Probability Estimation of Uncertain Process Trace Realizations

Marco Pegoraro, Bianka Bakullari, Merih Seran Uysal, Wil M.P. van der Aalst, RWTH Aachen University

International Workshop on Event Data and Behavioral Analytics, 1 November 2021

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Uncertain Data

- Uncertain event data: **events with quantified imprecision in their attributes**

- "Quantified" means we can obtain a **description** of the uncertain attribute value(s)
  - For categorical attributes: **a set of possible values**
  - For numerical attributes: **an interval of possible values**

- We can also have events that have been recorded, but might not have occurred (**indeterminate events**)

- Often obtained through **pre-processing** and **domain knowledge**

Chair of Process and Data Science | RWTH AACHEN UNIVERSITY

Flight passenger #3167 landed in Munich from a high risk area.

Case ID	Event ID	Activity	Timestamp	Notes
3167	e_1	passenger check in	$8:13:43$	
3167	e_2	test	$11:00$	
3167	e_3	communicate test result	$11:00$	reception not signed
3167	e_4	remove health data	$22:00:00$	

Two types of test are possible: **after landing** and **before departure**

Uncertainty in Event Logs

Case ID	Event ID	Activity	Timestamp	Event type
3167	e_1	passenger check in	$8:13:43$	!
3167	e_2	{test after landing : 0.9, test before departure: 0.1}	$[11:00:00, 11:59:59]$	!
3167	e_3	communicate test result	$[11:00:00, 11:59:59]$	!: 0.2 ?: 0.8
3167	e_4	remove health data	$22:00:00$	!

through domain knowledge or heuristics, we determine that the "test after landing" has 90% of probability of being the label of e_2

Chair of Process and Data Science

Uncertainty in Event Logs

Case ID	Event ID	Activity	Timestamp	Event type
3167	e_1	passenger check in	$8:13:43$	!
3167	e_2	{test after landing : 0.9, test before departure: 0.1}	$[11:00:00, 11:59:59]$	!
3167	e_3	communicate test result	$[11:00:00, 11:59:59]$	!: 0.2 ? : 0.8
3167	e_4	remove health data	$22:00:00$	!

we represent the timestamps of e_2 and e_3 as time intervals

Uncertainty in Event Logs

Case ID	Event ID	Activity	Timestamp	Event type
3167	e_1	passenger check in	$8:13:43$	$!$
3167	e_2	{test after landing : 0.9, test before departure: 0.1}	$[11:00:00, 11:59:59]$	$!$
3167	e_3	communicate test result	$[11:00:00, 11:59:59]$	$!: 0.2 \;\; ?: 0.8$
3167	e_4	remove health data	$22:00:00$	$!$

the “!“ symbol indicates the event actually occurred, the “?“ means the event did not occur but has been recorded

we determine the event did not occur with 80% probability

Uncertain Trace Realizations

- This trace corresponds to many possible **real-life scenarios** depending on the true value of its **uncertain attributes**

- Every sequence of activities possible in the uncertain trace is called a **realization**

Uncertain Trace Realizations

Case ID	Event ID	Activity	Timestamp	Event type
3167	e_1	passenger check in	$8:13:43$	$!$
3167	e_2	{test after landing : 0.9, test before departure: 0.1}	$[11:00:00, 11:59:59]$	$!$
3167	e_3	communicate test result	$[11:00:00, 11:59:59]$	$!: 0.2 \ \ ?: 0.8$
3167	e_4	remove health data	$22:00:00$	$!$

<passenger check in, test after landing, communicate test results, remove health data>

<passenger check in, test before departure, remove health data>

<passenger check in, communicate results, test before departure, remove health data>

<passenger check in, test after landing, remove health data>

…

Uncertain Trace Realizations

- Research question: **what is the probability of occurrence of each realization?**
- This information is essential in the context of process mining on uncertain event data
- We will see how to determine such probabilities
- We will see their importance in an example of application: **conformance checking**

Uncertain Trace Realizations

Case ID	Event ID	Activity	Timestamp	Event type
3167	e_1	a passenger check in	$8:13:43$	!
3167	e_2	b c {test after landing : 0.9, test before departure: 0.1}	$[11:00:00, 11:59:59]$	!
3167	e_3	d communicate test result	$[11:00:00, 11:59:59]$	!: 0.2 ?: 0.8
3167	e_4	e remove health data	$22:00:00$	!

Event seq. S_e	Realization S_a
$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$
	$\langle a, c, d, e \rangle$
$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$
	$\langle a, d, c, e \rangle$
$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$
	$\langle a, c, e \rangle$

Probability of Uncertain Trace Realizations

Event seq. S_e	Realization S_a
$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$
	$\langle a, c, d, e \rangle$
$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$
	$\langle a, d, c, e \rangle$
$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$
	$\langle a, c, e \rangle$

Probability of observing the realization s_a given we observed the event sequence s_e

$$P(s_a) = \sum_{s_e \in S_e} P(s_e) \cdot P(s_a \mid s_e)$$

probability of observing the event sequence s_e

Probability of Uncertain Trace Realizations

We assume independence!

Event seq. S_e	Realization S_a
$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$
	$\langle a, c, d, e \rangle$
$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$
	$\langle a, d, c, e \rangle$
$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$
	$\langle a, c, e \rangle$

Probability of observing the realization s_a given we observed the event sequence s_e

$$P(s_a) = \sum_{s_e \in S_e} P(s_e) \cdot P(s_a \mid s_e)$$

probability of observing the event sequence s_e

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Probability of Uncertain Trace Realizations

Event set		Event seq. s_e	Realization s_a	$P(s_e)$	$P(s_a \mid s_e)$	$P(s_a)$
$\{e_1, e_2, e_3, e_4\}$	S_1	$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$			
			$\langle a, c, d, e \rangle$			
		$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$			
			$\langle a, d, c, e \rangle$			
$\{e_1, e_2, e_4\}$	S_2	$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$			
			$\langle a, c, e \rangle$			

$$P(s_e) = \frac{1}{|S_i|} \cdot \prod_{e \in s_e} P(e \text{ is } !) \cdot \prod_{e \notin s_e} P(e \text{ is } ?)$$

Event ID	Activity	Timestamp	Event type
e_1	a	$8{:}13{:}43$	!
e_2	{b: 0.9, c: 0.1}	$[11{:}00{:}00, 11{:}59{:}59]$	!
e_3	d	$[11{:}00{:}00, 11{:}59{:}59]$	!: 0.2 ?: 0.8
e_4	e	$22{:}00{:}00$	!

Probability of Uncertain Trace Realizations

Event set	Event seq. s_e	Realization s_a	$P(s_e)$	$P(s_a \mid s_e)$	$P(s_a)$
$\{e_1, e_2, e_3, e_4\}$ **S₁**	$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$	0.1		
		$\langle a, c, d, e \rangle$			
	$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$			
		$\langle a, d, c, e \rangle$			
$\{e_1, e_2, e_4\}$ **S₂**	$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$			
		$\langle a, c, e \rangle$			

$$P(\langle e_1, e_2, e_3, e_4 \rangle) = \frac{1}{|S_1|} \cdot p(e_3 \; is \; !) = \frac{1}{2} \cdot 0.2 = 0.1$$

Event ID	Activity	Timestamp	Event type
e_1	a	$8:13:43$	!
e_2	{b: 0.9, c: 0.1}	$[11:00:00, 11:59:59]$	!
e_3	d	$[11:00:00, 11:59:59]$	!: 0.2 ?: 0.8
e_4	e	$22:00:00$	!

Probability of Uncertain Trace Realizations

Event set		Event seq. s_e	Realization s_a	$P(s_e)$	$P(s_a \mid s_e)$	$P(s_a)$
$\{e_1, e_2, e_3, e_4\}$	S_1	$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$	0.1		
			$\langle a, c, d, e \rangle$			
		$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$	0.1		
			$\langle a, d, c, e \rangle$			
$\{e_1, e_2, e_4\}$	S_2	$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$	0.8		
			$\langle a, c, e \rangle$			

$$P(\langle e_1, e_2, e_4 \rangle) = \frac{1}{|S_2|} \cdot p(e_3\ is\ ?) = \frac{1}{1} \cdot 0.8 = 0.8$$

Event ID	Activity	Timestamp	Event type
e_1	a	$8:13:43$	$!$
e_2	{b: 0.9, c: 0.1}	$[11:00:00, 11:59:59]$	$!$
e_3	d	$[11:00:00, 11:59:59]$	$!: 0.2\ \ ?: 0.8$
e_4	e	$22:00:00$	$!$

Probability of Uncertain Trace Realizations

Event set		Event seq. s_e	Realization s_a	$P(s_e)$	$P(s_a \vert s_e)$	$P(s_a)$
$\{e_1, e_2, e_3, e_4\}$	S$_1$	$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$	0.1	0.9	
			$\langle a, c, d, e \rangle$		0.1	
		$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$	0.1	0.9	
			$\langle a, d, c, e \rangle$		0.1	
$\{e_1, e_2, e_4\}$	S$_2$	$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$	0.8	0.9	
			$\langle a, c, e \rangle$		0.1	

$$P(\, s_a \mid s_e \,) = \prod_{i=1}^{|s_e|} P(e_i \text{ executes } a_i)$$

Event ID	Activity	Timestamp	Event type
e_1	a	$8{:}13{:}43$	!
e_2	{b: 0.9, c: 0.1}	$[11{:}00{:}00, 11{:}59{:}59]$	!
e_3	d	$[11{:}00{:}00, 11{:}59{:}59]$	!: 0.2 ?: 0.8
e_4	e	$22{:}00{:}00$	!

Probability of Uncertain Trace Realizations

Event set		Event seq. s_e	Realization s_a	$P(s_e)$	$P(s_a \mid s_e)$	$P(s_a)$
$\{e_1, e_2, e_3, e_4\}$	S₁	$\langle e_1, e_2, e_3, e_4 \rangle$	$\langle a, b, d, e \rangle$	0.1	0.9	**0.09**
			$\langle a, c, d, e \rangle$		0.1	**0.01**
		$\langle e_1, e_3, e_2, e_4 \rangle$	$\langle a, d, b, e \rangle$	0.1	0.9	**0.09**
			$\langle a, d, c, e \rangle$		0.1	**0.01**
$\{e_1, e_2, e_4\}$	S₂	$\langle e_1, e_2, e_4 \rangle$	$\langle a, b, e \rangle$	0.8	0.9	**0.72**
			$\langle a, c, e \rangle$		0.1	**0.08**

Event ID	Activity	Timestamp	Event type
e_1	a	$8{:}13{:}43$	!
e_2	{b: 0.9, c: 0.1}	$[11{:}00{:}00, 11{:}59{:}59]$	!
e_3	d	$[11{:}00{:}00, 11{:}59{:}59]$	!: 0.2 ?: 0.8
e_4	e	$22{:}00{:}00$	!

Conformance Checking on Uncertain Data

- Let's look at a specific domain of application: **conformance checking**

- If we have a reference model M, we can naturally define conformance checking on an uncertain trace as

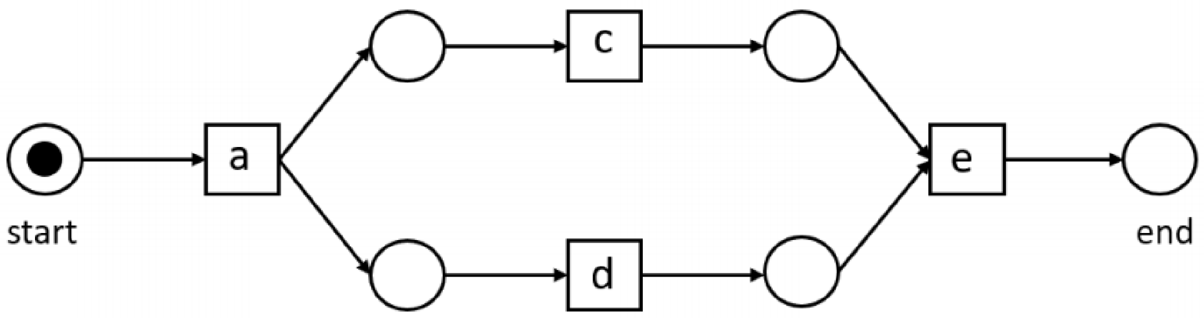
$$Conf = \sum_{s_a \in Realizations} P(s_a) \cdot conf(s_a, M)$$

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Conformance Checking on Uncertain Data

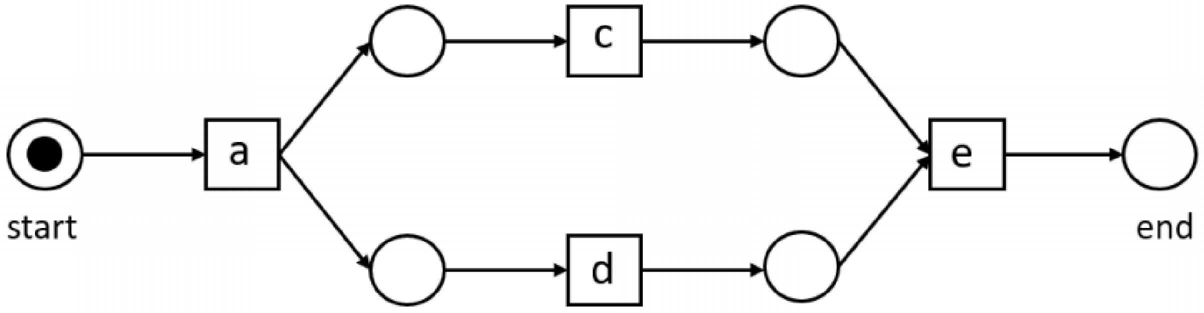
Reference model



Realization s_a	$P(s_a)$	$conf(s_a)$
$\langle a, b, d, e \rangle$	**0.09**	2
$\langle a, c, d, e \rangle$	**0.01**	0
$\langle a, d, b, e \rangle$	**0.09**	2
$\langle a, d, c, e \rangle$	**0.01**	0
$\langle a, b, e \rangle$	**0.72**	3
$\langle a, c, e \rangle$	**0.08**	1

$$Conf = \sum_{s_a \in Realizations} P(s_a) \cdot conf(s_a, M) = 2.6$$

Reference model



Realization s_a	$P(s_a)$	$conf(s_a)$
$\langle a, b, d, e \rangle$	**0.09**	2
$\langle a, c, d, e \rangle$	**0.01**	0
$\langle a, d, b, e \rangle$	**0.09**	2
$\langle a, d, c, e \rangle$	**0.01**	0
$\langle a, b, e \rangle$	**0.72**	3
$\langle a, c, e \rangle$	**0.08**	1

$$Conf = \sum_{s_a \in Realizations} P(s_a) \cdot conf(s_a, M) = 2.6$$

Much more representative than the unweighted average, 1.3

Chair of Process and Data Science

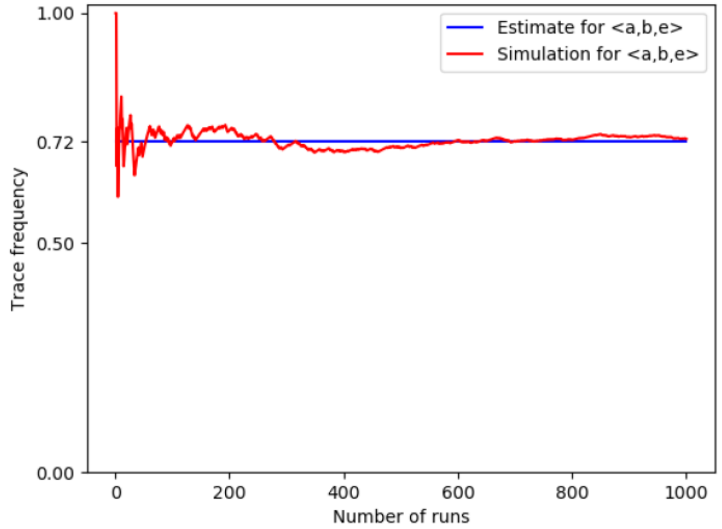
Evaluation

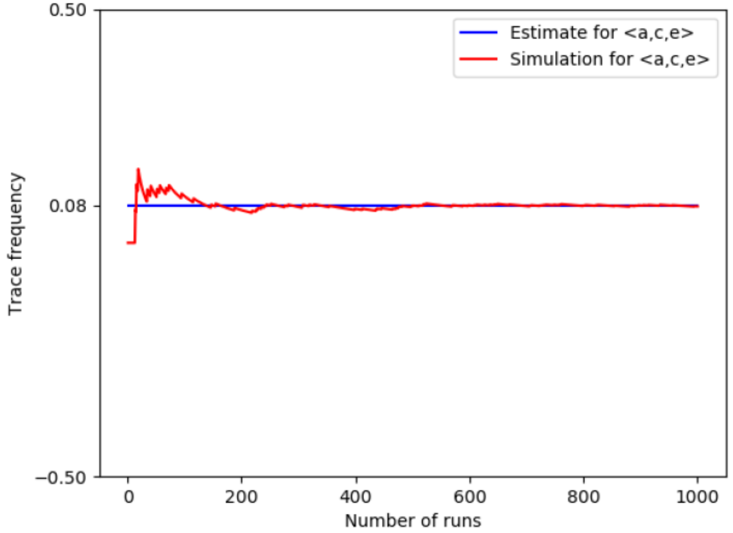
- To evaluate our probability estimation, we use a **Monte Carlo** method
- We generate realizations by sampling values for uncertain attributes in a trace
- We repeat the process, and we measure the **frequency** of each realization
- We then compare such frequency with our probability estimation

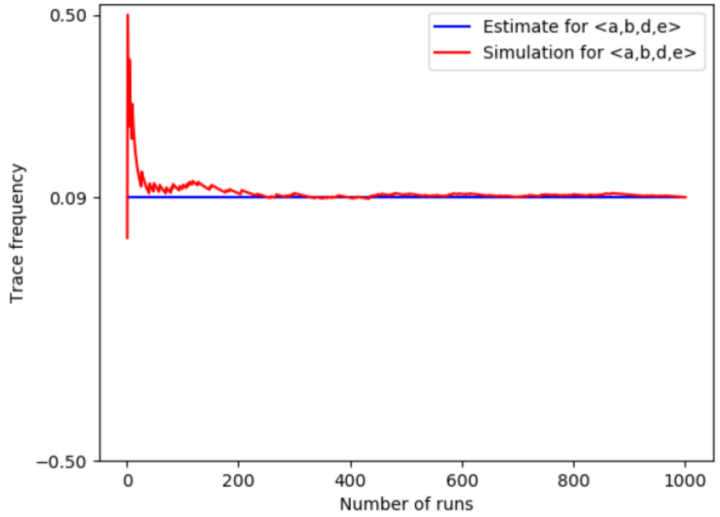
Chair of Process and Data Science

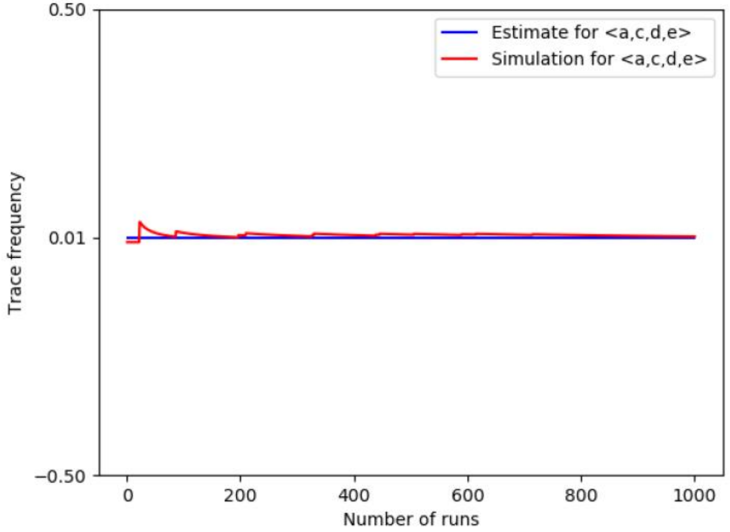
RWTH AACHEN UNIVERSITY

Evaluation









Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Conclusion

- In our work, we provide a method to reliably compute **probabilities of realizations** of uncertain traces

- The probability distribution of such realization gives important information
  - e.g., we can identify highly likely critical cases

- This information is an important complement to the insights provided, e.g., by **conformance checking over uncertain data**

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Future Work

- Addressing the problem of possible **dependencies** among uncertain attributes

- Extending existing approaches for **process discovery on uncertain data**

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Marco Pegoraro

pegoraro@pads.rwth-aachen.de

www.mpegoraro.net

@pegoraro_marco

https://www.researchgate.net/profile/Marco-Pegoraro-2

References

Pegoraro, Marco, Merih Seran Uysal, and Wil M.P. van der Aalst. "Discovering process models from uncertain event data." *International Conference on Business Process Management*. Springer, Cham, 2019.

Pegoraro, Marco, Merih Seran Uysal, and Wil M.P. van der Aalst. "Conformance checking over uncertain event data." *Information Systems* (2021): 101810.

Pegoraro, Marco, Merih Seran Uysal, and Wil M.P. van der Aalst. "Efficient time and space representation of uncertain event data." *Algorithms* 13.11 (2020): 285.

Pegoraro, Marco, Merih Seran Uysal, and Wil MP van der Aalst. "PROVED: A Tool for Graph Representation and Analysis of Uncertain Event Data." *International Conference on Applications and Theory of Petri Nets and Concurrency*. Springer, Cham, 2021.