗
ne	✓	✗	✓	✗

The experiments of Part II and their conclusions indicate several important challenges to be tackled by future research. Challenges related to process quality measurement itself — e.g. how to estimate system-quality in an unbiased manner? — as well as to the experimental set-up for the empirical evaluation — which types of models or how many systems to generate are questions in these kind of experiments to which current literature does not provide an answer.

3 Process Analytics

A process model, notwithstanding how superior in quality it might be, will always make abstraction of certain information — such as information on resources,

[1] Note that, in contrast to fitness and precision, there is no universally agreed-upon definition of generalization.

time, or other attributes — thereby partly sacrificing the realism one has about the process. While a model can indicate certain surprising or interesting patterns with regards to the process, the practitioner will want to have a means to further investigate this pattern, to understand why and how it came about, before he can decide whether — and which — corrective actions are required to improve the performance or compliance of the process.

As such, this dissertation is also motivated by the necessity a tool-set to analyse process data in a flexible and powerful way, able to focus on very specific segments or perspectives of the processes. Important in this respect is the capability to use proven data analytics techniques — from statistics to contemporary data mining tools — in order to truly unravel these patterns and confirm their reality. While many developments with respect to process analysis tools have been already made, important limitations can still be found which prevent these type of flexible and transparent inquiries.

The contribution of the second part of this dissertation is therefore the development of a tool-set, answering to specific requirements which are identified based on the inventory of state-of-the-art tools, both of open-source and commercial nature. In particular, the following characteristics are considered:

- **Flexibility** — the ability of the tool to analyse multiple perspectives of the process besides the omnipresent focus on control-flow. Also non-standard case and event attributes should receive their place in the analysis of the process.
- **Connectivity** — the ability to use existing tools and techniques. Being connected with these existing functionalities will prevent that process analysis will end isolated from the advances in the broader data science field.
- **Transparency** — abolishing the often obscuring characteristics of process analytics tools, such as hidden assumptions and ambiguous, behind-the-scenes pre- or post-processing steps. In order to bring about process realism, the tool should clearly document the workings of all the functionalities and allow for reproducible work-flows.

Based on these aspects, the framework `bupaR` is introduced. `bupaR` is an extensible set of R-packages for business process analysis, developed in order to support flexible, reproducible and extensible process analytics. As an evaluation of the framework, it is applied to two case studies. First, how can we use process data to better understand students' study trajectories and to better guide students? Secondly, how can we apply process analysis in a railway context, in order to achieve a smoother service for passengers?

Both case studies show that the framework clearly has added value, and that the answers to the questions asked can help to improve the processes under consideration. At the same time, unresolved challenges within process mining are also emphasised, such as the analysis of processes at the right level of granularity, and the assumption that process instances are independent of each other.

4 Conclusions

From both perspectives, process model and process data, recommendations are made for future research, and a call is made to give the *process realism* mindset a central place within process mining analyses. The research objective of the first part, with a focus on process model quality, was to analyse quality measures to examine their usefulness in terms of validity, sensitivity and feasibility, as well as their ability to quantify the quality of the model as a representation of the underlying process. So far, little research has been done concerning the evaluation and comparison of quality measures. The empirical analyses done in this dissertation elucidate this poor comprehension.

Secondly, reproducible analysis is an important requirement for tools in current times. It is relevant for both academics — allowing them to reproduce experiments — as for industry — to rerun analyses on new data, or using new assumptions. Reproducibility can be obtained using different approaches. One is to support the creation of graphical work-flows, such as RapidMiner and other graphical data analysis environments do. Another approach is through scripting, which was the approach taken by `bupaR`. Because of investments in documentation, tutorials, examples and a website, as well as through a straightforward API design, many actions have been taken to make `bupaR` as accessible as possible, thereby helping to spread the adoption of process mining to a broader audience.

References

1. van der Aalst, W.M.P., Reijers, H.A., Weijters, A.J., van Dongen, B.F., De Medeiros, A.A., Song, M., Verbeek, H.M.W.: Business process mining: An industrial application. Information Systems 32(5), 713–732 (2007)
2. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow mining: Discovering process models from event logs. Knowledge and Data Engineering, IEEE Transactions on 16(9), 1128–1142 (2004)
3. Adriansyah, A., Munoz-Gama, J., Carmona, J., van Dongen, B.F., van der Aalst, W.M.P.: Alignment based precision checking. In: Business Process Management Workshops. pp. 137–149. Springer (2012)
4. Garcia-Banuelos, L., Dumas, M., La Rosa, M., De Weerdt, J., Ekanayake, C.C.: Controlled automated discovery of collections of business process models. Information Systems 46, 85–101 (2014)
5. Leemans, S.J., Fahland, D., van der Aalst, W.M.P.: Discovering block-structured process models from event logs: A constructive approach. In: International conference on applications and theory of Petri nets and concurrency, pp. 311–329 (2013)
6. de Leoni, M., Maggi, F.M., van der Aalst, W.M.P.: An alignment-based framework to check the conformance of declarative process models and to preprocess event-log data. Information Systems 47, 258–277 (2015)
7. Senderovich, A., Weidlich, M., Yedidsion, L., Gal, A., Mandelbaum, A., Kadish, S., Bunnell, C.A.: Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. Information Systems 62 (2016)