Process Mining on Uncertain Event Data (Extended Abstract)

Marco Pegoraro*
Chair of Process and Data Science (PADS)
Department of Computer Science, RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany
*Corresponding author. Email: pegoraro@pads.rwth-aachen.de

***Abstract*—With the widespread adoption of process mining in organizations, the field of process science is seeing an increase in the demand for ad-hoc analysis techniques of non-standard event data. An example of such data are *uncertain event data*: events characterized by a described and quantified attribute imprecision. This paper outlines a research project aimed at developing process mining techniques able to extract insights from uncertain data. We set the basis for this research topic, recapitulate the available literature, and define a future outlook.**

I. Introduction

Since its inception, process mining has ultimately proved its value in commercial applications. An ever-increasing number of success stories has led to a vast demand of the most diverse process analysis techniques, often customized to meet the needs of specific domains. Among these, novel techniques have been introduced to mine non-standard types of data.

This paper presents a research direction aimed to mine one such type of anomalous (i.e, uncommon) type of event data: *uncertain data*. Such data is associated with a degree of imprecision that affects event attributes, which is described and quantified through sets of possible attribute labels, intervals of possible values, or probability distributions.

The remainder of the paper is structured as follows. Section II illustrates with examples the structure of uncertain event data. Section III shows the research principles in regard of process mining on uncertain data, and reports recent results on the topic. Finally, Section IV outlines open challenges, outlook, and future perspectives of this line of research.

II. Uncertain Data

In order to more clearly visualize the structure of the attributes in uncertain events, let us consider the following process instance, which is a simplified version of actually occurring anomalies, e.g., in the processes of the healthcare domain. An elderly patient enrolls in a clinical trial for an experimental treatment against myeloproliferative neoplasms, a class of blood cancers. This enrollment includes a lab exam and a visit with a specialist; then, the treatment can begin. The lab exam, performed on the 8th of July, finds a low level of platelets in the blood of the patient, a condition known as thrombocytopenia (TP). During the visit on the 10th of July, the patient reports an episode of night sweats on the night of the 5th of July, prior to the lab exam. The medic notes this but also hypothesizes that it might not be a symptom, since it can be caused either by the condition or by external factors (such as very warm weather). The medic also reads the medical records of the patient and sees that, shortly prior to the lab exam, the patient was undergoing a heparin treatment (a blood-thinning medication) to prevent blood clots. The thrombocytopenia, detected by the lab exam, can then be either primary (caused by the blood cancer) or secondary (caused by other factors, such as a concomitant condition). Finally, the medic finds an enlargement of the spleen in the patient (splenomegaly). It is unclear when this condition has developed: it might have appeared at any moment prior to that point. These events are recorded in the trace ID192-1 (shown in Table I) within the hospital's information system.

Such scenario, with no known probability, is known as *strong uncertainty*. In this trace, the rightmost column refers to event indeterminacy: in this case, e_1 has been recorded, but it might not have occurred in reality, and is marked with a "?" symbol. Event e_2 has more then one possible activity labels, either *PrTP* or *SecTP*. Lastly, event e_3 has an uncertain timestamp, and might have happened at any point in time between the 4th and 10th of July.

Uncertain events may also have probability values associated with them, a scenario defined as *weak uncertainty* (trace ID192-2 in Table I). In the example described above, suppose the medic estimates that there is a high chance (90%) that the thrombocytopenia is primary (caused by the cancer). Furthermore, if the splenomegaly is suspected to have developed three days prior to the visit, which takes place on the 10th of July, the timestamp of event e_3 may be described through a Gaussian curve with $\mu = 7$. Lastly, the probability that the event e_1 has been recorded but did not occur in reality may be known (for example, it may be 25%).

TABLE I: Two uncertain traces related to an example of healthcare process. The timestamps column shows only the day of the month.

Case ID	Event ID	Timestamp	Activity	Indeterminacy
ID192-1	e_1	5	*NightSweats*	?
ID192-1	e_2	8	*PrTP*, *SecTP*	
ID192-1	e_3	4–10	*Splenomeg*	
ID192-2	e_4	5	*NightSweats*	? : 25%
ID192-2	e_5	8	*PrTP:* 90%, *SecTP:* 10%	
ID192-2	e_6	$\mathcal{N}(7, 1)$	*Splenomeg*	

Table II summarizes the types of uncertain data subject of

our research.

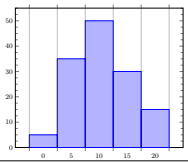
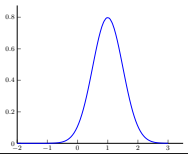
	Weak uncertainty (stochastic)	Strong uncertainty (non-deterministic)
Discrete data	Discrete probability distribution 	Set of possible values $\{x_1, x_2, x_3, \dots\} \subseteq X$
Continuous data	Probability density function 	Interval $\{x \in \mathbb{R} \mid a \leq x \leq b\}$

TABLE II: The four different types of uncertainty subject of this research project.

III. Research Approach

We will now illustrate the guiding principles of our research plans, through a series of assertions.

Assertion 1 (Uncertainty is not noise)**.** *Uncertain data contain information and value. We do not aim to analyze the data beyond the uncertainty, but the data within the uncertainty.*

Assertion 2 (Uncertainty should not be filtered or repaired)**.** *To extract information from uncertainty itself, existing approaches to filter or repair data are not applicable: information from uncertainty must be accounted for, and not altered.*

Assertion 3 (Uncertainty is behavior)**.** *The many possible values for event attributes entail numerous possible scenarios for the control-flow perspective of an uncertain trace—which can be represented as behavior. To fully analyze uncertain process instances, it is necessary to account for such behavior.*

The fundamental technique that enables the analysis of uncertain traces is their representation as dynamic objects that incorporate the intrinsic behavior of uncertain traces, such as graphs or Petri nets (*behavior graphs* or *behavior nets* [1], respectively). This leads to the schematic visible in Figure 1.

A number of mining techniques for uncertain event data are now present in literature. A taxonomy of uncertain event data is available [1], as well as a method to reliably compute the probability associated with each real-life scenario in an uncertain trace [2]. There exist approaches for conformance checking [3] and process discovery [4] over strongly uncertain event data. The key phase in uncertain data analysis of building graph representation has been optimized through efficient algorithms [5], [6]. Such techniques are available in the PROVED toolset [7], which employs an ad-hoc extension of the XES standard to represent uncertain data [8]. A real-life source of uncertain data, convolutional neural network sensing in video feeds of processes, has been described, as well as an additional taxonomy also involving process models [9].

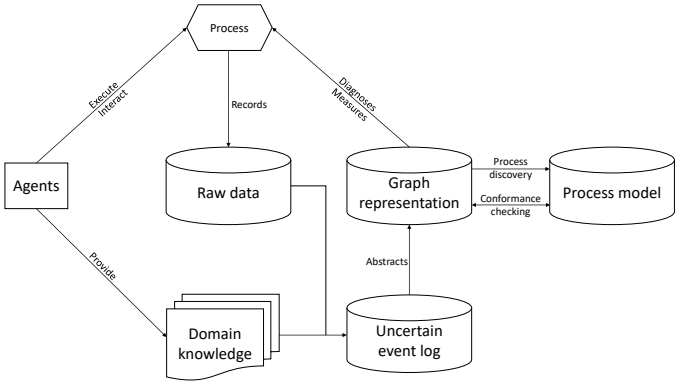


Fig. 1: The overall schema for process mining over uncertainty.

IV. Open Challenges and Conclusion

The field of process mining over uncertain data is still in its infancy. While some techniques to perform discovery and conformance checking over uncertainty do exist, the weakly uncertain case is still unexplored. The principle of the four quality metrics of logs and processes (fitness, precision, simplicity, precision), a cornerstone of process mining, needs to be (re)developed in the context of uncertain data.

Through analyzing uncertain event data without discarding any of the attributes in an uncertain event log, this research direction unlocks the extraction of process information formerly inaccessible. Insights from process mining analyses can, as a consequence, maintain quantified guarantees of reliability and accuracy even in presence of data affected by uncertainty.

Acknowledgments

I am very grateful to Prof. Wil van der Aalst, who advises my doctoral studies, and to Merih Seran Uysal, who supervises me in researching this topic. I thank the Alexander von Humboldt (AvH) Stiftung for supporting my research interactions.

References

[1] M. Pegoraro and W. M. P. van der Aalst, "Mining uncertain event data in process mining," in *International Conference on Process Mining (ICPM)*. IEEE, 2019, pp. 89–96.
[2] M. Pegoraro, B. Bakullari, M. S. Uysal, and W. M. P. van der Aalst, "Probability estimation of uncertain process trace realizations," in *International Workshop on Event Data and Behavioral Analytics (EdbA)*. Springer, 2021.
[3] M. Pegoraro, M. S. Uysal, and W. M. P. van der Aalst, "Conformance checking over uncertain event data," *Information Systems*, 2021.
[4] ——, "Discovering process models from uncertain event data," in *International Conference on Business Process Management (BPM)*. Springer, 2019, pp. 238–249.
[5] ——, "Efficient construction of behavior graphs for uncertain event data," in *International Conference on Business Information Systems*. Springer, 2020, pp. 76–88.
[6] ——, "Efficient time and space representation of uncertain event data," *Algorithms*, vol. 13, no. 11, p. 285, 2020.
[7] ——, "PROVED: A tool for graph representation and analysis of uncertain event data," in *International Conference on Applications and Theory of Petri Nets and Concurrency*. Springer, 2021, pp. 476–486.
[8] ——, "An XES extension for uncertain event data," in *International Conference on Business Process Management (BPM)*. Springer, 2021.
[9] I. Cohen and A. Gal, "Uncertain process data with probabilistic knowledge: Problem characterization and challenges," *CoRR/abs*, 2021. [Online]. Available: https://arxiv.org/abs/2106.03324