Quantifying Uncertainty for Explainable Process Mining (PhD Proposal)

Arvid Lepsien[1]

[1]Kiel University, Group Process Analytics, Hermann-Rodewald-Str. 3, 24118 Kiel, Germany

Abstract

Process mining has proven its usefulness in a wide range of practical applications, however, it generally requires event logs to be certain to guarantee trustworthy results. Typical approaches to uncertainty in event logs, like frequency-based filtering or other heuristics, restrict insight into processes because they possibly discard relevant information. Several alternative approaches to uncertainty in process mining have been proposed to provide more detailed insights, but these approaches each only address a single type of uncertainty. A great number of domains could benefit from process mining techniques that can handle the simultaneous occurrence of multiple types of uncertainty, e.g., probabilistic processes where event logs are extracted from unstructured data. The proposed PhD project aims to develop a holistic approach to uncertainty in process mining, adding a comprehensive perspective of uncertainty to the insights generated by process mining analyses. To achieve this, methods concerned with different types of uncertainty in process mining, namely data, correlation, and process uncertainty, will be investigated, and then combined into a harmonized framework, providing a foundation for improved decision-making.

Keywords

Process mining, Uncertainty, Root cause analysis, Unstructured data

1. Motivation

Process mining has proven its usefulness in a wide range of practical applications in order to uncover bottlenecks and inefficiencies in processes or to identify tasks for automation [1]. One future avenue for process mining should be to increase the *trustworthiness* of its results, i.e., to develop methods to provide confidence and trust in its automatically generated insights to end users [2]. In general, process mining requires event logs that are accurate in terms that the recorded events actually happened and the attributes of events were recorded correctly [3]. While this assumption might be appropriate for some settings (e.g., when event logs are extracted from process-aware information systems), it is more challenging to comply with this assumption in other settings, leading to reduced result trustworthiness. For instance, event logs sourced from unstructured data and unstructured processes can only indicate likelihoods when mapping low-level events onto activities and cases [4]. Therefore, to increase trustworthiness in process mining applications, the challenge to deal with is *uncertainty*.

13th International Workshop on Enterprise Modeling and Information Systems Architectures (EMISA), May 11-12, 2023, Stockholm, Sweden

ale@informatik.uni-kiel.de (A. Lepsien)
0000-0002-8105-382X (A. Lepsien)

Generally, three different types of uncertainty exist for processes discovered from unstructured data, namely data uncertainty, correlation uncertainty, and process uncertainty. Data uncertainty refers to the degree of noise in the data like inaccurate, imprecise, untrustworthy, and unknown data. Correlation uncertainty refers to the likelihood of event-activity mappings since there is often more than one possible solution. Then, process mining can also be affected by the uncertainty inherent in the analyzed process (i.e., probabilistic dependencies and contextual influences affected by randomness).

The reduction and quantification of uncertainty might improve process discovery results, improve conformance checking and predictive monitoring and even elevate process discovery techniques on unstructured data. The purpose of this PhD project is to develop a holistic framework to address uncertainty in process mining, especially process mining applied to unstructured data. This includes quantifying the impact of uncertainty, uncovering the sources of uncertainty in event log extraction, quantifying the random factors of uncertainty when mapping events onto activities, and providing a user-friendly communication of the uncertainty-aware process mining methods.

2. Related Work

In most process mining approaches, uncertainty is treated as an issue of event log quality, and addressed with frequency-based filtering or other heuristics. While this is a practical approach to reduce noise (i.e., erroneous recordings), it may also suppress outliers (i.e., correctly recorded, but unexpected behavior), which are highly relevant to analyze process deviations [5].

Recently, some approaches have been suggested that integrate uncertainty into process mining, instead of disregarding uncertain event data. Pegoraro [6] proposed a framework that adds a perspective on data uncertainty to process mining. An event log is annotated with (meta-)information related to the uncertainty of the events contained in an event log, which is used as additional input to uncertainty-aware algorithms for process discovery and conformance checking. Also, process models are annotated to represent the uncertainty of the event log they were discovered from. Qafari et al. [7] proposed an approach to identify the causes of process performance and compliance problems from event logs. Structural causal models are used to discover causal relations between distinct features (e.g., event or case attributes, the occurrence of an activity) and problematic process outcomes. Leemans et al. [8] developed a method to identify long-term dependencies between control-flow decisions in a process. Control-flow decisions are identified from a process model and event log of this process, then probabilistic causalities between control-flow decisions at the decision points are discovered, and finally, the size of each causal effect is estimated. Alman et al. [9] proposed a framework to extend declarative process mining methods with a process uncertainty perspective.

To sum up, no approach exists to quantify different types of uncertainty in process mining. Mostly, existing approaches are limited to either data or process uncertainty and focus on settings where structured event logs are available. The integration of multiple types of uncertainty into process mining in a combined approach is still an open challenge. Additionally, current event log extraction techniques are unable to provide explicit uncertainty information, leaving a blind spot with respect to correlation uncertainty. Thus, handling uncertainty is particularly challenging for process mining on unstructured data, where event logs need to be extracted first.

A large body of research is available on the quantification of uncertainty in domains other than process mining (e.g., deep learning [10], mechanical engineering [11], or climate modeling [11]), which can be built on to provide uncertainty quantification techniques for process mining, especially in order to address data and correlation uncertainty. For instance, Zhang et al. [12] provide a general framework guiding the application of existing uncertainty quantification methods. Abdar et al. [10] discuss applications of uncertainty quantification in deep learning. Similarly, to address process uncertainty, the PhD project can rely on a large body of causal inference techniques [13] to quantify probabilistic dependencies in processes.

3. Research Design

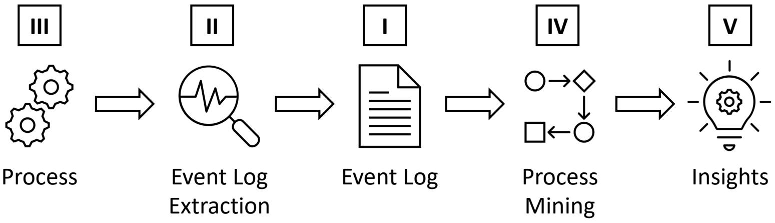

Figure 1: Overview of the general structure of process analytics pipelines for unstructured data. Roman numerals indicate the aspects addressed in each phase of the research project.

The goal of the PhD project is to address the various challenges outlined above. Particularly, a structured approach making uncertainty explicit in the process discovery phase and quantifying the degree of multiple types of uncertainty will be developed. The focus of this approach lies on the analysis of unstructured data since the level of uncertainty in the data is very high and all three types of uncertainty are significant. Fig. 1 shows the general structure of approaches to process mining on unstructured data (e.g., [14]), which are used to divide the PhD project into five phases. Phases (I)-(III) serve to enrich event logs with information on uncertainty and its impact on the trustworthiness of the insights, and phases (IV) and (V) are concerned with developing uncertainty-aware process mining methods and communicating the impact of uncertainty related to the gained insights. To support the generalizability of our approach, large and heterogeneous evaluation datasets with known ground-truth processes and characteristics (e.g., the amount of uncertainty) will be created by generating synthetic event data [15].

The first step addresses data uncertainty in event logs (I). For this, a taxonomy of uncertainty in event logs will be developed and used to generate synthetic event logs of certain processes with varying levels of data uncertainty. Then, uncertainty quantification [12, 10] methods will be adapted to assess the impact of data uncertainty related to the quality of insights gained from the process mining results. Next, the scope is widened to include correlation uncertainty by integrating uncertainty awareness into event log extraction (II). The applicability of the methods developed in (I) will be extended to unstructured data sources by developing extraction techniques that produce explicit uncertainty information. The quantification of data uncertainty can be extended to event log extraction techniques to enable the automatic identification of sources of data uncertainty. In the third step, process uncertainty is addressed (III). Process uncertainty can be isolated by generating high-quality event logs of uncertain, probabilistic processes. To provide a solution, (1) existing approaches for discovering (probabilistic) causalities in process mining will be reviewed, (2) causal inference techniques [13] will be adapted to address the gaps identified in this review, and (3) means to enrich event logs with quantitative information of the discovered causalities will be developed. The goal of the fourth step is to integrate the different views on uncertainty explicitly into common process mining tasks (IV). To do this, uncertainty-aware process mining techniques need to be developed, which can explicitly encode uncertainty to improve the quality compared to non-uncertainty-aware techniques and improve the quantification of uncertainty through measures. Finally, explainability will be addressed to improve the communication to non-experts between the technical design and the process mining results (V). The explainability of uncertainty provides a basis to gain additional insights for informed process management decisions. For this, different means to communicate process mining results (e.g., process models, reports) need to be offered for the uncertainty-enriched outputs developed in the previous phases.

4. Conclusion

In this PhD proposal, the challenges related to different types of uncertainty in process mining were described. The goal of the proposed PhD project is to address these challenges by suggesting a holistic framework to manage uncertainty in process mining. In order to achieve this, methods to enrich event logs with information on data, correlation, and process uncertainty will be developed, and this information will then be made explicit in the analysis results by integrating uncertainty into process mining methods and the communication of process mining results. The PhD project contributes to trustworthy process mining, laying the foundation for improved process mining-based decision-making.

Acknowledgments

I would like to thank my thesis advisor, Agnes Koschmider, for her invaluable guidance and support in writing this proposal, and Milda Aleknonytė-Resch for the helpful discussions about the research design and presentation of the proposal. This project has received funding from the German Federal Ministry for Economic Affairs and Climate Action under the Marispace-X project grant no. 68GX21002E.

References

[1] L. Reinkemeyer (Ed.), Process Mining in Action: Principles, Use Cases and Outlook, Springer, Cham, 2020. doi:10.1007/978-3-030-40172-6.

[2] A. Koschmider, N. Oppelt, M. Hundsdörfer, Confidence-driven communication of process mining on time series, Informatik Spektrum 45 (2022) 223–228. doi:10.1007/s00287-022-01470-3.

[3] W. van der Aalst, et al., Process Mining Manifesto, in: F. Daniel, K. Barkaoui, S. Dustdar (Eds.), BPM 2011 Workshops, volume 99 of *LNBIP*, Springer, Berlin, Heidelberg, 2012, pp. 169–194. doi:10.1007/978-3-642-28108-2_19.

[4] A. Koschmider, F. Mannhardt, T. Heuser, On the Contextualization of Event-Activity Mappings, in: F. Daniel, Q. Z. Sheng, H. Motahari (Eds.), BPM 2018 Workshops, volume 342 of *LNBIP*, Springer, Cham, 2019, pp. 445–457. doi:10.1007/978-3-030-11641-5_35.

[5] A. Koschmider, K. Kaczmarek, M. Krause, S. J. van Zelst, Demystifying Noise and Outliers in Event Logs: Review and Future Directions, in: A. Marrella, B. Weber (Eds.), BPM 2021 Workshops, volume 436 of *LNBIP*, Springer, Cham, 2022, pp. 123–135. doi:10.1007/978-3-030-94343-1_10.

[6] M. Pegoraro, Probabilistic and Non-deterministic Event Data in Process Mining: Embedding Uncertainty in Process Analysis Techniques, in: A. V. Looy, B. Weber, M. Rosemann (Eds.), CAiSE 2022 Doctoral Consortium, volume 3139 of *CEUR Workshop Proceedings*, CEUR-WS.org, Leuven, Belgium, 2022, pp. 37–46. URL: https://ceur-ws.org/Vol-3139/#paper05.

[7] M. S. Qafari, W. van der Aalst, Root Cause Analysis in Process Mining Using Structural Equation Models, in: A. Del Río Ortega, H. Leopold, F. M. Santoro (Eds.), BPM 2020 Workshops, LNBIP, Springer, Cham, 2020, pp. 155–167. doi:10.1007/978-3-030-66498-5_12.

[8] 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.), CAiSE 2022, LNCS, Springer, Cham, 2022, pp. 183–200. doi:10.1007/978-3-031-07472-1_11.

[9] A. Alman, F. M. Maggi, M. Montali, R. Peñaloza, Probabilistic declarative process mining, Information Systems 109 (2022) 102033. doi:10.1016/j.is.2022.102033.

[10] M. Abdar, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297. doi:10.1016/j.inffus.2021.05.008.

[11] J. O. Berger, L. A. Smith, On the Statistical Formalism of Uncertainty Quantification, Annual Review of Statistics and Its Application 6 (2019) 433–460. doi:10.1146/annurev-statistics-030718-105232.

[12] J. Zhang, J. Yin, R. Wang, Basic Framework and Main Methods of Uncertainty Quantification, Mathematical Problems in Engineering 2020 (2020) e6068203. doi:10.1155/2020/6068203.

[13] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, A. Zhang, A Survey on Causal Inference, ACM Transactions on Knowledge Discovery from Data 15 (2021) 74:1–74:46. doi:10.1145/3444944.

[14] A. Lepsien, J. Bosselmann, A. Melfsen, A. Koschmider, Process Mining on Video Data, in: J. Manner, D. Lübke, S. Haarmann, S. Kolb, N. Herzberg, O. Kopp (Eds.), ZEUS 2022, volume 3113 of *CEUR Workshop Proceedings*, CEUR-WS.org, Bamberg, Germany, 2022, pp. 56–62.

[15] Y. Zisgen, D. Janssen, A. Koschmider, Generating Synthetic Sensor Event Logs for Process Mining, in: J. De Weerdt, A. Polyvyanyy (Eds.), CAiSE Forum 2022, volume 452 of *LNBIP*, Springer, Cham, 2022, pp. 130–137. doi:10.1007/978-3-031-07481-3_15.