



Mining Uncertain Event Data in Process Mining

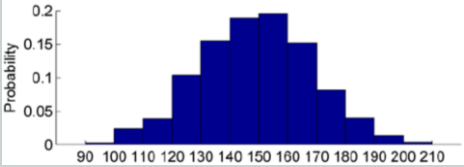
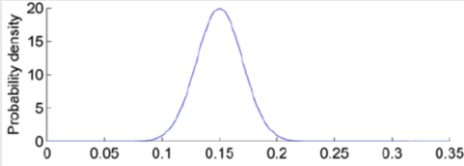
Marco Pegoraro and Wil M.P. van der Aalst

Uncertainty in event logs

An **uncertain event log** is an event log where some of the values include some sort of quantified uncertainty.

Uncertainty can be defined not only on the *attribute* level, but also on the *event* level.

Uncertainty - Taxonomy

	Weak uncertainty	Strong uncertainty
Discrete data	<p>Discrete probability distribution</p>  <p>A histogram showing a discrete probability distribution. The x-axis represents values from 90 to 210 in increments of 10. The y-axis represents probability from 0 to 0.2 in increments of 0.05. The distribution is bell-shaped and centered around 150.</p>	<p>Set of possible values</p> $\{x, y, z, \dots\}$
Continuous data	<p>Probability density function</p>  <p>A graph showing a probability density function. The x-axis represents values from 0 to 0.35 in increments of 0.05. The y-axis represents probability density from 0 to 20 in increments of 5. The curve is a smooth, bell-shaped Gaussian distribution centered at approximately 0.15.</p>	<p>Interval</p> $\{x \in \mathbb{R} a \leq x \leq b\}$

Uncertainty - Taxonomy

Uncertainty on the *attribute* level:

- **Case ID:** discrete
- **Activity:** discrete
- **Timestamp:** continuous

Uncertainty on the event level:

- **Indeterminate event:** an event that has been recorded, but it might not have happened. Discrete (binary)

Example of strongly uncertain trace

Case ID	Timestamp	Activity	Indet. event
{0, 1}	2011-12-05T00:00	A	!
0	2011-12-07T00:00	{B, C, D}	!
0	[2011-12-06T00:00, 2011-12-10T00:00]	D	?
0	2011-12-09T00:00	{A, C}	!
{0, 1, 2}	2011-12-11T00:00	E	?

Example of weakly uncertain trace

Case ID	Timestamp	Activity	Indet. event
{0:0.9, 1:0.1}	2011-12-05T00:00	A	!
0	2011-12-07T00:00	{B:0.7, C:0.3}	!
0	$\mathcal{N}(2011-12-08T00:00, 2)$	D	?:0.5
0	2011-12-09T00:00	{A:0.2, C:0.8}	!
{0:0.4, 1:0.6}	2011-12-11T00:00	E	?:0.7

Uncertainty in event logs

There can be many high-level sources of uncertainty in event data:

- **Incorrectness:** errors happened while recording data or manipulating the logs (e.g. while merging logs)
- **Coarseness:** variability of an attribute caused by imprecision of a measure (e.g. limitation of sensors)
- **Ambiguity:** the event data is recorded in a way that needs interpretation (e.g. event data recorded as free text)

Uncertainty in event logs

Very often, we have **coarseness on the timestamp attribute**.

Mainly because of two reasons:

- Data formats **too coarse** (e.g., timestamps recorded with the date but not the time)
- **Recording of events in batches** (e.g., a doctor that inputs data in an information system at the end of the round of visits)

Conformance checking in uncertain settings

- **Goal:** given a log with traces that contains uncertainty and a (non uncertain) model, calculate a measure of conformance for the **best and worst case scenario**
 - Search among possible realization of the uncertain trace the best and worst fitting
 - Provide an upper and lower bound for conformity cost in uncertain setting
 - We are going to use **alignments**
- **Setting:**
 - **Strong uncertainty on activities and timestamps**
 - **Strongly uncertain indeterminate events**

Running example

Setting:

- Strong uncertainty on activities and timestamps
- Strongly uncertain indeterminate events

Case ID	Timestamp	Activity	Indet. event
0	2011-12-05T00:00	A	!
0	2011-12-07T00:00	{B, C}	!
0	[2011-12-06T00:00 2011-12-10T00:00]	D	!
0	2011-12-09T00:00	{A, C}	!
0	2011-12-11T00:00	E	?

Realizations of a trace

- **Realizations of a trace:** all possible certain traces obtained by selecting an available value for the uncertain attributes.

Case ID	Timestamp	Activity	Indet. event
0	2011-12-05T00:00	A	!
0	2011-12-07T00:00	{B, C}	!
0	[2011-12-06T00:00 2011-12-10T00:00]	D	!
0	2011-12-09T00:00	{A, C}	!
0	2011-12-11T00:00	E	?

Realizations:

<A, B, C, D, E>

<A, B, D, C, E>

<A, C, D, C, E>

<A, C, D, A, E>

<A, D, C, C, E>

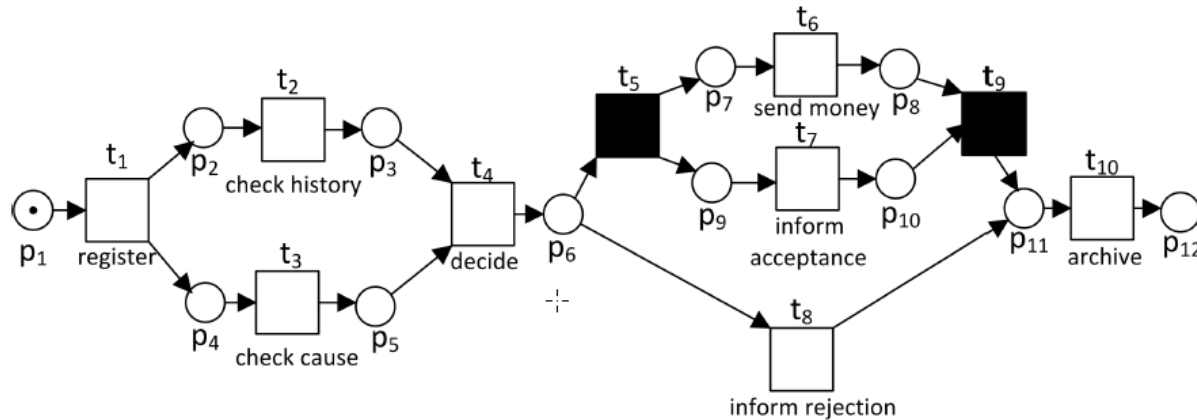
<A, D, B, C>

<A, D, C, A>

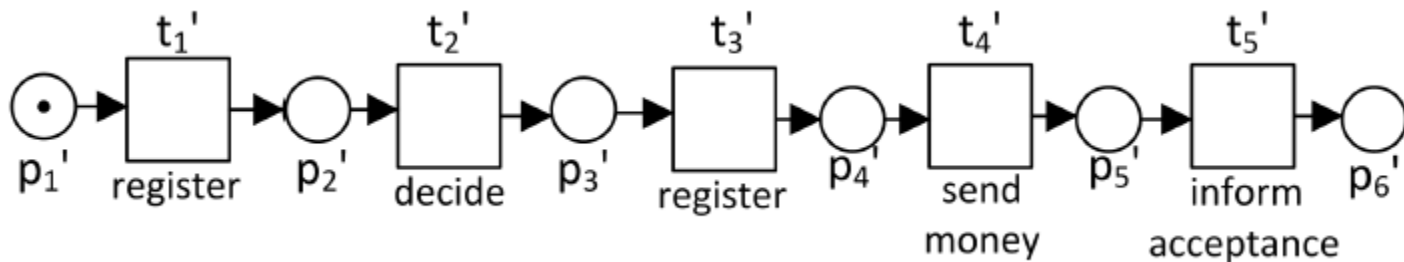
...

Alignments

To align a trace with a model, we need to firstly turn the trace into an **event net**, a sequence-shaped Petri net able to execute only that specific trace.

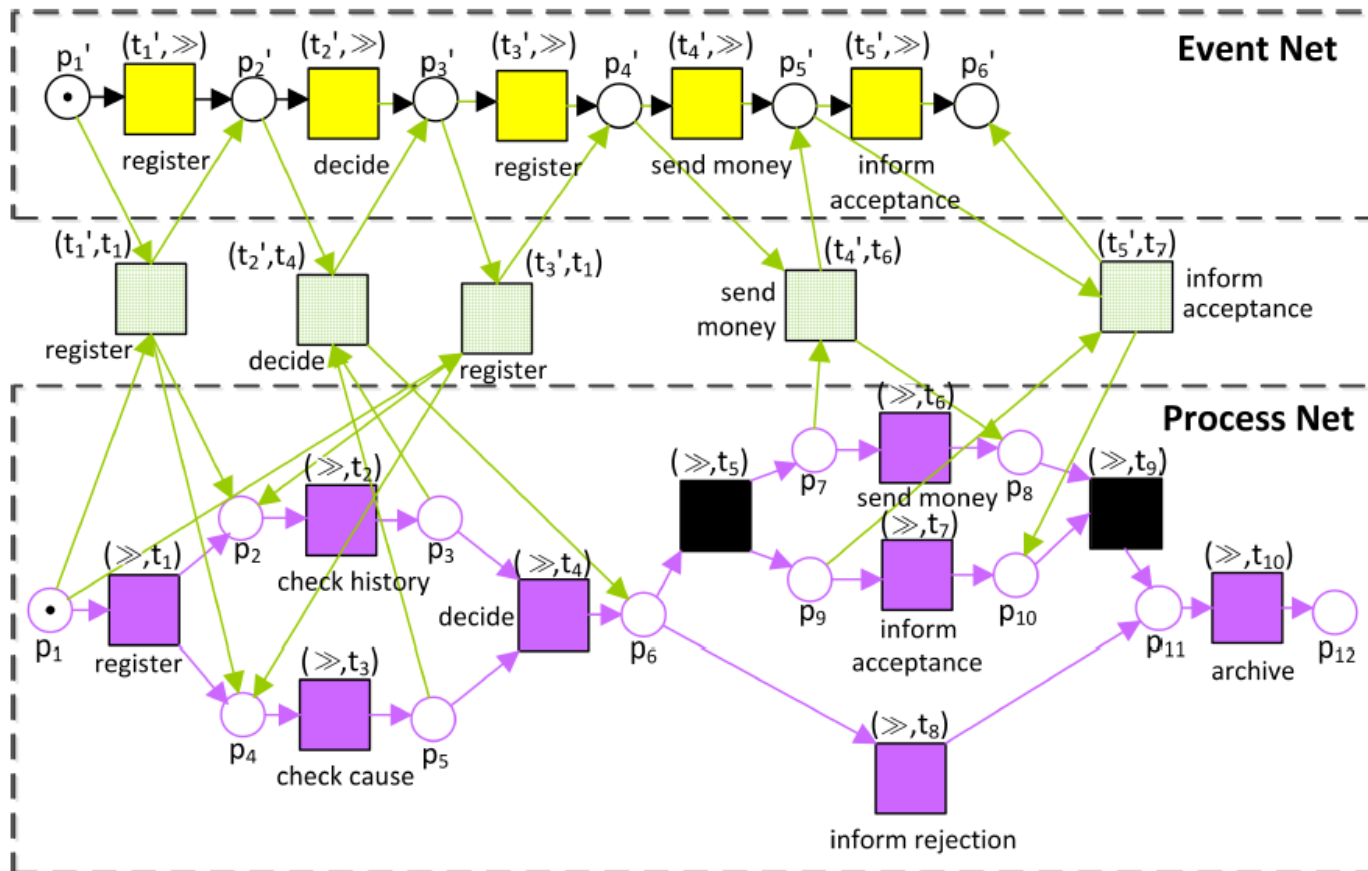


A process model



Event net of the trace $\langle \text{register}, \text{decide}, \text{register}, \text{send money}, \text{inform acceptance} \rangle$

Product net



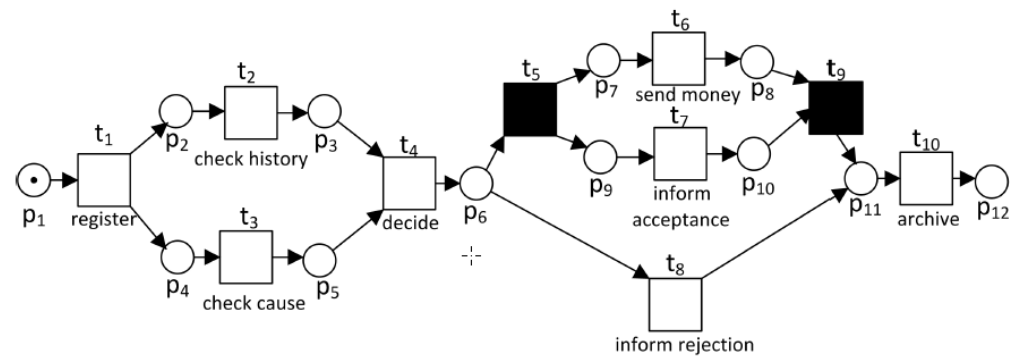
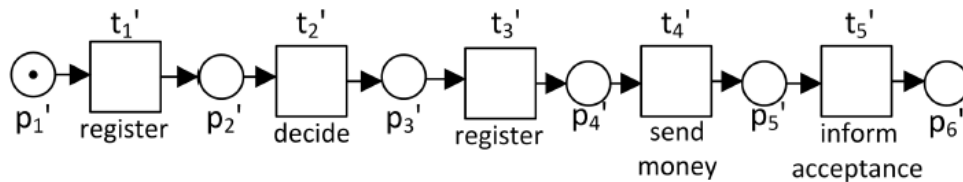
LEGEND

- Move on model
- Move on model (invisible transitions)
- Synchronous move
- Move on log

A. Adriansyah, doctoral thesis, 2014

Alignments

<i>register</i>	\gg	\gg	<i>decide</i>	<i>register</i>	\gg	<i>send money</i>	<i>inform acceptance</i>	\gg	\gg
<i>register</i>	<i>check history</i>	<i>check cause</i>	<i>decide</i>			<i>send money</i>	<i>inform acceptance</i>		<i>archive</i>
t_1	t_2	t_3	t_4	\gg	t_5	t_6	t_7	t_9	t_{10}



A. Adriansyah, doctoral thesis, 2014

Conformance checking in uncertain settings

Bruteforce approach

1. Generate all the realizations of an uncertain trace
2. Align all of them
3. Pick the ones with the minimum and maximum score

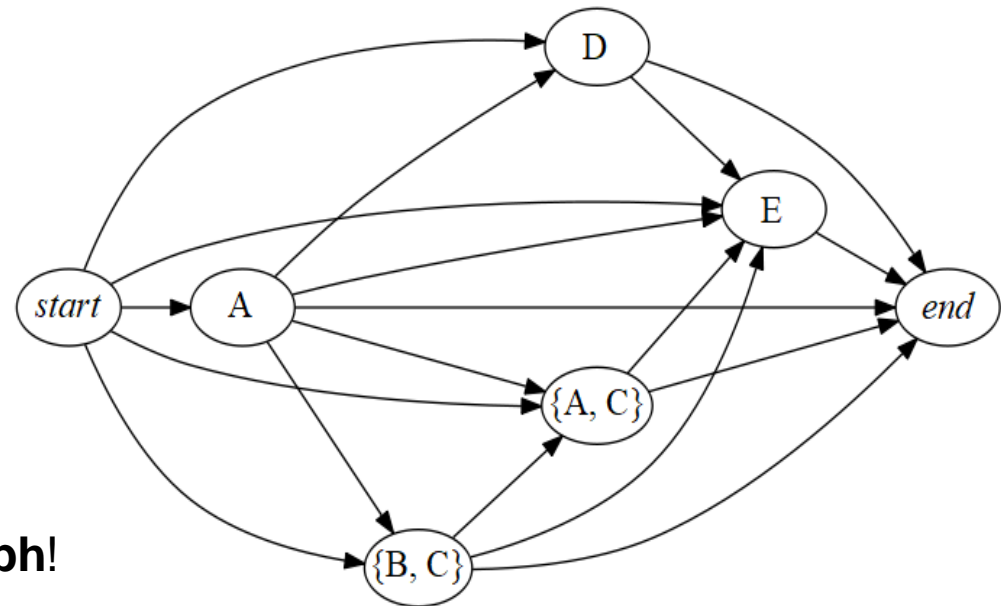
Very slow!

There is a quicker way to compute the **lower bound for conformance cost**

Process mining over uncertainty: behavior graph

1. Create a node for each uncertain event
2. Create two extra nodes *start* and *end*
3. $A \rightarrow B$ iff the event in node A has happened **before** the event in node B

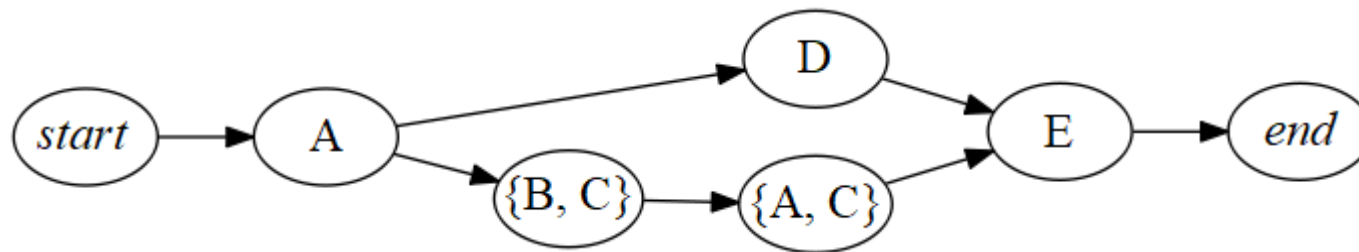
Case ID	Timestamp	Activity	Indet. event
0	2011-12-05T00:00	A	!
0	2011-12-07T00:00	{B, C}	!
0	[2011-12-06T00:00 2011-12-10T00:00]	D	!
0	2011-12-09T00:00	{A, C}	!
0	2011-12-11T00:00	E	?



Notice that a behaviour graph will always be a **directed acyclic graph!**

Process mining over uncertainty: reduced behavior graph

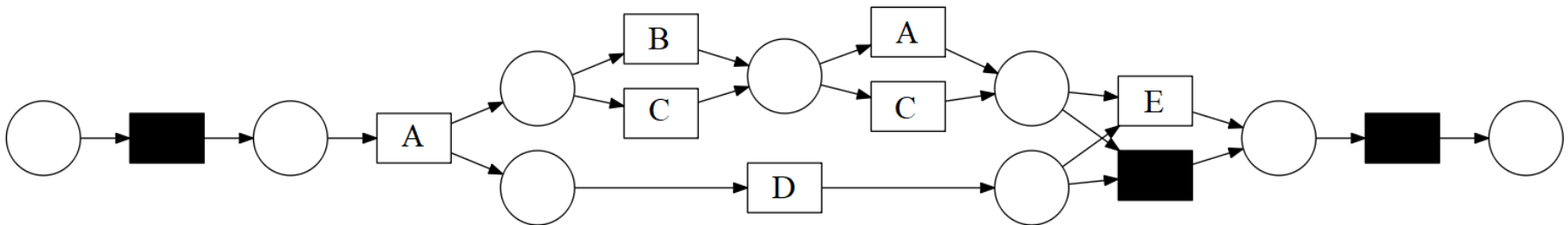
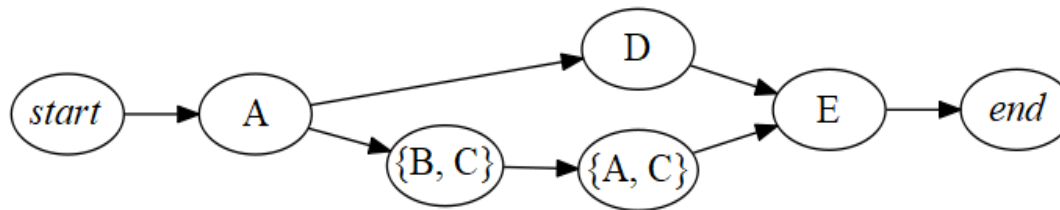
We then perform a **transitive reduction**.



Now, $A \rightarrow B$ if the event in node A happened **immediately before** the event in node B.

Process mining over uncertainty: behavior net

Case ID	Timestamp	Activity	Indet. event
0	2011-12-05T00:00	A	!
0	2011-12-07T00:00	{B, C}	!
0	[2011-12-06T00:00 2011-12-10T00:00]	D	!
0	2011-12-09T00:00	{A, C}	!
0	2011-12-11T00:00	E	?



Process mining over uncertainty: behavior net

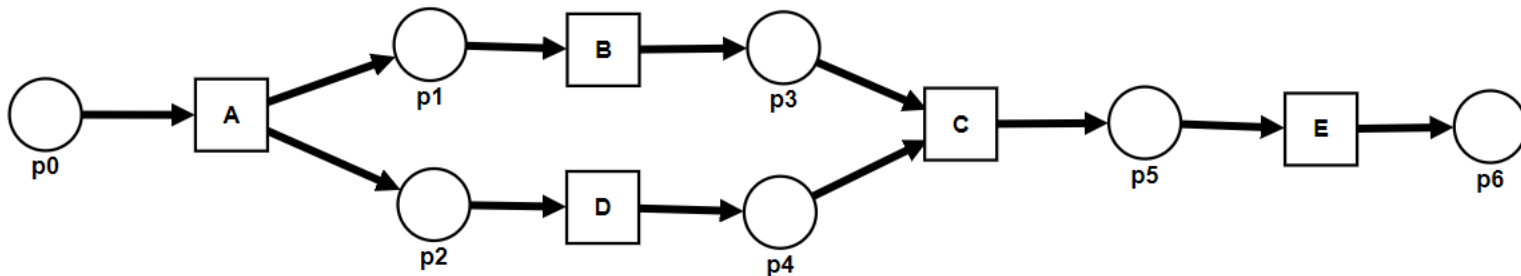
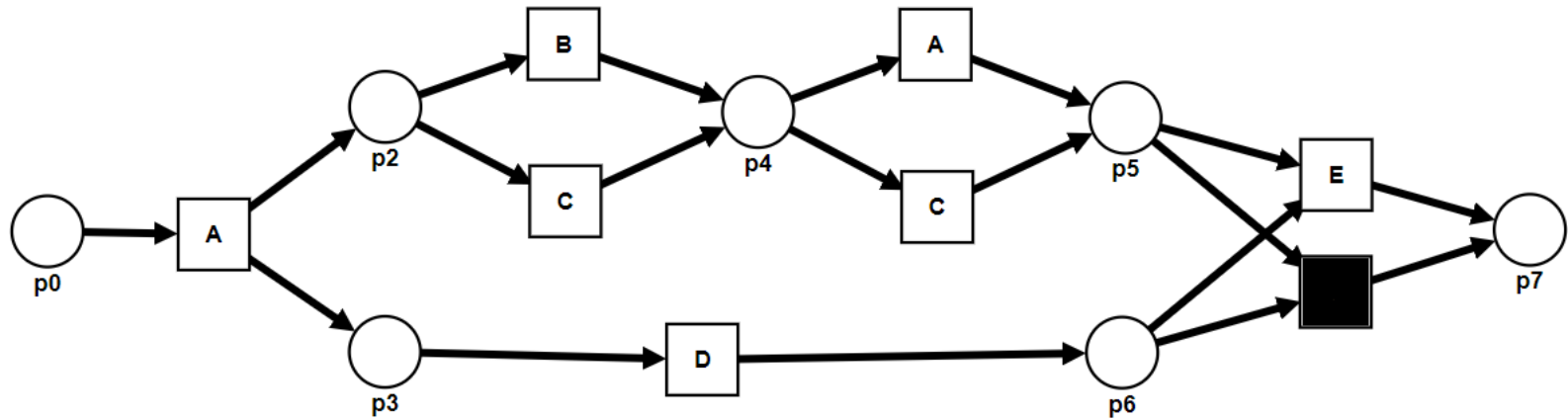
We can use the behavior net instead of the event net to compute alignments.

We obtain **two complete firing sequences**, one on the behavior net and one on the model.

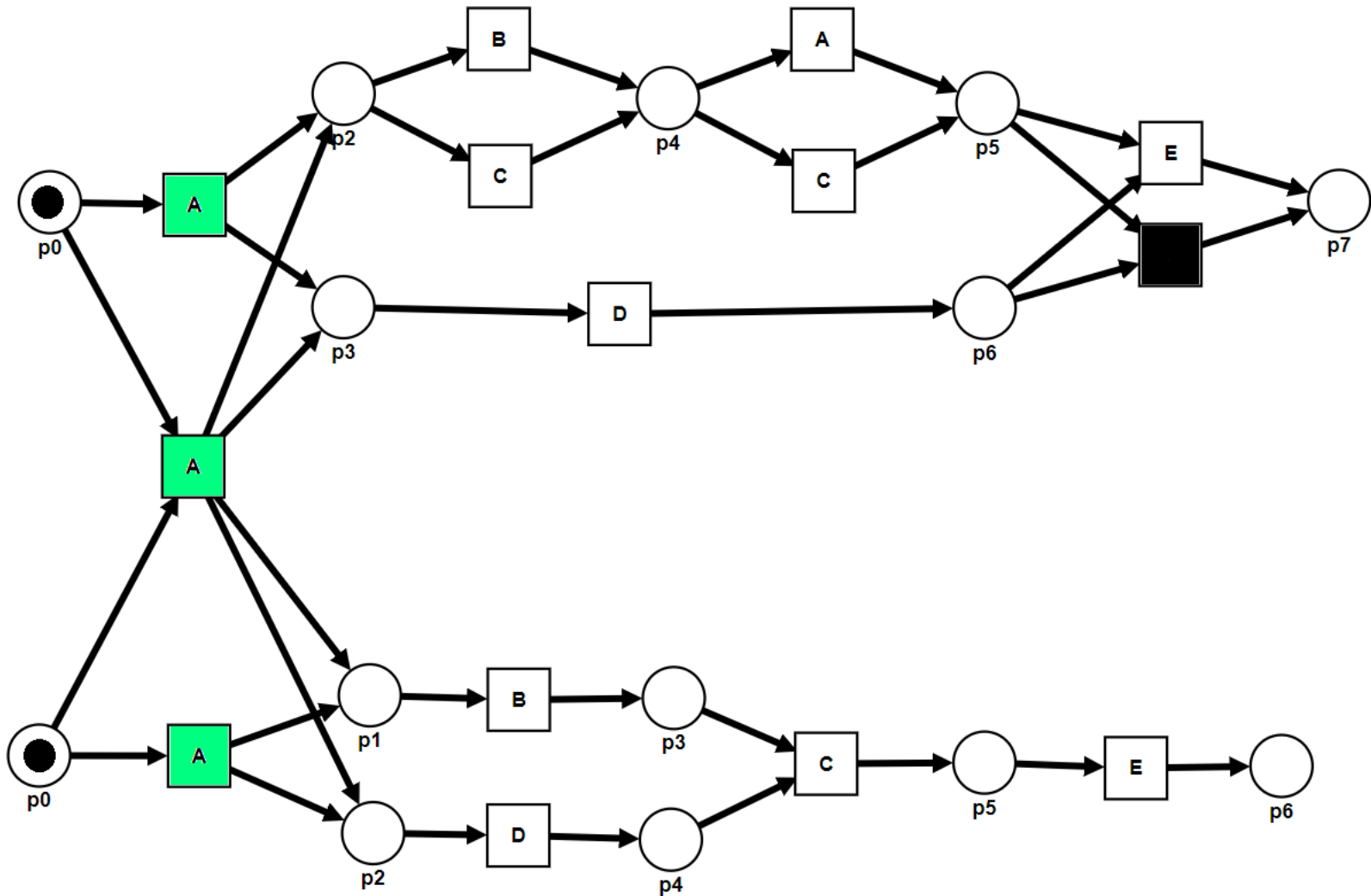
The firing sequence in the behavior net will be a **realization** of the uncertain trace.

Since the search returns the path through the product net with the minimal cost, the realization returned by the alignment will be the **lower bound** for conformance cost

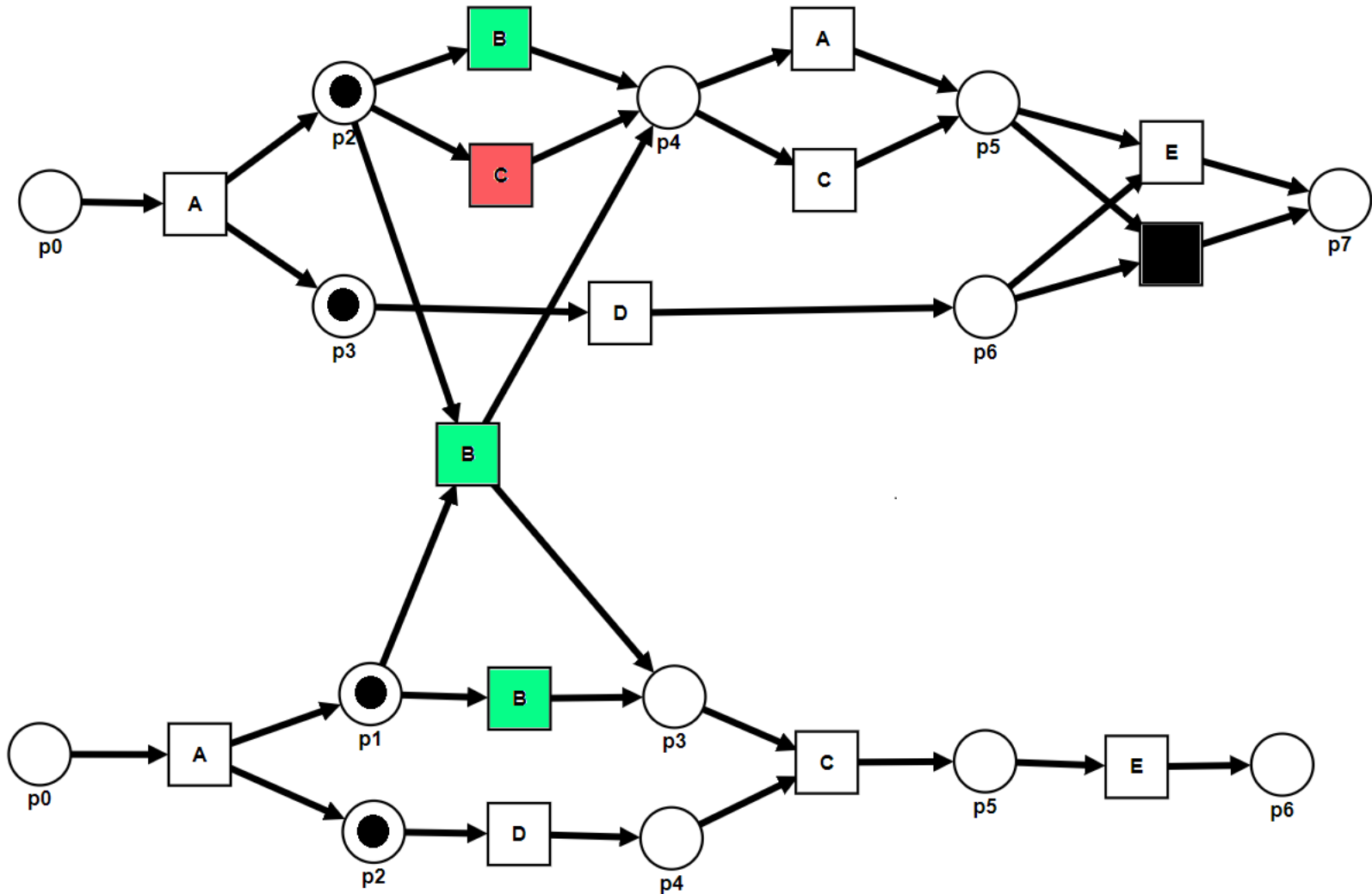
Process mining over uncertainty: alignments



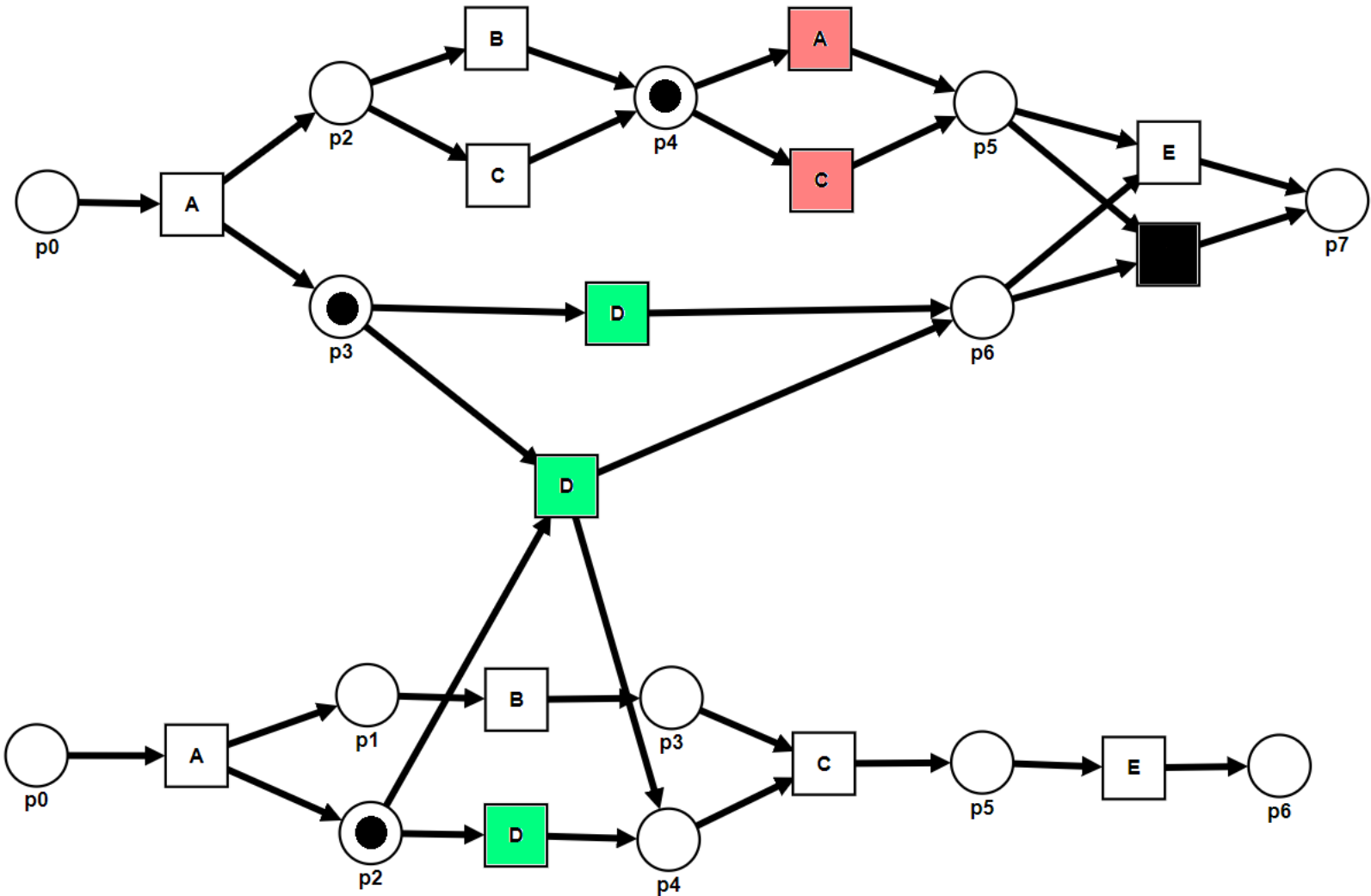
Process mining over uncertainty: alignments



Process mining over uncertainty: alignments



Process mining over uncertainty: alignments

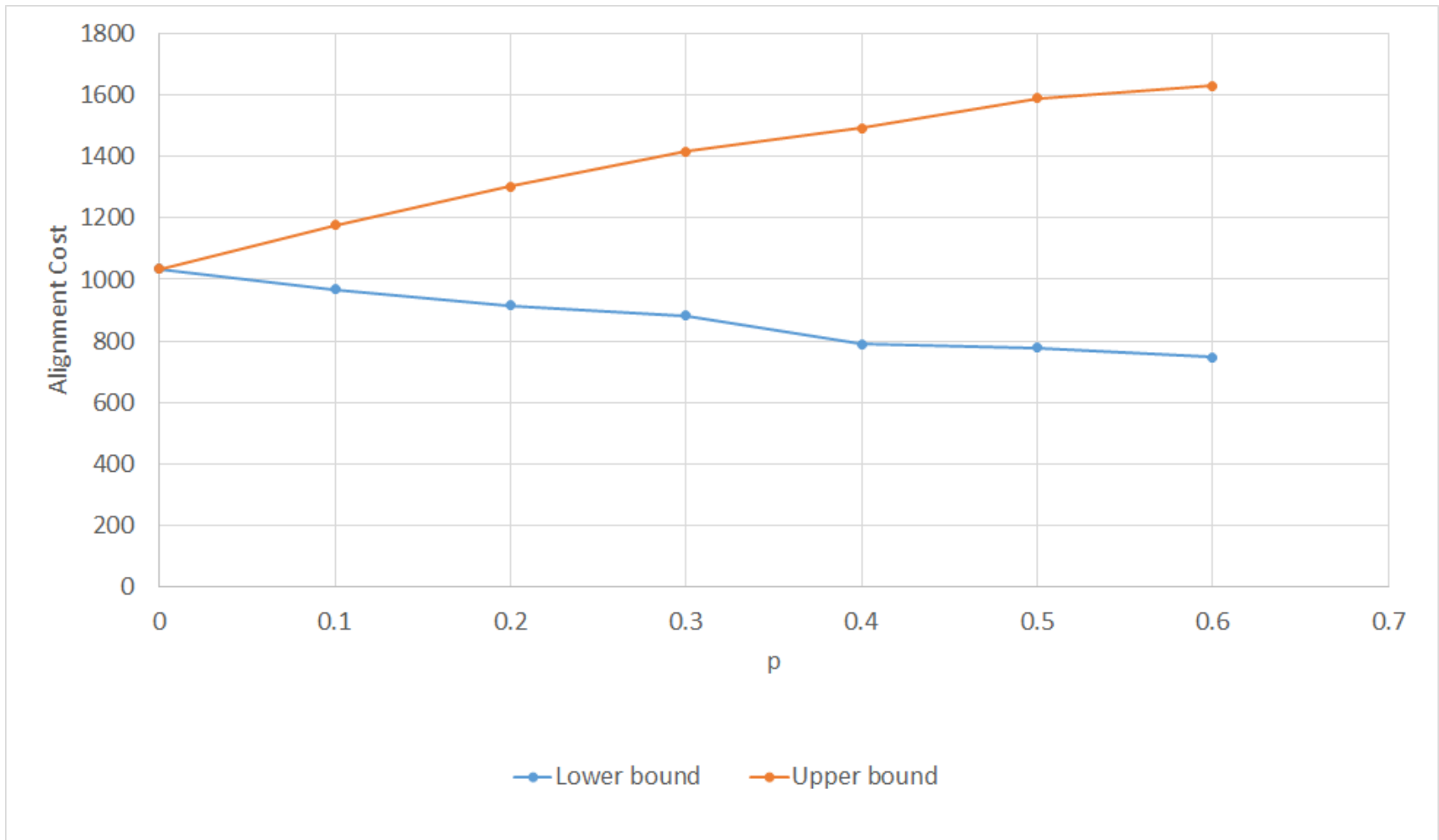


Experimental results

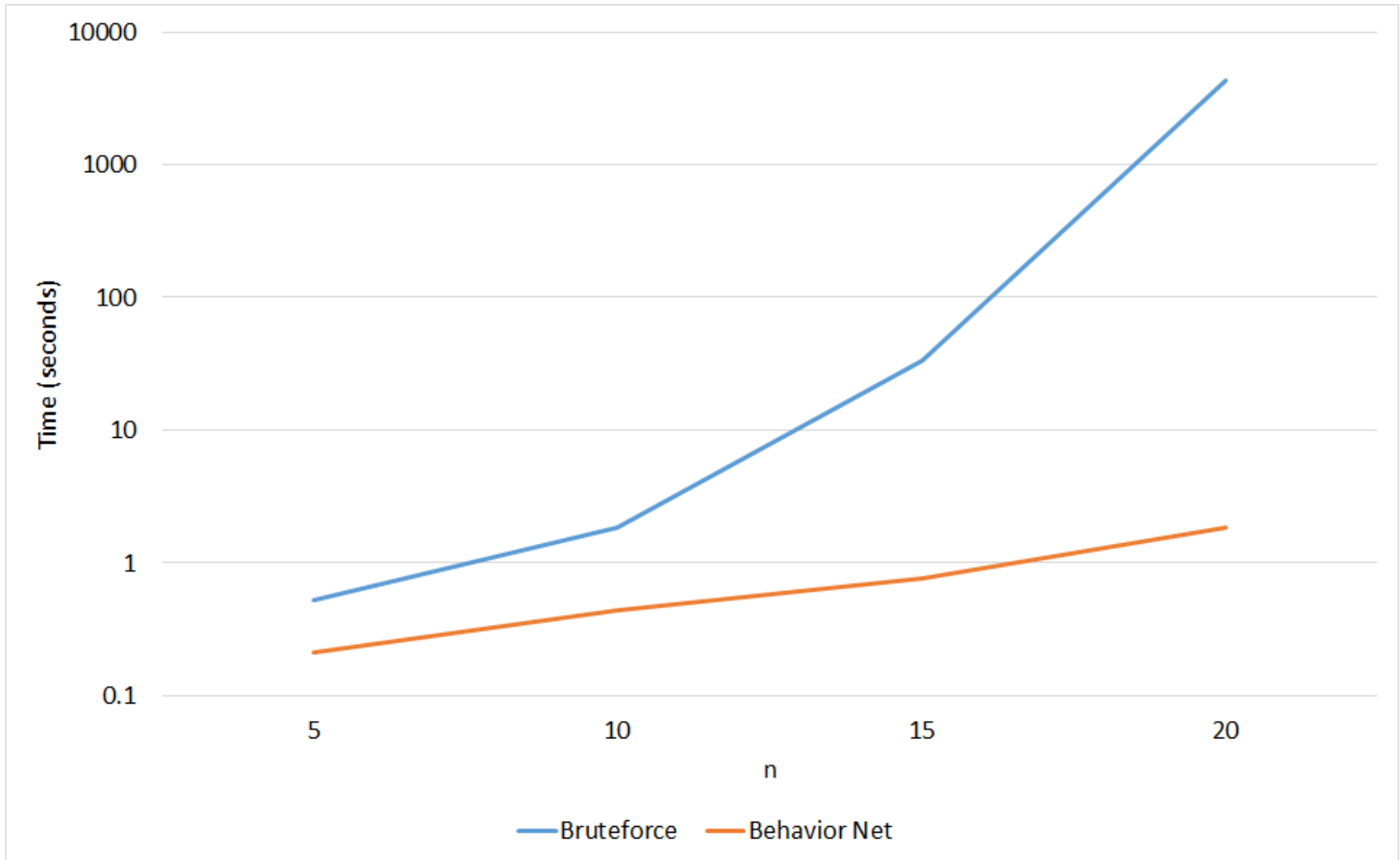
Two research questions:

- **Q1: do the upper and lower bounds for the conformance cost behave as expected in uncertain traces?**
- **Q2: does aligning the behavior net to calculate the lower bound for the conformance cost yield lower computing times?**

Experimental results: Q1



Experimental results: Q2



Future developments

- Discovery of uncertain models
- Optimization of the worst case scenario of alignment computation
- Extension to weak uncertainty

Contacts and references

Marco Pegoraro

pegoraro@pads.rwth-aachen.de

Twitter: @pegoraro_marco

Site: <http://mpegoraro.net/>



<http://pm4py.pads.rwth-aachen.de/>

Twitter: @pm4py



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

Twitter: @pads_rwth

Blog: <https://blog.rwth-aachen.de/pads/>