

Conformance Checking over Uncertain Event Data^{*}

Marco Pegoraro^[0000-0002-8997-7517], Merih Seran Uysal^[0000-0003-1115-6601],
and Wil M.P. van der Aalst^[0000-0002-0955-6940]

Process and Data Science Group (PADS)

Department of Computer Science, RWTH Aachen University, Aachen, Germany

{pegoraro,uysal,wvdaalst}@pads.rwth-aachen.de

<http://www.pads.rwth-aachen.de/>

Abstract. Nowadays, more and more process data are automatically recorded by information systems, and made available in the form of *event logs*. Process mining techniques enable process-centric analysis of data, including automatically discovering process models and checking if event data conform to a certain model. In this paper, we analyze the previously unexplored setting of uncertain event logs: logs where quantified uncertainty is recorded together with the corresponding data. We define a taxonomy of uncertain event logs and models, and we examine the challenges that uncertainty poses on process discovery and conformance checking. Finally, we show how upper and lower bounds for conformance can be obtained aligning an uncertain trace onto a regular process model.

Keywords: Process Mining · Uncertain Data · Partial Order.

1 Introduction

Over the last decades, the concept of *process* has become more and more central in formally describing the activities of businesses, companies and other similar entities, structured in specific steps and phases. A process is thus defined as a well-structured set of activities, potentially performed by multiple actors (*resources*), which contribute to the completion of a specific task or to the achievement of a specific goal. In this context, a very important notion is the concept of *case*, that is, a single instance of a process. For example, in a healthcare process a case may be a single hospitalization of a patient, or the patient himself; if the process belongs to a credit institution, a case may be a loan application from a customer, and so on. The case notion allows us to define a process as a procedure that precisely defines the steps needed to handle a case from inception to completion. This procedure is referred to as *process model*, and can be expressed in a number of different formalisms (transition systems, Petri nets, BPMN and UML

^{*} We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research interactions. We acknowledge Elisabetta Benevento for the valuable input. Please do not print this document unless strictly necessary.

diagrams, and many more). Consequently, the study and adoption of analysis techniques specifically customized to deal with process data and process models has enabled the bridging of business administration and data science and the development of dedicated disciplines like *business intelligence* and *business process management* (BPM).

The processes that govern the innards of business companies are increasingly supported by software tools. Performing specific activities is both aided and recorded by *process-aware information systems* (PAISs), which support the definition and management of processes. The information regarding the execution of processes can then be extracted from PAISs in the form of an *event log*, a database or file containing the digital footprint of the operations carried out in the context of the execution of a process and recorded as *events*. Event logs can vary in form, and contain differently structured information depending on the information system that enacted data collection in the organization. There are however some basic information regarding events that are very often recorded: these are the time in which the event occurred, the activity that has been performed, and the case identifier to which the event belongs. This last attribute allows to group events in clusters belonging to the same case, and these resulting clusters (usually organized in sequences sorted by timestamp) are called *process traces*. The discipline of *process mining* concerns the automatic analysis of event logs, with the goal of extracting knowledge regarding e.g. the structure of the process, the conformity of events to a specific normative process model, the performances in executing the process, the relationships between groups of actors in the process.

In this paper, we will consider the analysis of a specific class of event logs: the logs that contain *uncertain event data*. **Uncertain events are recordings of executions of specific activities in a process which are enclosed with an indication of uncertainty in the event attributes.** Specifically, we consider the case where the attributes of an event are **not recorded as a precise value but as a range or a set of alternatives.**

The recording of uncertain event data is a common occurrence in process management. The *Process Mining Manifesto* [2] describes a fundamental property of event data as *trustworthiness*, the assumption that the recorded data can be considered correct and accurate. In a general sense, uncertainty as defined here is an explicit absence of trustworthiness, with an indication of uncertainty recorded together with the event data. In the taxonomy of event data proposed in the Manifesto the logs at the two lower levels of quality frequently lack trustworthiness, and thus can be uncertain. This encompasses a wide range of processes, such as event logs of document and product management systems, error logs of embedded systems, worksheets of service engineers, and any process recorded totally or partially on paper. There are many possible causes behind the recording of uncertain event data, such as:

- *Incorrectness*. In some instances, the uncertainty is simply given by errors occurred while recording the data itself. Faults of the information system,

or human mistakes in a data entry phase can all lead to missing or altered event data that can be subsequently modeled as uncertain event data.

- *Coarseness*. Some information systems have limitations in their way of recording data - often tied to factors like the precision of the data format - such that the event data can be considered uncertain. A typical example is an information system that only records the date, but not the time, of the occurrence of an event: if two events are recorded in the same day, the order of occurrence is lost. This is an especially common circumstance in the processes that are, partially or completely, recorded on paper and then digitalized. Another factor that can lead to uncertainty in the time of recording is the information system being overloaded and, thus, delaying memorization of data. This type of uncertainty can also be generated by the limited sensibility of a sensor.
- *Ambiguity*. In some cases, the data recorded is not an identifier of a certain event attribute; in these instances, the data needs to be interpreted, either automatically or manually, in order to obtain a value for the event attribute. Uncertainty can arise if the meaning of the data is ambiguous and cannot be interpreted with precision. Example are data in the form of images, text, or video.

Aside from the causes, we can individuate other types of uncertain event logs based on the frequency of uncertain data. Uncertainty can be *infrequent*, when a specific attribute is only seldomly recorded together with explicit uncertainty; the uncertainty is rare enough that uncertain events can be considered outliers. Conversely, *frequent* uncertain behavior of the attribute is systematic, pervasive in a high number of traces, and thus not to be considered an outlier. The uncertainty can be considered part of the process itself. These concepts are not meant to be formal, and are laid out to distinguish between logs that are still processable regardless of the uncertainty, and logs where the uncertainty is too invasive to analyze them with existing process mining techniques.

In this paper, we propose a taxonomy of the different types of explicit uncertainty in process mining, together with a formal, mathematical formulation. As an example of practical application, we will consider the case of conformance checking [12], and we will apply it to uncertain data by assessing what are the upper and lower bounds on the conformance score for possible values of the attributes in an uncertain trace.

The main driving reasons behind this work is to provide the means to treat uncertainty as a relevant part of a process; thus, we aim not to filter it out but model it. In conclusion, there are two novel aspects regarding uncertain data that we intend to address in this work. The first is the *explicitness of uncertainty*: we work with the underlying assumption that the actual value of the uncertain attribute, while not directly provided, is described formally. This is the case when meta-information about the uncertainty in the attribute is available, either deduced from the features of the information system(s) that record the logs or included in the event log itself. Note that, as opposed to all previous work on the topic, the fact that uncertainty is explicit in the data means that the concept of uncertain behavior is completely separated from the

concept of infrequent behavior. The second is the goal of *modeling uncertainty*: we consider uncertainty part of the process. Instead of filtering or cleaning the log we introduce the uncertainty perspective in process mining by extending the currently available techniques to incorporate it.

The rest of this paper is organized as follows. Section 2 proposes a taxonomy of the different possible types of uncertain process data. Section 3 contains the formal definitions needed to manage uncertainty. Section 5 describes a practical application of process mining over uncertain event data, the case of conformance checking through alignments. Section 6 shows experimental results on computing conformance checking scores for synthetic uncertain data, as well as a case of application on real-life data. Section 7 discusses previous and related work on the management of uncertain data and on the topic of conformance checking. Finally, Section 8 concludes the paper and discusses about future work.

2 A Taxonomy of Uncertain Event Data

The goal of this section of the paper is to propose a categorization of the different types of uncertainty that can appear in process mining. In process management, a central concept is the distinction between the data perspective (the event log) and the behavioral perspective (the process model). The first one is a static representation of process instances, the second summarizes the behavior of a process. Both can be extended with a concept of explicit uncertainty: this concept also implies an extension of the process mining techniques that have currently been implemented.

In this paper we will focus on uncertainty in event data, while the concept of uncertainty applied to models will be examined in a future work. Specifically, as an example application we will consider computing the conformance score of uncertain process data on classical models.

We can individuate two different notions of uncertainty:

- *Strong uncertainty*: the possible values for the attributes are known, but the probability that the attribute will assume a certain instantiation is unknown or unobservable.
- *Weak uncertainty*: both the possible values of an attribute and their respective probabilities are known.

In the case of a discrete attribute, the strong notion of uncertainty consists on a set of possible values assumed by the attribute. In this case, the probability for each possible value is unknown. Vice-versa, in the weak uncertainty scenario we also have a discrete probability distribution defined on that set of values. In the case of a continuous attribute, the strong notion of uncertainty can be represented with an interval for the variable. Notice that an interval do not indicate a uniform distribution; there is no information on the likelihood of values in it. Vice-versa, in the weak uncertainty scenario we also have a probability density function defined on a certain interval. Figure 1 summarizes this concepts. This leads to very simple representations of explicit uncertainty.

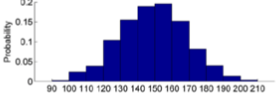
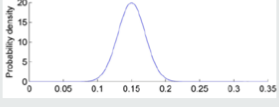
	Weak uncertainty	Strong uncertainty
Discrete data	Discrete probability distribution 	Set of possible values $\{x, y, z, \dots\}$
Continuous data	Probability density function 	Interval $\{x \in \mathbb{R} a \leq x \leq b\}$

Fig. 1. The four different types of uncertainty.

In this paper we consider only the control flow and time perspective of a process – namely, the attributes of the events that allow to discover a process model. These are the unique identifier of a process instance (case ID), the timestamp (often represented by the distance from a fixed origin point, e.g. the *Unix Epoch*), and the activity identifier of an event. Case IDs and activities are values chosen from a finite set of possible values; they are discrete variables. Timestamps, instead, are represented by numbers and thus are continuous variables.

We will also describe an additional type of uncertainty, which lays on the event level rather than the attribute level:

- *Indeterminate event*: the event may have not taken place even though it was recorded in the event log. Indeterminate events are indicated with a ? symbol, while determinate (regular) events are marked with a ! symbol.

Table 1. An example of strongly uncertain trace.

Case ID	Timestamp	Activity	Indet. event
{ID327, ID412}	2011-12-05T00:00	A	!
ID327	2011-12-07T00:00	{B, C, D}	!
ID327	[2011-12-06T00:00, 2011-12-10T00:00]	D	?
ID327	2011-12-09T00:00	{A, C}	!
{ID327, ID412, ID573}	2011-12-11T00:00	E	?

Table 2. An example of weakly uncertain trace.

Case ID	Timestamp	Activity	Indet. event
{ID327:0.9, ID412:0.1}	2011-12-05T00:00	A	!
ID327	2011-12-07T00:00	{B:0.7, C:0.3}	!
ID327	$\mathcal{N}(2011-12-08T00:00, 2)$	D	?:0.5
ID327	2011-12-09T00:00	{A:0.2, C:0.8}	!
{ID327:0.4, ID412:0.6}	2011-12-11T00:00	E	?:0.7

Examples of strongly and weakly uncertain traces are shown in Tables 1 and 2 respectively.

3 Preliminaries

Definition 1 (Power Set). *The power set of a set A is the set of all possible subsets of A , and is denoted with $\mathcal{P}(A)$. $\mathcal{P}_{NE}(A)$ denotes the set of all the non-empty subsets of A : $\mathcal{P}_{NE}(A) = \mathcal{P}(A) \setminus \{\emptyset\}$.*

Definition 2 (Multiset). *A multiset is an extension of the concept of set that keeps track of the cardinality of each element. $\mathcal{B}(A)$ is the set of all multisets over some set A . Multisets are denoted with square brackets, e.g. $b = [x, x, y]$.*

Definition 3 (Sequence, Subsequence and Permutation). *Given a set X , a finite sequence over X of length n is a function $s \in X^* : \{1, \dots, n\} \rightarrow X$, and it is written as $s = \langle s_1, s_2, \dots, s_n \rangle$. We denote with $\langle \rangle$ the empty sequence, the sequence with no elements and of length 0. Over the sequence s we define $|s| = n$, $s[i] = s_i$ and $x \in s \Leftrightarrow \exists 1 \leq i \leq n, s[i] = x$. The concatenation between two sequences is denoted with $\langle s_1, s_2, \dots, s_n \rangle \cdot \langle s'_1, s'_2, \dots, s'_m \rangle = \langle s_1, s_2, \dots, s_n, s'_1, s'_2, \dots, s'_m \rangle$. Given two sequences $s = \langle s_1, s_2, \dots, s_n \rangle$ and $s' = \langle s'_1, s'_2, \dots, s'_m \rangle$, s' is a subsequence of s if and only if there exists a sequence of strictly increasing natural numbers $\langle i_1, i_2, \dots, i_m \rangle$ such that $\forall 1 \leq j \leq m, s[i_j] = s'_j$. We indicate this with $s' \subseteq s$. A permutation of the set X is a sequence x_S that contains all elements of X without duplicates: $x_S \in X$, $X \subseteq x_S$, and for all $1 \leq i \leq |x_S|$ and for all $1 \leq j \leq |x_S|$, $x_S[i] = x_S[j] \rightarrow i = j$. We denote with \mathcal{S}_X all such permutations of set X .*

Definition 4 (Sequence Projection). *Let X be a set and $Q \subseteq X$ one of its subsets. $\downarrow_Q \in X^* \rightarrow Q^*$ is the sequence projection function and is defined recursively: $\langle \rangle \downarrow_Q = \langle \rangle$ and for $\sigma \in X^*$ and $x \in X$:*

$$(\langle x \rangle \cdot \sigma) \downarrow_Q = \begin{cases} \sigma \downarrow_Q & \text{if } x \notin Q \\ \langle x \rangle \cdot \sigma \downarrow_Q & \text{if } x \in Q \end{cases}$$

For example, $\langle y, z, y \rangle \downarrow_{\{x, y\}} = \langle y, y \rangle$.

Definition 5 (Applying Functions to Sequences). Let $f \in X \nrightarrow Y$ be a partial function. f can be applied to sequences of X using the following recursive definition: $f(\langle \rangle) = \langle \rangle$ and for $\sigma \in X^*$ and $x \in X$:

$$f(\langle x \rangle \cdot \sigma) = \begin{cases} f(\sigma) & \text{if } x \notin \text{dom}(f) \\ \langle f(x) \rangle \cdot f(\sigma) & \text{if } x \in \text{dom}(f) \end{cases}$$

Definition 6 (Transitive Relation and Correct Evaluation Order). Let X be a set of objects and R be a binary relation $R \subseteq X \times X$. R is transitive if and only if for all $x, x', x'' \in X$ we have that $(x, x') \in R \wedge (x', x'') \in R \rightarrow (x, x'') \in R$. A correct evaluation order is a permutation $s \in \mathcal{S}_X$ of the elements of the set X such that for all $1 \leq i < j \leq |s|$ we have that $(s[i], s[j]) \in R$.

Definition 7 (Strict Partial Order). Let S be a set of objects. Let $s, s' \in S$. A strict partial order (\prec, S) is a binary relation that have the following properties:

- Irreflexivity: $s \prec s$ is false.
- Transitivity: see Definition 6.
- Antisymmetry: $s \prec s'$ implies that $s' \prec s$ is false. Implied by irreflexivity and transitivity [15].

Definition 8 (Directed Graph). A directed graph $G \in \mathcal{U}_G$ is a tuple (V, E) where V is the set of vertices and $E \subseteq V \times V$ is the set of directed edges. The set \mathcal{U}_G is the graph universe. A path in a directed graph $G = (V, E)$ is a sequence of vertices p such that for all $1 < i < |p| - 1$ we have that $(p_i, p_{i+1}) \in E$. We denote with P_G the set of all such possible paths over the graph G . Given two vertices $v, v' \in V$, we denote with $p_G(v, v')$ the set of all paths beginning in v and ending in v' : $p_G(v, v') = \{p \in P_G \mid p[0] = v \wedge p[|p|] = v'\}$. v and v' are connected (and v' is reachable from v), denoted by $v \xrightarrow{G} v'$, if and only if there exists a path between them in G : $p_G(v, v') \neq \emptyset$. Conversely, $v \not\xrightarrow{G} v' \Leftrightarrow p_G(v, v') = \emptyset$. We omit the superscript G if it is clear from the context. A directed graph G is acyclic if there exists no path $p \in P_G$ satisfying $p[1] = p[|p|]$.

Definition 9 (Topological Sorting). Let $G = (V, E)$ be an acyclic directed graph. A topological sorting [18] $o_G \in \mathcal{S}_V$ is a permutation of the vertices of G such that for all $1 \leq i < j \leq |o_G|$ we have that $o_G[j] \not\xrightarrow{G} o_G[i]$. We denote with $\mathcal{O}_G \subseteq \mathcal{S}_V$ all such possible topological sortings over G .

Definition 10 (Transitive Reduction). A transitive reduction [7] $tr: \mathcal{G} \rightarrow \mathcal{G}$ of a graph $G = (V, E)$ is a graph $tr(G) = (V, E_r)$ with $E_r \subseteq E$ where every pair of vertices connected in $tr(G)$ is not connected by any other path: for all $(v, v') \in E_r$, $p_G(v, v') = \{\langle v, v' \rangle\}$. $tr(G)$ is the graph with the minimal number of edges that maintain the reachability between edges of G . The transitive reduction of a directed acyclic graph always exists and is unique [7].

Definition 11 (Dependency Graph). Let X be a set of objects and R be a transitive relation $R \subseteq X \times X$. A dependency graph [19] $\mathcal{D}(X, R) \in \mathcal{U}_G$ is the

directed graph $tr((X, R))$. Since R is transitive, for all $x, x' \in X$ we have that $x \mapsto x' \Leftrightarrow (x, x') \in R$, thus all the topological sortings $\mathcal{O}_{\mathcal{D}(X, R)}$ are also all possible correct evaluation orders of the objects in X for the relation R .

In general, and on a more abstract level, a dependency graph is a structure that explicitly expresses the property of adjunction between directed graphs and transitive relations, meaning that directed graphs define transitive relations and vice versa [28].

Let us now define the basic artifacts needed to perform process mining.

Definition 12 (Universes). Let \mathcal{U}_E be the set of all the event identifiers. Let \mathcal{U}_C be the set of all the case id identifiers. Let \mathcal{U}_A be the set of all the activity identifiers. Let \mathcal{U}_T be the totally ordered set of all the timestamp identifiers.

Definition 13 (Events and event logs). Let us denote with $\mathcal{E}_C = \mathcal{U}_E \times \mathcal{U}_C \times \mathcal{U}_A \times \mathcal{U}_T$ the universe of certain events. A certain event log is a set of events $L_C \subseteq \mathcal{E}_C$ such that every event identifier in L_C is unique.

Definition 14 (Simple certain traces and logs). Let $\{(e_1, c_1, a_1, t_1), (e_2, c_2, a_2, t_2), \dots, (e_n, c_n, a_n, t_n)\} \subseteq \mathcal{E}_C$ be a set of certain events and let $c_1 = c_2 = \dots = c_n$ and $t_1 < t_2 < \dots < t_n$. A simple certain trace is the sequence of activities $\langle a_1, a_2, \dots, a_n \rangle \in \mathcal{U}_A^*$ induced by such a set of events. \mathcal{T} denotes the universe of certain traces. $L \in \mathcal{B}(\mathcal{T})$ is a simple certain log. We will drop the qualifier “simple” if it is clear from the context.

As a preliminary application of process mining over uncertain event data we will consider conformance checking. Starting from an event log and a process model, conformance checking verifies if the event data in the log conforms to the model, providing a diagnostic of the deviations. Conformance checking serves many purposes, such as checking if process instances follow a specific normative model, assessing if a certain execution log has been generated from a specific model, or verifying the quality of a process discovery technique.

The conformance checking algorithm that we are applying in this paper is based on *alignments*. Introduced by Adriansyah [5], conformance checking through alignments finds deviations between a trace and a Petri net model of a process by creating a correspondence between the sequence of activities executed in the trace and the firing of the transitions in the Petri net. The following definitions are partially from [3].

Definition 15 (Petri Net). A Petri net is a tuple $N = (P, T, F)$ with P the set of places, T the set of transitions, $P \cap T = \emptyset$, and $F \subseteq (P \times T) \cup (T \times P)$ the flow relation. A Petri net $N = (P, T, F)$ defines a directed graph (V, E) with vertices $V = P \cup T$ and edges $E = F$. A marking $M \in \mathcal{B}(P)$ is a multiset of places.

A marking defines the state of a Petri net, and indicates how many *tokens* each place contains. For any $x \in P \cup T$, $\overset{N}{\bullet}x = \{x' \mid (x', x) \in F\}$ denotes the set

of input nodes and $x \bullet^N = \{x' \mid (x, x') \in F\}$ denotes the set of output nodes. We omit the superscript N if it is clear from the context.

A transition $t \in T$ is *enabled* in marking M of net N , denoted as $(N, M)[t]$, if each of its input places $\bullet t$ contains at least one token. An enabled transition t may *fire*, i.e., one token is removed from each of the input places $\bullet t$ and one token is produced for each of the output places $t \bullet$. Formally: $M' = (M \setminus \bullet t) \uplus t \bullet$ is the marking resulting from firing enabled transition t in marking M of Petri net N . $(N, M)[t](N, M')$ denotes that t is enabled in M and firing t results in marking M' .

Let $\sigma_T = \langle t_1, t_2, \dots, t_n \rangle \in T^*$ be a sequence of transitions. $(N, M)[\sigma_T](N, M')$ denotes that there is a set of markings M_0, M_1, \dots, M_n such that $M_0 = M$, $M_n = M'$, and $(N, M_i)[t_{i+1}](N, M_{i+1})$ for $0 \leq i < n$. A marking M' is *reachable* from M if there exists a σ_T such that $(N, M)[\sigma_T](N, M')$.

Definition 16 (Labeled Petri Net). A labeled Petri net $N = (P, T, F, l)$ is a Petri net (P, T, F) with labeling function $l \in T \rightarrow \mathcal{U}_A$ where \mathcal{U}_A is some universe of activity labels. Let $\sigma = \langle a_1, a_2, \dots, a_n \rangle \in \mathcal{U}_A^*$ be a sequence of activities. $(N, M)[\sigma \triangleright (N, M')]$ if and only if there is a sequence $\sigma_T \in T^*$ such that $(N, M)[\sigma_T](N, M')$ and $l(\sigma_T) = \sigma$.

If $t \notin \text{dom}(l)$, it is called *invisible*. To indicate invisible transitions we use the placeholder symbol τ ; by definition $\tau \notin \text{dom}(l)$. An occurrence of visible transition $t \in \text{dom}(l)$ corresponds to observable activity $l(t)$.

Definition 17 (System Net). A system net is a triplet $SN = (N, M_{init}, M_{final})$ where $N = (P, T, F, l)$ is a labeled Petri net, $M_{init} \in \mathcal{B}(P)$ is the initial marking, and $M_{final} \in \mathcal{B}(P)$ is the final marking. \mathcal{U}_{SN} is the universe of system nets. Over a system net we define the following:

- $T_v(SN) = \text{dom}(l)$ is the set of visible transitions in SN ,
- $A_v(SN) = \text{rng}(l)$ is the set of corresponding observable activities in SN ,
- $T_v^u(SN) = \{t \in T_v(SN) \mid \forall t' \in T_v(SN) \ l(t) = l(t') \Rightarrow t = t'\}$ is the set of unique visible transitions in SN (i.e., there are no other transitions having the same visible label),
- $A_v^u(SN) = \{l(t) \mid t \in T_v^u(SN)\}$ is the set of corresponding unique observable activities in SN ,
- $\phi(SN) = \{\sigma \mid (N, M_{init})[\sigma \triangleright (N, M_{final})]\}$ is the set of visible traces starting in M_{init} and ending in M_{final} , and
- $\phi_f(SN) = \{\sigma_T \mid (N, M_{init})[\sigma_T](N, M_{final})\}$ is the corresponding set of complete firing sequences.

Figure 2 shows a system net with initial and final markings $M_{init} = \{\text{start}\}$ and $M_{final} = \{\text{end}\}$. Given a system net, $\phi(SN)$ is the set of all possible *visible* activity sequences, i.e. the labels of complete firing sequences starting in M_{init} and ending in M_{final} projected onto the set of observable activities. Given the set of activity sequences $\phi(SN)$ obtainable via complete firing sequences on a certain system net, we can define a perfectly fitting event log as a set of traces which activity projection is contained in $\phi(SN)$.

Definition 18 (Perfectly Fitting Log). Let $L \in \mathcal{B}(\mathcal{T})$ be a certain event log and let $SN = (N, M_{init}, M_{final}) \in \mathcal{U}_{SN}$ be a system net. L is perfectly fitting SN if and only if $\{\sigma \in L\} \subseteq \phi(SN)$.

These definitions allow us to build *alignments* in order to compute the fitness of trace on a certain model. An alignment is a correspondence between a sequence of activities (extracted from the trace) and a sequence of transitions with the relative labels (fired in the model while replaying the trace). The first sequence indicates the “moves in the log” and the second indicates the “moves in the model”. If a move in the model cannot be mimicked by a move in the log, then a “ \gg ” (“no move”) appears in the top row; conversely, if a move in the log cannot be mimicked by a move in the model, then a “ \gg ” (“no move”) appears in the bottom row. “no moves” not corresponding to invisible transitions point to deviations between the model and the log. A *move* is a pair $(x, (y, t))$ where the first element refers to the log and the second element to the model. A “ \gg ” in the first element of the pair indicates a move on the model, in the second element it indicates a move on the log.

Definition 19 (Legal Moves). Let $L \in \mathcal{B}(\mathcal{T})$ be a certain event log, let $A \subseteq \mathcal{U}_A$ be the set of activity labels appearing in the event log, and let $SN = (N, M_{init}, M_{final}) \in \mathcal{U}_{SN}$ be a system net with $N = (P, T, F, l)$. $A_{LM} = \{(x, (x, t)) \mid x \in A \wedge t \in T \wedge l(t) = x\} \cup \{(\gg, (x, t)) \mid t \in T \wedge l(t) = x\} \cup \{(x, \gg) \mid x \in A\}$ is the set of legal moves.

An alignment is a sequence of legal moves such that after removing all “ \gg ” symbols, the top row corresponds to a trace in the log and the bottom row corresponds to a firing sequence starting in M_{init} and ending M_{final} . Notice that if $t \notin \text{dom}(l)$ is an invisible transition, the activation of t is indicated by a “ \gg ” on the log in correspondence of t and the placeholder label τ . Hence, the middle row corresponds to a visible path when ignoring the τ steps. Figure 2 shows a system net with two examples of alignments, σ_1 of a fitting trace and σ_2 of a non-fitting trace.

Definition 20 (Alignment). Let $\sigma \in L$ be a certain trace and $\sigma_T \in \phi_f(SN)$ a complete firing sequence of system net SN . An alignment of σ and σ_T is a sequence $\gamma \in A_{LM}^*$ such that the projection on the first element (ignoring “ \gg ”) yields σ and the projection on the last element (ignoring “ \gg ” and transition labels) yields σ_T .

A trace and a model can have several possible alignments. In order to select the most appropriate one, we introduce a function that associate a *cost* to undesired moves - the ones associated with deviations.

Definition 21 (Cost of Alignment). Cost function $\delta \in A_{LM} \rightarrow \mathbb{N}$ assigns costs to legal moves. The cost of an alignment $\gamma \in A_{LM}^*$ is the sum of all costs: $\delta(\gamma) = \sum_{(x,y) \in \gamma} \delta(x, y)$.

Moves where log and model agree have no costs, i.e., $\delta(x, (x, t)) = 0$ for all $x \in A$. Moves on model only have no costs if the transition is invisible, i.e., $\delta(\gg, (\tau, t)) = 0$ if $l(t) = \tau$. $\delta(\gg, (x, t)) > 0$ is the cost when the model makes an “ x move” without a corresponding move of the log (assuming $l(t) = x \neq \tau$). $\delta(x, \gg) > 0$ is the cost for an “ x move” only on the log. In this paper we often use a standard cost function δ_S that assigns unit costs: $\delta_S(x, (x, t)) = 0$, $\delta_S(\gg, (\tau, t)) = 0$, and $\delta_S(\gg, (x, t)) = \delta_S(x, \gg) = 1$ for all $x \in A$.

Definition 22 (Optimal Alignment). Let $L \in \mathcal{B}(\mathcal{T})$ be a certain event log and let $SN \in \mathcal{U}_{SN}$ be a system net with $\phi(SN) \neq \emptyset$.

- For $\sigma \in L$, we define: $\Gamma_{\sigma, SN} = \{\gamma \in A_{LM}^* \mid \exists_{\sigma_T \in \phi_f(SN)} \gamma \text{ is an alignment of } \sigma \text{ and } \sigma_T\}$.
- An alignment $\gamma \in \Gamma_{\sigma, SN}$ is optimal for trace $\sigma \in L$ and system net SN if for any $\gamma' \in \Gamma_{\sigma, SN}$: $\delta(\gamma') \geq \delta(\gamma)$.
- $\lambda_{SN} \in \mathcal{T} \rightarrow A_{LM}^*$ is a deterministic mapping that assigns any trace σ to an optimal alignment, i.e., $\lambda_{SN}(\sigma) \in \Gamma_{\sigma, SN}$ and $\lambda_{SN}(\sigma)$ is optimal.
- costs(L, SN, δ) = $\sum_{\sigma \in L} \delta(\lambda_{SN}(\sigma))$ are the misalignment costs of the whole event log.

$\sigma \in L$ is a (perfectly) fitting trace for the system net SN if and only if $\delta(\lambda_{SN}(\sigma)) = 0$. L is a (perfectly) fitting event log for the system net SN if and only if costs(L, SN, δ) = 0.

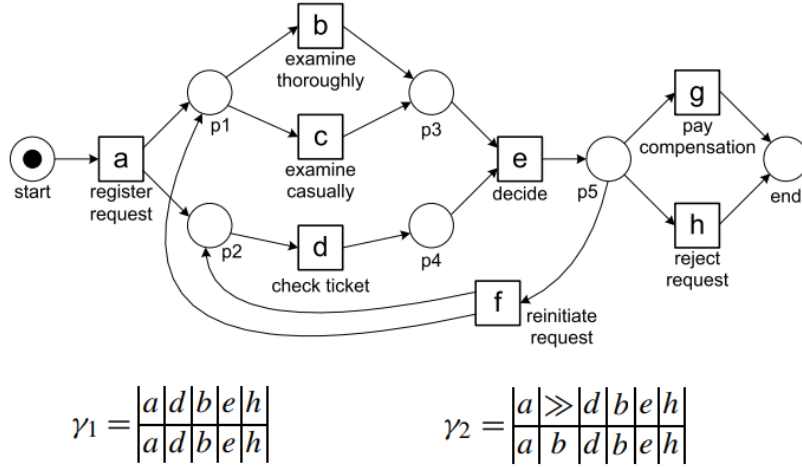


Fig. 2. Example of alignments on a system net. The alignment γ_1 shows that the trace $\langle a, d, b, e, h \rangle$ is perfectly fitting the net. The alignment γ_2 shows that the trace $\langle a, b, d, b, e, h \rangle$ is misaligned with the net in one point, indicated by “ \gg ”.

The technique to compute the optimal alignment [5] is as follows. Firstly, it creates an *event net*, a sequence-structured system net able to replay only the

trace to align. The transitions in the event net have labels corresponding to the activities in the trace. Then, a *product net* should be computed; it is the union of the event net and the model together with synchronous transitions added. These additional transitions are paired with transitions in the event net and in the process model that have the same label; they are then connected with arcs from the input places and to the output places of those transitions. The product net is able to represent moves on log, moves on model and synchronous moves by means of firing transitions: the transitions of the event net correspond to moves on log, the transitions of the process model correspond to moves on model, the added synchronous transitions correspond to synchronous moves. The union of the initial and final markings of the event net and the process model constitute respectively the initial and final marking of the product net: every complete firing sequence on the product net corresponds to a possible alignment. Lastly, the product net is translated to a state space, and a state space exploration via the A^* algorithm is performed in order to find the complete firing sequence that yields the lowest cost.

Let us define formally the construction of the event net and the product net:

Definition 23 (Event Net). *Let $\sigma \in \mathcal{T}$ be a certain trace. The event net $en : \mathcal{T} \rightarrow \mathcal{U}_{SN}$ of σ is a system net $en(\sigma) = (P, T, F, l, M_{init}, M_{final})$ such that:*

- $P = \{p_i \mid 1 \leq i \leq |\sigma| + 1\}$,
- $T = \{t_i \mid 1 \leq i \leq |\sigma|\}$,
- $F = \bigcup_{1 \leq i \leq |\sigma|} \{(p_i, t_i), (t_i, p_{i+1})\}$
- $l : T \rightarrow \mathcal{U}_A$ such that for all $1 \leq i \leq |\sigma|$, $l(t_i) = \sigma[i]$,
- $M_{init} = \{p_1\}$,
- $M_{final} = \{p_{|\sigma|+1}\}$.

Definition 24 (Product of two Petri Nets [32]). *Let $S_1 = (P_1, T_1, F_1, l_1, M_{init_1}, M_{final_1})$ and $S_2 = (P_2, T_2, F_2, l_2, M_{init_2}, M_{final_2})$ be two system nets. The product net of S_1 and S_2 is the system net $S = S_1 \otimes S_2 = (P, T, F, l, M_{init}, M_{final})$ such that:*

- $P = P_1 \cup P_2$,
- $T \subseteq (T_1 \cup \{\gg\} \times T_2 \cup \{\gg\})$ such that $T = \{(t_1, \gg) \mid t_1 \in T_1\} \cup \{(\gg, t_2) \mid t_2 \in T_2\} \cup \{(t_1, t_2) \in (T_1 \times T_2) \mid l_1(t_1) = l_2(t_2) \neq \tau\}$,
- $F \subseteq (P \times T) \cup (T \times P)$ such that

$$F = \{(p, (t, \gg)) \mid p \in P_1 \wedge t \in T_1 \wedge (p, t) \in F_1\} \cup$$

$$\{((t, \gg), p) \mid t \in T_1 \wedge p \in P_1 \wedge (t, p) \in F_1\} \cup$$

$$\{(p, (t, \gg)) \mid p \in P_2 \wedge t \in T_2 \wedge (p, t) \in F_2\} \cup$$

$$\{((t, \gg), p) \mid t \in T_2 \wedge p \in P_2 \wedge (t, p) \in F_2\} \cup$$

$$\{(p, (t_1, t_2)) \mid p \in P_1 \cup P_2 \wedge (t_1, t_2) \in T \cap (T_1 \times T_2)\} \cup$$

$$\{((t_1, t_2), p) \mid p \in P_1 \cup P_2 \wedge (t_1, t_2) \in T \cap (T_1 \times T_2)\}$$
- $l : T \rightarrow \mathcal{U}_A$ such that for all $(t_1, t_2) \in T$, $l((t_1, t_2)) = l_1(t_1)$ if $t_2 = \gg$, $l((t_1, t_2)) = l_2(t_2)$ if $t_1 = \gg$, and $l((t_1, t_2)) = l_1(t_1)$ otherwise,
- $M_{init} = M_{init_1} \uplus M_{init_2}$,
- $M_{final} = M_{final_1} \uplus M_{final_2}$.

4 Uncertainty in Process Mining

Definition 25 (Determinate and indeterminate event qualifiers). Let $\mathcal{U}_O = \{!, ?\}$, where the “!” symbol denotes determinate events, and the “?” symbol denotes indeterminate events.

Definition 26 (Uncertain events). Let us denote $\mathcal{E}_S = \mathcal{U}_E \times \mathcal{P}_{NE}(\mathcal{U}_C) \times \mathcal{P}_{NE}(\mathcal{U}_A) \times \mathcal{P}_{NE}(\mathcal{U}_T) \times \mathcal{U}_O$ the universe of strongly uncertain events. $\mathcal{E}_W = \{(e, f) \in \mathcal{U}_E \times (\mathcal{U}_C \times \mathcal{U}_A \times \mathcal{U}_T \not\rightarrow [0, 1]) \mid \sum_{(a,c,t) \in \text{dom}(f)} f(c, a, t) \leq 1\}$ is the universe of weakly uncertain events. Over a strongly uncertain event $(e, c_s, a_s, t_s, o) \in \mathcal{E}_S$ we define the following projection functions: $\pi_c^{\mathcal{E}_S}(e) = c_s$, $\pi_a^{\mathcal{E}_S}(e) = a_s$, $\pi_t^{\mathcal{E}_S}(e) = t_s$ and $\pi_o^{\mathcal{E}_S}(e) = o$. We omit the superscript \mathcal{E}_S from the projection functions if it is clear from the context.

Now that the definitions of strongly and weakly uncertain events are structured, let us aggregate them in uncertain event logs.

Definition 27 (Event logs). A strongly uncertain event log is a set of events $L_S \subseteq \mathcal{E}_S$ such that every event identifier in L_S is unique. A weakly uncertain event log is a set of events $L_W \subseteq \mathcal{E}_W$ such that every event identifier in L_W is unique.

A weakly uncertain event log $L_W \subseteq \mathcal{E}_W$ has a corresponding strongly uncertain event log $\overline{L_W} = L_S \subseteq \mathcal{E}_S$ such that $L_S = \{(e, c_s, a_s, t_s, o) \in \mathcal{E}_S \mid \exists (e', f) \in L_W e = e' \wedge c_s = \{c \in \mathcal{U}_C \mid \exists_{a,t}((c, a, t) \in \text{dom}(f) \wedge f(c, a, t) > 0)\} \wedge a_s = \{a \in \mathcal{U}_A \mid \exists_{c,t}((c, a, t) \in \text{dom}(f) \wedge f(c, a, t) > 0)\} \wedge t_s = \{t \in \mathcal{U}_T \mid \exists_{c,a}((c, a, t) \in \text{dom}(f) \wedge f(c, a, t) > 0)\} \wedge (o = ! \Leftrightarrow \sum_{(c,a,t) \in \text{dom}(f)} f(c, a, t) = 1 \wedge (o = ? \Leftrightarrow \sum_{(c,a,t) \in \text{dom}(f)} f(c, a, t) < 1)\}$.

Definition 28 (Realization of an event log). $L_C \subseteq \mathcal{E}_C$ is a realization of $L_S \subseteq \mathcal{E}_S$ if and only if:

- For all $(e, c, a, t) \in L_C$ there is a distinct $(e', c_s, a_s, t_s, o) \in L_S$ such that $e' = e$, $a \in a_s$, $c \in c_s$ and $t \in t_s$;
- For all $(e, c_s, a_s, t_s, o) \in L_S$ with $o = !$ there is a distinct $(e', c, a, t) \in L_C$ such that $e' = e$, $a \in a_s$, $c \in c_s$ and $t \in t_s$.

$\mathcal{R}_L(L_S)$ is the set of all such realizations of the log L_S .

Note that these definitions allow us to transform a weakly uncertain log into a strongly uncertain one, and a strongly uncertain one in a set of certain logs.

The types of uncertainty in the specific scenario we consider in this paper includes:

- Strong uncertainty on the activity;
- Strong uncertainty on the timestamp;

- Strong uncertainty on indeterminate events.

All three can happen concurrently. Table 3 shows such a trace, which we will use as running example. It is worth noticing that the specific case of uncertainty on the case ID causes a problem; since an event can have many possible case IDs, it can belong to different traces. In data format where the events are already aggregated into traces, such as the very common XES standard, this means that the information related to a trace can be *non-local* to the trace itself, but can be stored in some other points of the log. We will focus on the problem of uncertainty on the case ID attribute in a future work.

Firstly, we will lay down some simplified notation in order to model the problem at hand in a more compact way.

Definition 29 (Simple uncertain events, traces and logs). $e_U^S \in \mathcal{U}_E \times \mathcal{P}_{NE}(\mathcal{U}_A) \times \mathcal{U}_T \times \mathcal{U}_T \times \mathcal{U}_O$ is a simple uncertain event. Let us denote with $\mathcal{E}_U^S = \mathcal{U}_E \times \mathcal{P}_{NE}(\mathcal{U}_A) \times \mathcal{U}_T \times \mathcal{U}_T \times \mathcal{U}_O$ the universe of all simple uncertain events. $\sigma_U \in \mathcal{P}(\mathcal{E}_U^S)$ is a simple uncertain trace if for all $(e, a_s, t_{min}, t_{max}, o) \in \sigma_U$, $t_{min} < t_{max}$ and all the event identifiers are unique. \mathcal{T}_U denotes the universe of simple uncertain traces. $L_U \in \mathcal{B}(\mathcal{T}_U)$ is a simple uncertain log if all the event identifiers in the log are unique. For $e_U^S = (e, a_s, t_{min}, t_{max}, o) \in \sigma_U$ we define the following projection functions: $\pi_a^{L_U}(e_U^S) = a_s$, $\pi_{t_{min}}^{L_U}(e_U^S) = t_{min}$, $\pi_{t_{max}}^{L_U}(e_U^S) = t_{max}$ and $\pi_o^{L_U}(e_U^S) = o$. We omit the superscript L_U from the projection functions if it is clear from the context.

These simplified traces and logs can be related to the more general framework described in the previous section through the following transformation: let $L_S \subseteq \mathcal{E}_S$ be a strongly uncertain log and let $g: \mathcal{U}_E \rightarrow \mathcal{U}_C$ be a function mapping events onto cases such that $dom(g) = \{e \mid (e, c_s, a_s, t_s, u) \in L_S\}$ and for all $(e, c_s, a_s, t_s, u) \in L_S$, $g(e) \in c_s$. Thus, for $c \in rng(g)$, $g^{-1}(c) = \{e \in \mathcal{U}_E \mid g(e) = c\}$. The simple uncertain event log defined by g on L_S is $L_U = [\{(e, \pi_a(e), \min(\pi_t(e)), \max(\pi_t(e)), \pi_o(e)) \mid e \in g^{-1}(c) \mid c \in rng(g)\}]$.

In order to more easily work with timestamps in simple uncertain events, let us frame their time relationship as a strict partial order.

Definition 30 (Strict partial order over simple uncertain events). Let $e, e' \in \mathcal{E}_U^S$ be two simple uncertain events. (\prec, \mathcal{E}_U^S) is a strict partial order defined on the universe of strongly uncertain events \mathcal{E}_U^S as:

$$e \prec e' \Leftrightarrow \pi_{t_{max}}(e) < \pi_{t_{min}}(e')$$

Theorem 1 $((\prec, \mathcal{E}_U^S)$ is a strict partial order).

Proof. All properties characterizing strict partial orders are fulfilled by (\prec, \mathcal{E}_U^S) . For all $e, e', e'' \in \mathcal{E}_U^S$ we have:

- Irreflexivity: this property is always verified, since $\pi_{t_{max}}(e) < \pi_{t_{min}}(e)$ is false (see Definition 26).

- Transitivity: since $\pi_{t_{max}}(e) < \pi_{t_{min}}(e') \leq \pi_{t_{max}}(e') < \pi_{t_{min}}(e'')$ and \mathcal{U}_T is totally ordered, we have that $\pi_{t_{max}}(e) < \pi_{t_{min}}(e'')$ and this property is always verified.

□

Lemma 1 (Uncomparable events share possible timestamp values). *Let $e, e' \in \mathcal{E}_U^S$ be two strongly uncertain events. e and e' are uncomparable with respect to the strict partial order (\prec, \mathcal{E}_U^S) (i.e., neither $e \prec e'$ nor $e' \prec e$ are true) if and only if e and e' share some possible values of their timestamp.*

Proof.

(\Rightarrow) From Definition 30, it follows that two events $e, e' \in \mathcal{E}_U^S$ are comparable if and only if either $\pi_{t_{max}}(e) < \pi_{t_{min}}(e')$ or $\pi_{t_{max}}(e') < \pi_{t_{min}}(e)$. If both are false, then $\pi_{t_{min}}(e') \leq \pi_{t_{max}}(e)$ and $\pi_{t_{min}}(e) \leq \pi_{t_{max}}(e')$. If we assume that $\pi_{t_{min}}(e) \leq \pi_{t_{min}}(e')$ then $\pi_{t_{min}}(e) \leq \pi_{t_{min}}(e') \leq \pi_{t_{max}}(e)$, while if $\pi_{t_{min}}(e) > \pi_{t_{min}}(e')$ then $\pi_{t_{min}}(e') < \pi_{t_{min}}(e) \leq \pi_{t_{max}}(e')$. In both cases, there are values common to both uncertain timestamps.

(\Leftarrow) If the two events share timestamp values, it follows that at least one of the extremes of one event is encompassed by the extremes of the other. Assume that e encompasses at least one of the extremes of e' (the other case is symmetric): then either $\pi_{t_{min}}(e) \leq \pi_{t_{min}}(e') \leq \pi_{t_{max}}(e)$ or $\pi_{t_{min}}(e) \leq \pi_{t_{max}}(e') \leq \pi_{t_{max}}(e)$. In the first case, considering that \mathcal{U}_T is totally ordered and that $\pi_{t_{min}}(e') \leq \pi_{t_{max}}(e')$, we have that both $\pi_{t_{min}}(e') \leq \pi_{t_{max}}(e)$ and $\pi_{t_{min}}(e) \leq \pi_{t_{max}}(e')$ are true, and e and e' are uncomparable. The second case is proved analogously. □

Definition 31 (Realizations of simple uncertain traces). *Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace. An order-realization $\sigma_O \in \mathcal{S}(\sigma_U)$ is a permutation of the events in σ_U such that for all $1 \leq i < j \leq |\sigma_O|$ we have that $\sigma_O[j] \not\prec \sigma_O[i]$, i.e. σ_O is a correct evaluation order for σ_U over (\prec, \mathcal{E}_U^S) , and the (total) order in which events are sorted in σ_O is a linear extension of the strict partial order (\prec, \mathcal{E}_U^S) . We denote with $\mathcal{R}_O(\sigma_U)$ the set of all such order-realizations of the trace σ_U .*

Given an order-realization $\sigma_O \in \mathcal{R}_O(\sigma_U)$, $\sigma_A \in (\mathcal{U}_A \cup \{\tau\})^$ is an activity-realization of σ_O if and only if $|\sigma_A| = |\sigma_O| = n$ and for all $1 \leq i \leq n$ we have that*

$$\sigma_A[i] \in \begin{cases} \pi_a(\sigma_O[i]) & \text{if } \pi_o(\sigma_O) = ! \\ \pi_a(\sigma_O[i]) \cup \{\tau\} & \text{if } \pi_o(\sigma_O) = ? \end{cases}$$

We denote with $\overline{\mathcal{R}_A}(\sigma_O)$ the multiset of activity-realizations obtainable from the order-realization σ_O , and with $\mathcal{R}_A(\sigma_U) = \biguplus_{\sigma_O \in \mathcal{R}_O(\sigma_U)} \overline{\mathcal{R}_A}(\sigma_O)$ the multiset of all activity-realizations of the trace σ_U .

Given an activity-realization $\sigma_A \in \mathcal{R}_A(\sigma_U)$, the realization $\overline{\overline{\mathcal{R}}}(\sigma_A) \in \mathcal{T}$ is the certain trace obtained by removing all occurrences of τ from σ_A : $\overline{\overline{\mathcal{R}}}(\sigma_A) = \sigma_A \downarrow \mathcal{U}_A$. Given $\sigma_O \in \mathcal{R}_O(\sigma_U)$, we denote with $\overline{\mathcal{R}}(\sigma_O) = \biguplus_{\sigma_A \in \overline{\mathcal{R}_A}(\sigma_O)} \overline{\overline{\mathcal{R}}}(\sigma_A)$ the multiset of all realizations obtainable from σ_O , and with $\mathcal{R}(\sigma_U) = \biguplus_{\sigma_O \in \mathcal{R}_O(\sigma_U)} \overline{\mathcal{R}}(\sigma_O)$ the multiset of all realizations obtainable from σ_U .

Simple uncertain traces and log carry less information than their certain counterpart. Nevertheless, it is possible to extend existing process mining algorithms to extract the information in a simple uncertain log to design a process model that describes its possible behavior, or verify that it conforms to a given normative model.

5 Conformance Checking on Uncertain Event Data

Depending on the possible values for a_s , t_{min} , t_{max} , and u there are multiple possible realizations of a trace. This means that, given a model, a simple uncertain trace could be fitting for certain realizations, but non-fitting for others. The question we are interested in answering is: given a simple uncertain trace and a Petri net process model, is it possible to find an upper and lower bound for the conformance score? More formally, when usually we are interested in the optimal alignments (the ones with the minimal cost), we are now interested in the minimum and maximum cost of alignments in the realization set of a simple uncertain trace.

Definition 32 (Upper and Lower Bound on Alignment Cost for a Trace).

Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace, and let $SN \in \mathcal{U}_{SN}$ be a system net. The upper bound for the alignment cost is a function $\delta_{max}: \mathcal{T}_U \rightarrow \mathbb{N}$ such that $\delta_{max}(\sigma_U) = \max_{\sigma \in \mathcal{R}(\sigma_U)} \delta(\lambda_{SN}(\sigma))$. The lower bound for the alignment cost is a function $\delta_{min}: \mathcal{T}_U \rightarrow \mathbb{N}$ such that $\delta_{min}(\sigma_U) = \min_{\sigma \in \mathcal{R}(\sigma_U)} \delta(\lambda_{SN}(\sigma))$.

A simple way to compute the upper and lower bounds for the cost of any uncertain trace is using a bruteforce approach: enumerating the possible realizations of the trace, then searching for the costs of optimal alignments for all the realizations, and picking the minimum and maximum as bounds. We now present a technique which improves the performance of calculating the lower bound for conformance cost over using a bruteforce method.

This technique is best illustrated by following a running example. Let us consider the following process instance. In a hospital, a medic visits a patient that is feeling feverish. The body temperature is taken at 11:00, and the thermometer reaches 40.3°C, a high fever. An antipyretic is then administered at 11:10, and the patient rests for some time on one of the beds of the emergency room. At 12:00, a nurse discovers a rash on the back of the patient's left arm. It is unclear when the rash developed; together with the fever, it might indicate a bacterial infection, but at the same time it is a known side effect of the administered antipyretic for patients with drug sensitivity. The medics decide to admit the patient in the infectious diseases ward at 13:00. Later on, two facts are discovered by the medics: first, the thermometer used on the patient gives very inaccurate readings, so the fever might have been way less severe. Second, the nurse did not record in the patient's folder which dosage of antipyretic was administered - either the 2g dose or the 4g dose. These events generate the trace of Table 3 in the information system of the hospital.

Table 3. The uncertain trace of an instance of healthcare process used as running example. For sake of readability, on the timestamps column only the time is shown.

Case ID	Event ID	Timestamp	Activity	Indet. event
ID327	e_1	11:00	<i>HighFever</i>	?
ID327	e_2	11:10	$\{Apyr2, Apyr4\}$!
ID327	e_3	[10:00, 12:00]	<i>Rash</i>	!
ID327	e_4	13:00	<i>Adm</i>	!

We will produce a version of the event net that embeds the possible behaviors of the uncertain trace. We define a *behavior net*, a Petri net that can replay all and only the realizations of an uncertain trace. As an intermediate step in order to obtain such a Petri net, we first build the *behavior graph*, a dependency graph representing the uncertain trace. This graph contains a vertex for each uncertain event in the trace and contains an edge between two vertices if the corresponding uncertain events may happen one directly after the other.

Definition 33 (Behavior Graph). Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace. A behavior graph $bg: \mathcal{T}_U \rightarrow \mathcal{U}_G$ is the dependency graph representing the time relationship between simple uncertain events in σ_U : $bg(\sigma_U) = \mathcal{D}(\sigma_U, (\prec, \mathcal{E}_U^S))$

The behavior graph provides a structured representation of the uncertainty on the timestamp: when a specific vertex has two or more outbound edges, the events corresponding to the destination vertices can occur in any order, concurrently with each other. We can see the result on the example trace in Figures 3 and 4.

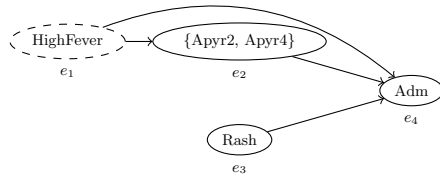


Fig. 3. The behavior graph of the trace in Table 3 before applying the transitive reduction. The dashed node represents an indeterminate event.

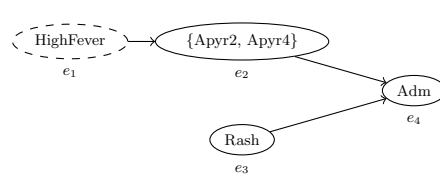


Fig. 4. The same behavior graph after the transitive reduction.

Theorem 2 (Correctness of behavior graphs). Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace and $bg(\sigma_U) = (V, E)$ be its behavior graph. The behavior graph $bg(\sigma_U)$ is acyclic; additionally, the set of all topological sortings of the behavior graph corresponds to the set of order-realizations of σ_U : $\mathcal{O}_{bg(\sigma_U)} = \mathcal{R}_O(\sigma_U)$.

Proof. From Theorem 1 we know that (\prec, \mathcal{E}_U^S) is a strict partial order. Let $p = \langle p_1, p_2, \dots, p_m \rangle \in P_{bg}$ be a path in the behavior graph: if p was a cycle, that means that according to Definition 33 we have $p_1 \prec p_2 \prec \dots \prec p_m \prec p_1$. Since (\prec, \mathcal{E}_U^S) is transitive we have that $p_1 \prec p_m$ and $p_m \prec p_1$, which would violate the antisymmetry property in Definition 7 and would contradict Theorem 1. Thus the behavior graph is necessarily acyclic.

The result $\mathcal{O}_{bg(\sigma_U)} = \mathcal{R}_O(\sigma_U)$ immediately follows from Definitions 11, 31 and 33, and from Theorem 1. \square

Lemma 2 (Semantics of behavior graphs). *Events connected by paths in a given behavior graph have a precedence relationship; events not connected by any paths share possible values for their timestamps and thus might have happened in any order.*

Proof. Immediately follows from Theorems 1 and 2, and from Lemma 1. \square

We then obtain a *behavior net* by replacing every vertex in the behavior graph with one or more transitions in an XOR configuration, each representing an activity contained in the π_a set of the corresponding uncertain event. The edges of the behavior graph become connection through places in the behavior net.

Definition 34 (Behavior Net). *Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace, and let $bg(\sigma_U) = (V, E)$ be the corresponding behavior graph. A behavior net $bn: \mathcal{T}_U \rightarrow \mathcal{U}_{SN}$ is a system net $bn(\sigma_U) = (P, T, F, l, M_{init}, M_{final})$ such that:*

- $P = E \cup$
 $\{(\text{START}, v) \mid v \in V \wedge \nexists_{v' \in V} (v', v) \in E\} \cup$
 $\{(v, \text{END}) \mid v \in V \wedge \nexists_{v' \in V} (v, v') \in E\}$
- $T = \{(v, a) \mid v \in V \wedge a \in \pi_a(v)\} \cup \{(v, \tau) \mid v \in V \wedge \pi_o(v) = ?\}$
- $F = \{((v, a), (v, w)) \mid (v, a) \in T \wedge (v, w) \in E\} \cup$
 $\{((v, w), (w, a)) \mid (v, w) \in E \wedge (w, a) \in T\} \cup$
 $\{((\text{START}, v), (v, a)) \mid (v, a) \in T \wedge (\text{START}, v) \in P\} \cup$
 $\{((v, a), (v, \text{END})) \mid (v, a) \in T \wedge (v, \text{END}) \in P\}$
- $l = \{((v, a), a) \mid (v, a) \in T \wedge a \neq \tau\}$
- $M_{init} = \{(\text{START}, v) \in P\}$
- $M_{final} = \{(v, \text{END}) \in P\}.$

In Figure 5 we can see the behavior net corresponding to the uncertain trace in Table 3. It is important to notice that every set of edges in the behavior graph with the same source vertex generates an AND split in the behavior net, and a set of edges with the same destination vertex generates an AND join. At the same time, the transitions whose labels correspond to different possible activities in an uncertain event will appear in an XOR construct inside the behavior net.

Thus, every set of events which timestamps allow for overlap will be represented in the behavior net by transitions inside an AND construct, and will then allow to execute in the net all the possible sequences of events obtained choosing a possible value for the uncertain timestamp attribute. In the same fashion, an

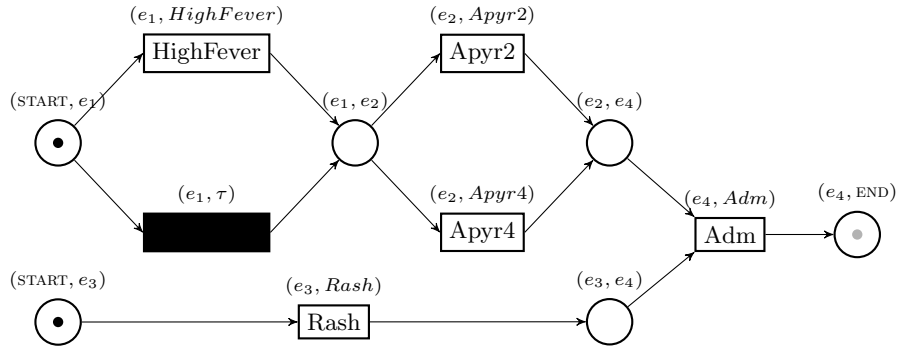


Fig. 5. The behavior net corresponding to the uncertain trace in Table 3. The labels show the objects involved in the construction of Definition 34. The initial marking is displayed; the gray “token slot” represents the final marking.

event with uncertainty on the activity will be represented by a number of transitions in an XOR construct, that allows to replay any possible choice for the activity attribute. It follows that, by construction, for a certain simple uncertain trace σ_U we have that $\phi(bn(\sigma_U)) = \mathcal{R}(\sigma_U)$.

We can use the behavior net of an uncertain trace σ_U in lieu of the event net to compute alignments with a model $SN \in \mathcal{U}_{SN}$; the search algorithm returns an optimal alignment, a sequence of moves $(x, (y, t))$ with $x \in \mathcal{U}_A$, $y \in \mathcal{U}_A$ and t transition of the model SN . After removing all “ \gg ” symbols, the sequence of first elements of the moves will describe a complete firing sequence σ_{bn} of the behavior net. Since σ_{bn} is complete, $\sigma_{bn} \in \phi(bn(\sigma_U))$ and, thus, $\sigma_{bn} \in \mathcal{R}(\sigma_U)$. It follows that σ_{bn} is a realization of σ_U , and the search algorithm ensures that σ_{bn} is a realization with optimal conformance cost for the model SN : $\delta(\lambda_{SN}(\sigma_{bn})) = \min_{\sigma \in \mathcal{R}(\sigma_U)} \lambda_{SN}(\sigma) = \delta_{min}(\sigma_U)$.

Theorem 3 (Correctness of behavior nets). *Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace and let $bg(\sigma_U) = (V, E)$ be its behavior graph. The corresponding behavior net $bn(\sigma_U) = (P, T, F, l, M_{init}, M_{final})$ can replay all and only the realizations of σ_U : $\phi(bn(\sigma_U)) = \mathcal{R}(\sigma_U)$.*

Proof. Let $(v, v') \in E$ be an edge of the behavior graph, which also defines a place in the behavior net: $(v, v') = p_{v, v'} \in P$. Let us denote with \mathbb{T}_v the set of transitions in the behavior net generated from the vertex v : $\mathbb{T}_v = \{(v', a) \in T \mid v' = v\}$.

(\subseteq) Let $\sigma = \langle a_1, a_2, \dots, a_n \rangle \in \phi(bn(\sigma_U))$ be any certain trace accepted by $bn(\sigma_U)$. Let $\sigma_T = \langle t_1, t_2, \dots, t_n \rangle \in \phi_f(bn(\sigma_U))$ be a complete firing sequence of $bn(\sigma_U)$ such that $l(\sigma_T)|_{\mathcal{U}_A} = \sigma$ and $\sigma = \overline{\mathcal{R}}(l(\sigma_T))$. Let $\langle v_1, v_2, \dots, v_n \rangle$ be a sequence of vertices in $bg(\sigma_U)$ such that $t_1 = (v_1, a_1), t_2 = (v_2, a_2), \dots, t_n = (v_n, a_n)$ and $t_1 \in \mathbb{T}_{v_1}, t_2 \in \mathbb{T}_{v_2}, \dots, t_n \in \mathbb{T}_{v_n}$. Let \mathcal{V} be the set of all such sequences; by the flow relation in Definition 34 there must exist a sequence

$\sigma_O = \langle v_1, v_2, \dots, v_n \rangle \in \mathcal{V}$ such that $((v_1, a_1), (v_1, v_2)) \in F, ((v_1, v_2), (v_2, a_2)) \in F, ((v_2, a_2), (v_2, v_3)) \in F, ((v_2, v_3), (v_3, a_3)) \in F, \dots, ((v_{n-1}, a_{n-1}), (v_{n-1}, v_n)) \in F, ((v_{n-1}, v_n), (v_n, a_n)) \in F$. This implies that $(v_1, v_2) \in E, (v_2, v_3) \in E, \dots, (v_{n-1}, v_n) \in E$. From Definition 33 we then have that $v_1 \not\prec v_2 \not\prec \dots \not\prec v_n$. Furthermore, since there exist a \mathbb{T}_v for all $v \in V$ and for all $1 \leq i \leq n$ exactly one transition $t_i \in \mathbb{T}_{v_i}$ has to fire to complete the firing sequence, we have that for all $v \in V$, $v \in \sigma_O$ and is unique. Thus, $\sigma_O \in \mathcal{S}_V$. Because all vertices in σ_O are sorted by a linear extension of (\prec, \mathcal{E}_V^S) , we also have that $\sigma_O \in \mathcal{O}_{bg(\sigma_U)}$. By Definition 33, we then have that σ_O is an order-realization of σ_U : $\sigma_O \in \mathcal{R}_O(\sigma_U)$. Since, by construction, $l(t_i) \in \pi_a(v_i)$ if $\pi_o(v_i) = !$ and $l(t_i) \in \pi_a(v_i) \cup \{\tau\}$ if $\pi_o(v_i) = ?$, we have that $l(\sigma_T) \in \overline{\mathcal{R}}_A(\sigma_O)$ and thus $\sigma \in \mathcal{R}(\sigma_U)$. Since this construction is valid for any $\sigma \in \phi(bn(\sigma_U))$, every complete firing sequence of the behavior net is a realization of σ_U : $\phi(bn(\sigma_U)) \subseteq \mathcal{R}(\sigma_U)$.

(\supseteq) Let $\sigma_O \in \mathcal{R}_O(\sigma_U)$ be any order-realization of σ_U , and let $n = |\sigma_U|$. Since $\sigma_O[1] \prec \sigma_O[2] \prec \dots \prec \sigma_O[n]$ (by Definition 31), there exists a path $p \in P_{bg(\sigma_U)}$ such that $p = \langle v_1, v_2, \dots, v_n \rangle = \langle \sigma_O[1], \sigma_O[2], \dots, \sigma_O[n] \rangle$ (by Theorem 2). Let $p_{1,2} = (v_1, v_2)$, $p_{2,3} = (v_2, v_3)$, and so on. Let $t_1 \in \mathbb{T}_{v_1}, t_2 \in \mathbb{T}_{v_2}, \dots, t_n \in \mathbb{T}_{v_n}$ and let $\sigma_T = \langle t_1, t_2, \dots, t_n \rangle$. By the construction in Definition 34, in $bn(\sigma_U) = N$ we have that

$$(N, M_{init})[t_1](N, M_{1,2})[t_2](N, M_{2,3})[t_3], \dots, [t_{n-1}](N, M_{n-1,n})[t_n](N, M_{final})$$

where:

$$\begin{aligned} M_{1,2} &= (M_{start} \setminus \{(start, v_1)\}) \uplus \{p_{1,2}\} \\ M_{2,3} &= (M_{1,2} \setminus \{p_{1,2}\}) \uplus \{p_{2,3}\} \\ &\dots \\ M_{n-1,n} &= (M_{n-2,n-1} \setminus \{p_{n-2,n-1}\}) \uplus \{p_{n-1,n}\} \\ M_{final} &= (M_{n-1,n} \setminus \{p_{n-1,n}\}) \uplus \{(v_n, end)\} \end{aligned}$$

This construction implies that $(N, M_{init})[\sigma_T] \triangleright (N, M_{final})$ and therefore $\sigma_T \in \phi_f(bn(\sigma_U))$.

The definition of the labeling function in the behavior net is such that, for all $1 \leq i \leq n$, we have that $(v_i, a) \in \mathbb{T}_{v_i} \Leftrightarrow a \in \pi_a(v_i)$. By Definition 31, the labeling of the sequence $\langle t_1, t_2, \dots, t_n \rangle$ is then an activity-realization of σ_O : $\sigma_A = l(\sigma_T) \in \overline{\mathcal{R}}_A(\sigma_O)$. Therefore, the projection on the universe of activities of the firing sequence of the net is a valid realization of σ_A : $l(\sigma_T)|_{\mathcal{U}_A} = \overline{\mathcal{R}}(\sigma_A)$. Since this construction is valid for any $\sigma_O \in \mathcal{R}_O(\sigma_U)$, the behavior net can replay any realization of σ_U : $\mathcal{R}(\sigma_U) \subseteq \phi(bn(\sigma_U))$. \square

Theorem 4 (Correctness of uncertain alignments). *Let $\sigma_U \in \mathcal{T}_U$ be a simple uncertain trace and let $SN \in \mathcal{U}_{SN}$ be a system net. Computing an alignment using the product net between SN and the behavior net $bn(\sigma_U)$ yields the alignment with the lowest cost among all realizations of σ_U : $\delta(\lambda_{SN}(\sigma_{bn})) = \min_{\sigma \in \mathcal{R}(\sigma_U)} \lambda_{SN}(\sigma) = \delta_{min}(\sigma_U)$.*

Proof. Recall from Definition 22 that $\lambda_{SN} \in \mathcal{T} \rightarrow A_{LM}^*$ is a deterministic mapping that assigns any trace σ to an optimal alignment. Adriansyah [5] details how to compute such a function λ_{SN} through a state-based \mathbb{A}^* search over a state space defined by the reachable markings of the product net $SN \otimes en(\sigma)$ between a reference system net SN and the event net a certain trace $\sigma \in \mathcal{T}$. As per Definition 20, this search retrieves an alignment which is optimal with respect to a certain cost function δ and, ignoring “ \gg ”, is composed by a complete firing sequence of the system net $\sigma_T \in \phi_f(SN)$ and the only complete firing sequence of the event net $en(\sigma)$, which corresponds to σ by construction. Given a system net $SN \in \mathcal{U}_{SN}$, an uncertain trace $\sigma_U \in \mathcal{T}_U$ and its respective behavior net $bn(\sigma_U)$, the same search algorithm for λ_{SN} over $SN \otimes bn(\sigma_U)$ yields an optimal alignment containing a complete firing sequence for the reference system net $\sigma_T \in \phi_f(SN)$ and a complete firing sequence for the behavior net of the uncertain trace $\sigma \in \phi(bn(\sigma_U))$. Since λ_{SN} minimizes the cost and $\sigma \in \mathcal{R}(\sigma_U)$ is a valid realization of σ due to Theorem 3, the resulting alignment has the minimal cost possible over all the possible realizations of the uncertain trace. \square

6 Experiments

The framework here illustrated for computing conformance bounds for uncertain event data rises some research questions that need to be addressed in a practical and empirical manner.

- *Q1*: how do conformance bounds behave, when computed on uncertain data?
- *Q2*: what is the impact of different deviating behavior and different type of uncertain behavior on the conformance score of uncertain event logs?
- *Q3*: what is the impact on efficiency of computing uncertain alignments utilizing the behavior net as opposed to the baseline method of enumerating and aligning all realizations?
- *Q4*: how does computing uncertain alignments utilizing the behavior net impact different types of uncertain behavior?
- *Q5*: are uncertain alignments applicable to real-life data to obtain a best case and worst case scenario for the execution of process instances?

The technique to compute conformance for strongly uncertain traces and to create the behavior net hereby described has been implemented in the Python programming language, thanks to the facilities for log importing, model creation and manipulation, and alignments provided by the library PM4Py [10]. Uncertainty has been represented in the XES standard through meta-attributes and constructs such as lists, such that any XES importer can read an uncertain log file. The algorithm was designed to be fully compatible with any event log in the XES format (both including and non including uncertainty); the meta-attributes for uncertainty were designed to be backward compatible with other process mining algorithms – meta-attributes describing the possible values for an uncertain activity or the interval of an uncertain timestamp can also specify a “fallback value” that other process mining software will read as (certain) activity or timestamp value.

6.1 Qualitative and Quantitative Experiments on Synthetic Data

The first four research questions listed above have been addressed by tests on synthetic uncertain event logs. To this end, we implemented the following software components necessary to the experiments:

- a *noise generator*, to introduce deviations in a controlled way in an event log. This component allows to alter the activity label, swap the order of events or add redundant events to an event log with a given probability or frequency.
- an *uncertainty generator*, to alter the XES attributes present in the log by appending additional meta-information which is then interpreted as uncertainty. The component introduces uncertainty information in an event log, with the possibility to add any of the strongly uncertain attributes described in the taxonomy of Section 2. This also allows for exporting the generated uncertain event log through the XES exporter of the PM4Py library.
- a number of smaller extensions to PM4Py functionalities, also useful for other process mining applications. Examples are the generation of all possible process variants (language) of a PM4Py Petri net, and a memoized version of alignments, which allows to trade off space in memory in order to speed up the computation of the conformance of an event log and a model.

In order to answer to *Q1* and *Q2*, we set up an experiment with the goal to inspect the bounds for conformance score as increasingly more uncertainty is added to an event log. We ran the tests on synthetic event logs where we added simulated uncertainty. In this way we can control the amounts and types of uncertainty in event data.

Every iteration of this experiment is as follows:

1. We generate a random Petri net with a fixed dimension ($n = 10$ transitions) through the ProM plugin “*Generate block-structured stochastic Petri nets*”.
2. We play out an event log of 100 traces from the Petri net.
3. We randomly alter the activity label of a specific percentage d_a of events.
4. We randomly swap a specific percentage d_s of events with their successor.
5. We randomly duplicate a specific percentage d_d of events.
6. we randomly introduce uncertainty on activity label in a specific percentage u_a of events.
7. we randomly introduce uncertainty on timestamps in a specific percentage u_t of events.
8. we randomly transform a specific percentage u_i of events in indeterminate events.
9. We measure upper and lower bound for conformance score with increasing percentage p of uncertainty.

In terms of amount of deviation to be considered in each configuration, we aimed at recreating a situation where there is significant deviating behavior with respect to the normative model; for each kind of deviation considered, we introduced anomalous behavior in 30% of events. Thus, we consider four different settings for the addition of deviating behavior to events logs: *Activity labels* =

$\{d_a = 30\%, d_s = 0\%, d_d = 0\%\}$, $Swaps = \{d_a = 0\%, d_s = 30\%, d_d = 0\%\}$, $Extra$
 $events = \{d_a = 0\%, d_s = 0\%, d_d = 30\%\}$ and $All = \{d_a = 30\%, d_s = 30\%, d_d =$
 $30\%\}$.

We consider four different settings for the addition of uncertain behavior to events logs: *Activities* = $\{u_a = p, u_t = 0\%, u_i = 0\%\}$, *Timestamps* = $\{u_a = 0\%, u_t = p, u_i = 0\%\}$, *Indeterminate events* = $\{u_a = 0\%, u_t = 0\%, u_i = p\}$ and *All* = $\{u_a = p, u_t = p, u_i = p\}$. We test all four different configurations of deviation against each of the four configurations of uncertainty, with increasing values of p , for a total of 16 separate experiments.

Figure 6 summarizes our findings. The scatter plots on this figure represent the average of 10 runs as described above.

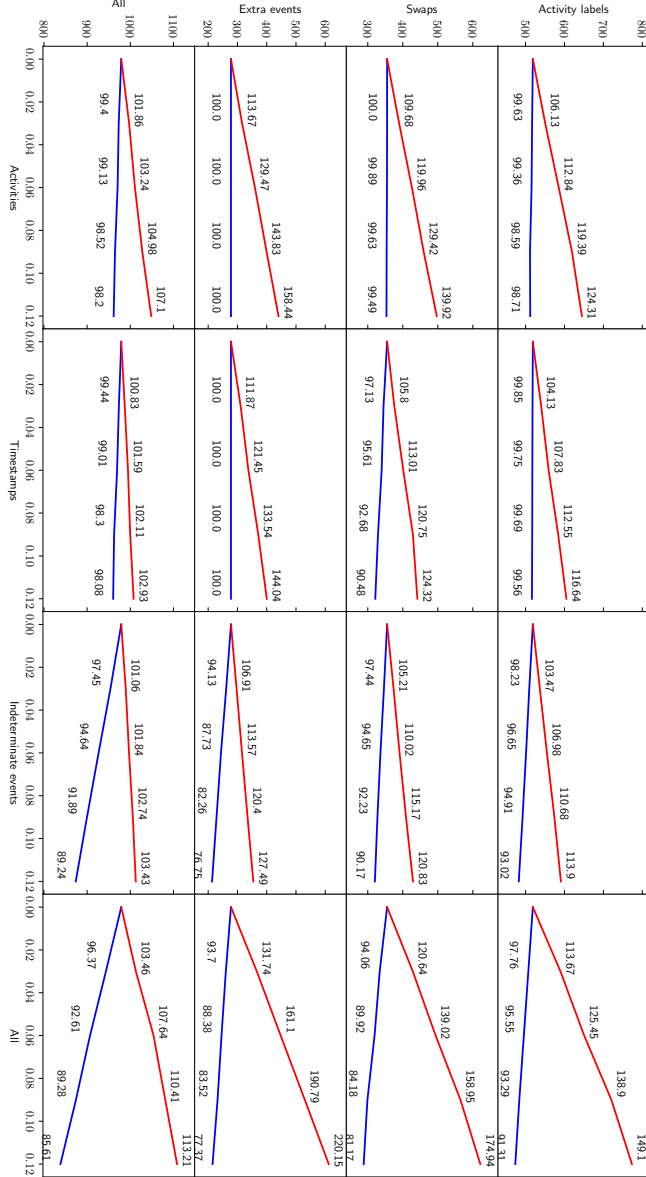


Fig. 6. Upper (red) and lower (blue) bound for conformance cost for synthetic event logs with increasing uncertainty. Every plot shows a different configuration of deviation added to the log and types of uncertainty simulated in the event data. The x-axis shows the percentage of uncertainty p added to the logs; the y-axis shows the amount of deviations, computed with alignments. The labels inside the graph indicate the relative change in deviation score with respect to $p = 0$, in percentage.

We can observe that, in general, all plots show the expected behavior: the upper and lower bound for conformance coincide at percentage of uncertain events $p = 0$ for all experiments, to then diverge while p increases. Another general observation is the fact that, when adding only one type of deviation and only the “matching” type of uncertainty, the lower bound decreases faster than the other configurations. A number of additional observations can be made looking at individual configurations for deviation or uncertainty, or at specific scatter plots.

When only uncertainty on activity labels is added to the event log, we see a deterioration of the upper bound for conformance cost, but the lower bound does not improve – in fact, it is essentially constant. This can be attributed to the fact that, since to generate uncertainty on activity label we sample from the set of labels randomly, the chances of observing a realization of a trace where an uncertain activity label matches the alteration introduced by the deviations are small. Uncertainty on timestamps makes the lower bound decrease only when the introduced deviations are swaps: as expected, the possibility of changing the order of pairs of events does not have a sensible improvement in the lower bound for deviation when extra events are added or activity labels of existing events are altered. Conversely, the possibility to “skip” some critical events has a positive effect on the lower bound of all possible configurations for deviations: in fact, when marking some events as indeterminate in a log where extra events were added as deviations, the average conformance cost drops by 23.25% at $p = 12\%$, the largest drop on all the experiments. The experiment with all three types of uncertainty and extra events as deviations essentially displays the same effect (improvement in lower bound is slightly lower, but not significantly, with a decrease in deviation of 22.63% at $p = 12\%$). On the experiments where all types of deviations were added at once we can see that, as could be anticipated, the differences in deviation scores on the two bounds become smaller in relative terms (because of the very high amount of deviations $p = 0\%$), but larger in absolute terms. As per the previous experiments, the largest contributor in decreasing the conformance cost of the lower bound is the addition of indeterminate events, which by itself decreases the deviation cost by 10.76% at $p = 12\%$. In general, the vast variability in measuring the conformance of an uncertain log indicates that, if all types of uncertainty can occur with high frequency in a process, the business owner should act on the uncertainty sources, since they will be a major obstacle in obtaining accurate measurements of process conformance. Vice versa, in the case of limited occurrences of uncertainty in event data the algorithm here proposed is able to provide actionable bounds for conformance score, together with descriptions of best and worst case scenarios of process conformance for a given trace.

The second experiment we setup answers questions $Q3$ and $Q4$, and concerns the performance of calculating the lower bound of the cost via the behavior net versus the bruteforce method of separately listing all the realizations of an uncertain trace, evaluating all of them through alignments, then picking the best value. We used a constant percentage of uncertain events of $p = 5\%$ and

logs of 100 traces for this test, with progressively increasing values of n . We ran 4 different experiments, each with one of the four configurations for uncertain behavior *Activities*, *Timestamps*, *Indeterminate events* and *All* illustrated above.

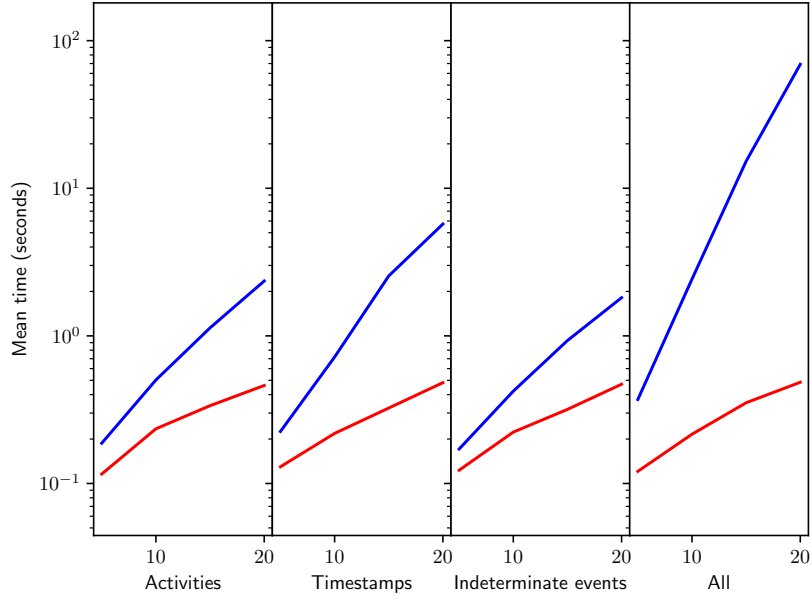


Fig. 7. Effect on time performance of calculating the lower bound for conformance cost with the brute force method vs. the behavior net on four different configurations for uncertain events.

Figure 7 summarizes the results. As the diagram shows, the difference in time between the two methods tends to diverge quickly even on a logarithmic scale. The largest model we could test was $n = 20$, a Petri net with 20 transitions, which is comparatively tiny in practical terms; however, even at these small scales the brute force method takes roughly 3 orders of magnitude more than the time needed by the behavior net, when all the types of uncertainty are added with $p = 5\%$. This shows a very large improvement in the computing time for the lower bound computation; thus, the best case scenario for the conformance cost of an uncertain trace can be obtained efficiently thanks to the structural properties of the behavior net. This graph also shows the dramatic impact on the number of realizations of a behavior net – and thus on the time of brute force computation of alignments – when the effects of different kinds of uncertainty are compounded.

6.2 Applications on Real-Life Data

As illustrated in Section 1, uncertainty in event data can be originated by a number of diverse causes in real-world applications. One prominent source of uncertainty is missing data: attribute values not recorded in an event log can on occasions be described by uncertainty, through domain knowledge provided by process owners or experts. Then, as described in this paper, it is possible to obtain a detailed analysis of the deviations of a best and worst case scenario for the conformance to a process model.

To seek to answer research question Q_4 through a direct application of conformance checking over uncertainty, let us consider a process related to the medical procedures performed in the Intensive Care Units (ICU) of a hospital. Figure 8 shows a ground truth model for the process:

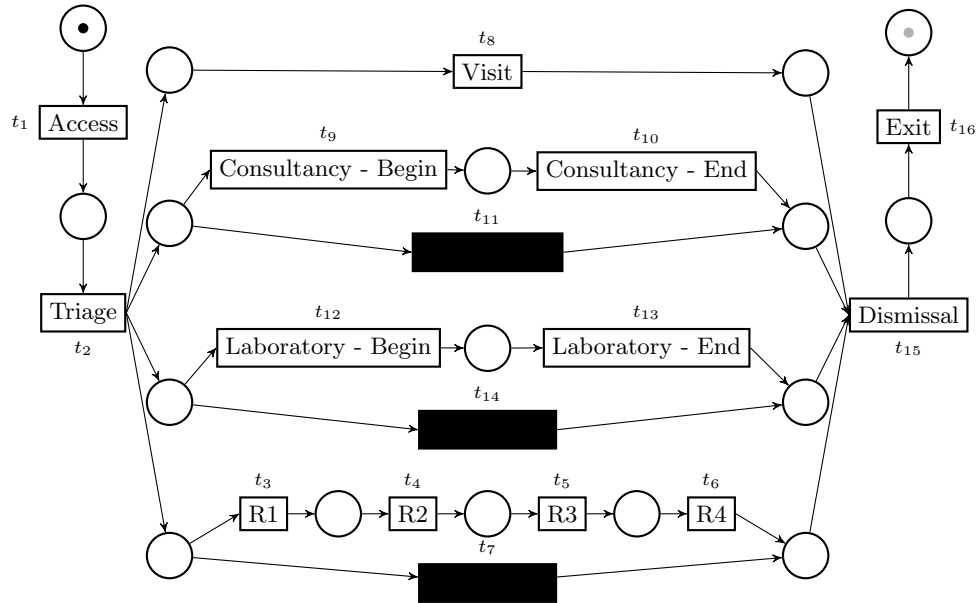


Fig. 8. The Petri net that models the process related to the treatment of patients in the ICU ward of an Italian hospital. The activities R1 through R4 are abbreviations for the four phases of a radiology exam: respectively, *Radiology - Submitted Request*, *Radiology - Accepted Request*, *Radiology - Exam*, *Radiology - Results*.

An execution log containing events that concern this ICU process is available. Throughout the process, some anomalies with attribute values can be spotted – namely, a number of anomalies in regard of the timestamp attributes. Tables 4 and 5 are two examples of traces with anomalous timestamp behavior. We can see that in the trace of Table 4 the event *Triage* has an imprecise timestamp – only

the day has been recorded. This can be modeled with an uncertain timestamp encompassing a range of 24 hours. The column *Preprocessed Timestamp* shows the results of this preprocessing step.

Table 4. Events relative to one case of the ICU process. The timestamp of the “Triage” event is imprecise: through domain knowledge, we are able to represent it with uncertainty.

Event ID	Raw Timestamp	Preprocessed Timestamp	Activity
e_1	20-02-2017 23:59:31	20-02-2017 23:59:31	<i>Access</i>
e_2	21-02-2017 00:02:58	21-02-2017 00:02:58	<i>Visit</i>
e_3	21-02-2017 00:06:30	21-02-2017 00:06:30	<i>Consultancy - Begin</i>
e_4	21-02-2017 00:29:12	21-02-2017 00:29:12	<i>R1</i>
e_5	21-02-2017 00:41:00	21-02-2017 00:41:00	<i>R2</i>
e_6	21-02-2017 00:41:00	21-02-2017 00:41:00	<i>R3</i>
e_7	21-02-2017 01:02:00	21-02-2017 01:02:00	<i>R4</i>
e_8	21-02-2017 01:56:26	21-02-2017 01:56:26	<i>Consultancy - End</i>
e_9	21-02-2017 02:01:37	21-02-2017 02:01:37	<i>Dismissal</i>
e_{10}	21-02-2017 02:02:36	21-02-2017 02:02:36	<i>Exit</i>
e_{11}	21-02-2017	[21-02-2017 00:00:00, 21-02-2017 23:59:59]	<i>Triage</i>

Some of the events in the trace of Table 5 are missing the timestamp value entirely. In this case, we can resort to domain knowledge provided by the process owners: it is known that events related to the *Radiology* exams happen after the *Triage* event, and before the *Dismissal* event. This allows to represent the timestamp with ranges of possible values. Notice that such a small interval of time, obtainable from the domain knowledge available, is preferable to larger possible intervals (e.g., 27-08-2017 00:00:00 to 27-08-2017 23:59:59), since it minimizes the amount of possible overlaps in time with other events in the trace and, in turn, there exists a smaller number of possible realization of the uncertain trace, granting a faster conformance checking. As before, the results of modeling timestamp uncertainty are shown in the column *Preprocessed Timestamp*.

Once represented the data in the event log through the formalization shown in this paper, it is possible to apply conformance checking over uncertainty. The technique of alignments illustrated here provides two results, corresponding to the lower and upper bound for the conformance score. The two traces provided have the same best case scenario alignment, which is shown in Table 6; aligning through the behavior net of these traces has allowed the algorithm to select a value for the uncertain timestamps of the traces (translated in a specific ordering) such that the deviations between data and model is the smallest possible. For both traces, the best case scenario has a cost equal to 0, thus, no deviations occur in that case.

Let us now look at the worst case scenarios. One of the alignments with the worst possible score for the trace in Table 4 is shown in Table 7. In this scenario, the deviations are one move on model (the *Triage* activity should have occurred

Table 5. Events relative to one case of the ICU process. Some of the timestamp attributes are missing: through domain knowledge, we are able to represent them with uncertainty within a small interval of time. In bold (resp., italic), the minimum (resp., maximum) value for uncertain timestamps recovered from the context of the trace.

Event ID	Raw Timestamp	Preprocessed Timestamp	Activity
e_1	27-08-2017 11:47:46	27-08-2017 11:47:46	<i>Access</i>
e_2	27-08-2017 11:47:53	27-08-2017 11:47:53	<i>Triage</i>
e_3	27-08-2017 12:14:25	27-08-2017 12:14:25	<i>Visit</i>
e_4	27-08-2017 12:33:24	27-08-2017 12:33:24	<i>R1</i>
e_5	27-08-2017 13:04:11	27-08-2017 13:04:11	<i>Consultancy - Begin</i>
e_6	27-08-2017 <i>13:04:53</i>	27-08-2017 13:04:53	<i>Dismissal</i>
e_7	27-08-2017 13:08:07	27-08-2017 13:08:07	<i>Exit</i>
e_8	NULL	[27-08-2017 11:47:53 , 27-08-2017 <i>13:04:53</i>]	<i>Consultancy - End</i>
e_9	NULL	[27-08-2017 11:47:53 , 27-08-2017 <i>13:04:53</i>]	<i>R2</i>
e_{10}	NULL	[27-08-2017 11:47:53 , 27-08-2017 <i>13:04:53</i>]	<i>R3</i>
e_{11}	NULL	[27-08-2017 11:47:53 , 27-08-2017 <i>13:04:53</i>]	<i>R4</i>

Table 6. A valid alignment for both traces of Tables 4 and 5. This alignment has a deviation cost equal to 0, and corresponds to the best case scenario for conformance between the process model and both uncertain traces.

Access	Triage	Visit	Consultancy - Begin	R1	R2	R3	R4	Consultancy - End	⋈	Dismissal	Exit
Access	Triage	Visit	Consultancy - Begin	R1	R2	R3	R4	Consultancy - End	τ	Dismissal	Exit
t_1	t_2	t_8	t_9	t_3	t_4	t_5	t_6	t_{10}	t_{14}	t_{15}	t_{16}

after the *Access* but did not), and one move on log (the activity *Triage* occurs in the data at an unexpected moment in the process).

Table 7. A valid alignment for the trace of Table 4. This alignment has a deviation cost equal to 2 (1 move on log and 1 move on model), and corresponds to the worst case scenario for conformance between the process model and the uncertain trace.

Access	⋈	Visit	Consultancy - Begin	R1	R2	R3	R4	Consultancy - End	⋈	Dismissal	Exit	Triage
Access	Triage	Visit	Consultancy - Begin	R1	R2	R3	R4	Consultancy - End	τ	Dismissal	Exit	⋈
t_1	t_2	t_8	t_9	t_3	t_4	t_5	t_6	t_{10}	t_{14}	t_{15}	t_{16}	

The worst case scenario for the trace in Table 5 is illustrated in Table 8. In this case, the deviation is equal to 6, given by the wrong order of the event related to the *Radiology* exam. Note that, in this example, we assume that every deviation has a unit cost, but the alignment technique allows to define different costs for different types of deviations based on impact in the process. For example, a patient that exits the hospital without official dismissal might have a worse impact than an unauthorized laboratory exam. For simplicity, in this case we assume that all types of deviation have a unit cost.

Through the means provided by uncertain alignments, the process owner can utilize the results to gain insights and decide actions in regard of the process.

Table 8. A valid alignment for the trace of Table 5. This alignment has a cost equal to 6 (3 moves on log and 3 moves on model), and corresponds to the worst case scenario for conformance between the process model and the uncertain trace.

Access	Triage	Visit	Consultancy - Begin	>>	>>	>>	R4	R3	R2	R1	Consultancy - End	>>	Dismissal	Exit
Access	Triage	Visit	Consultancy - Begin	R1	R2	R3	R4	τ	τ	τ	Consultancy - End	τ	Dismissal	Exit
t_1	t_2	t_8	t_9	t_3	t_4	t_5	t_6				t_{10}	t_{14}	t_{15}	t_{16}

The potential violation shown in the worst case scenario for the trace in Table 4 can be investigated, as well as the source of said uncertainty; the process owner can, furthermore, decide whether the consequences and the likelihood of the worst case scenario are indicative of a need for a process restructuration, or whether the risk of such potential violation of the normative process model are not critical for the process execution.

7 Related Work

7.1 Conformance Checking

The discipline of conformance checking, a subfield of process mining, concerns itself with defining metrics to compare how well an event log matches a given process model. The input for this task consists in an execution log and a process model (most commonly a labeled Petri net) and the output is a measurement of the distance – that is, the deviation – between the model and the log, or the traces that compose the log. The two main goals of conformance checking are measuring the quality of a process discovery algorithm by comparing the discovered process model with the source event log, to verify the extent to which the model fits the log; and comparing an execution log with a normative process model (often defined partially or completely by hand) in order to verify the deviations between the rules governing the process and the tasks carried out in reality. Often, the conformance metric defined between logs (or traces) and models includes not only a distance in absolute terms, but also an indication of where and what deviated from the norm in the process. Conformance checking was introduced by Rozinat and van der Aalst [24], who obtained a conformance measure by tracking counts of tokens during replay of traces in a Petri net. Currently, state-of-the-art approaches are based on alignments, introduced by Adriansyah et al. [6].

7.2 Event Data Uncertainty

As mentioned, the occurrence of data containing uncertainty – in a broad sense – is common both in more classic disciplines like statistics and Data Mining [17] and in process mining [2]; and logs that show an explicit uncertainty in the control flow perspective can be classified in the lower levels of the quality ranking proposed in the process mining manifesto.

Within process mining there exist various techniques to deal with a kind of uncertainty different, albeit closely related, from the one that we analyze here: missing or incorrect data. This can be considered as a form of non-explicit uncertainty: no measure or indication on the nature of the uncertainty is given in the event log. The work of Suriadi et al. [29] provides a taxonomy of this type of issues in event logs, laying out a series of data patterns that model errors in process data. In these cases, and if this behavior is infrequent enough to allow the event log to remain meaningful, the most common way for existing process mining techniques to deal with missing data is by filtering out the affected traces and performing discovery and conformance checking on the resulting filtered event log. A case study illustrating such situation is e.g. the work of Benevento et al. [9]. While filtering out missing values is straightforward, various methodologies of event log filtering have been proposed in the past to solve the problem of incorrect event attributes: the filtering can take place thanks to a reference model, which can be given as process specification [31], or from information discovered from the frequent and well-formed traces of the same event log; for example extracting an automaton from the frequent traces [14], computing conditional probabilities of frequent sequences of activities [25], or discovering a probabilistic automaton [33]. In the latter cases, the noise is identified as infrequent behavior.

Some previous work attempt to repair the incorrect values in an event log. Conforti et al. [13] propose an approach for the restoration of incorrect timestamps based on a log automaton, that repairs the total ordering of events in a trace based on correct frequent behavior. Fani Sani et al. [26] define outlier behavior as the unexpected occurrence of an event, the absence of an event that is supposed to happen, and the incorrect order of events in the trace; then, they propose a repairing method based on probabilistic analysis of the context of an outlier (events preceding or following the anomalous event). Again, both of these methods define anomalous/incorrect behavior on the basis of the frequency of occurrence.

The definition of uncertainty on activity labels as defined in the taxonomy of Section 2 has not been, to the best of our knowledge, been previously employed in the field of process mining. There are, however, related examples of anomalies or uncertainties on activity labels of events: for instance, the problem of matching event identifiers to normative activity labels [8]. In this case, an event is associated with only one activity label, but this association is not known. There are a number of techniques to estimate the correct association, including some that consider the data perspective, together with the control flow perspective [27]. On this setting, van der Aa et al. [30] proposed a technique to estimate bounds of conformance scores for event logs with unknown or partially known event-to-activity mapping. Another related domain is the many-to-one abstraction from low level events to a higher order of activity labels, which can be performed via clustering events in minimal conflict groups [16] or representing low level patterns with data Petri nets which then discovers high level activities by matching patterns through alignments [21].

A kind of anomaly in event data which is even more related to uncertainty as discussed in this paper is incompleteness in the order of events in a trace. This occurs when total ordering among events is lost or not available, and only a partial order is known. In the field of concurrent and distributed system, the absence of a total order among logged activities has historically been relevant by virtue of being both caused by, and a necessary condition for, the presence of concurrency in a system (refer e.g. to Beschastnikh et al. [11]). An important concept at the base of this paper is the representation of uncertainties in the timestamp dimension through directed acyclic graphs, which express these partial orders. This intuition was first presented by Lu et al. [20], also in the context of conformance checking, in order to produce partially ordered alignments. More recently, van der Aa et al. [1] proposed a technique to resolve such order uncertainty, through estimates based on probabilistic inference aided by a normative process model.

In process mining, a notion well known for a long time is the fact that in many cases the definition of case is not part of the normative information immediately accessible to the process analyst, so there needs to be a decision on which attribute or attributes constitutes the case of the process. In some cases, multiple definitions of cases are possible and analysis on a subset of them is desirable. This specific setting, which can be interpreted as uncertainty on the case notion, has a long history both in terms of mathematical formalization and in terms of implementation and definition of data standards. For a thorough introduction to this subfield of process mining we refer to [4].

This paper presents an extended version of the preliminary analysis on uncertain event data in process mining shown in [22]. We elaborate on this previous work adding an extended formalization, proving theorems on uncertainty in process mining, and reporting on new experiments. The framework for uncertain data proposed in this paper has also been expanded by providing an algorithm capable of process discovery on uncertain event data [23], as well as an improved algorithm that allows to preprocess uncertain traces in quadratic time, enabling fast uncertainty analysis.

8 Conclusion

As the need of quickly and effectively analyze process data has arisen in the recent past and is growing to this day, many new types of information regarding events are recorded; this calls for new techniques able to provide an adequate interpretation of the new data. Not only more and more event data is available to the analyst, but these data are accessible in association with a wealth of information and meta-information about the process, the resources that executed activities, data about the outcome of those actions, and many other types of knowledge about the nature of events, activities, and the process as a whole. In this paper we presented a new paradigm for process mining applied to event data: explicit uncertainty. We described the possible form it can assume, building a taxonomy of different types of uncertainty, and we provided examples of how

uncertainty can originate in a process, and how uncertainty information can be inferred from the available data and from domain knowledge provided by process experts. We then designed a formal mathematical infrastructure to define the various flavors of uncertainty shown in the taxonomy. Then, in order to assess the practical applications of the uncertainty framework, we applied it to a well consolidated technique for conformance checking: aligning data to a reference Petri net. This application of uncertainty analysis is integrated by theorems that prove the correctness of the techniques developed and illustrated here within the framework previously described. The results can provide insights on the possible violations of process instances recorded with uncertainty against a normative model. The behavior net provides an efficient way to compute the lower bound for the conformance cost – i.e. the best case scenario for conformity of uncertain process data – with a large improvement on time performance with respect to a brute-force procedure.

The approaches shown here can be extended in a number of ways. From a performance perspective, to improve the usability of alignments over uncertainty the computation of the upper bound of the conformance cost should either be optimized, or replaced by an approximate algorithm. Another direction for future work is extending the conformance checking technique to logs with weak uncertainty, weighting the deviation by means of the probability distributions attached to activities, timestamps and indetermineness. Additionally, investigation on real-life data is an important milestone for this line of research, and it is vital to analyze in depth a complete use case in real life of process mining in presence of uncertain event data.

References

1. van der Aa, H., Leopold, H., Weidlich, M.: Partial order resolution of event logs for process conformance checking. *Decision Support Systems* p. 113347 (2020)
2. Van der Aalst, W., Adriansyah, A., De Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., Van Den Brand, P., Brandtjen, R., Buijs, J., et al.: Process mining manifesto. In: *International Conference on Business Process Management*. pp. 169–194. Springer (2011)
3. Van der Aalst, W.M.: Decomposing petri nets for process mining: A generic approach. *Distributed and Parallel Databases* **31**(4), 471–507 (2013)
4. van der Aalst, W.M.: Object-centric process mining: Dealing with divergence and convergence in event data. In: *International Conference on Software Engineering and Formal Methods*. pp. 3–25. Springer (2019)
5. Adriansyah, A.: Aligning observed and modeled behavior (2014)
6. Adriansyah, A., van Dongen, B.F., van der Aalst, W.M.: Towards robust conformance checking. In: *International Conference on Business Process Management*. pp. 122–133. Springer (2010)
7. Aho, A.V., Garey, M.R., Ullman, J.D.: The transitive reduction of a directed graph. *SIAM Journal on Computing* **1**(2), 131–137 (1972)
8. Baier, T., Mendling, J.: Bridging abstraction layers in process mining by automated matching of events and activities. In: *Business process management*, pp. 17–32. Springer (2013)

9. Benevento, E., Dixit, P.M., Sani, M.F., Aloini, D., van der Aalst, W.M.: Evaluating the effectiveness of interactive process discovery in healthcare: A case study. In: International Conference on Business Process Management. pp. 508–519. Springer (2019)
10. Berti, A., van Zelst, S.J., van der Aalst, W.M.P.: Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science. In: ICPM Demo Track (CEUR 2374). p. 1316 (2019)
11. Beschastnikh, I., Brun, Y., Ernst, M.D., Krishnamurthy, A., Anderson, T.E.: Mining temporal invariants from partially ordered logs. In: Managing Large-scale Systems via the Analysis of System Logs and the Application of Machine Learning Techniques, pp. 1–10 (2011)
12. Carmona, J., van Dongen, B., Solti, A., Weidlich, M.: Conformance Checking: Relating Processes and Models. Springer (2018)
13. Conforti, R., La Rosa, M., ter Hofstede, A.: Timestamp repair for business process event logs (2018), <http://hdl.handle.net/11343/209011>, [preprint]
14. Conforti, R., La Rosa, M., ter Hofstede, A.H.: Filtering out infrequent behavior from business process event logs. *IEEE Transactions on Knowledge and Data Engineering* **29**(2), 300–314 (2017)
15. Flaška, V., Ježek, J., Kepka, T., Kortelainen, J.: Transitive closures of binary relations. i. *Acta Universitatis Carolinae. Mathematica et Physica* **48**(1), 55–69 (2007)
16. Günther, C.W., van der Aalst, W.M.: Mining activity clusters from low-level event logs. Beta, Research School for Operations Management and Logistics (2006)
17. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
18. Kalvin, A.D., Varol, Y.L.: On the generation of all topological sortings. *Journal of Algorithms* **4**(2), 150–162 (1983)
19. Liu, X., Smolka, S.A.: Simple linear-time algorithms for minimal fixed points. In: International Colloquium on Automata, Languages, and Programming. pp. 53–66. Springer (1998)
20. Lu, X., Fahland, D., van der Aalst, W.M.: Conformance checking based on partially ordered event data. In: International conference on business process management. pp. 75–88. Springer (2014)
21. Mannhardt, F., De Leoni, M., Reijers, H.A., Van Der Aalst, W.M., Toussaint, P.J.: From low-level events to activities-a pattern-based approach. In: International conference on business process management. pp. 125–141. Springer (2016)
22. Pegoraro, M., van der Aalst, W.M.: Mining uncertain event data in process mining. In: 2019 International Conference on Process Mining (ICPM). pp. 89–96. IEEE (2019)
23. Pegoraro, M., Uysal, M.S., van der Aalst, W.M.: Discovering process models from uncertain event data. In: International Conference on Business Process Management. pp. 238–249. Springer (2019)
24. Rozinat, A., Van der Aalst, W.M.: Conformance checking of processes based on monitoring real behavior. *Information Systems* **33**(1), 64–95 (2008)
25. Sani, M.F., van Zelst, S.J., van der Aalst, W.M.: Improving process discovery results by filtering outliers using conditional behavioural probabilities. In: International Conference on Business Process Management. pp. 216–229. Springer (2017)
26. Sani, M.F., van Zelst, S.J., van der Aalst, W.M.: Repairing outlier behaviour in event logs. In: International Conference on Business Information Systems. pp. 115–131. Springer (2018)

27. Senderovich, A., Rogge-Solti, A., Gal, A., Mendling, J., Mandelbaum, A.: The road from sensor data to process instances via interaction mining. In: International Conference on Advanced Information Systems Engineering. pp. 257–273. Springer (2016)
28. Spivak, D.I.: Category theory for the sciences. MIT Press (2014)
29. Suriadi, S., Andrews, R., ter Hofstede, A.H., Wynn, M.T.: Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems* **64**, 132–150 (2017)
30. Van Der Aa, H., Leopold, H., Reijers, H.A.: Efficient process conformance checking on the basis of uncertain event-to-activity mappings. *IEEE Transactions on Knowledge and Data Engineering* **32**(5), 927–940 (2019)
31. Wang, J., Song, S., Lin, X., Zhu, X., Pei, J.: Cleaning structured event logs: A graph repair approach. In: Data Engineering (ICDE), 2015 IEEE 31st International Conference on. pp. 30–41. IEEE (2015)
32. Winskel, G.: Petri nets, algebras, morphisms, and compositionality. *Information and Computation* **72**(3), 197–238 (1987)
33. van Zelst, S.J., Sani, M.F., Ostovar, A., Conforti, R., La Rosa, M.: Filtering spurious events from event streams of business processes. In: International Conference on Advanced Information Systems Engineering. pp. 35–52. Springer (2018)