Probability Estimation of Uncertain Process Trace Realizations

Marco Pegoraro ✉[1], Bianka Bakullari [1], Merih Seran Uysal [1], and Wil M.P. van der Aalst [1]

[1] *Chair of Process and Data Science (PADS), Department of Computer Science, RWTH Aachen University, Aachen, Germany*
{pegoraro, bianka.bakullari, uysal, vwdaalst}@pads.rwth-aachen.de

Abstract

Process mining is a scientific discipline that analyzes event data, often collected in databases called event logs. Recently, *uncertain event logs* have become of interest, which contain non-deterministic and stochastic event attributes that may represent many possible real-life scenarios. In this paper, we present a method to reliably estimate the probability of each of such scenarios, allowing their analysis. Experiments show that the probabilities calculated with our method closely match the true chances of occurrence of specific outcomes, enabling more trustworthy analyses on uncertain data.

Keywords: Process Mining · Uncertain Data · Partial Order.

Colophon

This document is an Author Accepted Manuscript (AAM) corresponding to the following scholarly paper:

Pegoraro, Marco et al. "Probability Estimation of Uncertain Process Trace Realizations". In: *International Workshop on Event Data and Behavioral Analytics (EdbA)*. Springer, 2021

Please, cite this document as shown above.

Publication chronology:

- 2021-06-15: abstract submitted to the International Conference on Process Mining (ICPM) 2021, main track
- 2021-07-01: full text submitted to the International Conference on Process Mining (ICPM) 2021, main track
- 2021-08-16: notification of rejection
- 2021-08-17: abstract submitted to the International Workshop on Event Data and Behavioral Analytics (EdbA) 2021
- 2021-08-20: full text submitted to the International Workshop on Event Data and Behavioral Analytics (EdbA) 2021
- 2021-09-16: notification of acceptance
- 2021-09-22: camera-ready version submitted
- 2021-11-01: presented
- 2022-03-24: proceedings published

The published version referred above is ©Springer.

Correspondence to:
Marco Pegoraro, Chair of Process and Data Science (PADS), Department of Computer Science,
RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany
Website: http://mpegoraro.net/ · Email: pegoraro@pads.rwth-aachen.de · ORCID: 0000-0002-8997-7517

Content: 16 pages, 7 figures, 4 tables, 11 references. Typeset with pdfLaTeX, Biber, and BibLaTeX.
Please do not print this document unless strictly necessary.

1 Introduction

Process mining is a discipline that focuses on extracting insights about processes in a data-driven manner. For instance, on the basis of the recorded information on historical process executions, process mining allows to automatically extract a model of the behavior of process instances, or to measure the compliance of the process data with a prescribed normative model of the process. In process mining, the central focus is on the *event log*, a collection of data that tracks past process instances. Every activity performed in a process is recorded in the event log, together with information such as the corresponding process case and the timestamp of the activity, in a sequence of events called a *trace*.

Recently, research on novel forms of event data have garnered the attention of the scientific community. Among these there are *uncertain event logs*, which contain data affected by imprecision [8]. This data contains meta-information describing the nature and entity of the uncertainty. Such meta-information can be obtained from the inherent precision with which the data has been recorded (e.g., timestamps only indicating the date have a possible "true value" range of 24 hours), from the precision of the tools involved in supporting the process (e.g., the absolute error of sensors), or from the domain knowledge provided by a process expert. An uncertain trace corresponds to multiple possible real-life scenarios, each of which might have very diverse implications on features of cases such as compliance to a model. It is then important to be able to assess the risk of occurrence of specific outcomes of uncertain traces, which enables to estimate the impact of such traces on indicators such as cost and conformance.

In this paper, we present a method to obtain a complete probability distribution over the possible instantiations of uncertain attributes in a trace. As a possible example of application, we frame our results in the context of conformance checking, and show the impact of assessing probability estimates for uncertain traces on insights about the compliance of an uncertain trace to a process model. We validate our method with experiments based on a Monte Carlo simulation, which shows that the probability estimates are reliable and reflect the true chances of occurrence of a specific outcome.

The remainder of the paper is structured as follows. Section 2 examines relevant related work. Section 3 illustrates a motivating running example for our technique. Section 4 presents preliminary definitions of different types of uncertainty in process mining. Section 5 illustrates a method for computing probabilities of realizations for uncertain process traces. Section 6 validates our method through experimental results. Finally, Section 7 concludes the paper.

2 Related Work

The analysis of uncertain data in process mining is a very recent research direction. The specific formulation and definition of uncertain data utilized in this paper has been introduced in 2019 [8], in the context of an analysis approach consisting in computing bounds for the conformance score of uncertain traces through alignments [5]. Subsequently, that work has been extended with an inductive mining approach for process discovery over uncertainty [10] and a taxonomy of different types of uncertain data, with their characteristics [9].

Uncertain data, as formulated in our present and previous work, is closely related to a considerably more studied data anomaly in process mining: partially ordered event data. In fact, uncertain data as described here is a generalization of partially ordered traces. Lu et al. [7] proposed a conformance checking approach based on alignments to measure conformance of partially ordered traces. More recently, Van der Aa et al. [1] illustrated a method for inferring a linear extension, i.e., a compliant total order, of events in partially ordered traces, based on examples of correct orderings extracted from other traces in the log. Busany et al. [4] estimated probabilities for partially ordered events in IoT event streams.

An associated topic, which draws from disciplines such as pattern and sequence mining and is antithetical to the analysis of partially ordered data, is the inference of partial orders from fully sequential data as a way to model its behavior. This goes under the name of *episode mining*, which can be performed with many techniques both on batched data and with online streams of events [11, 6, 2].

In this paper, we present a method to estimate the likelihood of any scenario in an uncertain setting, which covers partially ordered traces as well as other types of uncertainty illustrated in the taxonomy [9]. Furthermore, we will cover both the nondeterministic case (*strong uncertainty*) and the probabilistic case (*weak uncertainty*).

3 Running Example

In this section, we will provide a running example of uncertain process instance related to a sample process. We will then apply our probability estimation method to this uncertain trace, to illustrate its operation. The example we analyze here is a simplified generalization of a remote credit card fraud investigation process. This process is visualized by the Petri net in Figure 1.

Firstly, the credit card owner alerts the credit card company of a possibly fraudulent transaction. The customer may either notify the company by calling their hotline (*alert hotline*) or arrange an urgent meeting with personnel of the bank that issued the credit card (*alert bank*). In both scenarios, his credit is frozen (*freeze credit*) to prevent further

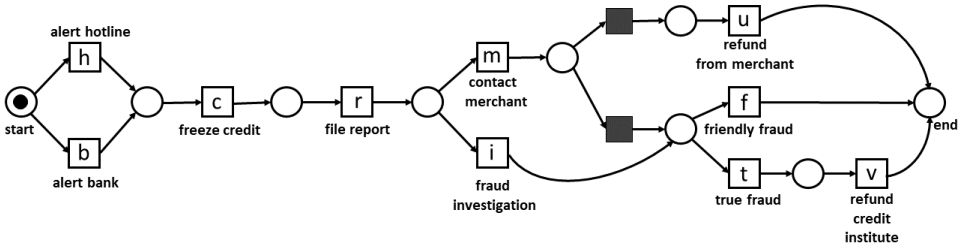

Figure 1: A Petri net model of the credit card fraud investigation process. This net allows for 10 possible traces.

fraud. All information provided by the customer about the transaction is summarized when filing the formal report (*file report*). As a next step, the credit card company tries to contact the merchant that charged the credit card. If this happens (*contact merchant*), the credit card company clarifies whether there has been just a mistake (e.g., merchant charging not delivering a product, or a billing mistake) on the merchant's side. In such cases, the customer gets a *refund from merchant* and the case is closed. Another outcome might be the discovery of a *friendly fraud*, which is when a cardholder makes a purchase and then disputes it as fraud even though it was not. If contacting the merchant is impossible, a *fraud investigation* is initiated. In this case, fraud investigators will usually start with the transaction data and look for timestamps, geolocation, IP addresses, and other elements that can be used to prove whether or not the cardholder was involved in the transaction. The outcome might be either friendly fraud or *true fraud*. True fraud can also happen when both the merchant and the cardholder are affected by the fraud. In this case, the cardholder receives a refund from the credit institute (activity *refund credit institute*) and the case is closed.

Note that for simplicity, we have used single letters to represent the activity labels in the Petri net transitions. Some possible traces in this process are for example: $\langle h, c, r, m, u \rangle$, $\langle b, c, r, m, f \rangle$, $\langle h, c, r, i, f \rangle$ and $\langle b, c, r, i, t, v \rangle$.

Suppose that the credit card company wants to perform conformance checking to identify deviant process instances. However, some traces in the information system of the company are affected by uncertainty, such as the one in Table 1.

Suppose that in the first half of October 2020, the company was implementing a new system for automatic event data generation. During this time, the event data regarding the credit card fraud investigation process often had to be inserted manually by the employees. Such manual recordings were subject to inaccuracies, leading to imprecise or missing data affecting the cases during this period. The process instance from Table 1 is one of the affected instances. Here, events e_2, e_3, e_5, e_6 are uncertain. The timestamp of event e_2 is not precise enough, so the possible timestamp lies between 06-10-2020 00:00

Table 1: Example of an uncertain case from the credit card fraud investigation process.

Case ID	Event ID	Activity	Timestamp	Ind.
5167	e_1	h (alert hotline)	05-10-2020 23:00	
5167	e_2	c (freeze credit)	06-10-2020	
5167	e_3	r (file report)	U(05-10-2020 20:00, 06-10-2020 10:00)	
5167	e_4	i (fraud investigation)	09-10-2020 10:00	
5167	e_5	$\{f : 0.3$ (friendly fraud), $t : 0.7$ (true fraud)$\}$	14-10-2020 09:00	
5167	e_6	v (refund credit institute)	15-10-2020 10:00	?

and 06-10-2020 23:59. Event e_3 has happened some time between 20:00 on October 5th and 10:00 on October 6th. Event e_5 has two possible activity labels: f with probability 0.3 and t with probability 0.7. Refunding the customer (event e_6) has been recorded in the system, but the customer has not received the money yet, which is why the event is indeterminate: this is indicated with a question mark (?) in the rightmost column, and indicates an event that has been recorded, but for which is unclear if it actually occurred in reality.

The credit card company is interested in understanding if and how the data in this uncertain trace conforms with the normative process model, and the entity of the actual compliance risk; they are specifically interested in knowing whether a severely non-compliant scenario is highly likely. In the remainder of the paper, we will describe a method able to estimate the probability of all possible outcome scenarios.

4 Preliminaries

Let us now present some preliminary definitions regarding uncertain event data.

Definition 1 (Uncertain attributes). *Let* $\mathbb{U}$ *be the* universe of attribute domains*, and the set* $\mathcal{D} \in \mathbb{U}$ *be an* attribute domain*. Any* $\mathcal{D} \in \mathbb{U}$ *is a discrete set or a totally ordered set. A* strongly uncertain attribute *of domain* $\mathcal{D}$ *is a subset* $d_S \subseteq \mathcal{D}$ *if* $\mathcal{D}$ *is a discrete set, and it is a closed interval* $d_S = [d_{min}, d_{max}]$ *with* $d_{min} \in \mathcal{D}$ *and* $d_{max} \in \mathcal{D}$ *otherwise. We denote with* $S_\mathcal{D}$ *the set of all such strongly uncertain attributes of domain* $\mathcal{D}$*. A* weakly uncertain attribute $f_\mathcal{D}$ *of domain* $\mathcal{D}$ *is a function* $f_\mathcal{D} \colon \mathcal{D} \not\to [0,1]$ *such that* $0 < \sum_{x \in \mathcal{D}} f_\mathcal{D}(x) \leq 1$ *if* $\mathcal{D}$ *is finite,* $0 < \int_{-\infty}^{\infty} f_\mathcal{D}(x)\,dx \leq 1$ *otherwise. We denote with* $W_\mathcal{D}$ *the set of all such weakly uncertain attributes of domain* $\mathcal{D}$*. We collectively denote with* $U_\mathcal{D} = S_\mathcal{D} \cup W_\mathcal{D}$ *the set of* uncertain attributes *of domain* $\mathcal{D}$*.*

It is easy to see how a "certain" attribute x, with a value not affected by any uncer-

tainty, can be represented through the definitions in use here: if its domain is discrete, it can be represented with the singleton $\{x\}$; otherwise, it can be represented with the degenerate interval $[x, x]$.

Definition 2 (Uncertain events). *Let* $\mathbb{U}_I$ *be the* universe of event identifiers. *Let* $\mathbb{U}_C$ *be the* universe of case identifiers. *Let* $A \in \mathbb{U}$ *be the discrete domain of all the* activity identifiers. *Let* $T \in \mathbb{U}$ *be the totally ordered domain of all the* timestamp identifiers. *Let* $O = \{?\} \in \mathbb{U}$*, where the "?" symbol is a placeholder denoting* event indeterminacy. *The* universe of uncertain events *is denoted with* $E = \mathbb{U}_I \times \mathbb{U}_C \times U_A \times U_T \times U_O$.

The activity label, timestamp and indeterminacy attribute values of an uncertain event are drawn from U_A, U_T and U_O; in accordance with Definition 1, each of these attributes can be strongly uncertain (set of possible values or interval) or weakly uncertain (probability distribution). The indeterminacy domain is defined on a single element "?": thus, strongly uncertain indeterminacy may be $\{?\}$ (indeterminate event) or $\varnothing$ (no indeterminacy). In weakly uncertain indeterminacy, the "?" element is associated to a probability value.

Definition 3 (Projection functions). *For an uncertain event* $e = (i, c, a, t, o) \in E$*, we define the following projection functions:* $\pi_a(e) = a$, $\pi_t(e) = t$, $\pi_o(e) = o$. *We define* $\pi_a^{set}(e) = a$ *if* a *is strongly uncertain, and* $\pi_a^{set}(e) = \{x \in U_A \mid f_A(x) > 0\}$ *with* $a = f_A$ *otherwise. If the timestamp* $t = [t_{min}, t_{max}]$ *is strongly uncertain, we define* $\pi_{t_{min}}(e) = t_{min}$ *and* $\pi_{t_{max}}(e) = t_{max}$. *If the timestamp* $t = f_T$ *is weakly uncertain, we define* $\pi_{t_{min}}(e) = \operatorname{argmin}_x(f_T(x) > 0)$ *and* $\pi_{t_{max}}(e) = \operatorname{argmax}_x(f_T(x) > 0)$.

Definition 4 (Uncertain traces and logs). $\tau \subset E$ *is an* uncertain trace *if all the event identifiers in* τ *are unique and all events in* τ *share the same case identifier* $c \in \mathbb{U}_C$. T *denotes the universe of uncertain traces.* $L \subset T$ *is an* uncertain log *if all the event identifiers in* L *are unique.*

Definition 5 (Realizations of uncertain traces). *Let* $e, e' \in E$ *be two uncertain events.* $\prec_E$ *is a strict partial order defined on the universe of strongly uncertain events* E *as* $e \prec_E e' \Leftrightarrow \pi_{t_{max}}(e) < \pi_{t_{min}}(e')$. *Let* $\tau \in T$ *be an uncertain trace. The sequence* $\varrho = \langle e_1, e_2, \ldots, e_n \rangle \in E^*$*, with* $n \leq |\tau|$*, is an* order-realization *of* τ *if there exists a total function* $f : \{1, 2, \ldots, n\} \to \tau$ *such that:*

- *for all* $1 \leq i < j \leq n$ *we have that* $\varrho[j] \not\prec_E \varrho[i]$,
- *for all* $e \in \tau$ *with* $\pi_o(e) = \varnothing$ *there exists* $1 \leq i \leq n$ *such that* $f(i) = e$.

We denote with $\mathcal{R}_Q(\tau)$ *the set of all such order-realizations of the trace* τ.

Given an order-realization $\varrho = \langle e_1, e_2, \ldots, e_n \rangle \in \mathcal{R}_Q(\tau)$*, the sequence* $\sigma \in U_A^*$ *is a* realization *of* ϱ *if* $\sigma \in \{\langle a_1, a_2, \ldots, a_n \rangle \mid \forall_{1 \leq i \leq n}\, a_i \in \pi_a^{set}(i)\}$. *We denote with*

$\mathcal{R}_{\mathcal{A}}(\varrho) \subseteq U_A{}^*$ the set of all such realizations of the order-realization ϱ. We denote with $\mathcal{R}_{\backslash}(\tau) \subseteq U_A{}^*$ the union of the realizations obtainable from all the order-realizations of τ: $\mathcal{R}_{\backslash}(\tau) = \bigcup_{\varrho \in \mathcal{R}_{\mathcal{Q}}(\tau)} \mathcal{R}_{\mathcal{A}}(\varrho)$. We will say that an order-realization $\varrho \in \mathcal{R}_{\mathcal{Q}}(\tau)$ enables *a sequence $\sigma \in U_A{}^*$ if $\sigma \in \mathcal{R}_{\mathcal{A}}(\varrho)$.*

Detailing an algorithm to generate all realizations of an uncertain trace is beyond the scope of this paper. The literature illustrates a conformance checking method over uncertain data which employs a *behavior net*, a Petri net able to replay all and only the realizations of an uncertain trace [8]. Exhaustively exploring all complete firing sequences of a behavior net, e.g., through its reachability graph, provides all realizations of the corresponding uncertain trace.

Given the above formalization, we can now define more clearly the research question that we are investigating in this paper. Given an uncertain trace $\tau \in T$ and one of its realizations $\sigma \in \mathcal{R}_{\backslash}(\tau)$, our goal is to obtain a procedure to reliably compute $P(\sigma \mid \tau)$ = "*probability of σ given that we observe τ*". In other words, provided that σ corresponds to a scenario (i.e., a realization) for the uncertain trace τ, we are interested in calculating the probability that σ is the actual scenario occurred in reality, which caused the recording of the uncertain trace τ in the event log. In the next section, we will illustrate how to calculate such probabilities of uncertain traces realizations.

5 Method

Before we show how we can obtain probability estimates for all realizations of an uncertain trace, it is important to state an assumption: the information on uncertainty related to a particular attribute in some event is independent of the possible values of the same attribute present in other events, and it is independent of the uncertainty information on other attributes of the same event. Note that in the examples of uncertainty sources given in Section 1 (data coarseness and sensor errors), this independence assumption often holds.

Additionally, we need to consider the fact that strongly uncertain attributes do not come with known probability values: their description only specifies the values that attributes might acquire, but not the likelihood of each possible value. As a consequence, estimating probability for specific realizations in a strongly uncertain environment is only possible with a-priori assumptions on how probability distributes among the attribute value. At times, it might be possible to assume the distribution in an informed way—for instance, on the basis of features of the information system hosting the data, of the sensors recording events and attributes, or other tools involved in the management of the process.

In case no indication is present, a reasonable assumption—which we will hold for the remainder of the paper—is that any possible value of a strongly uncertain attribute

is equally likely. Formally, with $e = (i, c, a, t, o) \in E$ let $\tau_s \colon E \to E$ be a function such that $\tau_s(e) = (i, c, a', t', o')$, where $a' = \{(x, \frac{1}{|\pi_a^{set}(e)|}) \mid x \in \pi_a^{set}(e)\}$ if $a \in S_A$ and $a' = a$ otherwise; $t' = U(\pi_{t_{min}}(e), \pi_{t_{max}}(e))$ if $t \in S_T$ and $t' = t$ otherwise; $o' = 0.5$ if $o = \{?\}$ and $o' = o$ otherwise.

First, observe that the probability $P(\sigma \mid \tau)$ that an activity sequence $\sigma \in {U_A}^*$ is indeed a realization of the trace $\tau \in T$, and thus $\sigma \in \mathcal{R}_{\curlywedge}(\tau)$, increases with the number of order-realizations enabling it. Furthermore, for each such order-realizations, one can construct a probability function $P_O(\varrho \mid \tau)$ reflecting the likelihood of the sequence ϱ itself given the trace τ, and a probability function $P_A(\sigma \mid \varrho)$ reflecting the likelihood that the realization corresponding to ϱ is indeed σ. The value of $P_O(\varrho \mid \tau)$ is affected by the uncertainty information in timestamps and indeterminate events, while the value of $P_A(\sigma \mid \varrho)$ is aggregated from the uncertainty information in the activity labels.

Given a realization σ of an uncertain process instance and the set of its enablers, its probability is computed as following:

$$P(\sigma \mid \tau) = \sum_{\varrho \in E^*} P_O(\varrho \mid \tau) \cdot P_A(\sigma \mid \varrho)$$

Note that, if ϱ does not enable σ, $P_A(\sigma \mid \varrho) = 0$. For any uncertain trace $\tau \in T$, it holds that $\sum_{\sigma \in \mathcal{R}_{\curlywedge}(\tau)} P(\sigma \mid \tau) = 1$, since both $P_O(\cdot)$ and $P_A(\cdot)$ are each constructed to be (independent) probability distributions.

We will now compute $P_A(\sigma \mid \varrho)$ using the information on the activity labels uncertainty. Let us write f_A^e as a shorthand for $\pi_a(e)$. If there is uncertainty in activities, then for each event $e \in \varrho$ and activity label $a \in \pi_a^{set}(e)$, the probability that e executes a is given by $f_A^e(a)$. Thus, for every $\varrho = \langle e_1, ..., e_n \rangle \in \mathcal{R}_Q(\tau)$ and $\sigma = \langle a_1, ..., a_n \rangle \in \mathcal{R}_Q(\tau)$, the value P_A can be aggregated from these distributions in the following way:

$$P_A(\sigma \mid \varrho) = \prod_{i=1}^{n} f_A^i(a_i)$$

Through the value of P_A, we can assess the likelihood that any given order-realization executes a particular realization. The next step is to estimate the probability of each order-realization ϱ from the set $\mathcal{R}_Q(\tau)$. The probability of observing ϱ needs to be aggregated from the probability that the corresponding set of events appears in the given particular order, which is determined by the timestamp intervals and, if applicable, the distributions over them; and the probability that the order-realization contains the corresponding specific set of events, which is determined by the uncertainty information on the indeterminacy. Multiplying the two values obtained above to yield a probability

estimate for the order-realization reflects our independence assumption. Let us firstly focus on uncertainty on timestamps, which causes the events to be partially ordered.

We will write $f^e_T(t)$ as a shorthand for $\pi_t(e)(t)$. For every event e, the value of $f^e_T(t)$ yields the probability that event e happened on timestamp t. This value is always 0 for all $t < \pi_{t_{min}}(e)$ and $t > \pi_{t_{max}}(e)$ (see $\pi_{t_{min}}$ and $\pi_{t_{max}}$ in Definition 3). Given the continuous domain of timestamps, $P_O(\cdot)$ is assessed by using integrals. For a trace $\tau \in T$ and an order-realization $\varrho = \langle e_1, ..., e_n \rangle \in \mathcal{R}_Q(\tau)$, let $a_i = \pi_{t_{min}}(i)$ and $b_i = \pi_{t_{max}}(i)$ for all $1 \leq i \leq n$. Then, we define:

$$
\begin{aligned}
I(\varrho) = &\int_{a_1}^{min\{b_1,...,b_n\}} f^{e_1}_T(x_1) \int_{max\{a_2,x_1\}}^{min\{b_2,...,b_n\}} f^{e_2}_T(x_2) \cdots \\
&\int_{max\{a_i,x_{i-1}\}}^{min\{b_i,...,b_n\}} f^{i}_T(x_i) \cdots \int_{max\{a_n,x_{n-1}\}}^{b_n} f^{e_n}_T(x_n)\, d_{x_n} \dots d_{x_1} \\
= &\int_{a_1}^{min\{b_1,...,b_n\}} \int_{max\{a_2,x_1\}}^{min\{b_2,...,b_n\}} \cdots \int_{max\{a_i,x_{i-1}\}}^{min\{b_i,...,b_n\}} \cdots \int_{max\{a_n,x_{n-1}\}}^{b_n} \prod_{i=1}^{n} f^{i}_T(x_i)\, d_{x_n} \dots d_{x_1}
\end{aligned}
$$

This chain of integrals allows us to compute the probability of a specific order among all the events in an uncertain trace. Now, to compute the probability of each realization from $\mathcal{R}_\varrho$, accounting for indeterminate events, we combine both the probability of the events having appeared in a particular order and the probability that the sequence contains exactly those events. For simplicity, we will use a function that acquires the value 1 if an event is not indeterminate. Let us define $f^e_O \colon O \to [0, 1]$ such that $f^e_O(?) = \pi_o(e)(?)$ if $\pi_o(e) \neq \varnothing$ and $f^e_O(?) = 1$ otherwise. More precisely, given $\tau \in T$ and $\varrho \in \mathcal{R}_Q(\tau)$, we compute:

$$
P_O(\varrho \mid \tau) = I(\varrho) \cdot \prod_{\substack{e \in \tau \\ e \in \varrho}} (1 - f^e_O(?)) \cdot \prod_{\substack{e \in \tau \\ e \notin \varrho}} f^e_O(?)
$$

We now have at our disposal all the necessary tools to compute a probability distribution over the trace realizations of any uncertain process instance in any possible uncertainty scenario. Let us then apply this method to compute the probabilities of all realizations of the trace τ in Table 1, and to analyze its conformance to the process in Figure 1.

Each order-realization of τ enables two realizations, because event e_5 has two possible activity labels. Since for events $e \in \tau \setminus \{e_5\}$, we have f^e_A equal to 1 for their corresponding unique activity label, the probability that an order-realization $\varrho \in \mathcal{R}_Q(\tau)$ has some realization $\sigma \in \mathcal{R}_A(\varrho)$ only depends on whether the trace σ contains activity f or t. Thus, for traces $\sigma^{1'}$, $\sigma^{2'}$, $\sigma^{3'}$, $\sigma^{4'}$, $\sigma^{5'}$, $\sigma^{6'}$ and their unique enabling sequences,

Table 2: The possible order-realizations of the process instance from Table 1 and their probabilities.

Order-realization ϱ	$I(\varrho)$	$P_O(\varrho)$
ϱ^1:$\langle e_1, e_2, e_3, e_4, e_5, e_6\rangle$	0.140	0.074
ϱ^2:$\langle e_1, e_3, e_2, e_4, e_5, e_6\rangle$	0.780	0.390
ϱ^3:$\langle e_3, e_1, e_2, e_4, e_5, e_6\rangle$	0.072	0.036
ϱ^4:$\langle e_1, e_2, e_3, e_4, e_5\rangle$	0.149	0.074
ϱ^5:$\langle e_1, e_3, e_2, e_4, e_5\rangle$	0.780	0.390
ϱ^6:$\langle e_3, e_1, e_2, e_4, e_5\rangle$	0.072	0.036

Table 3: The set of possible realizations of the example from Table 1, their enablers, their probabilities, and their conformance scores. The conformance score is equal to the cost of the optimal alignment between the trace and the Petri net in Figure 1.

Realization σ	ϱ	$P(\sigma \mid \tau)$	*conf*
$\sigma^{1'}$:$\langle h, c, r, i, f, v\rangle$	ϱ^1	$P_O(\varrho^1)\cdot P_A(\sigma^{1'}\mid\varrho^1) = 0.022$	1
$\sigma^{1''}$:$\langle h, c, r, i, t, v\rangle$	ϱ^1	$P_O(\varrho^1)\cdot P_A(\sigma^{1''}\mid\varrho^1) = 0.052$	0
$\sigma^{2'}$:$\langle h, r, c, i, f, v\rangle$	ϱ^2	$P_O(\varrho^2)\cdot P_A(\sigma^{2'}\mid\varrho^2) = 0.117$	3
$\sigma^{2''}$:$\langle h, r, c, i, t, v\rangle$	ϱ^2	$P_O(\varrho^2)\cdot P_A(\sigma^{2''}\mid\varrho^2) = 0.273$	2
$\sigma^{3'}$:$\langle r, h, c, i, f, v\rangle$	ϱ^3	$P_O(\varrho^3)\cdot P_A(\sigma^{3'}\mid\varrho^3) = 0.011$	3
$\sigma^{3''}$:$\langle r, h, c, i, t, v\rangle$	ϱ^3	$P_O(\varrho^3)\cdot P_A(\sigma^{3''}\mid\varrho^3) = 0.025$	2
$\sigma^{4'}$:$\langle h, c, r, i, f\rangle$	ϱ^4	$P_O(\varrho^4)\cdot P_A(\sigma^{4'}\mid\varrho^4) = 0.022$	0
$\sigma^{4''}$:$\langle h, c, r, i, t\rangle$	ϱ^4	$P_O(\varrho^4)\cdot P_A(\sigma^{4''}\mid\varrho^4) = 0.052$	1
$\sigma^{5'}$:$\langle h, r, c, i, f\rangle$	ϱ^5	$P_O(\varrho^5)\cdot P_A(\sigma^{5'}\mid\varrho^5) = 0.117$	2
$\sigma^{5''}$:$\langle h, r, c, i, t\rangle$	ϱ^5	$P_O(\varrho^5)\cdot P_A(\sigma^{5''}\mid\varrho^5) = 0.273$	3
$\sigma^{6'}$:$\langle r, h, c, i, f\rangle$	ϱ^6	$P_O(\varrho^6)\cdot P_A(\sigma^{6'}\mid\varrho^6) = 0.011$	2
$\sigma^{6''}$:$\langle r, h, c, i, t\rangle$	ϱ^6	$P_O(\varrho^6)\cdot P_A(\sigma^{6''}\mid\varrho^6) = 0.025$	3

we always have $P_A(\sigma^{i'} \mid s^i_\varrho) = f^{e_5}_A(f) = 0.3$, where $i \in \{1, \ldots, 6\}$. Similarly, for traces $\sigma^{1''}, \sigma^{2''}, \sigma^{3''}, \sigma^{4''}, \sigma^{5''}, \sigma^{6''}$ and their unique enabling sequences, we always have $P_A(\sigma^{i''} \mid \varrho^i) = f^{e_5}_A(t) = 0.7$, where $i \in \{1, \ldots, 6\}$. Next, we calculate the $P_O(\cdot)$ values for the 6 possible order-realizations in $\mathcal{R}_Q(\tau)$, which are displayed in Table 2.

One can notice that the I values only depend on the ordering of the first three events, which are also the only ones with overlapping timestamps. Since the indeterminate event e_6 does not overlap with any other event, pairs of sequences where the first three events have the same order also have the same probability. This reflects our assumption that the occurrence and non-occurrence of e_6 are both equally possible. Table 3 displays the calculations for the computation of the $P(\sigma \mid \tau)$ values for all realizations. Now we can compute the expected conformance score for the uncertain process instance $\tau = \{e_1, \ldots, e_6\}$. We can do so by computing alignments [5] for each realization of τ:

$$\begin{aligned}\overline{conf}(\tau) &= \sum_{\sigma \in \mathcal{R}_(\tau)} P(\sigma \mid \tau) \cdot conf(\sigma, \mathcal{M}) \\ &= 0.022 \cdot 1 + 0.05 \cdot 0 + 0.117 \cdot 3 + 0.273 \cdot 2 + 0.011 \cdot 3 + 0.025 \cdot 2 \\ &+ 0.022 \cdot 0 + 0.052 \cdot 1 + 0.117 \cdot 2 + 0.273 \cdot 3 + 0.011 \cdot 2 + 0.025 \cdot 3 \\ &= 2.204.\end{aligned}$$

Given the information on uncertainty available for the trace, this conformance score is a more realistic estimate of the real conformance score compared to taking the best, worst or average scores with values 0, 3 and 1.75 respectively.

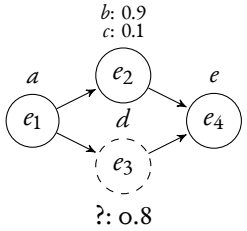

Figure 2: The behavior graph of the uncertain trace considered as example for validation.

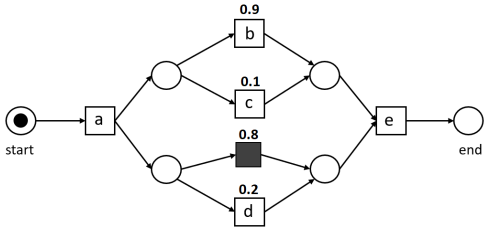

Figure 3: The behavior net obtained from the behavior graph in Figure 2.

Table 4: The set of realizations of the trace from Figure 2, their enablers, and their probabilities.

Realization σ	ϱ	$P(\sigma\|\tau)$
$\sigma^1{:}\langle a, b, e\rangle$	$\varrho^1{:}\langle e_1, e_2, e_4\rangle$	$P_O(\varrho^1)\cdot P_A(\sigma^1\|\varrho^1) = 0.8\cdot 0.9 = 0.72$
$\sigma^2{:}\langle a, b, d, e\rangle$	$\varrho^2{:}\langle e_1, e_2, e_3, e_4\rangle$	$P_O(\varrho^2)\cdot P_A(\sigma^2\|\varrho^2) = (0.5\cdot 0.2)\cdot 0.9 = 0.09$
$\sigma^3{:}\langle a, d, b, e\rangle$	$\varrho^3{:}\langle e_1, e_3, e_2, e_4\rangle$	$P_O(\varrho^3)\cdot P_A(\sigma^3\|\varrho^3) = (0.5\cdot 0.2)\cdot 0.9 = 0.09$
$\sigma^4{:}\langle a, c, e\rangle$	$\varrho^4{:}\langle e_1, e_2, e_4\rangle$	$P_O(\varrho^4)\cdot P_A(\sigma^4\|\varrho^4) = 0.8\cdot 0.1 = 0.08$
$\sigma^5{:}\langle a, c, d, e\rangle$	$\varrho^5{:}\langle e_1, e_2, e_3, e_4\rangle$	$P_O(\varrho^5)\cdot P_A(\sigma^5\|\varrho^5) = (0.5\cdot 0.2)\cdot 0.1 = 0.01$
$\sigma^6{:}\langle a, d, c, e\rangle$	$\varrho^6{:}\langle e_1, e_3, e_2, e_4\rangle$	$P_O(\varrho^6)\cdot P_A(\sigma^6\|\varrho^6) = (0.5\cdot 0.2)\cdot 0.1 = 0.01$

6 Validation of Probability Estimates

In this section, we compute the probability estimates for the realizations of an uncertain trace, and then show a validation of those estimates by Monte Carlo simulation on the behavior net of the trace. The process instance of our example has strong uncertainty in timestamps and weak uncertainty in activities and indeterminacy. It consists of 4 events: e_1, e_2, e_3 and e_4, where e_2 and e_3 have overlapping timestamps. Event e_2 executes b (resp., c) with probability 0.9 (resp., 0.1). There is a probability of 0.2 that e_3 did not occur. Figure 2 shows the corresponding behavior graph, an uncertain event data visualization that represents the time relationships between events with a directed acyclic graph [8]. Lastly, Table 4 list all the possible realizations, their probabilities, and the order-realizations enabling them.

We now validate our obtained probability estimates quantitatively by means of a Monte Carlo simulation approach. First, we construct the behavior net [9] corresponding to the uncertain process instance, which is shown in Figure 3. The set of replayable traces in this behavior net is exactly the set of realizations for the uncertain instance. Then, we simulate realizations on the behavior net, dividing the accumulated count of each realization by the number of runs, and compare those values to our probability estimates. Here, we use the *stochastic simulator* of the PM4Py library [3]. In every step

of the simulation, the stochastic simulator chooses one enabled transition to fire according to a stochastic map, assigning a weight to each transition in the Petri net (here, the behavior net).

To simulate uncertainty in activities, events and timestamps, we do the following: possible activities executed by the same event appearing in an XOR-split in the behavior net are weighted so to reflect the probability values of the activity labels. Indeterminacy is equivalently modeled as an XOR-choice between a visible transition and a silent one in the behavior net, so to model a "skip". If there are two or more possible activities for an indeterminate event, then the sum of the weights of the visible transitions in relation to the weight of the silent transition should be the same as in the distribution given in the event type uncertainty information. Whenever there are events with overlapping timestamps, these appear in an AND-split in the behavior net. The (enabled) path of the AND-split which is taken first signals which event is executed at that moment.

Let $bn(\tau) = (P, T)$ be the behavior net of trace τ. Let $(e, a) \in T$ be a visible transition related to some event $e \in \tau$. We weight (e, a) the following way:

$$weight((e, a)) = \begin{cases} f^e_A(a) & \text{if } \pi_o(e) = \varnothing, \\ (1 - f^e_O(?)) \cdot f^e_A(a) & \text{otherwise.} \end{cases}$$

If $e \in \tau$ is an indeterminate event, then $weight((e,)) = f^e_O(?)$.

Note that according to the weight assignment function, if e is determinate, then $\sum_{a \in \pi^{set}_a(e)} weight((e, a)) = 1$. Otherwise, $\sum_{a \in \pi^{set}_a(e)} weight((e, a)) = 1 - f^e_O(?) = 1 - weight((e, \tau))$. By construction of the behavior net, any transition related to an event in τ can only fire in accordance with the partial order of uncertain timestamps. Additionally, all transitions representing events with overlapping timestamps appear in an AND construct. By definition of our weight function, whenever the transitions of some $e \in \tau$ are enabled (in an XOR construct), the probability of firing one of them is $1/k$, where k is the number of events from τ for which none of the corresponding transitions have fired yet. This way, there is always a uniform distribution over the set of enabled transitions representing overlapping events. Assigning the weights according to this distribution allows to decorate the behavior net with probabilities that reflect the chances of occurrence of every possible value in uncertain attributes.

Applying the stochastic simulator n times yields n realizations. For each of the 6 possible realizations for the uncertain process instance, we obtain a probability measurement by dividing its simulated frequency by n. Figures 4 through 7 show how for greater n, this measurement converges to the probability estimates shown in Table 4, which were computed with our method.

To conclude, the Monte Carlo simulation shows that our estimated probabilities for realizations match their relative frequencies when one simulates the behavior net of the corresponding uncertain trace.

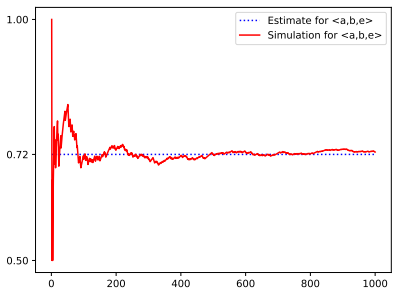

Figure 4: Plot showing how the frequency of trace $\langle a, b, e \rangle$ converges to the expected value of 0.72 over 1000 runs.

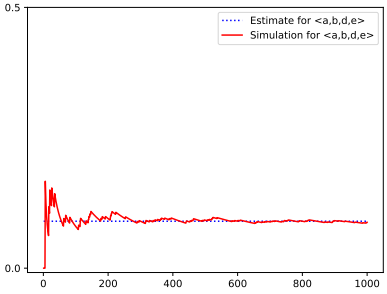

Figure 5: Plot showing how the frequency of trace $\langle a, b, d, e \rangle$ converges to the expected value of 0.09 over 1000 runs.

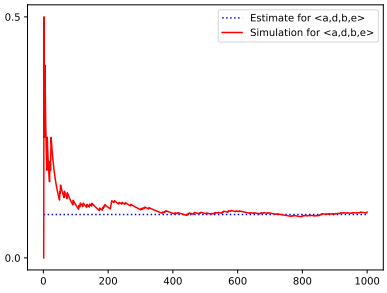

Figure 6: Plot showing how the frequency of trace $\langle a, d, b, e \rangle$ converges to the expected value of 0.09 over 1000 runs.

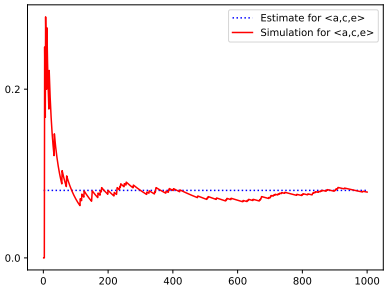

Figure 7: Plot showing how the frequency of trace $\langle a, c, e \rangle$ converges to the expected value of 0.08 over 1000 runs.

7 Conclusion

Uncertain traces inherently contain behavior, allowing for many realizations; these, in turn, correspond to diverse possible real-life scenarios, that may have different consequences on the management and governance of a process. In this paper, we presented a method to quantify the probability of each realization of an uncertain trace. This enables process analysts to weigh the impact of specific insights gathered with uncertainty-aware process mining techniques, such as conformance checking using alignments. As a consequence, information from process analysis techniques can be associated with a quantification of risk or opportunity for specific scenarios, making them more trustworthy.

Multiple avenues for future work on this topic are possible. These include inferring probabilities for uncertain traces from sections of the log not affected by uncertainty,

adopting certain traces or fragments of traces as ground truth. Moreover, inferring probabilities by examining evidence against a ground truth can also be achieved with a normative model that includes information concerning the probability of error or noise in specific parts of the process.

Acknowledgements

We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research interactions.

References

[1] van der Aa, Han, Henrik Leopold, and Matthias Weidlich. "Partial order resolution of event logs for process conformance checking". In: *Decision Support Systems* 136 (2020), p. 113347. DOI: 10.1016/j.dss.2020.113347.

[2] Ao, Xiang, Ping Luo, Chengkai Li, et al. "Online Frequent Episode Mining". In: *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*. Ed. by Gehrke, Johannes, Wolfgang Lehner, Kyuseok Shim, et al. IEEE Computer Society, 2015, pp. 891–902. DOI: 10.1109/ICDE.2015.7113342.

[3] Berti, Alessandro, Sebastiaan J. van Zelst, and Wil M. P. van der Aalst. "Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science". In: *ICPM Demo Track (CEUR 2374)*. 2019, pp. 13–16.

[4] Busany, Nimrod, Han van der Aa, Arik Senderovich, et al. "Interval-based Queries over Lossy IoT Event Streams". In: *Transanctions on Data Science* 1.4 (2020), 27:1–27:27. DOI: 10.1145/3385191.

[5] van Dongen, Boudewijn F., Josep Carmona, Thomas Chatain, et al. "Aligning Modeled and Observed Behavior: A Compromise Between Computation Complexity and Quality". In: *Advanced Information Systems Engineering - 29th International Conference, CAiSE 2017, Essen, Germany, June 12-16, 2017, Proceedings*. Ed. by Dubois, Eric and Klaus Pohl. Vol. 10253. Lecture Notes in Computer Science. Springer, 2017, pp. 94–109. DOI: 10.1007/978-3-319-59536-8_7.

[6] Leemans, Maikel and Wil M. P. van der Aalst. "Discovery of Frequent Episodes in Event Logs". In: *Proceedings of the 4th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2014), Milan, Italy, November 19-21, 2014*. Ed. by Accorsi, Rafael, Paolo Ceravolo, and Barbara Russo. Vol. 1293. CEUR Workshop Proceedings. CEUR-WS.org, 2014, pp. 31–45. URL: http://ceur-ws.org/Vol-1293/paper3.pdf.

[7] Lu, Xixi, Dirk Fahland, and Wil M. P. van der Aalst. "Conformance Checking Based on Partially Ordered Event Data". In: *Business Process Management Workshops - BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers*. Ed. by Fournier, Fabiana and Jan Mendling. Vol. 202. Lecture Notes in Business Information Processing. Springer, 2014, pp. 75–88. DOI: 10.1007/978-3-319-15895-2_7.

[8] Pegoraro, Marco and Wil M. P. van der Aalst. "Mining Uncertain Event Data in Process Mining". In: *International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24-26, 2019*. IEEE, 2019, pp. 89–96. DOI: 10.1109/ICPM.2019.00023.

[9] Pegoraro, Marco, Merih Seran Uysal, and Wil M. P. van der Aalst. "Conformance Checking over Uncertain Event Data". In: *Information Systems* (2021), p. 101810. DOI: 10.1016/j.is.2021.101810.

[10] Pegoraro, Marco, Merih Seran Uysal, and Wil M. P. van der Aalst. "Discovering Process Models from Uncertain Event Data". In: *Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1-6, 2019, Revised Selected Papers*. Ed. by Francescomarino, Chiara Di, Remco M. Dijkman, and Uwe Zdun. Vol. 362. Lecture Notes in Business Information Processing. Springer, 2019, pp. 238–249. DOI: 10.1007/978-3-030-37453-2_20.

[11] Zhu, Huisheng, Peng Wang, Xianmang He, et al. "Efficient Episode Mining with Minimal and Non-overlapping Occurrences". In: *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*. Ed. by Webb, Geoffrey I., Bing Liu, Chengqi Zhang, et al. IEEE Computer Society, 2010, pp. 1211–1216. DOI: 10.1109/ICDM.2010.25.