Discovering Process Models from Uncertain Event Data

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

The 15th International Workshop on Business Process Intelligence BPI@BPM 2019
September 2, 2019, Vienna

Question:

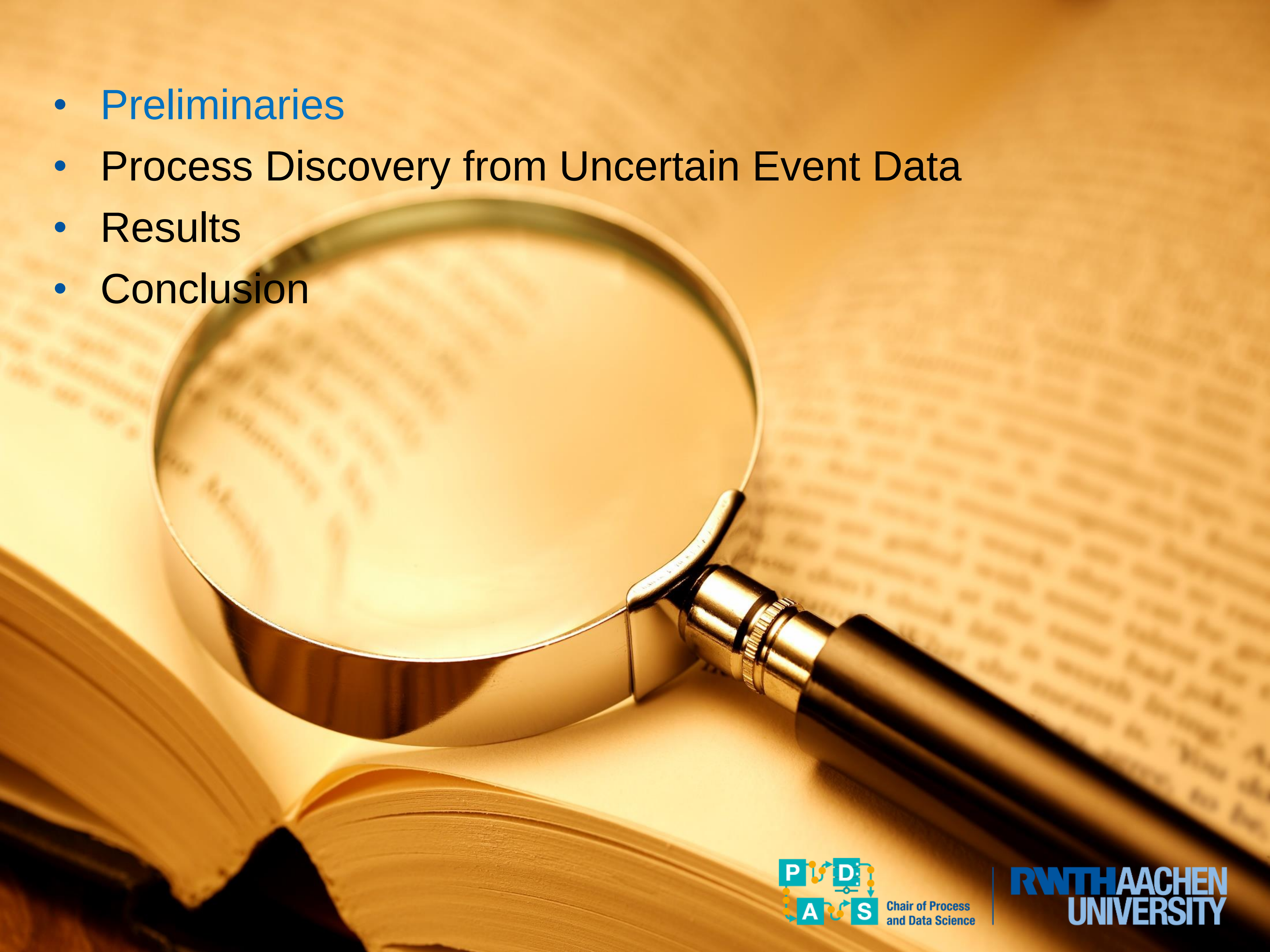
How to discover process models from uncertain event data?

Question:

How to discover process models from uncertain event data?

Proposal:

Utilize occurrences of activities and directly-follows relationships in uncertain event log to obtain a process model.



- Preliminaries
- Process Discovery from Uncertain Event Data
- Results
- Conclusion


UN CERTAIN

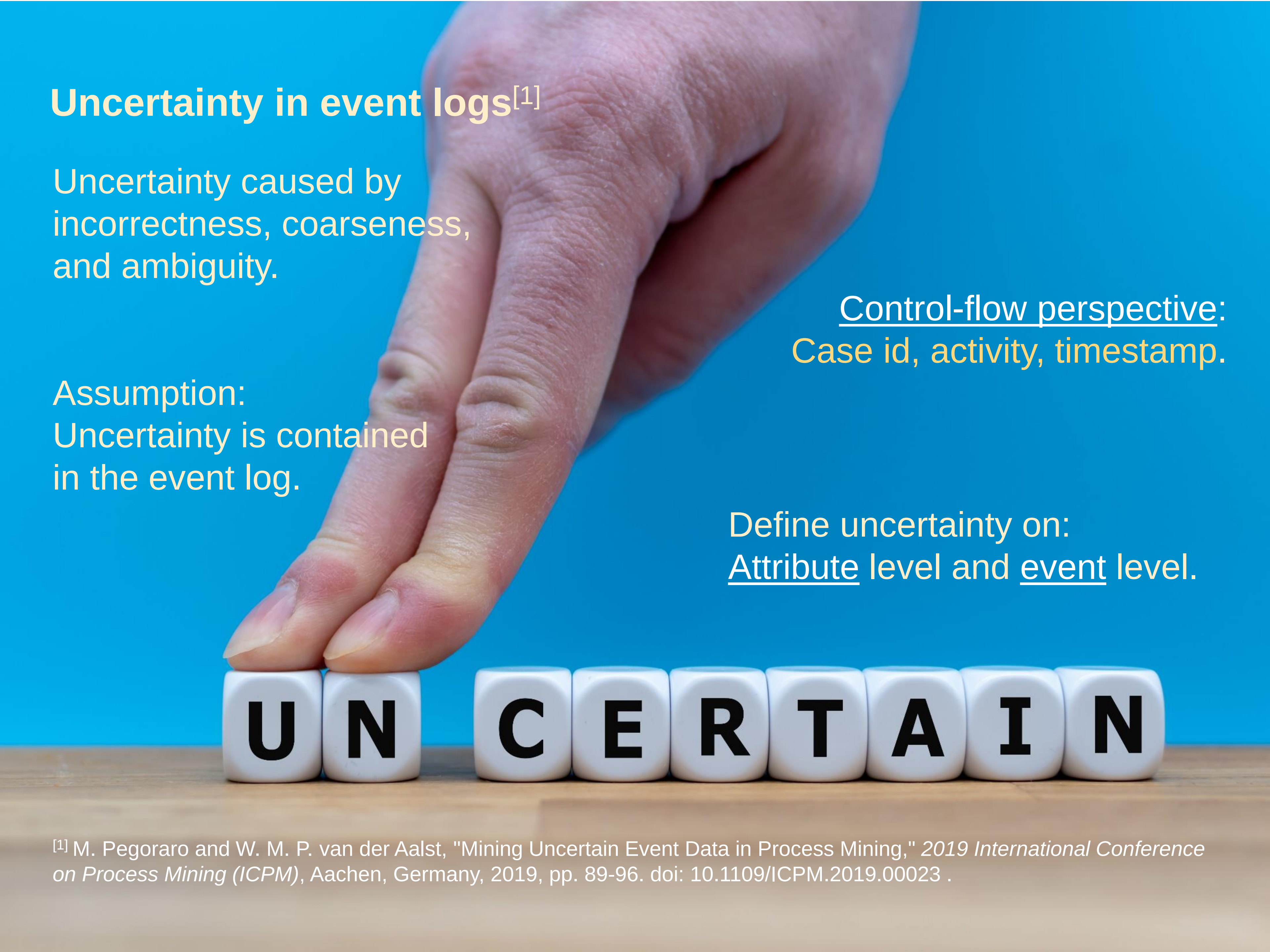
Uncertainty in event logs[1]

Uncertainty caused by incorrectness, coarseness, and ambiguity.

Control-flow perspective:
Case id, activity, timestamp.

Assumption:
Uncertainty is contained in the event log.

Define uncertainty on:
Attribute level and event level.



[1] M. Pegoraro and W. M. P. van der Aalst, "Mining Uncertain Event Data in Process Mining," *2019 International Conference on Process Mining (ICPM)*, Aachen, Germany, 2019, pp. 89-96. doi: 10.1109/ICPM.2019.00023 .

Did the event happen?

Representation of uncertainty

Event ID	Case ID	Timestamp	Activity	Did it happen?
e_1	0	05.12.2011	A	yes
e_2	0	07.12.2011	{B, C}	yes
e_3	0	**[06.12.2011, 10.12.2011]**	D	yes
e_4	0	09.12.2011	{A, C}	yes
e_5	0	11.12.2011	E	maybe

Continuous attributes: Represent uncertainty by an interval.
Example: “Timestamp”.

Discrete attributes: Represent uncertainty by a set of possible values.
Example: “Activity”.

Representation of uncertainty

Event ID	Case ID	Timestamp	**Activity**	Did it happen?
e_1	0	05.12.2011	A	yes
e_2	0	07.12.2011	**{B, C}**	yes
e_3	0	[06.12.2011, 10.12.2011]	D	yes
e_4	0	09.12.2011	**{A, C}**	yes
e_5	0	11.12.2011	E	maybe

Continuous attributes: Represent uncertainty by an interval.
Example: “Timestamp”.

Discrete attributes: Represent uncertainty by a set of possible values.
Example: “Activity”.

Example of uncertain trace

Event ID	Case ID	Timestamp	Activity	Did it happen?
e_1	0	05.12.2011	A	yes
e_2	0	07.12.2011	{B, C}	yes
e_3	0	**[06.12.2011, 10.12.2011]**	D	yes
e_4	0	09.12.2011	{A, C}	yes
e_5	0	11.12.2011	E	maybe

The exact timestamp of e_3 belongs to this interval

Example of uncertain trace

Event ID	Case ID	Timestamp	Activity	Did it happen?
e_1	0	05.12.2011	A	yes
e_2	0	07.12.2011	**{B, C}**	yes
e_3	0	[06.12.2011, 10.12.2011]	D	yes
e_4	0	09.12.2011	**{A, C}**	yes
e_5	0	11.12.2011	E	maybe

The events e_2 and e_4 have a set of possible activity labels

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

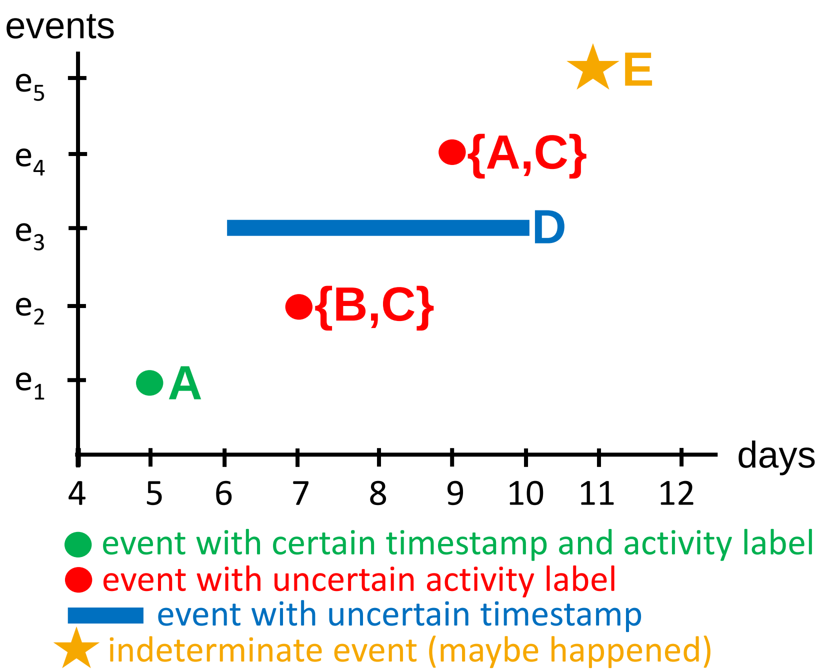
Example of uncertain trace

Event ID	Case ID	Timestamp	Activity	Did it happen?
e_1	0	05.12.2011	A	yes
e_2	0	07.12.2011	{B, C}	yes
e_3	0	[06.12.2011, 10.12.2011]	D	yes
e_4	0	09.12.2011	{A, C}	yes
e_5	0	11.12.2011	E	**maybe**

The event e_5 has been recorded but maybe it did not happen

Realizations of an uncertain trace

Event ID	Case ID	Timestamp	Activity	Did it happen?
e_1	0	05.12.2011	A	yes
e_2	0	07.12.2011	{B, C}	yes
e_3	0	[06.12.2011, 10.12.2011]	D	yes
e_4	0	09.12.2011	{A, C}	yes
e_5	0	11.12.2011	E	maybe



Example 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>

...

- Preliminaries
- Process Discovery from Uncertain Event Data
- Results
- Conclusion

Recall question: How to discover process models from uncertain event data?

Method:

Use minimum and maximum number of times an activity or a directly-follows relationship can appear in the realizations of the event log.

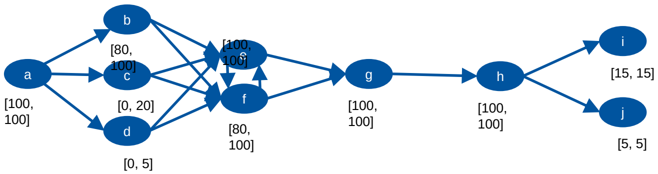
“Big picture”

<a, b, e, f, g, h>^{80}
<a, {b, c}, [e, f], g, h, i>^{15}
<a, {b, c, d}, [e, f], g, h, j>^{5}

Uncertain event log

Count activities and directly-follows relationships

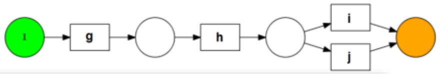
Uncertain Directly-Follows Graph (UDFG)



Filter on nodes/edges

Directly-Follows Graph

Inductive miner directly-follows [2]



Petri net

[2] Sander J.J. Leemans, Dirk Fahland, and Wil M.P. van der Aalst. "Scalable process discovery and conformance checking." Software & Systems Modeling 17.2 (2018): 599-631.

Formalism for uncertain event logs

Curly braces { } indicate uncertainty over activities. {b, c} indicates a single event that can be b or c.

<a, **{b, c}**, [e, f], g, h, i>

Formalism for uncertain event logs

Curly braces { } indicate uncertainty over activities. {b, c} indicates a single event that can be b or c.

Underlined events are indeterminate.
$\underline{f}$ indicates that the event may or may not have happened.

<a, {b, c}, [e, $\underline{f}$], g, h, i>

Formalism for uncertain event logs

Curly braces { } indicate uncertainty over activities. {b, c} indicates a single event that can be b or c.

Underlined events are indeterminate. f indicates that the event may or may not have happened.

<a, {b, c}, **[e, f]**, g, h, i>

Square brackets [ ] indicate uncertainty over timestamps. [e, f] indicates that the order between e and f is unknown.

Activity occurrences

Consider the following event log L:
<a, b, e, f, g, h>^{80}
<a, {b, c}, [e, f], g, h, i>^{15}
<a, {b, c, d}, [e, f], g, h, j>^{5}

Count the minimum and maximum number of times we observe an activity:

	<a, b, e, f, g, h>^{80}	<a, {b, c}, [e, f], g, h, i>^{15}	<a, {b, c, d}, [e, f], g, h, j>^{5}	L
a	[80, 80]	[15, 15]	[5, 5]	[100, 100]
b	[80, 80]	[0, 15]	[0, 5]	[80, 100]

min_a and ***max_a*** : Minimum and maximum number of times that ***a*** appears in L.

Directly-follows relationship occurrences

Consider the following event log L:

<a, b, e, f, g, h>^{80}

<a, {b, c}, [e, f], g, h, i>^{15}

<a, {b, c, d}, [e, f], g, h, j>^{5}

Count the minimum and maximum number of times we observe the directly-follows relation:

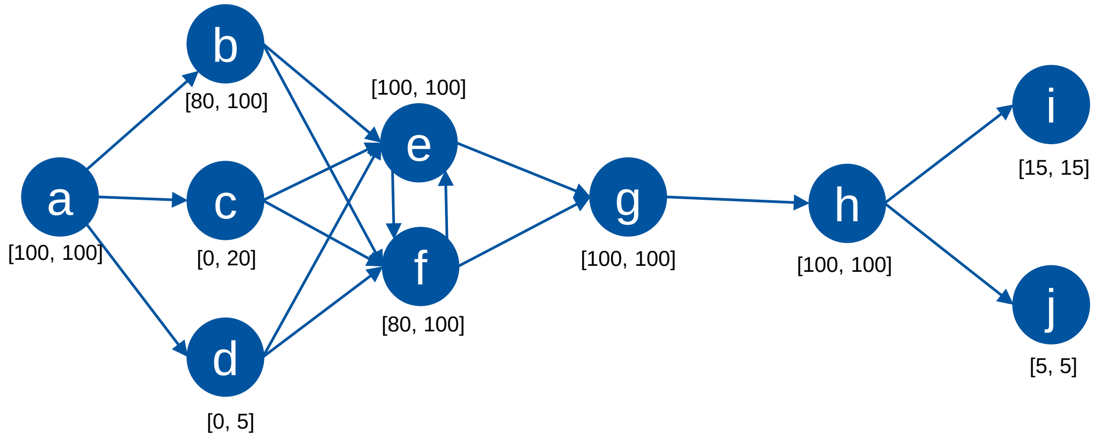
	<a, b, e, f, g, h>^{80}	<a, {b, c}, [e, f], g, h, i>^{15}	<a, {b, c, d}, [e, f], g, h, j>^{5}	L
$a \rightarrow b$	[80, 80]	[0, 15]	[0, 5]	[80, 100]
$a \rightarrow c$	[0, 0]	[0, 15]	[0, 5]	[0, 20]

$min_{a \rightarrow b}$ and ***$max_{a \rightarrow b}$***: Minimum and maximum number of times the ***$a \rightarrow b$*** appears in L.

Uncertain Directly-Follows Graph (1)

An **Uncertain Directly-Follows Graph (UDFG)** is a graph labeled with the intervals.

Example of UDFG showing the labels on activities (nodes):



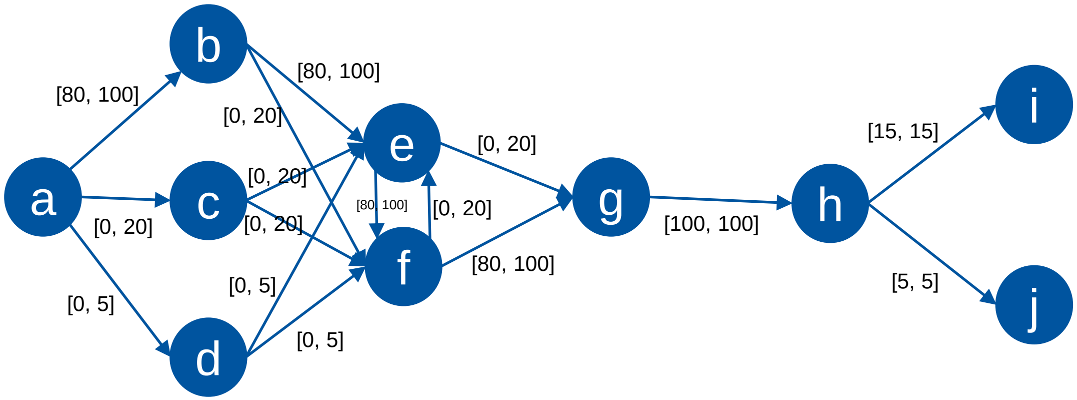
<a, b, e, f, g, h>^{80}
<a, {b, c}, [e, f], g, h, i>^{15}
<a, {b, c, d}, [e, f], g, h, j>^{5}

Chair of Process and Data Science

Uncertain Directly-Follows Graph (2)

An **Uncertain Directly-Follows Graph (UDFG)** is a graph labeled with the intervals.

Example of UDFG showing the labels on directly-follows relationships (edges):



<a, b, e, f, g, h>^{80}
<a, {b, c}, [e, f], g, h, i>^{15}
<a, {b, c, d}, [e, f], g, h, j>^{5}

Filtering approach

1. Determine:
   - Activity filtering parameters ***act****min* and ***act****max* and
   - Relationship filtering parameters ***rel****min* and ***rel****max*.

2. Keep the activities (vertices of the UDFG) for which it holds:

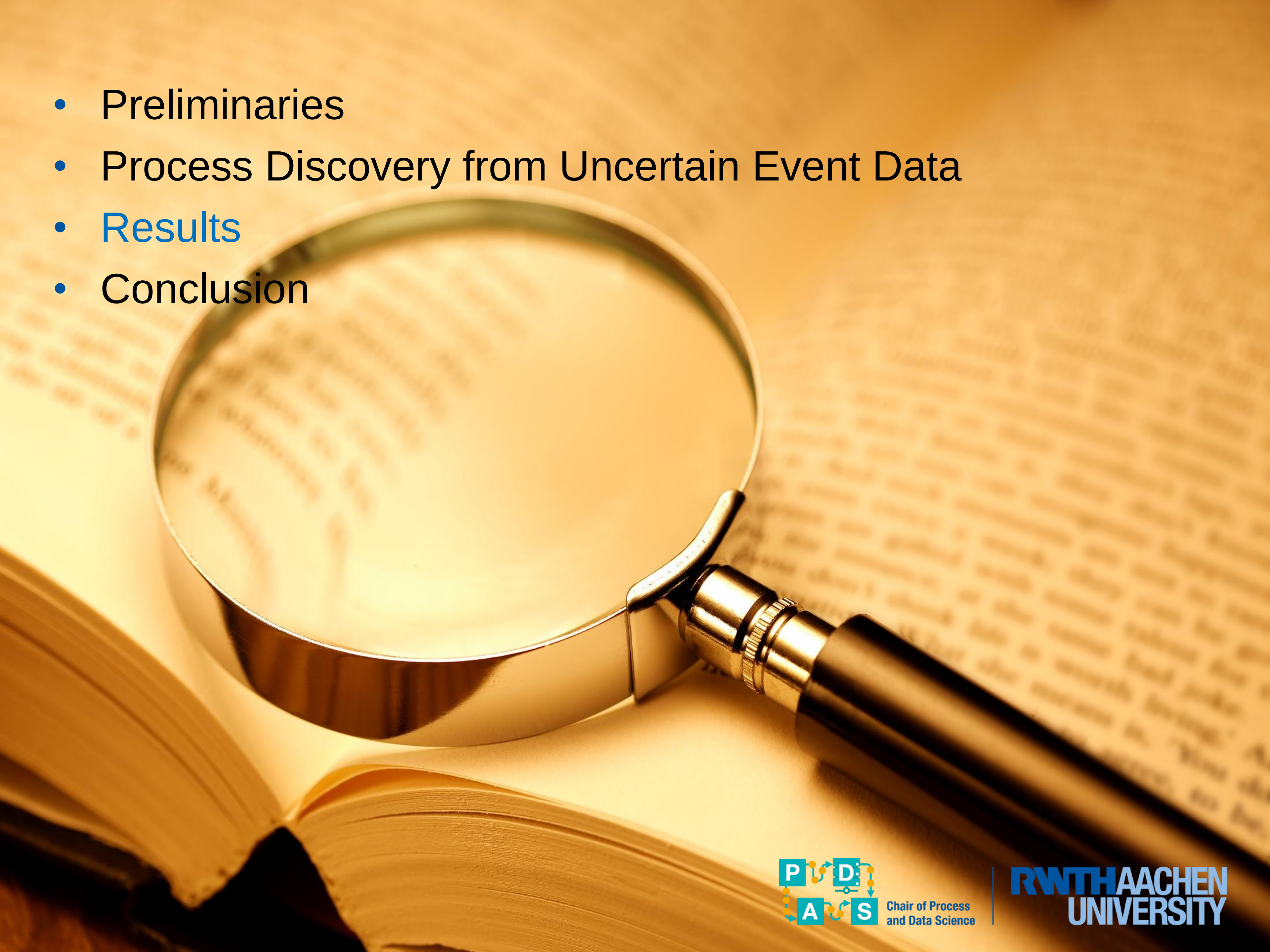
$$act_{min} \leq \frac{min_a}{max_a} \leq act_{max}$$

3. Keep the directly-follows relations (edges of the UDFG) for which it holds:

$$rel_{min} \leq \frac{min_{a \to b}}{max_{a \to b}} \leq rel_{max}$$

4. Perform Inductive Miner-directly-follows[2] approach on the filtered UDFG to obtain a process model.

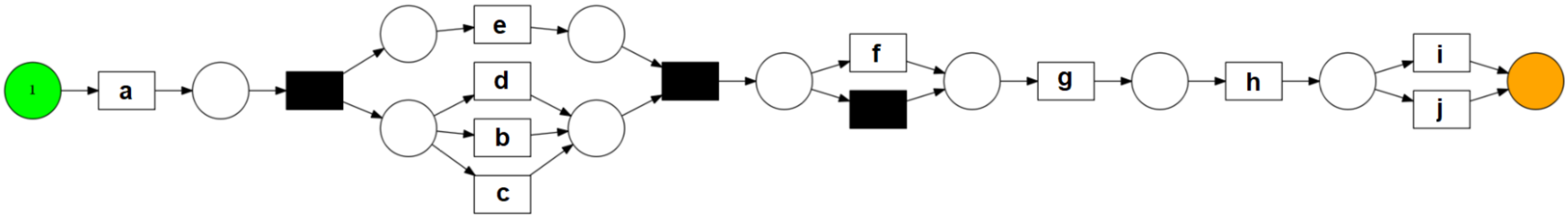
[2] Sander J.J. Leemans, Dirk Fahland, and Wil M.P. Van der Aalst. "Scalable process discovery and conformance checking." Software & Systems Modeling 17.2 (2018): 599-631.



- Preliminaries
- Process Discovery from Uncertain Event Data
- Results
- Conclusion

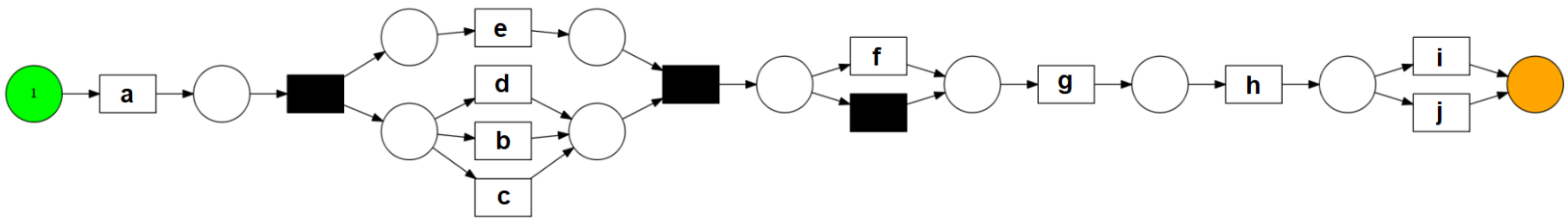
Results

<a, b, e, f, g, h>^{80}
<a, [{ b, c }, e], f, g, h, i>^{15}
<a, [{ b, c, d }, e], f, g, h, j>^{5}
Resulting petri net arising from unfiltered UDFG
i.e. $act_{min} = 0$; $act_{max} = 1$; $rel_{min} = 0$; $rel_{max} = 1$:

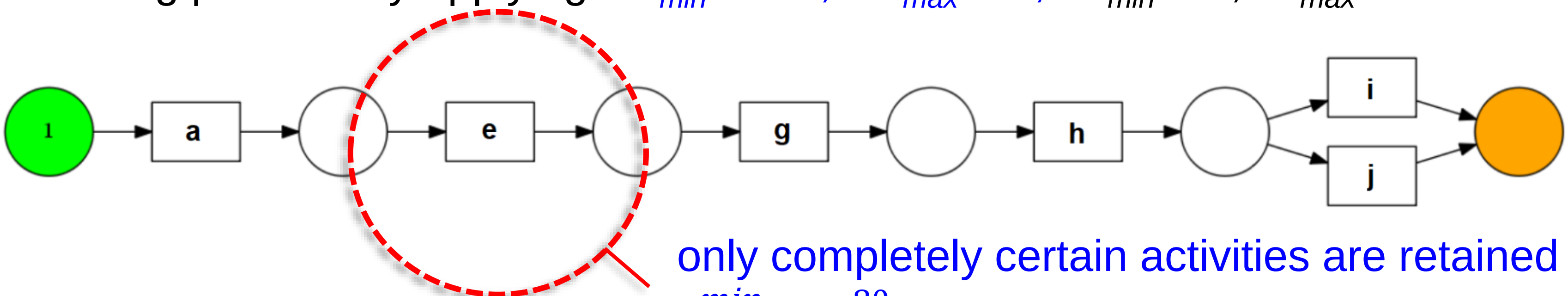


Results

<a, b, e, f, g, h>^{80}
<a, [{ b, c }, e], f, g, h, i>^{15}
<a, [{ b, c, d }, e], f, g, h, j>^{5}
Resulting petri net arising from unfiltered UDFG:



Resulting petri net by applying act_{min} = 0.9; act_{max} = 1; rel_{min} = 0; rel_{max} = 1 to UDFG:

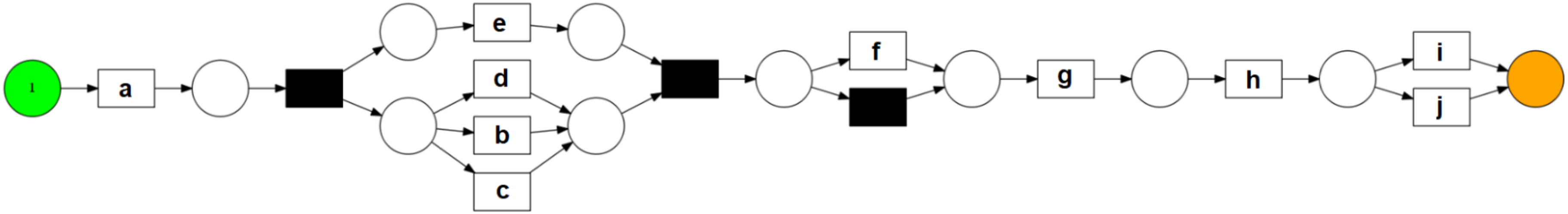


only completely certain activities are retained

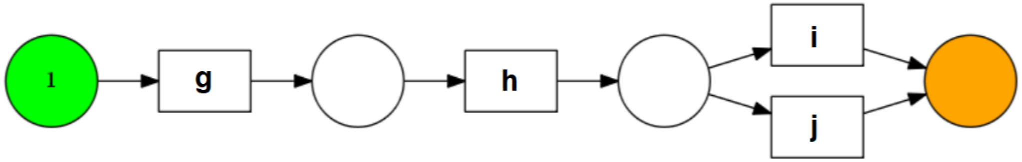
$\frac{min_b}{max_b} = \frac{80}{100} = 0.8 < act_{min}$ = 0.9, thus remove b

Results

<a, b, e, f, g, h>^{80}
<a, [{ b, c }, e], f, g, h, i>^{15}
<a, [{ b, c, d }, e], f, g, h, j>^{5}
Resulting petri net arising from unfiltered UDFG:



Resulting petri net by applying act_{min} = 0; act_{max} = 1; rel_{min} = 0.9; rel_{max} = 1 to UDFG:

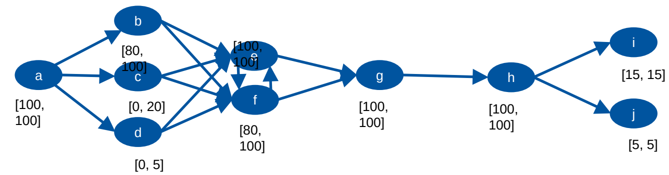


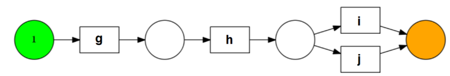
only the absolutely certain parts of the process are retained

$\frac{min_{a \to b}}{max_{a \to b}} = \frac{80}{100} = 0.8 < rel_{min} = 0.9$, thus remove a→b

- Preliminaries
- Process Discovery from Uncertain Event Data
- Examples and Results
- Conclusion

Conclusion





Summary:

- Uncertain event data
- Uncertain Directly-Follows-Graph (UDFG)
- Process discovery over a UDFG
- Keep or filter out the behavior of an uncertain event log

Future work:

- Computational cost analysis and performance optimization
- Definition of metrics and measures over uncertain event data
- Extensive experiments on real-world data

Thanks for your attention!

Any questions?

contact: pegoraro@pads.rwth-aachen.de

Dr.-Ing. Merih Seran Uysal
uysal@pads.rwth-aachen.de
www.pads.rwth-aachen.de

References

Marco Pegoraro and Wil M.P. van der Aalst, "Mining Uncertain Event Data in Process Mining," *2019 International Conference on Process Mining (ICPM)*, Aachen, Germany, 2019, pp. 89-96. doi: 10.1109/ICPM.2019.00023 .

Sander J.J. Leemans, Dirk Fahland, and Wil M.P. van der Aalst. "*Scalable process discovery and conformance checking.*" Software & Systems Modeling 17.2 (2018): 599-631.