Enhancing the Applicability of the eST-Miner: Efficient Precision-Guided Implicit Place Avoidance

Felix C. Groß, Lisa L. Mannel, Wil M.P. van der Aalst

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Introduction

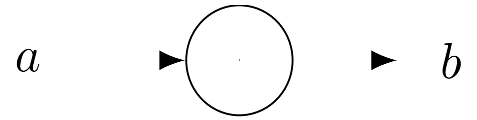
Idea: eST-Miner

Input: L event log, $\tau \in [0,1]$ noise threshold

Output: Petri net

1. Start with an “empty” Petri net P that consists of
   – A dedicated source place
   – A dedicated target place
   – A transition for each activity contained in the event log
2. Consider ***all possible places*** individually
3. Evaluate using token-based replay and add the ***fitting*** ones to P

$$L = [\langle \blacktriangleright, a, \blacksquare \rangle, \langle \blacktriangleright, b, \blacksquare \rangle]$$



Candidate Tree: $CT_{\{\blacktriangleright, a, b, \blacksquare\}}$

$(\blacktriangleright|a)$ $(\blacktriangleright|b)$ $(\blacktriangleright|\blacksquare)$ $(a|a)$ $(a|b)$ $(a|\blacksquare)$ $(b|a)$ $(b|b)$ $(b|\blacksquare)$

$(\blacktriangleright|a,b)$ $(\blacktriangleright|a,\blacksquare)$ $(\blacktriangleright|b,\blacksquare)$ $(a|a,b)$ $(a|a,\blacksquare)$ $(a|b,\blacksquare)$ $(b|a,b)$ $(b|a,\blacksquare)$ $(b|b,\blacksquare)$ $(\blacktriangleright,a|a)$ $(\blacktriangleright,b|a)$ $(a,b|a)$ $(\blacktriangleright,a|b)$ $(\blacktriangleright,b|b)$ $(a,b|b)$ $(\blacktriangleright,a|\blacksquare)$ $(\blacktriangleright,b|\blacksquare)$ $(a,b|\blacksquare)$

$(\blacktriangleright|a,b,\blacksquare)$ $(a|a,b,\blacksquare)$ $(b|a,b,\blacksquare)$ $(\blacktriangleright,a|a,b)$ $(\blacktriangleright,a|a,\blacksquare)$ $(\blacktriangleright,a|b,\blacksquare)$ $(\blacktriangleright,b|a,b)$ $(\blacktriangleright,b|a,\blacksquare)$ $(\blacktriangleright,b|b,\blacksquare)$ $(a,b|a,b)$ $(a,b|a,\blacksquare)$ $(a,b|b,\blacksquare)$ $(\blacktriangleright,a,b|a)$ $(\blacktriangleright,a,b|b)$ $(\blacktriangleright,a,b|\blacksquare)$

$(\blacktriangleright,a|a,b,\blacksquare)$ $(\blacktriangleright,b|a,b,\blacksquare)$ $(\blacktriangleright,\blacksquare|a,b,\blacksquare)$ $(\blacktriangleright,a,b|a,b)$ $(\blacktriangleright,a,b|a,\blacksquare)$ $(\blacktriangleright,a,b|b,\blacksquare)$

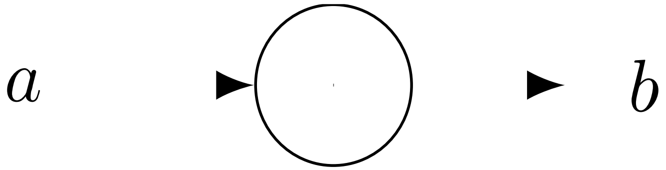
$(\blacktriangleright,a,b|a,b,\blacksquare)$

⇒ Traverse the candidate tree in BFS (evaluate each place using TBR)

Traversal Strategy

$$L = [\underbrace{\langle \blacktriangleright, a, \blacksquare \rangle}_{\sigma_1}, \underbrace{\langle \blacktriangleright, b, \blacksquare \rangle}_{\sigma_2}]$$

$$p = (a|b)$$

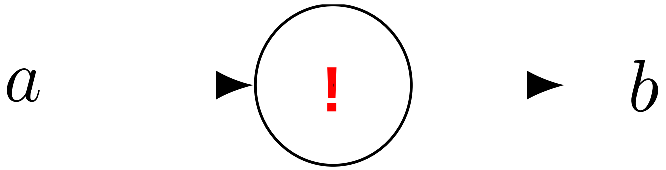


Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Traversal Strategy

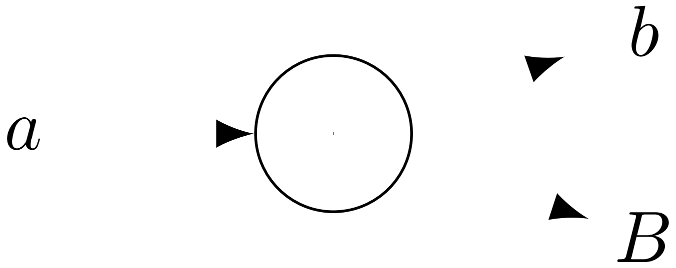
$$L = [\underbrace{\langle \blacktriangleright, a, \blacksquare \rangle}_{\sigma_1}, \underbrace{\langle \blacktriangleright, b, \blacksquare \rangle}_{\sigma_2}]$$

$p = (a|b)$



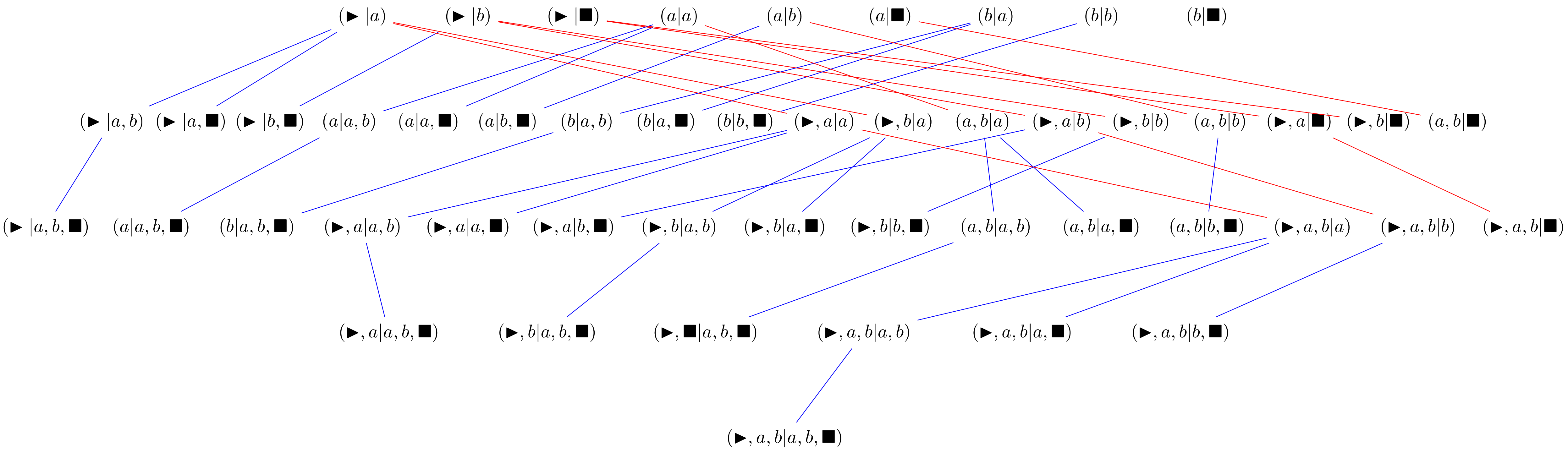
p is ***underfed*** concerning σ_2

⇒ all places of the form

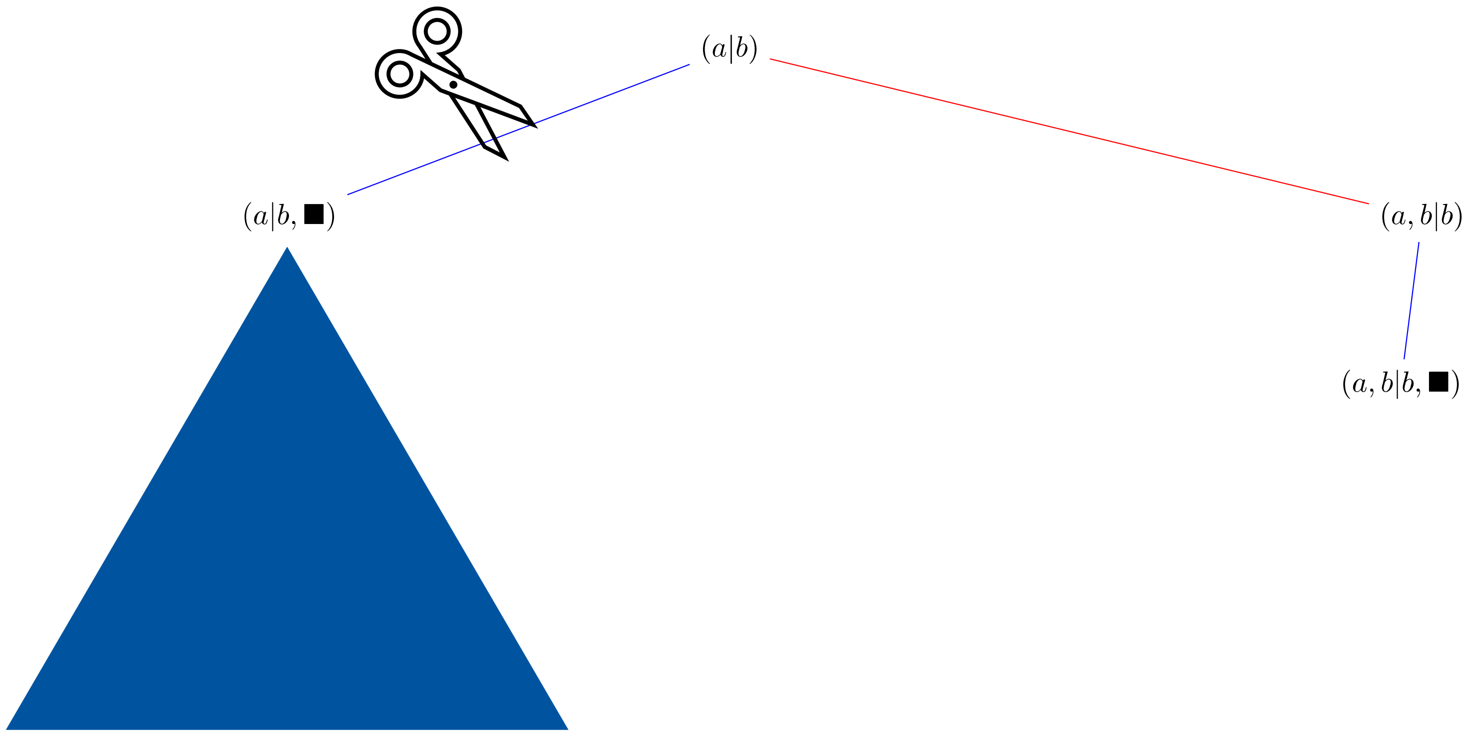


are underfed concerning σ_2 as well

Traversal Strategy



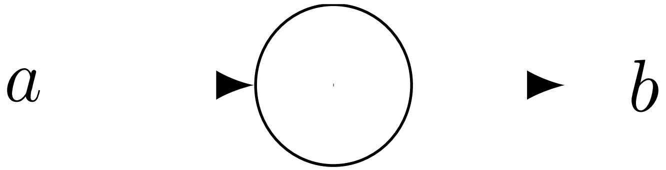
Traversal Strategy



Traversal Strategy

$$L = [\underbrace{\langle \blacktriangleright, a, \blacksquare \rangle}_{\sigma_1}, \underbrace{\langle \blacktriangleright, b, \blacksquare \rangle}_{\sigma_2}]$$

$$p = (a|b)$$

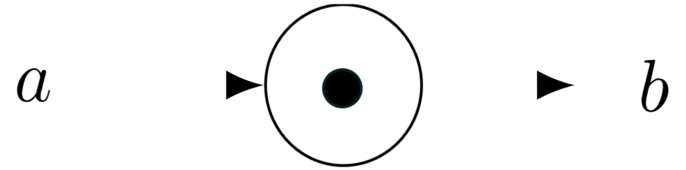


Traversal Strategy

$$L = [\langle \blacktriangleright, a, \blacksquare \rangle, \langle \blacktriangleright, b, \blacksquare \rangle]$$

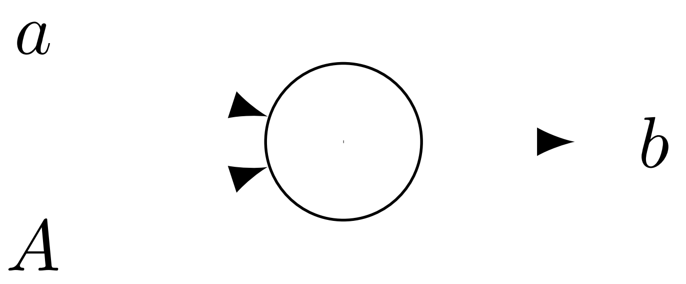
σ_1 σ_2

$p = (a|b)$



p is ***overfed*** concerning σ_1

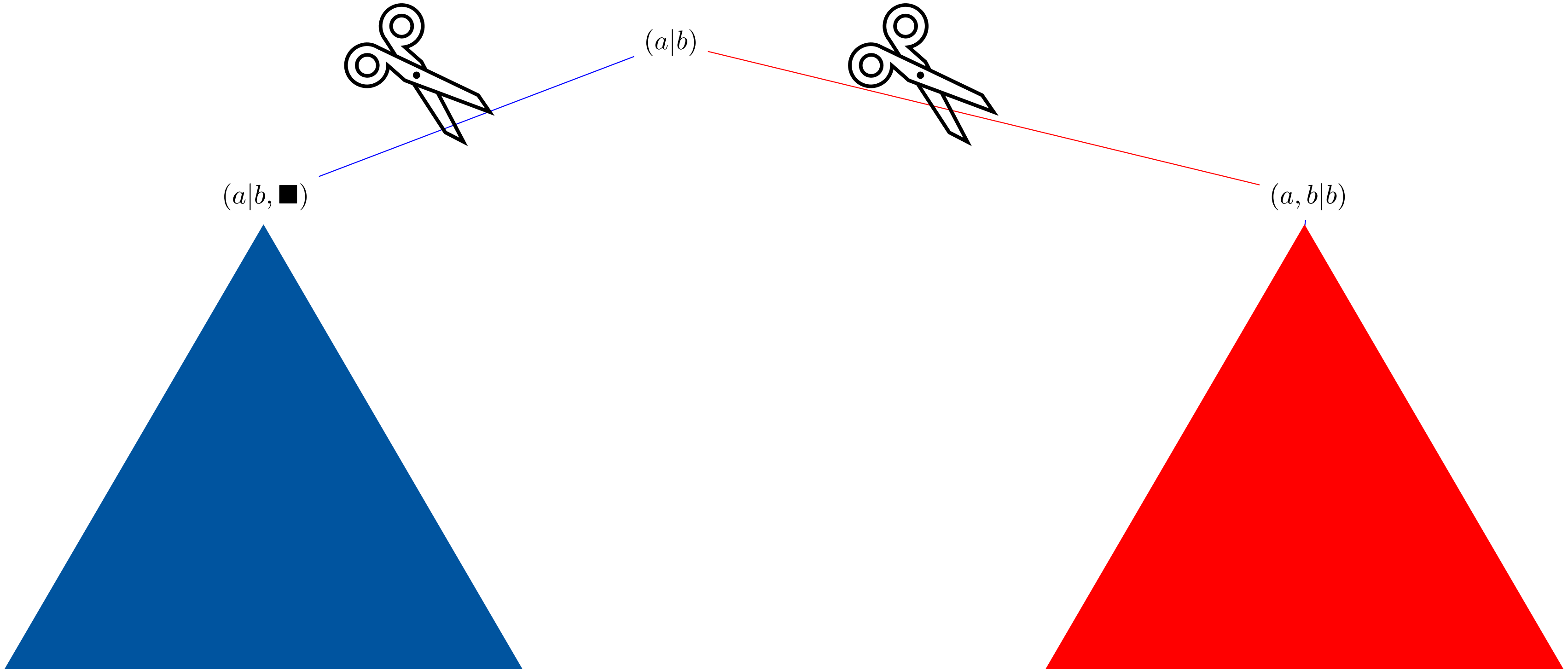
⇒ all places of the form



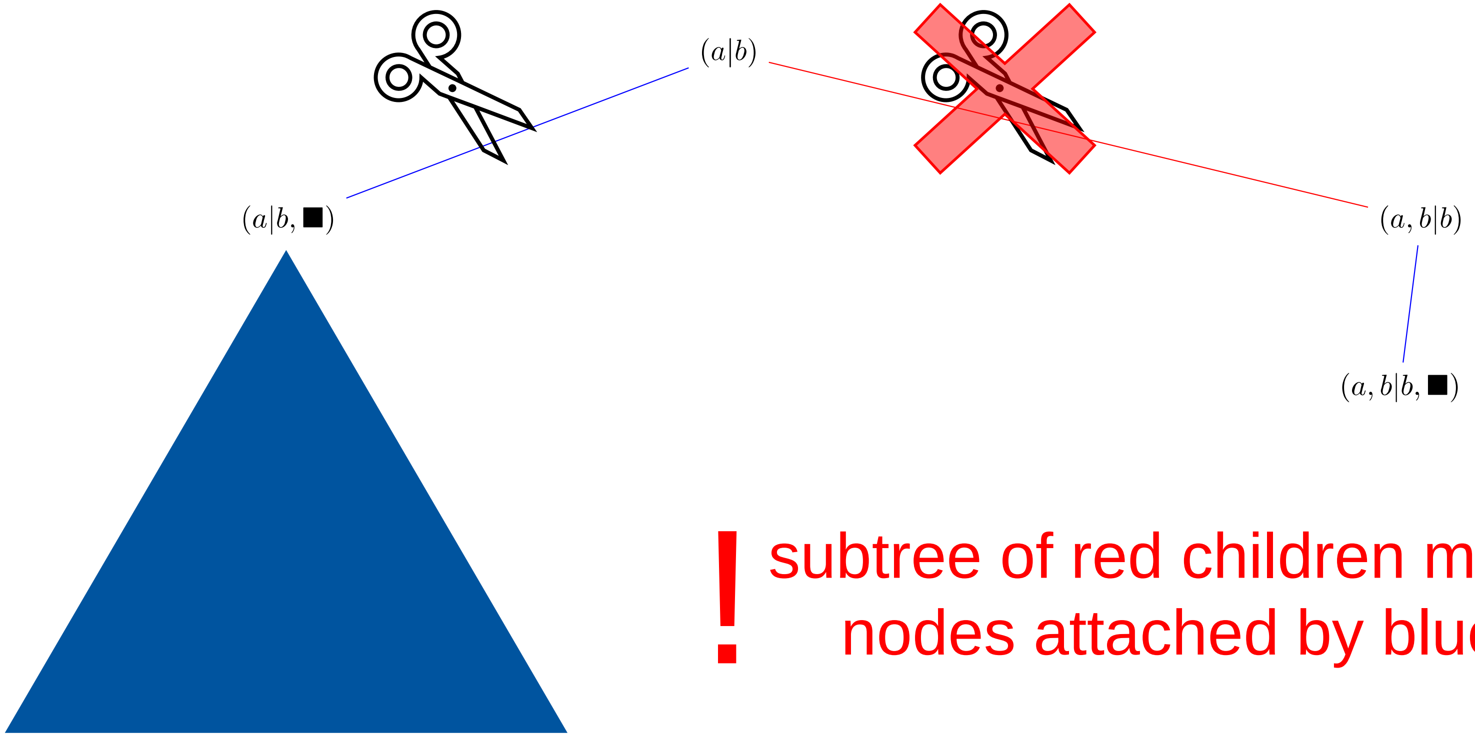
are overfed concerning σ_1 as well

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Traversal Strategy



Traversal Strategy



! subtree of red children may contain nodes attached by blue edges !

Handling Noise

local noise threshold $\tau \in [0,1]$

⬇

store places that fit at least $\tau \cdot |L|$ traces

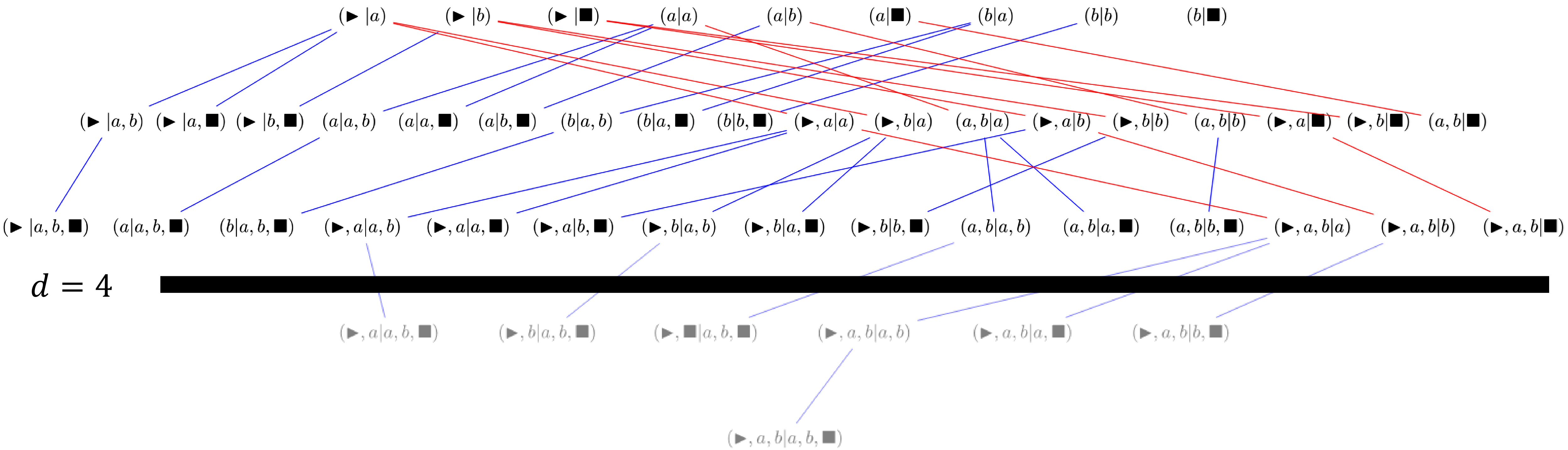
- $\tau = 1$ resulting Petri net can perfectly replay L
- $\tau < 1$ fitting places may contradict each other $\Rightarrow$ ***deadlocks***

δ-Variant:

- Only inserts places that do not decrease the fraction of replayable traces by more than $\delta \in [0,1]$
- guarantees that the resulting Petri net can replay at least $\tau \cdot |L|$ traces

Limiting the Search Space

maximum depth d #connecting activities

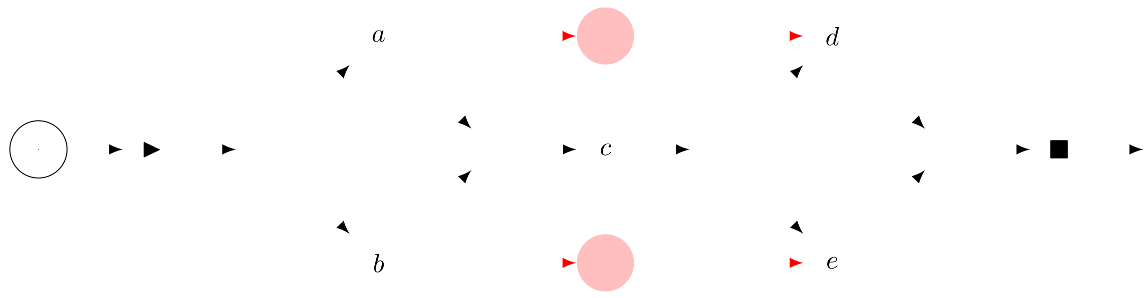


- Improves runtime significantly
- Tradeoff between simplicity, fitness, precision and generalization

Why bother?

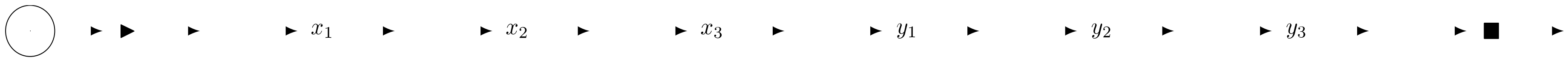
- Find complex control-flow structures (e.g., long-term dependencies)

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$



- Filter noise and infrequent behavior

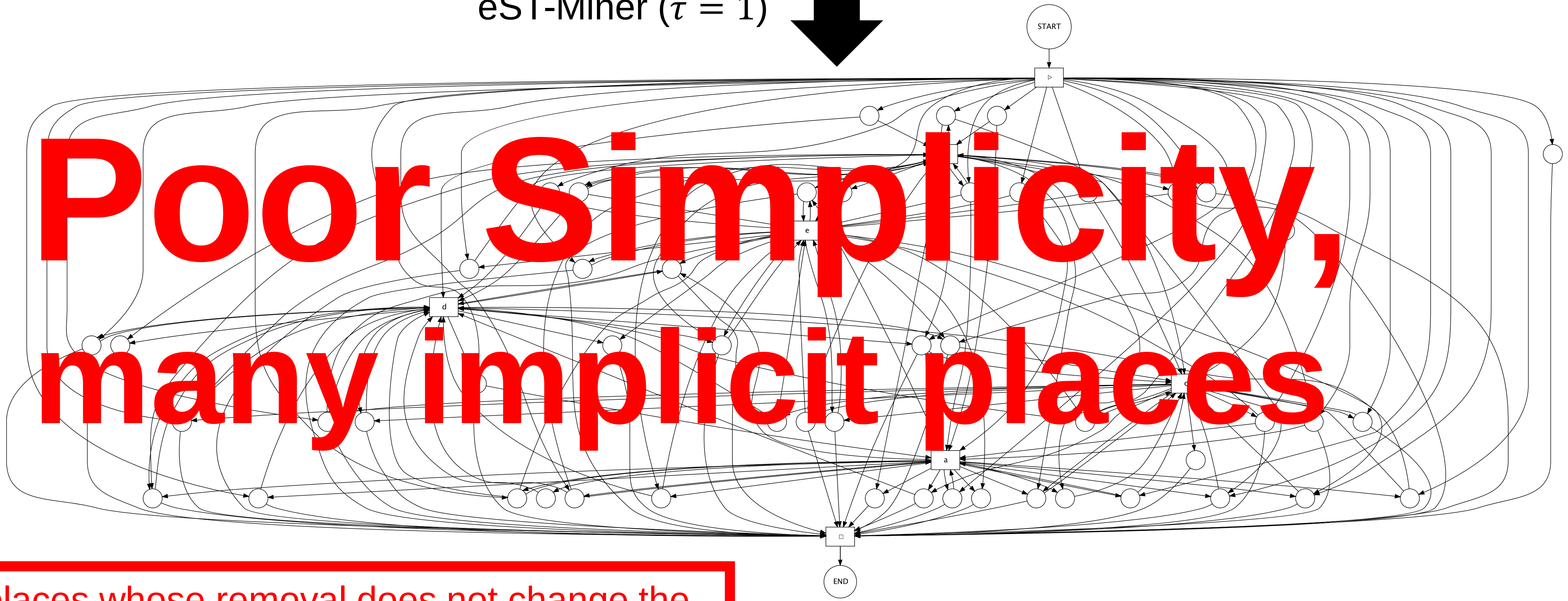
$$\begin{aligned} L = [&\langle \blacktriangleright, x_1, x_2, x_3, \cancel{y_3, y_2, y_1}, \blacksquare \rangle^1, \langle \blacktriangleright, \cancel{x_3, x_2, x_1}, y_1, y_2, y_3, \blacksquare \rangle^1, \\ &\langle \blacktriangleright, x_1, x_2, x_3, \cancel{y_3, y_1, y_2}, \blacksquare \rangle^1, \langle \blacktriangleright, \cancel{x_3, x_1, x_2}, y_1, y_2, y_3, \blacksquare \rangle^1, \\ &\langle \blacktriangleright, x_1, x_2, x_3, \cancel{y_2, y_3, y_1}, \blacksquare \rangle^1, \langle \blacktriangleright, \cancel{x_2, x_3, x_1}, y_1, y_2, y_3, \blacksquare \rangle^1, \\ &\langle \blacktriangleright, x_1, x_2, x_3, \cancel{y_2, y_1, y_3}, \blacksquare \rangle^1, \langle \blacktriangleright, \cancel{x_2, x_1, x_3}, y_1, y_2, y_3, \blacksquare \rangle^1, \\ &\langle \blacktriangleright, x_1, x_2, x_3, \cancel{y_1, y_3, y_2}, \blacksquare \rangle^1, \langle \blacktriangleright, \cancel{x_1, x_3, x_2}, y_1, y_2, y_3, \blacksquare \rangle^1] \end{aligned}$$



Motivation

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

eST-Miner ($\tau = 1$)



Poor Simplicity, many implicit places

places whose removal does not change the language of the net

How the eST-Miner currently deals with implicit places:

(user-definable) noise threshold parameter τ

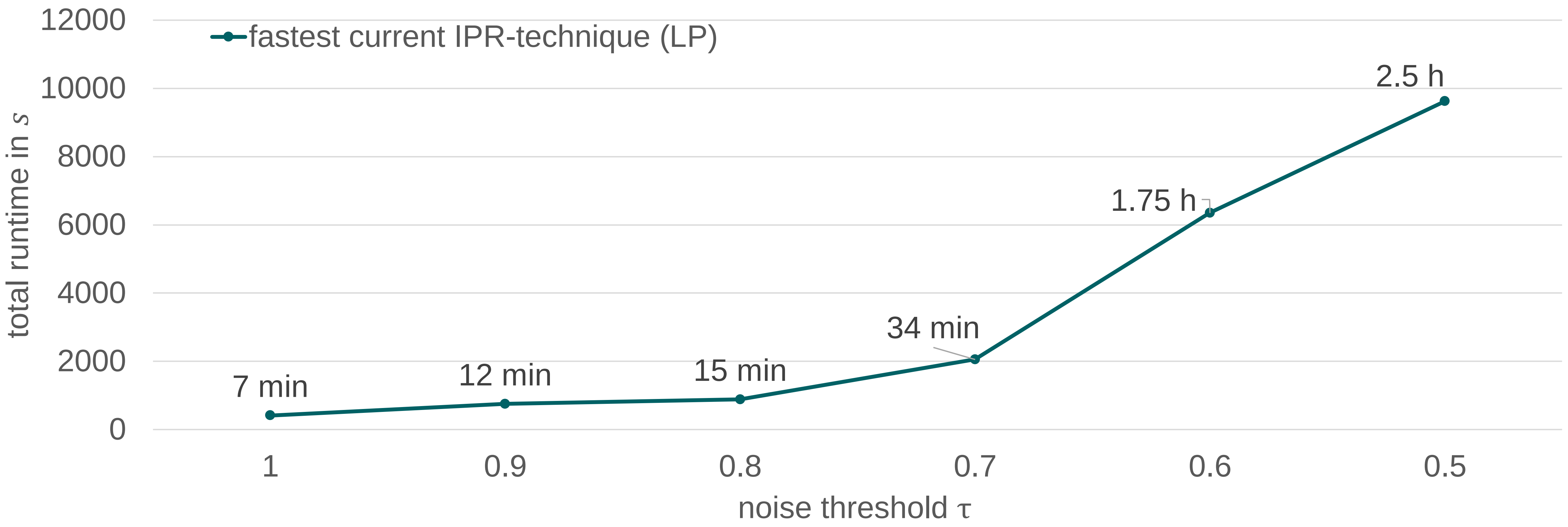
$\tau = 1.0$ → region-based

$\tau \leq 1.0$ → LP-based

Current Applicability of the eST-Miner

max. depth $d = 7$

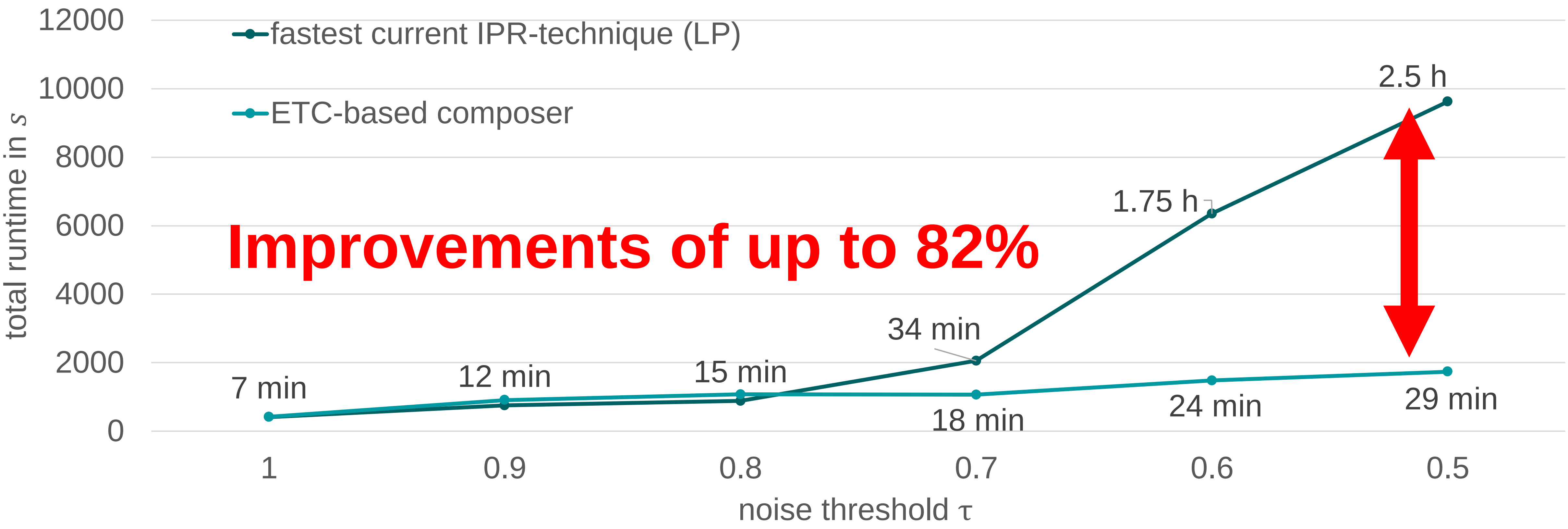
event log	#activities	#traces	#trace variants
Sepsis	18	1050	846



Current Applicability of the eST-Miner

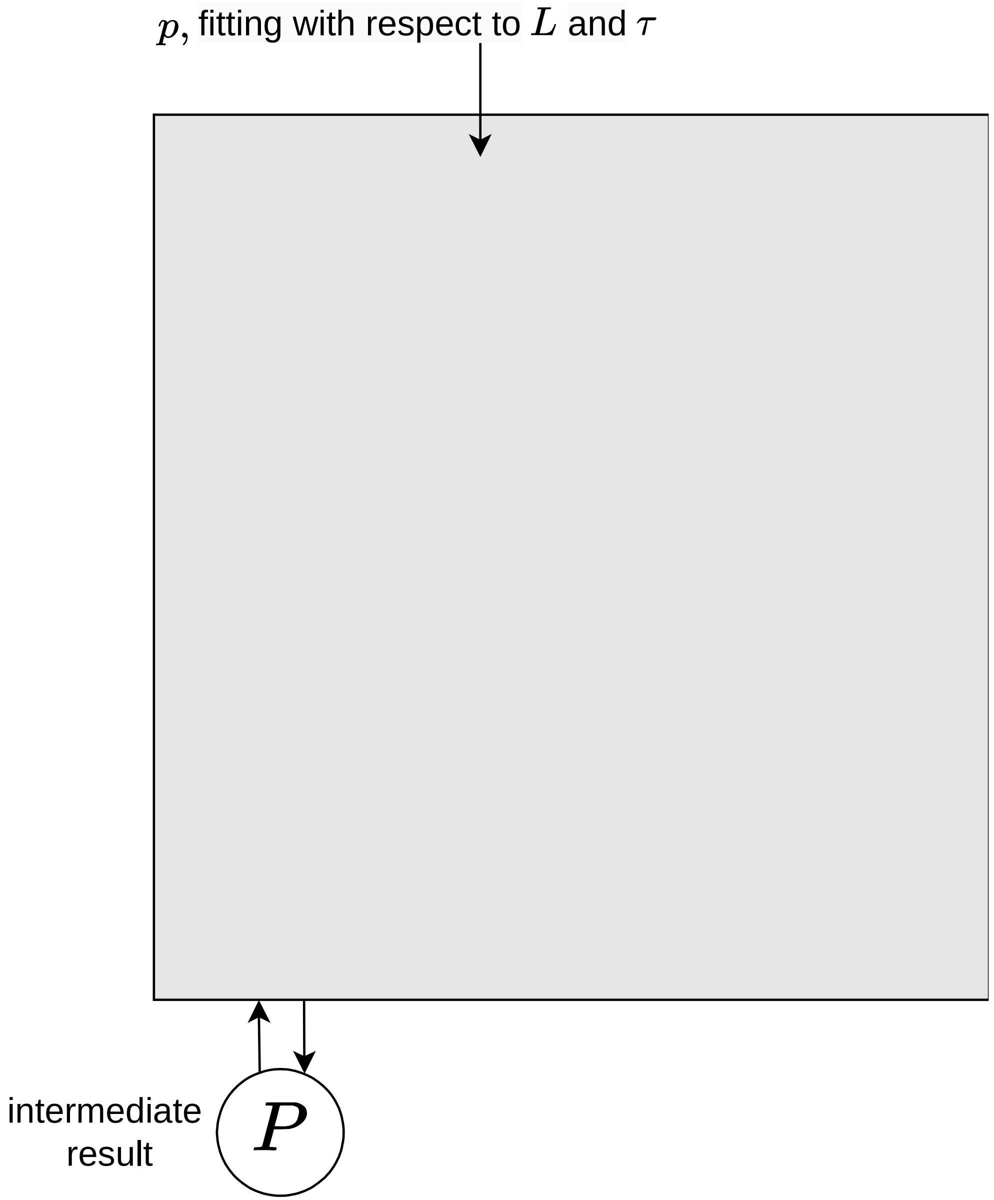
max. depth $d = 7$

event log	#activities	#traces	#trace variants
Sepsis	18	1050	846

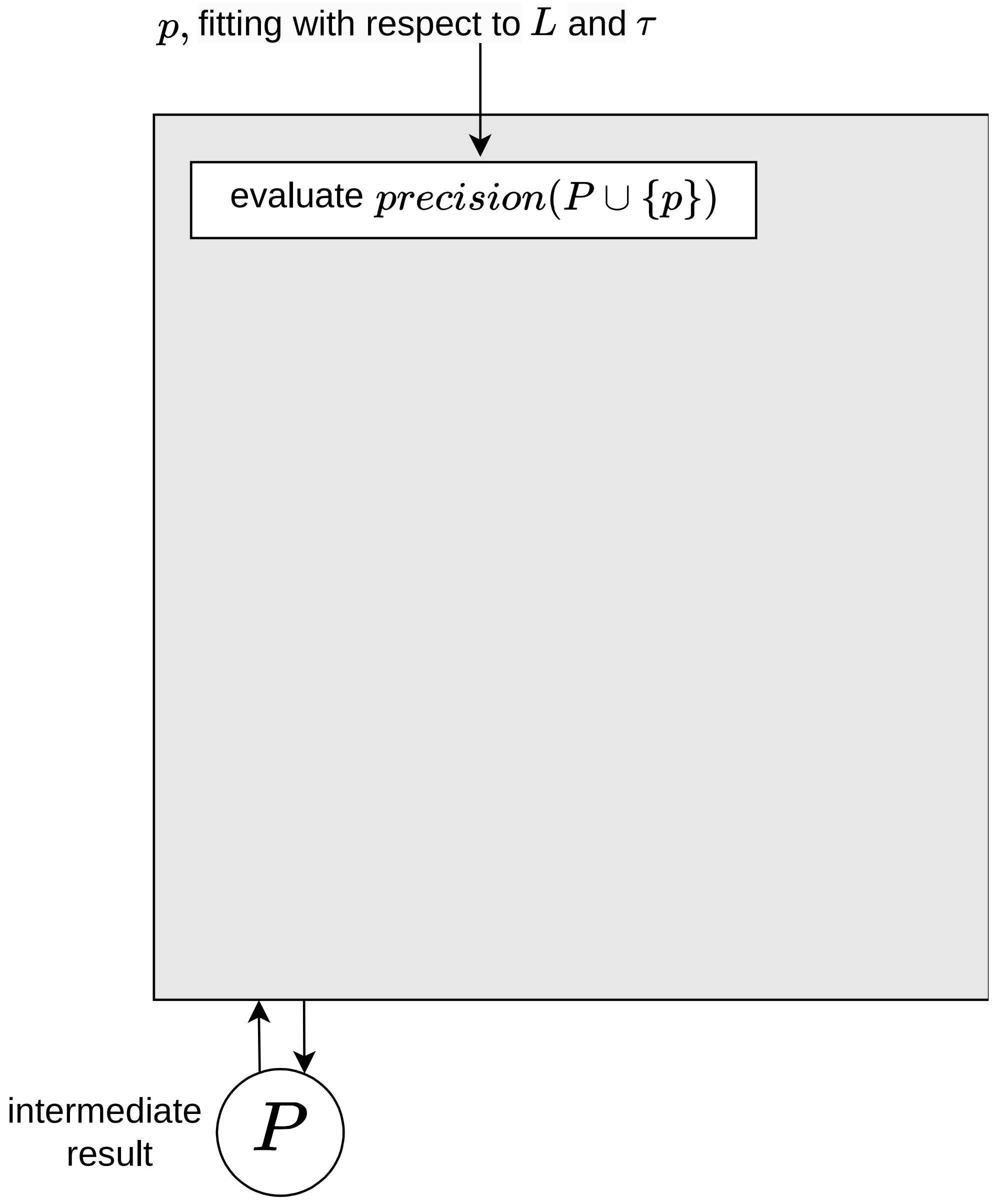


Precision-Guided Implicit Place Avoidance

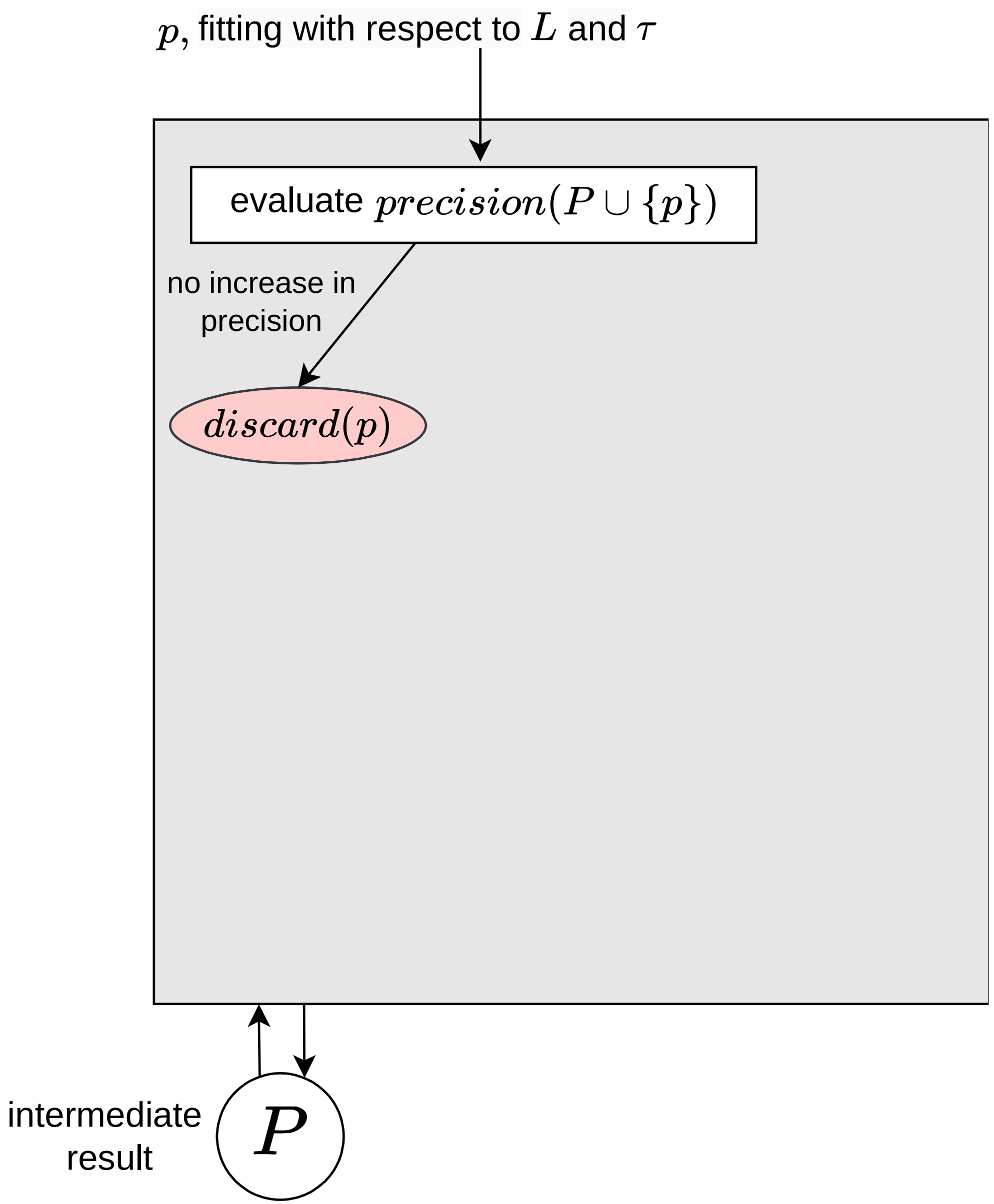
Algorithmic Framework



Algorithmic Framework

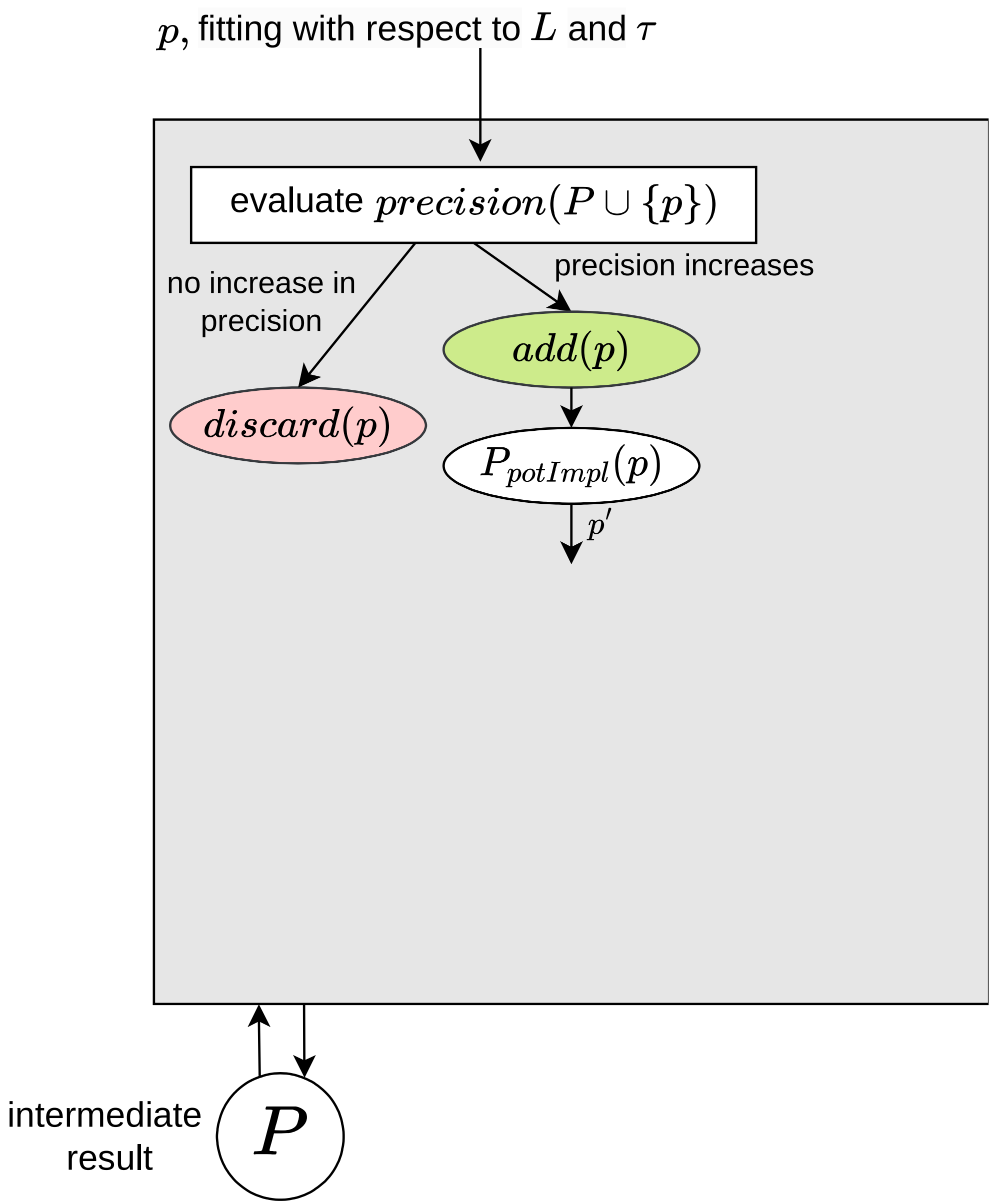


Algorithmic Framework



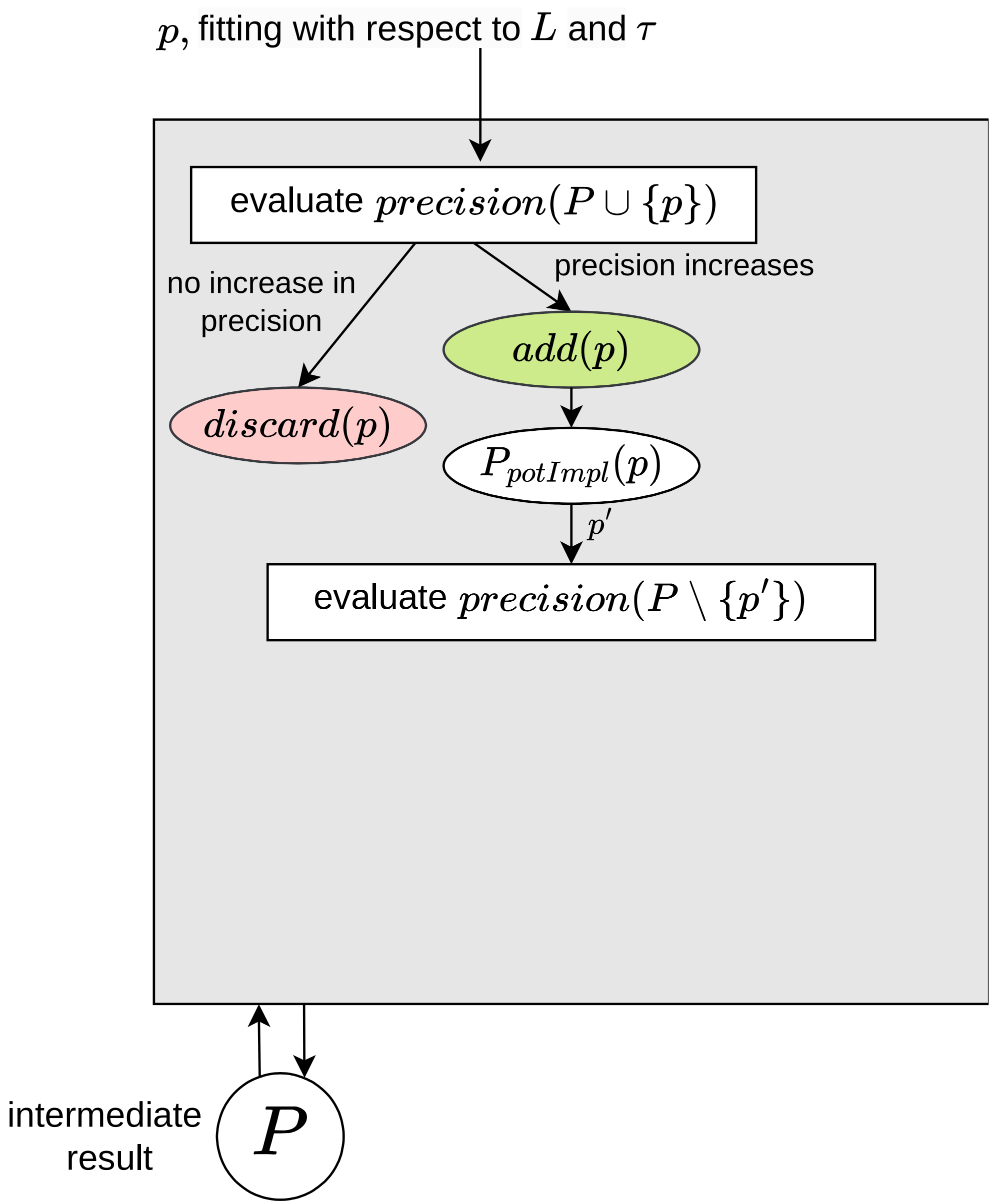
Algorithmic Framework

$$P_{potImpl}\big((I|O)\big) = \{(I'|O') \in P \mid O \cap O' \neq \emptyset\}$$



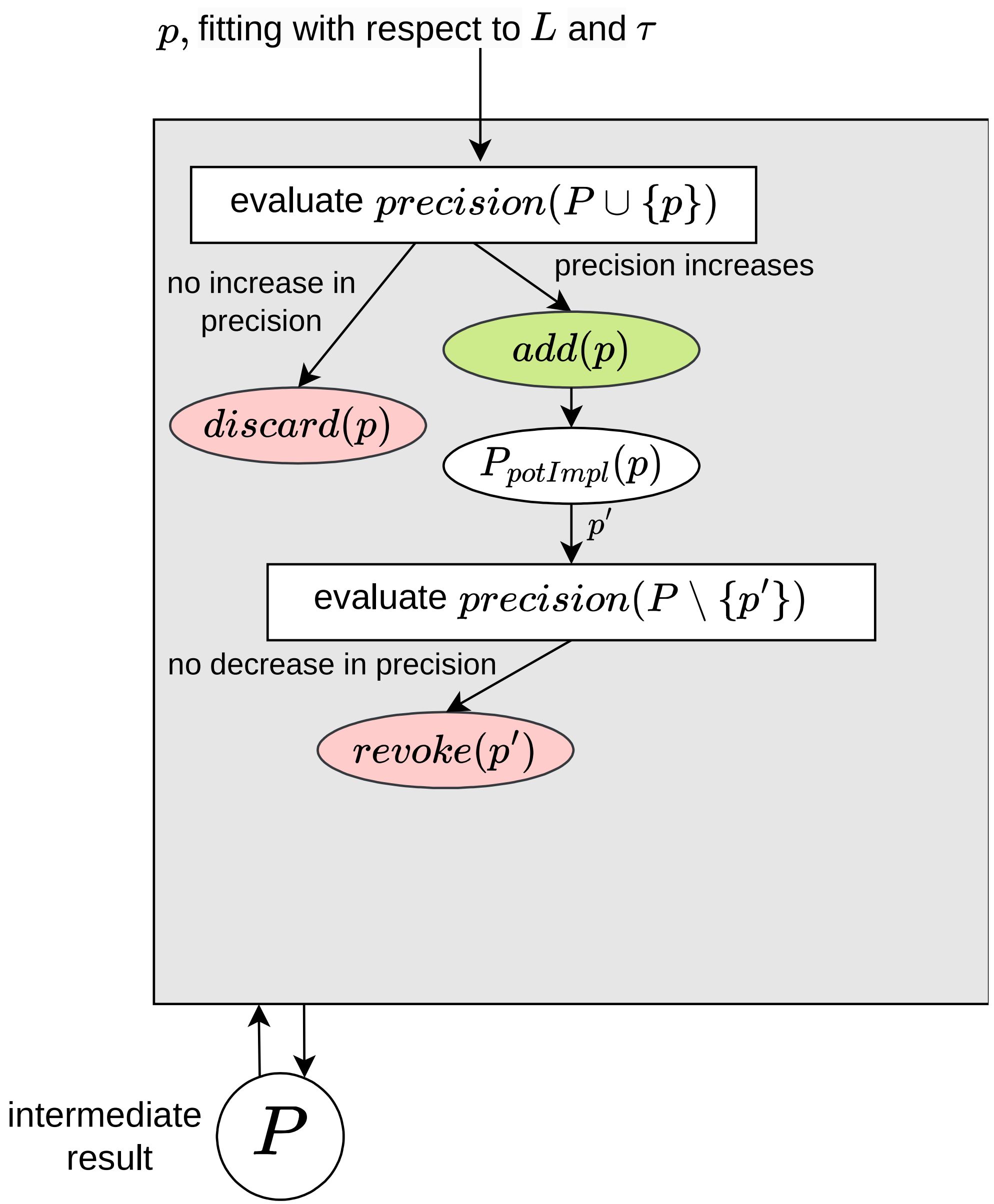
Algorithmic Framework

$$P_{potImpl}((I|O)) = \{(I'|O') \in P \mid O \cap O' \neq \emptyset\}$$



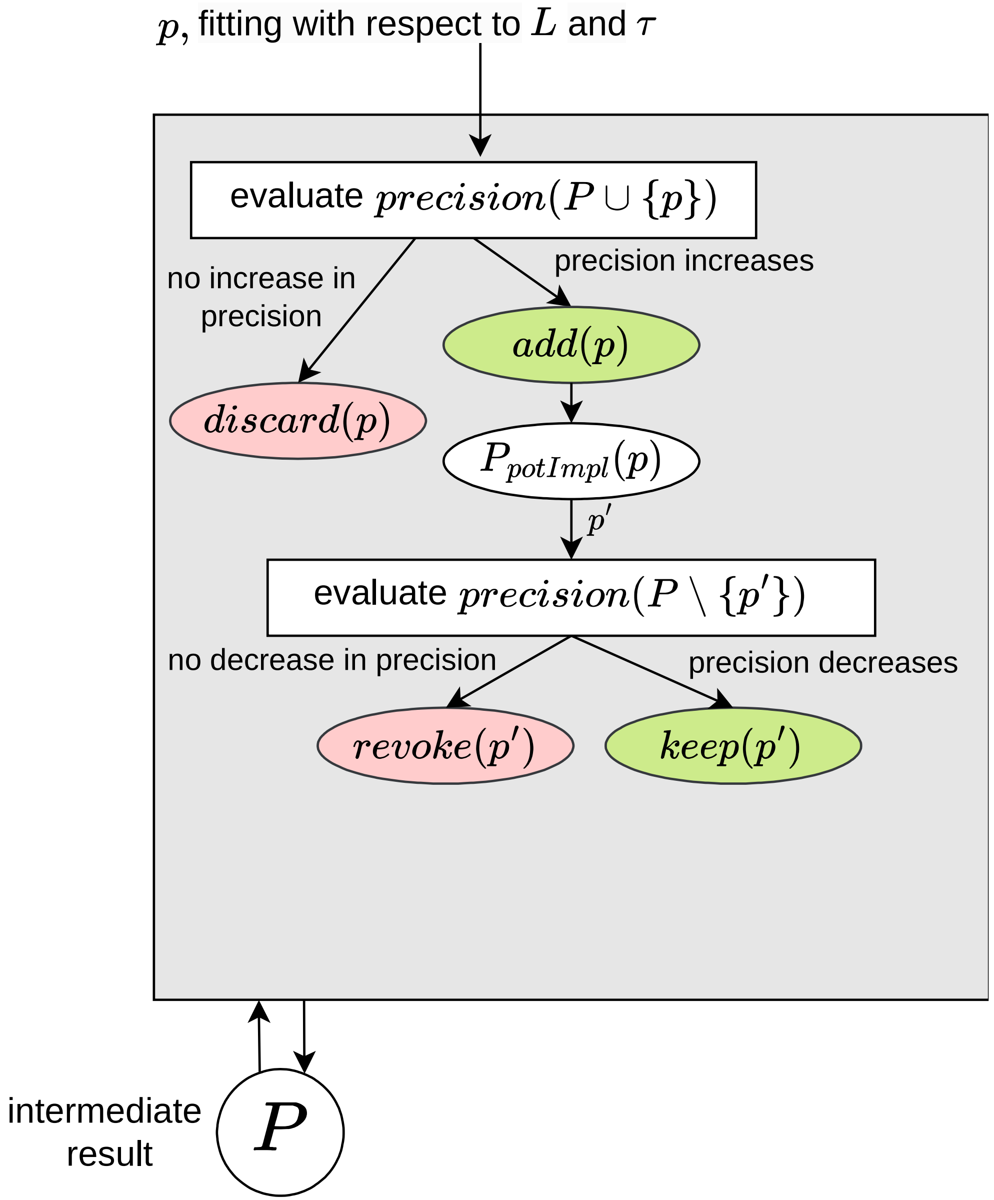
Algorithmic Framework

$$P_{potImpl}\big((I|O)\big) = \{(I'|O') \in P \mid O \cap O' \neq \emptyset\}$$



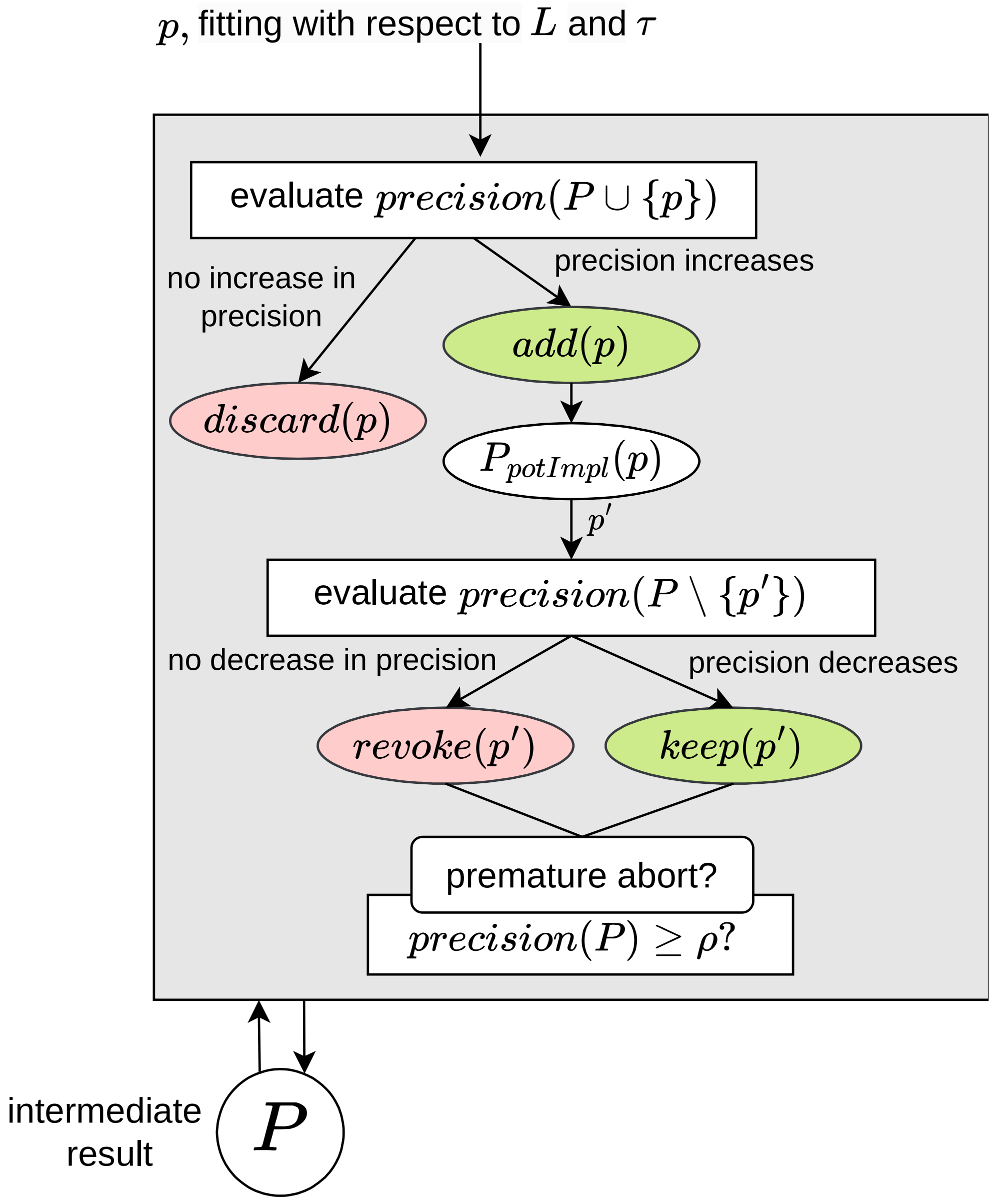
Algorithmic Framework

$$P_{potImpl}((I|O)) = \{(I'|O') \in P \mid O \cap O' \neq \emptyset\}$$



Algorithmic Framework

$$P_{potImpl}\big((I|O)\big) = \{(I'|O') \in P \mid O \cap O' \neq \emptyset\}$$



Example

Event log $L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

Noise threshold $\tau = 1$

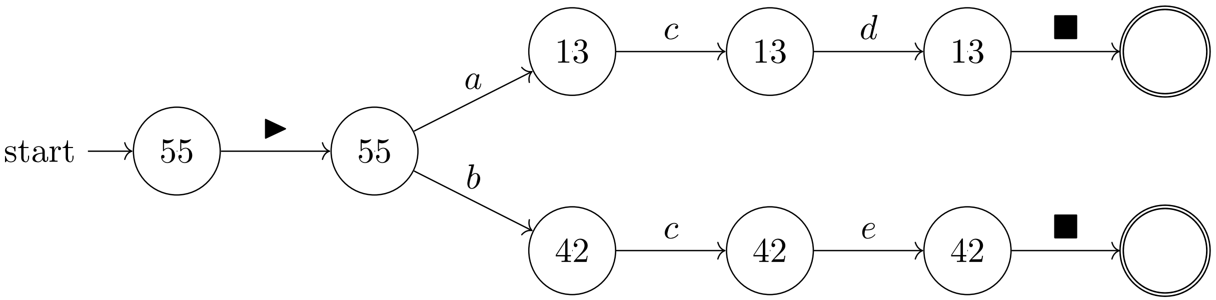
Precision abort threshold $\rho = 1$

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Example

$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

1. Build Prefix Automaton and Activity Mapping



$\cdot$	$A(\cdot)$	$E(\cdot)$
a		
b		
c		
d		
e		
$\blacksquare$		

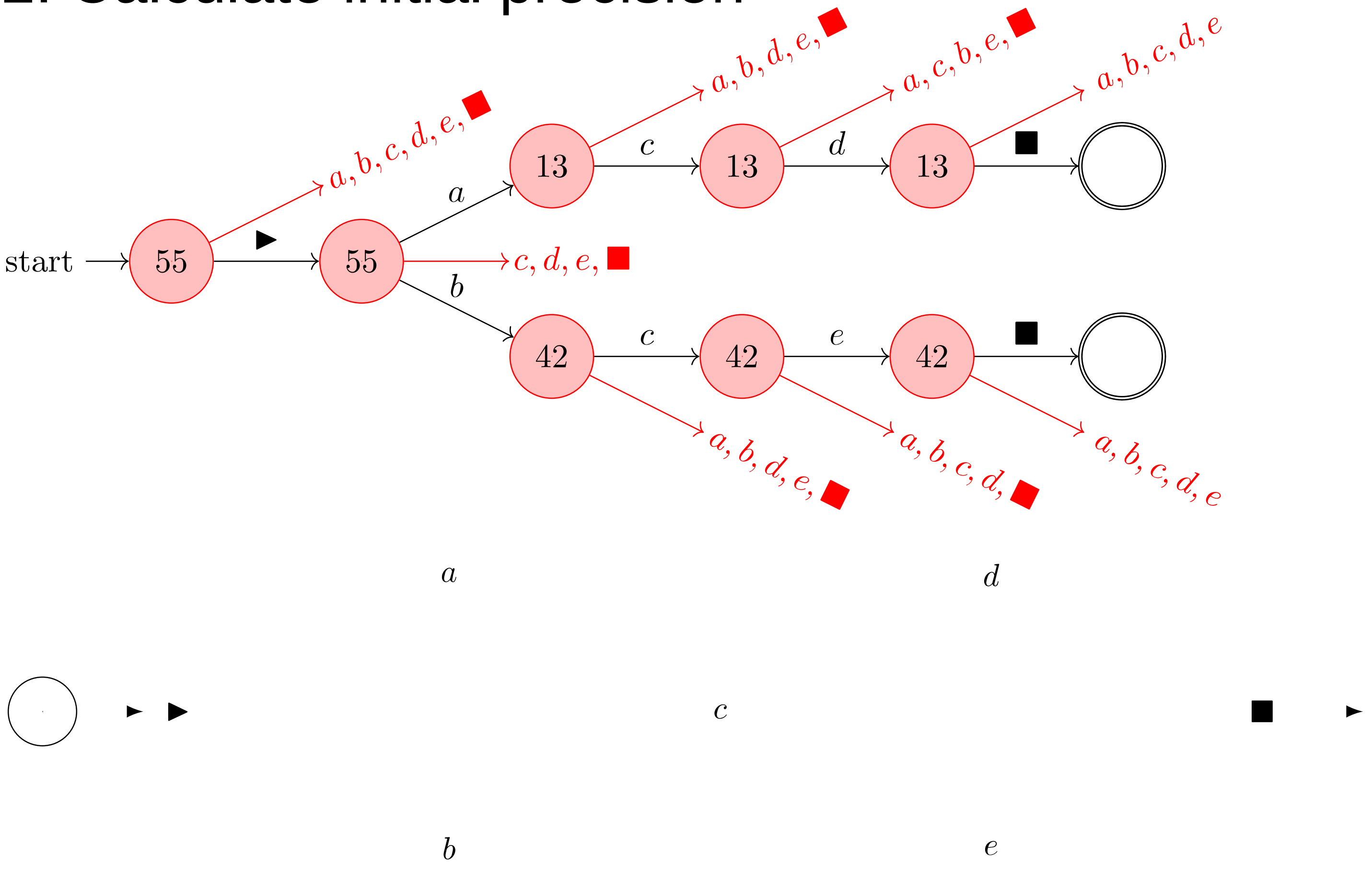
$A(\cdot)$: number of times $\cdot$ is allowed (when replaying L)

$E(\cdot)$: number of times allowing $\cdot$ is escaping (when replaying L)

Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

2. Calculate initial precision



$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	275	220
d	275	262
e	275	233
■	275	220

$$precision = 1 - \frac{1375}{1650 + 55} \approx 0.194$$

$A(\blacktriangleright)$

Escaping Activities

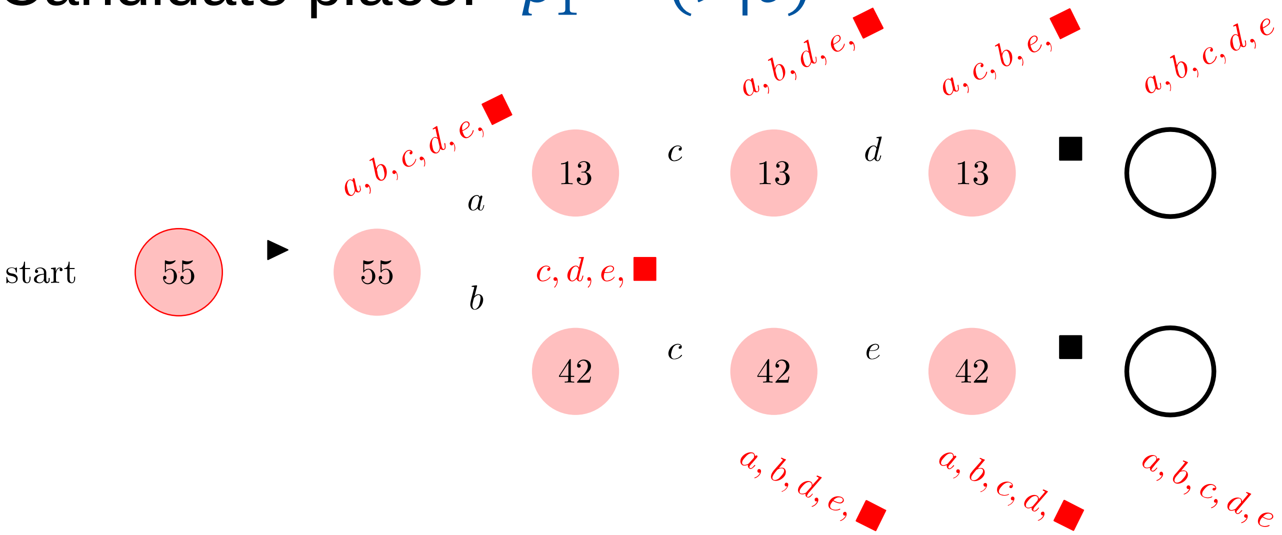
Allowed Activities

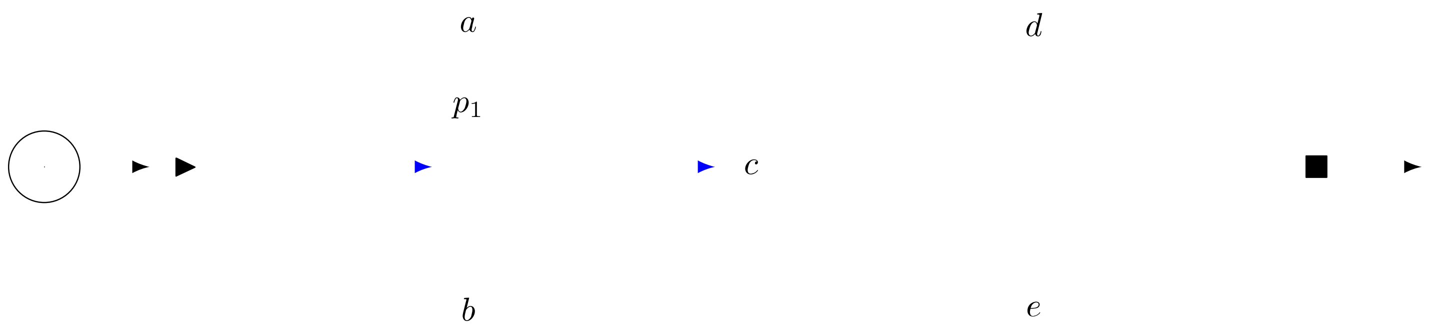
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	275	220
d	275	262
e	275	233
$\blacksquare$	275	220

Candidate place: $p_1 = (\blacktriangleright | c)$

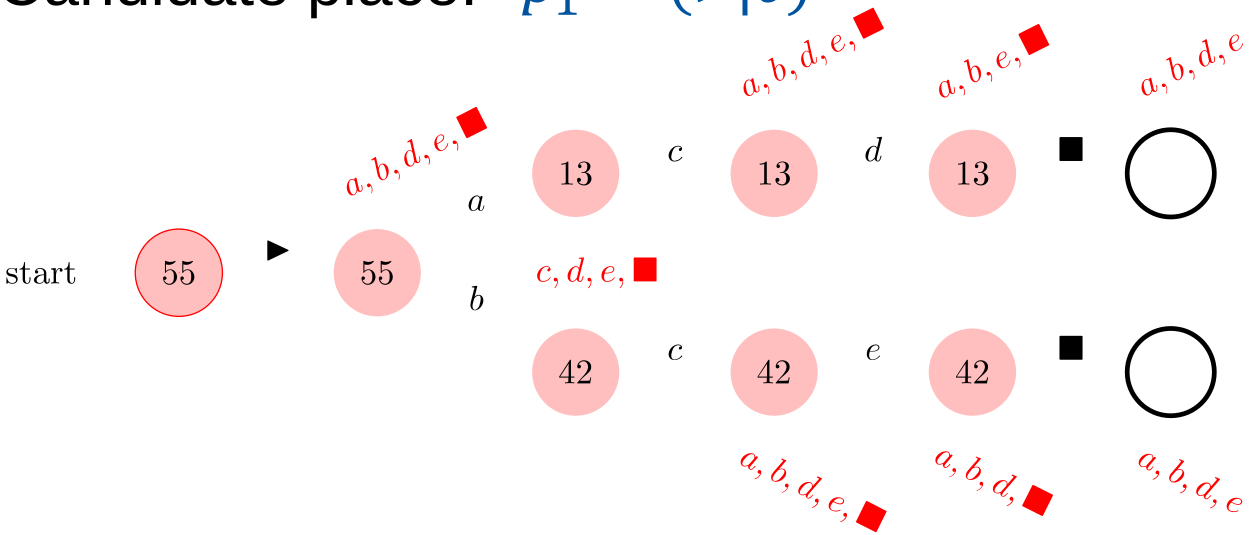




Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_1 = (\blacktriangleright | c)$

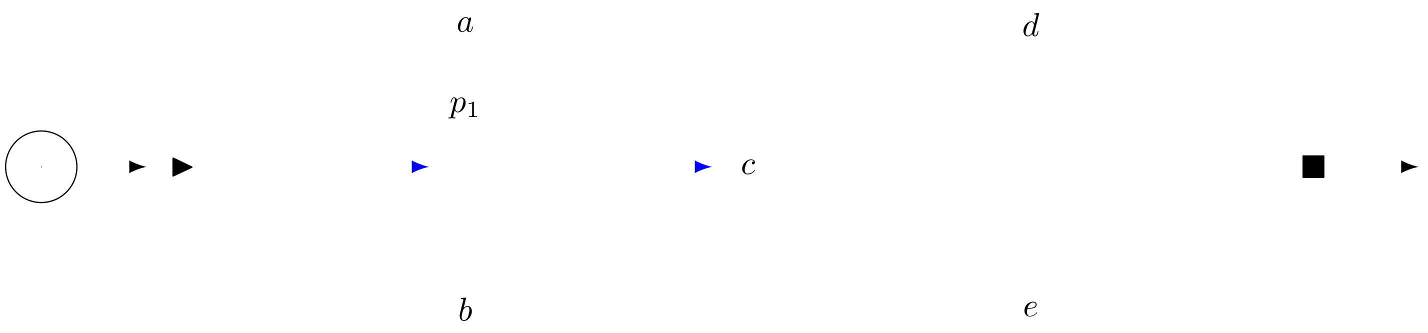


$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	~~275~~ 110	~~220~~ 55
d	275	262
e	275	233
■	275	220

$$precision(P) = 1 - \frac{1210}{1485+55} \approx 0.213 \; (> 0.194)$$

$\Rightarrow add(p_1)$

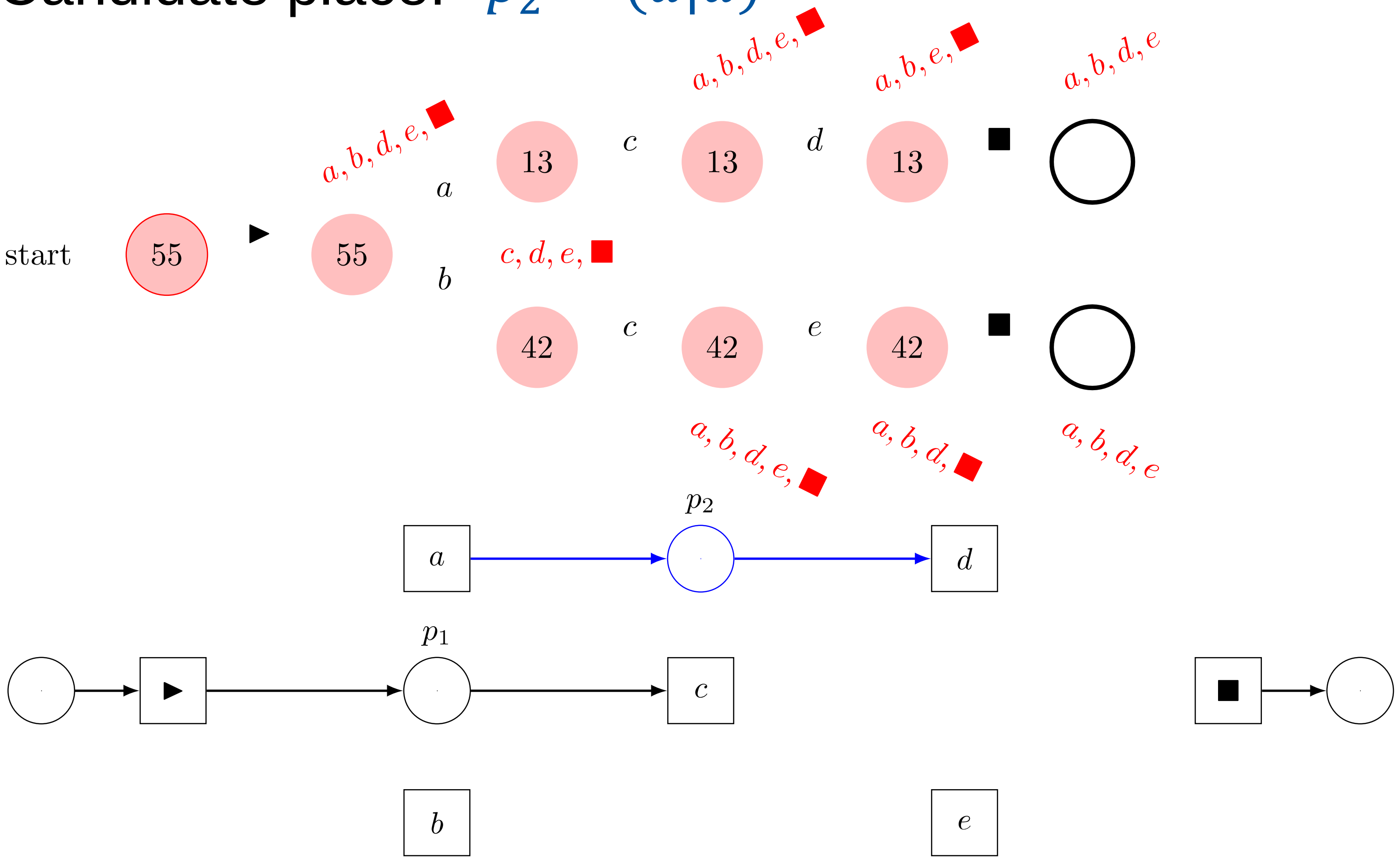
$P_{PotImpl}(p_1) = \emptyset$



Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_2 = (a|d)$

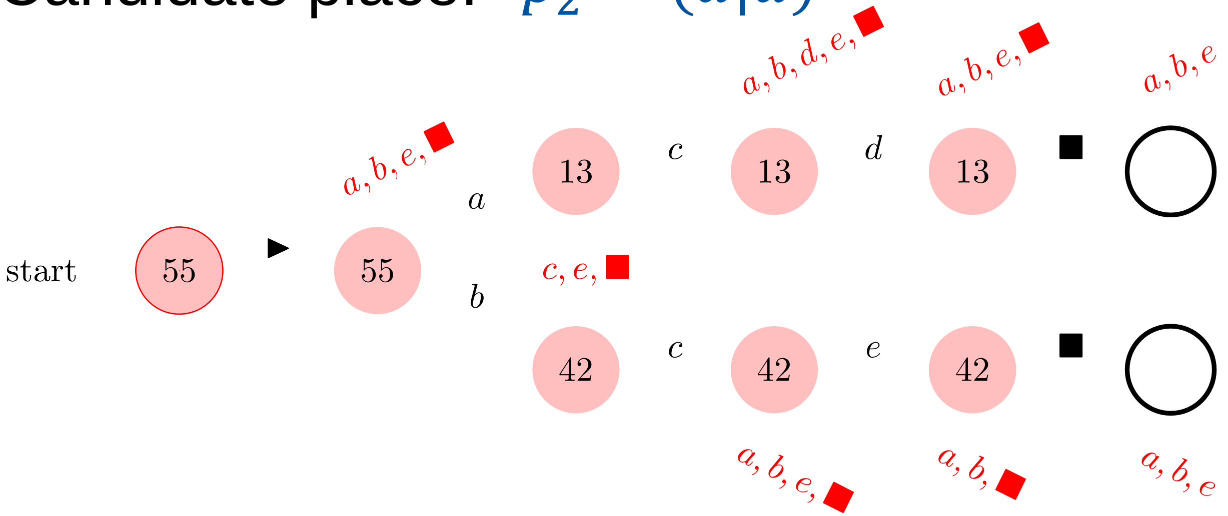


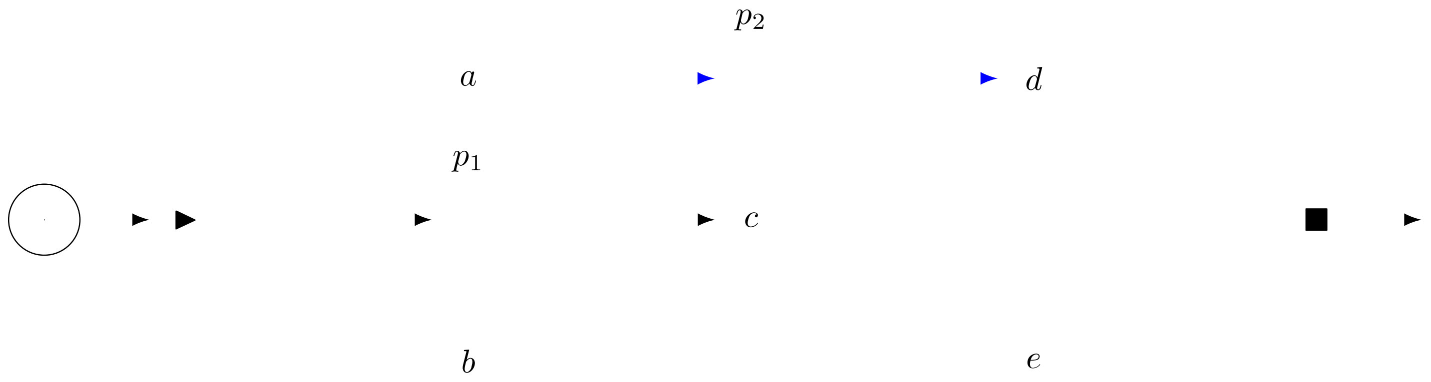
·	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	275	262
e	275	233
$\blacksquare$	275	220

Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare\rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare\rangle^{42}]$$

Candidate place: $p_2 = (a|d)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	~~275~~ 26	~~262~~ 13
e	275	233
■	275	220

$$precision(P) = 1 - \frac{961}{1236+55} \approx 0.256\ (> 0.213)$$

$\Rightarrow add(p_2)$

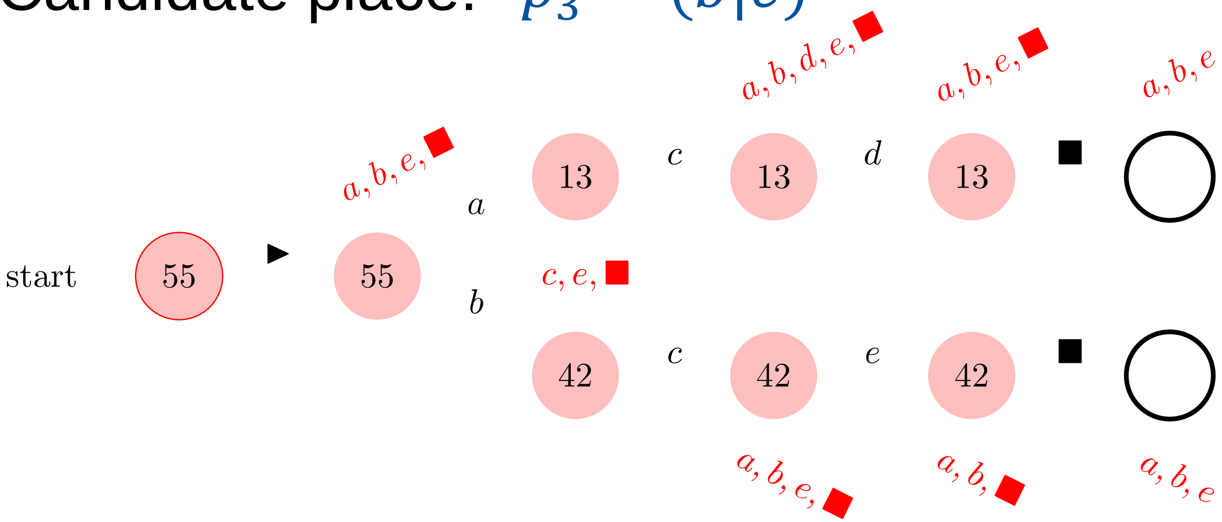
$$P_{PotImpl}(p_2) = \emptyset$$

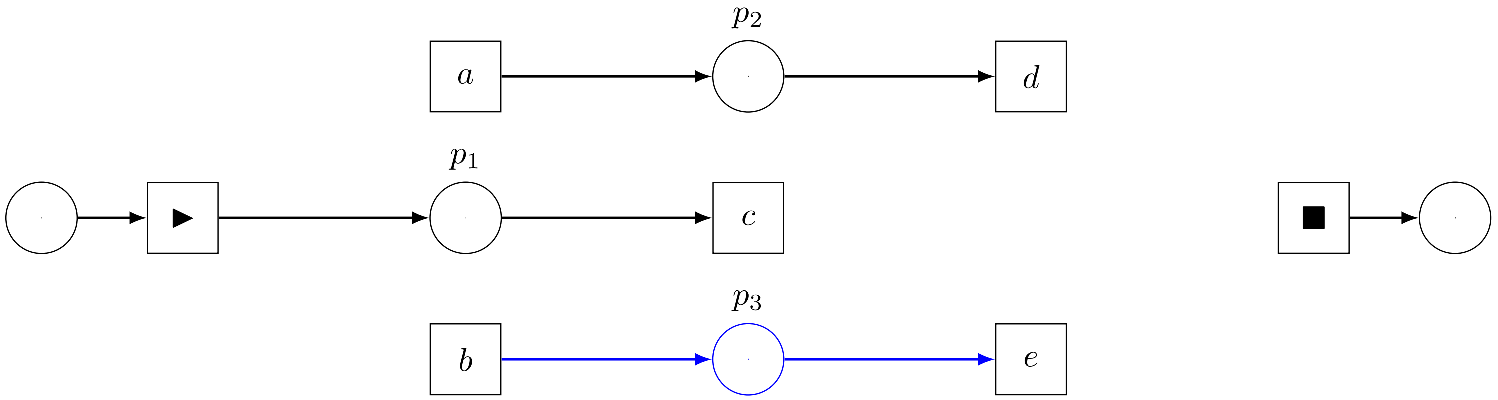
Example

$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	275	233
$\blacksquare$	275	220

Candidate place: $p_3 = (b|e)$



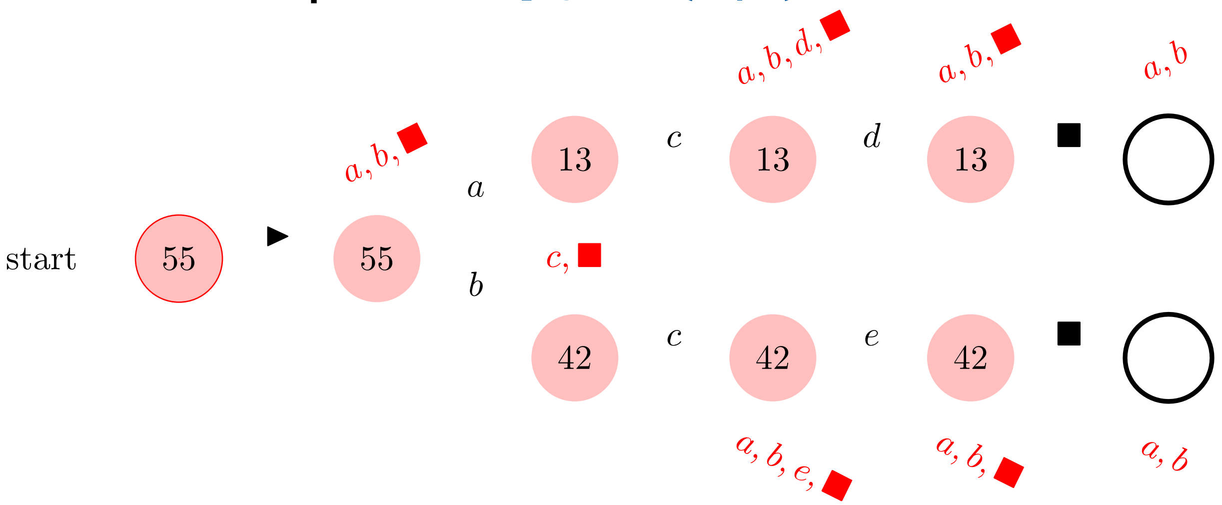


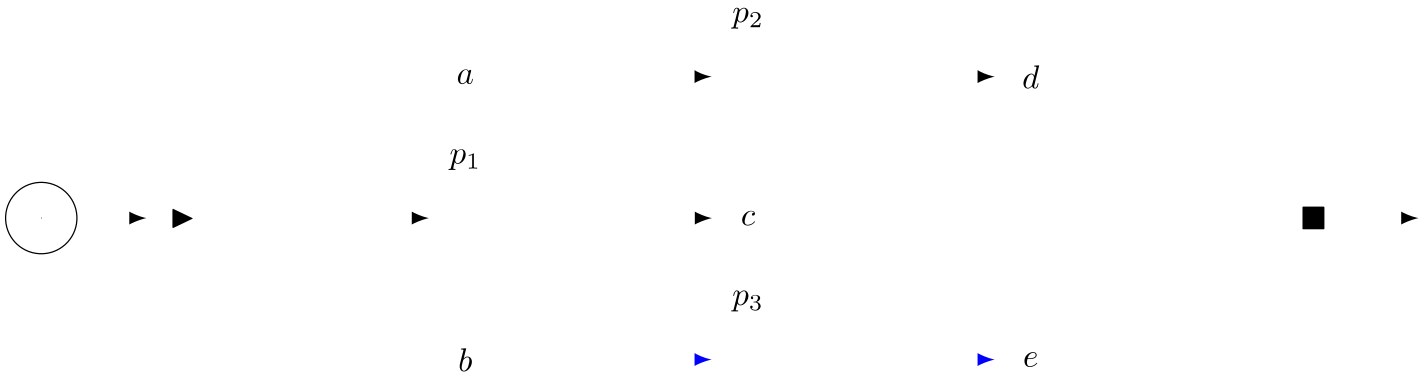
Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_3 = (b|e)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	~~275~~ 84	~~233~~ 42
$\blacksquare$	275	220

$$precision(P) = 1 - \frac{770}{1045+55}$$
$$= 0.3\ (> 0.256)$$

$\Rightarrow add(p_3)$

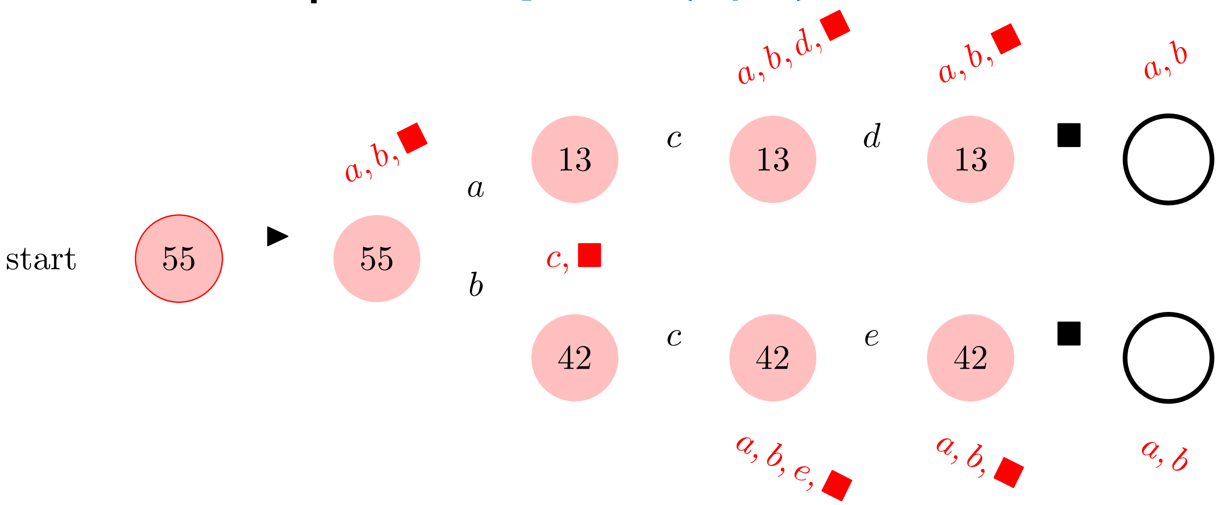
$P_{PotImpl}(p_3) = \emptyset$

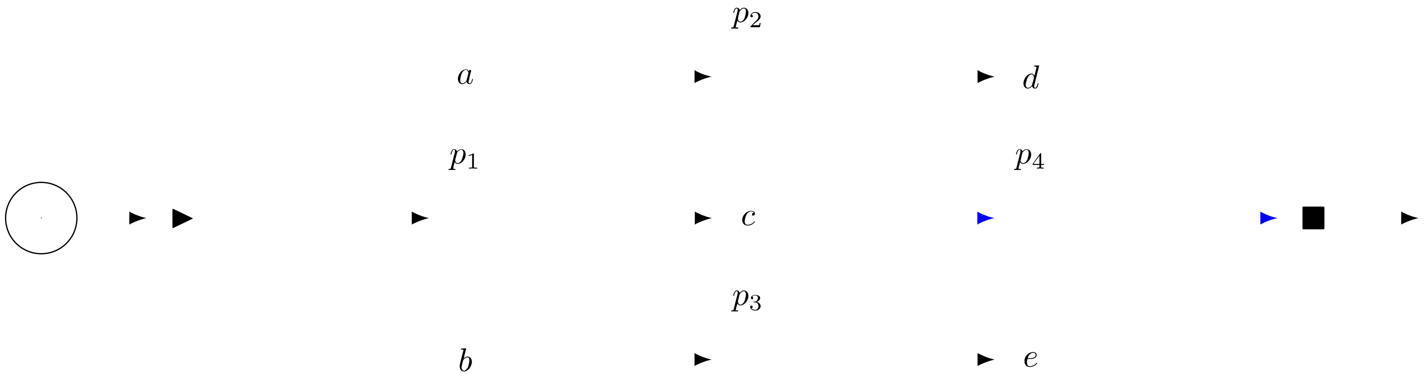
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_4 = (c|\blacksquare)$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	84	42
$\blacksquare$	275	220

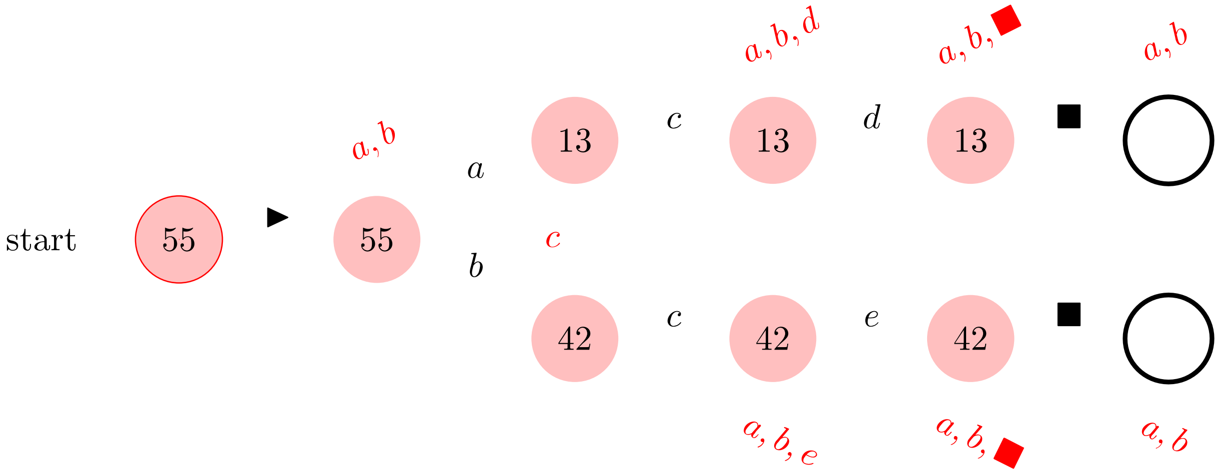


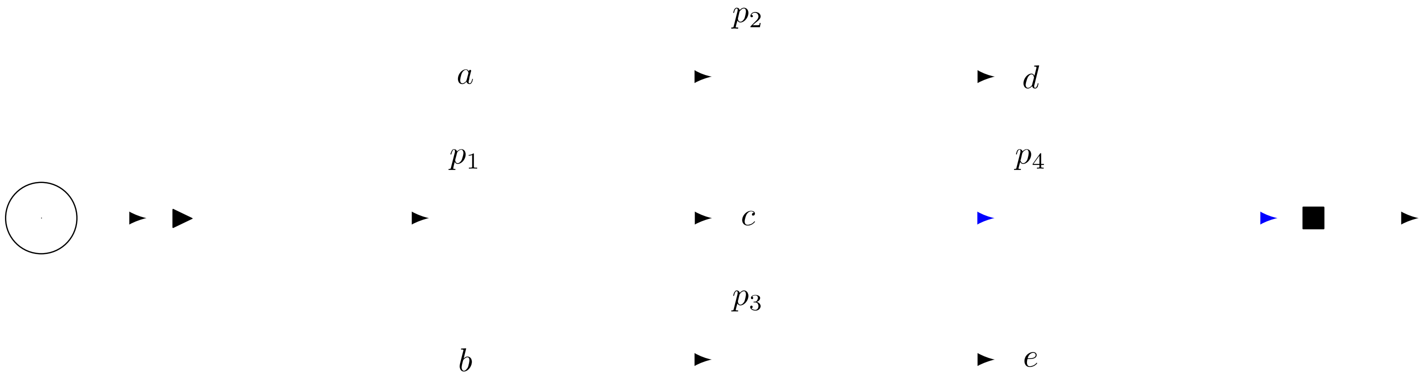


Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_4 = (c|\blacksquare)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	84	42
$\blacksquare$	~~275~~ 110	~~220~~ 55

$$precision(P) = 1 - \frac{605}{880+55} \approx 0.353\ (> 0.3)$$

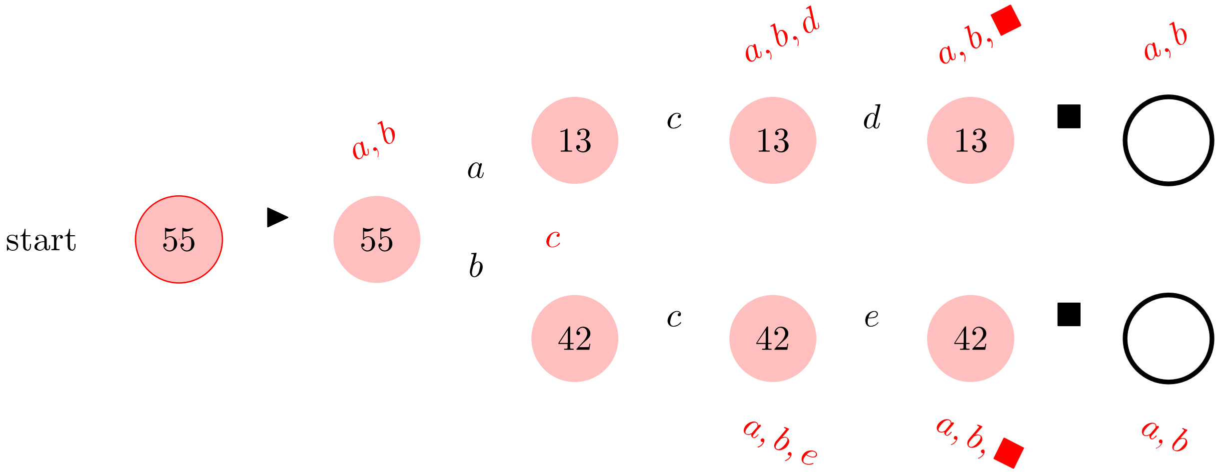
$\Rightarrow add(p_4)$

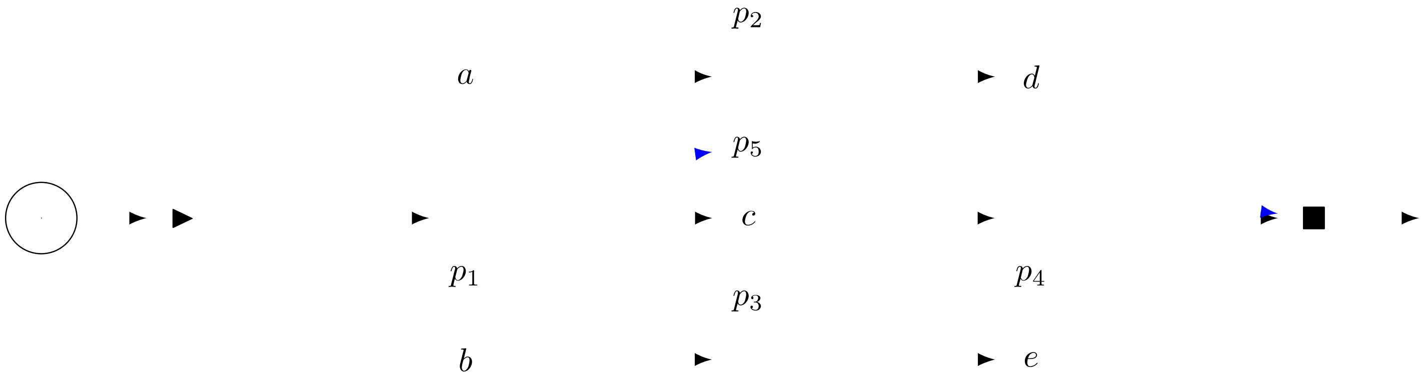
$P_{PotImpl}(p_4) = \emptyset$

Example

$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

Candidate place: $p_5 = (\blacktriangleright | \blacksquare)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	84	42
$\blacksquare$	110	55

$precision(P) = 0.353$

$\Rightarrow discard(p_5)$

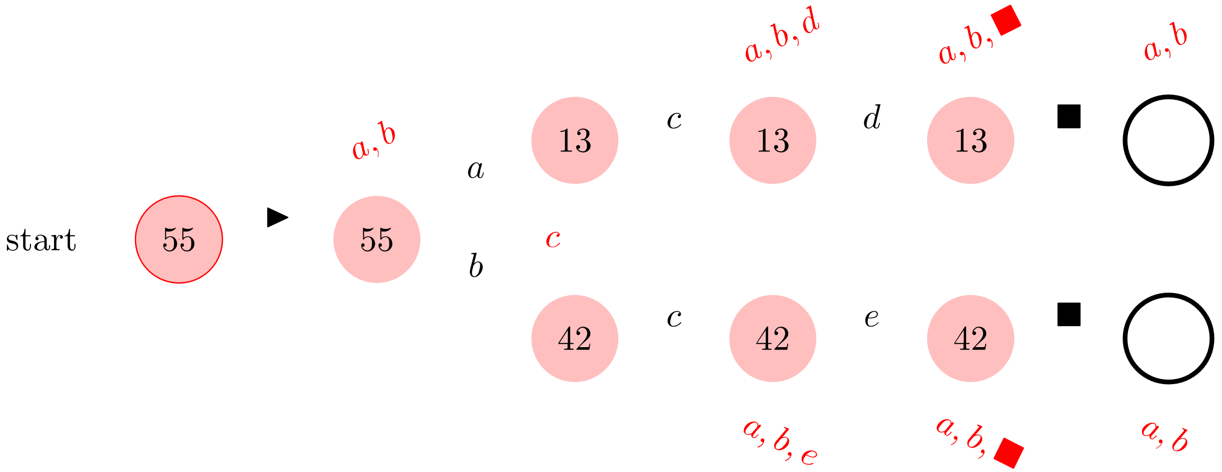
Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

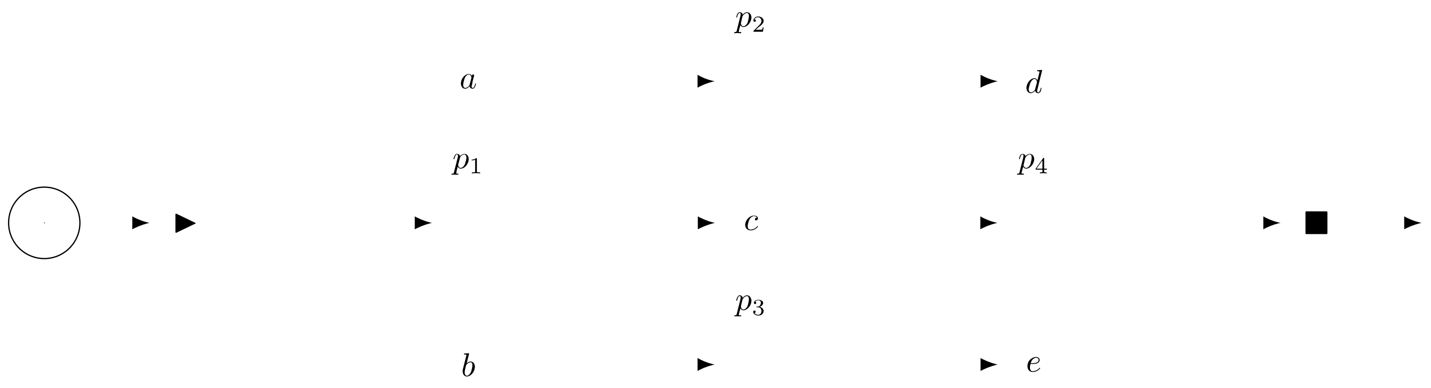
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	84	42
$\blacksquare$	110	55

Candidate place: $p_5 = (\blacktriangleright | \blacksquare)$



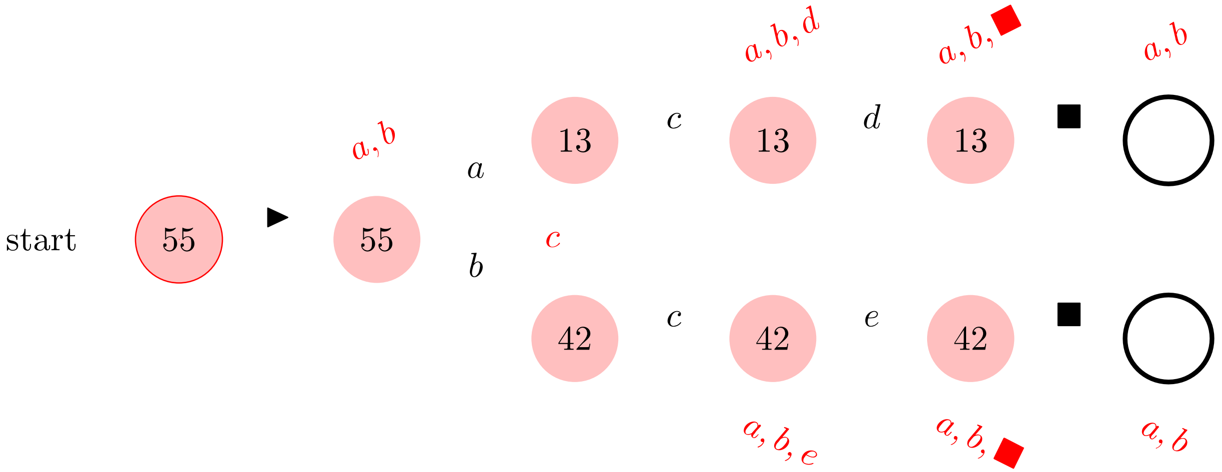


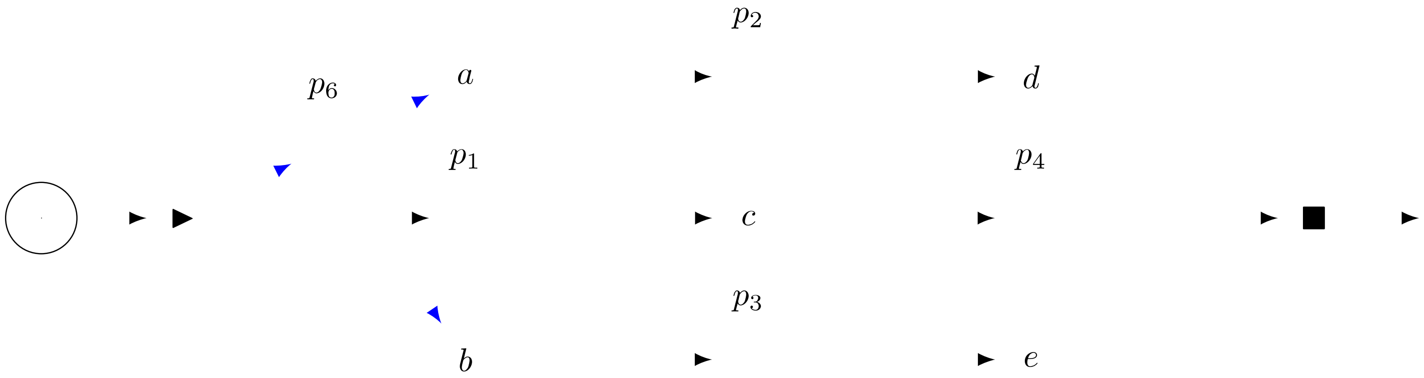
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	275	220
b	275	220
c	110	55
d	26	13
e	84	42
$\blacksquare$	110	55

Candidate place: $p_6 = (\blacktriangleright | a, b)$

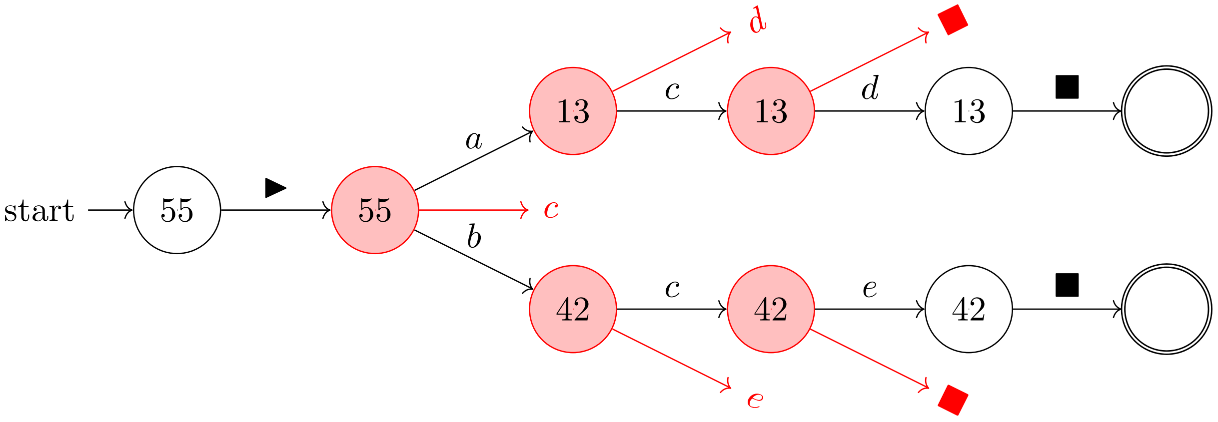


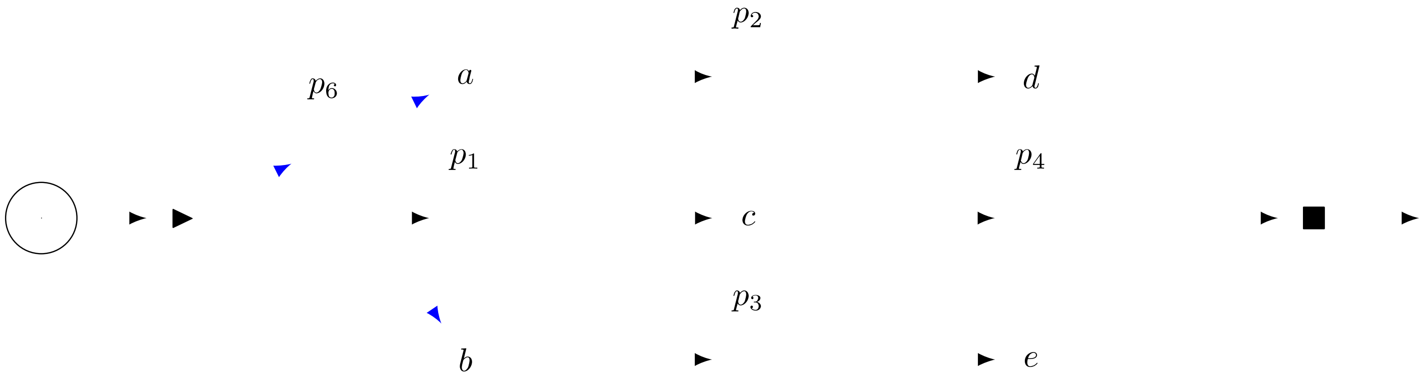


Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_6 = (\blacktriangleright | a, b)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	~~275~~ 55	~~220~~ 0
b	~~275~~ 55	~~220~~ 0
c	110	55
d	26	13
e	84	42
$\blacksquare$	110	55

$$precision(P) = 1 - \frac{165}{440+55} \approx 0.667\ (> 0.353)$$

$\Rightarrow add(p_6)$

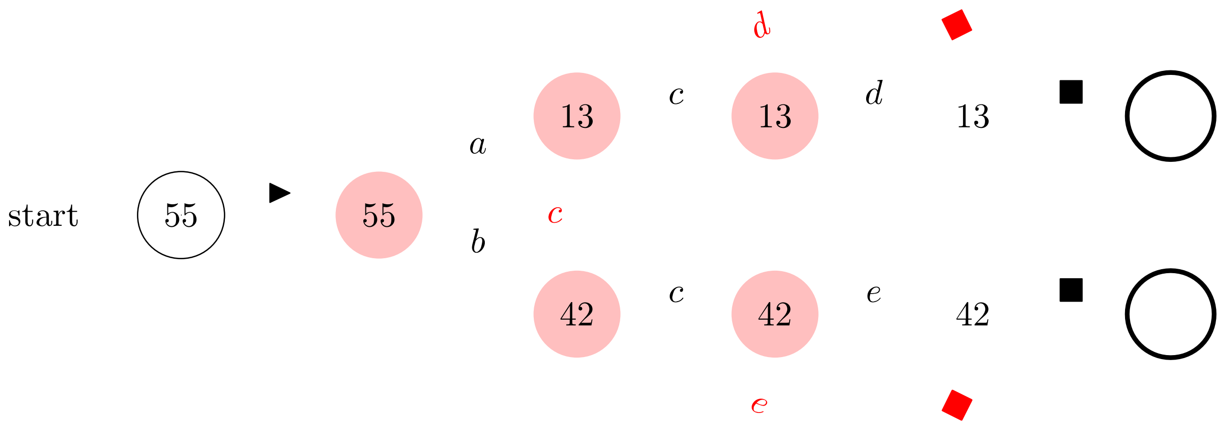
$P_{PotImpl}(p_6) = \emptyset$

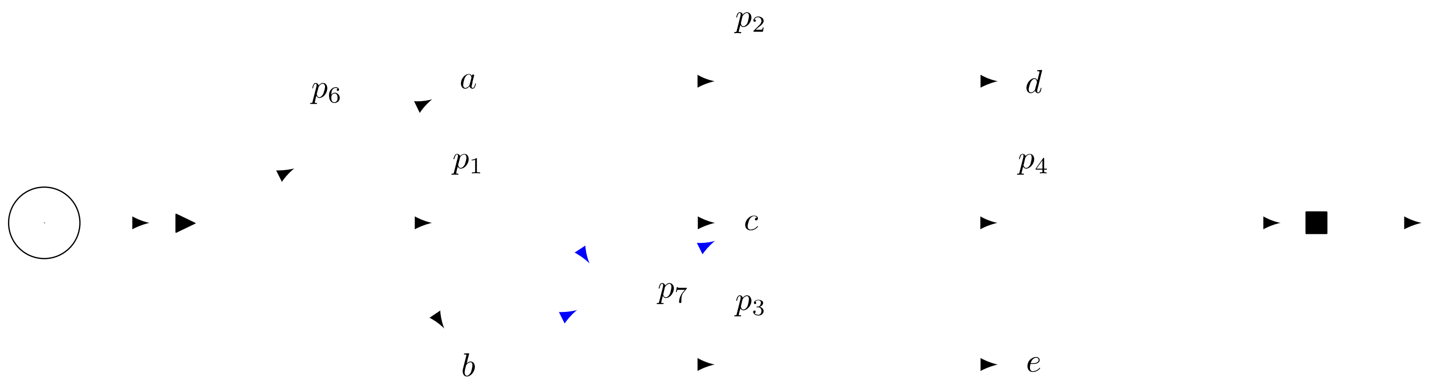
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_7 = (a, b | c)$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	110	55
d	26	13
e	84	42
$\blacksquare$	110	55

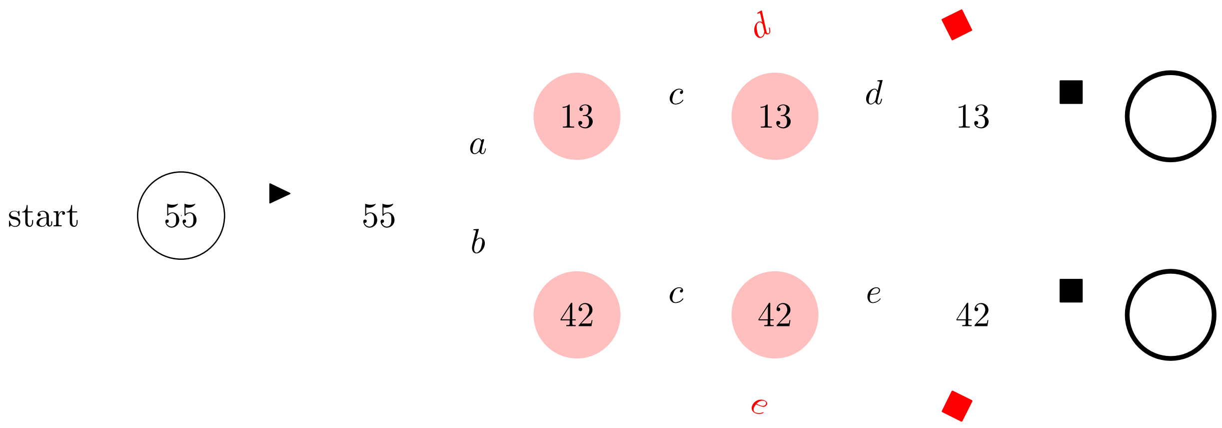


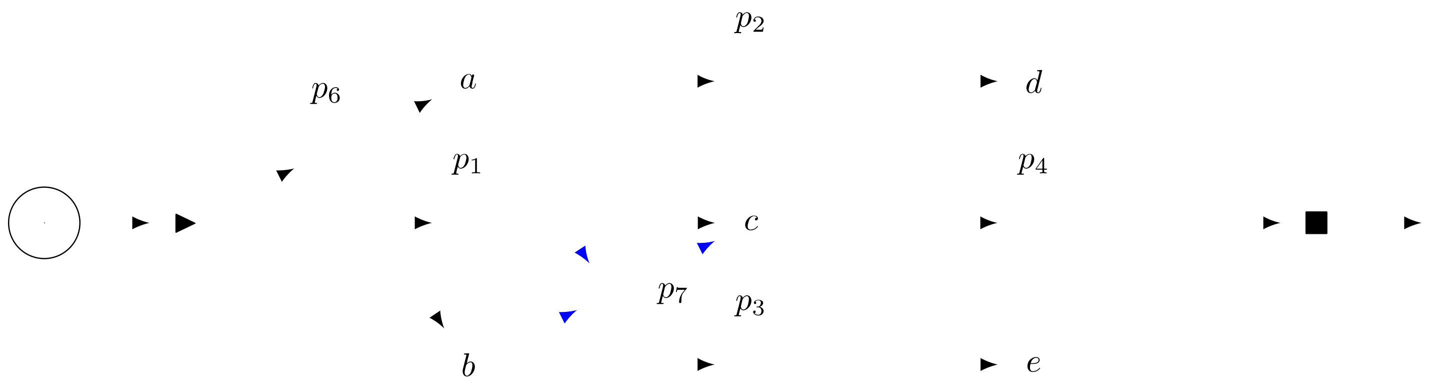


Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_7 = (a, b|c)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	~~110~~ 55	~~55~~ 0
d	26	13
e	84	42
$\blacksquare$	110	55

$$precision(P) = 1 - \frac{110}{385+55} = 0.75\ (> 0.667)$$

$\Rightarrow add(p_7)$

$$P_{PotImpl}(p_7) = \{(\blacktriangleright|c)\}$$

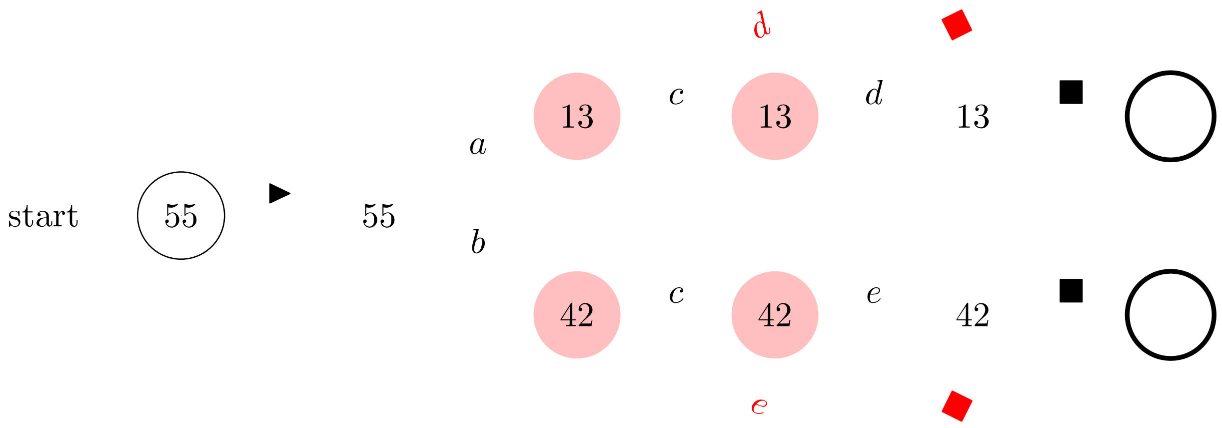
Example

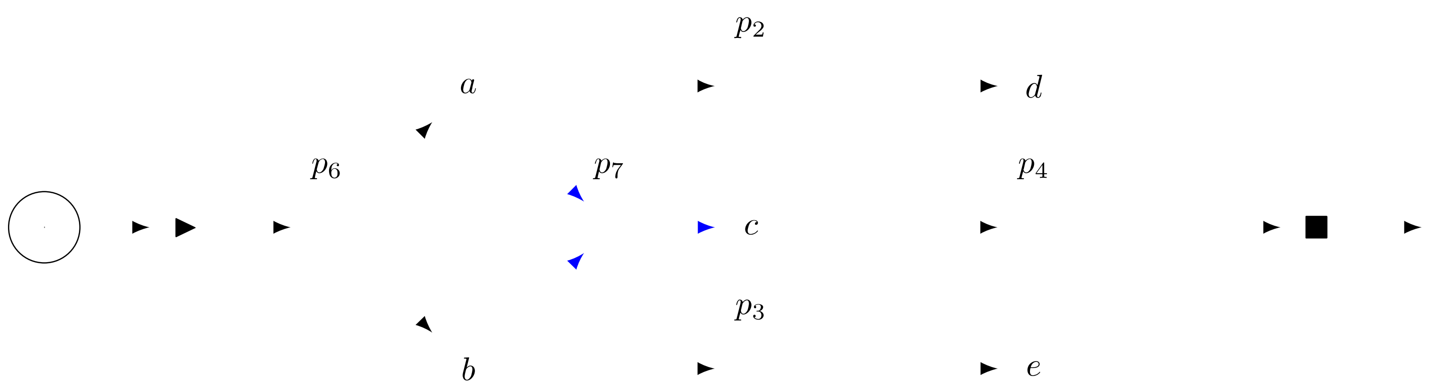
$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

Candidate place: $p_7 = (a, b | c)$

Pot. implicit place:

$p_1 = (\blacktriangleright | c)$





·	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	26	13
e	84	42
■	110	55

$precision(P) = 0.75$

$\Rightarrow revoke(p_1)$

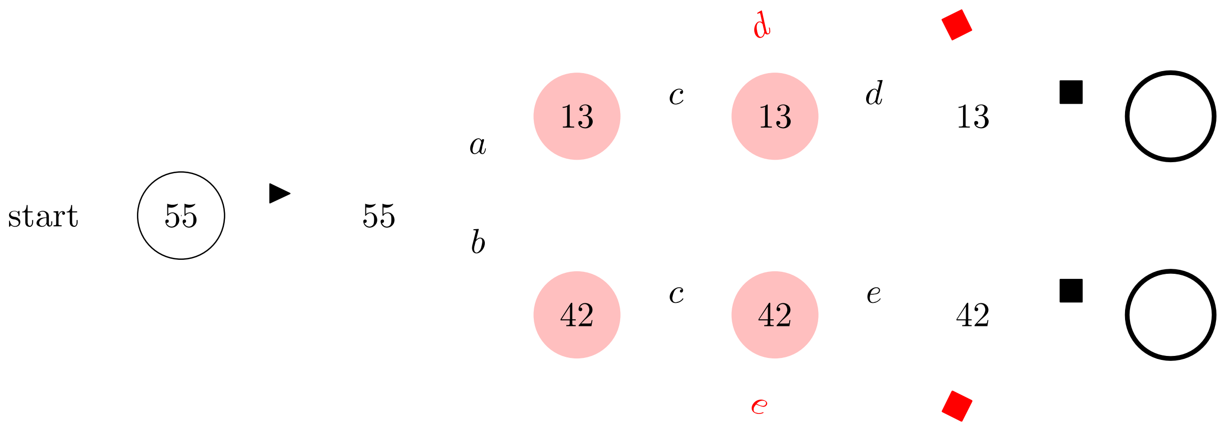
Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

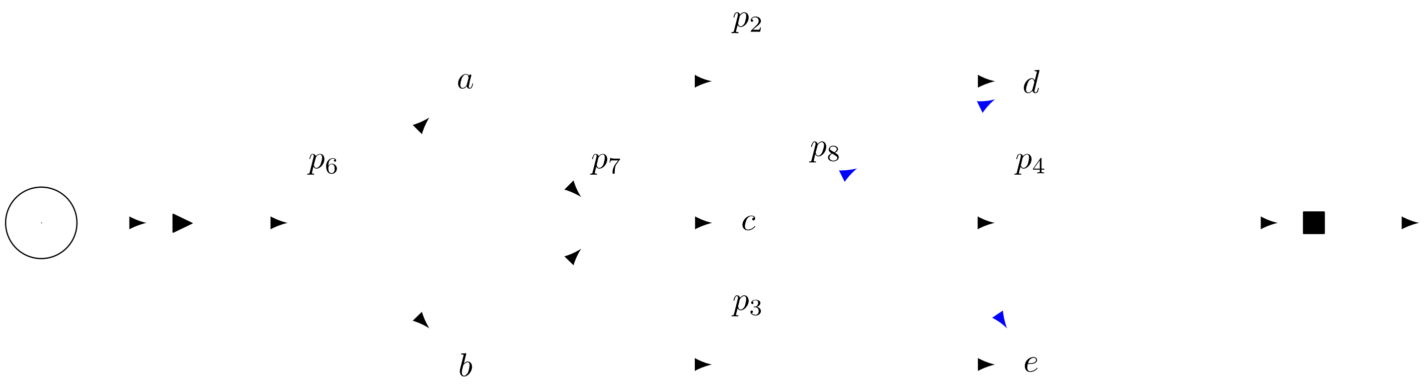
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare\rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare\rangle^{42}]$$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	26	13
e	84	42
$\blacksquare$	110	55

Candidate place: $p_8 = (c|d, e)$



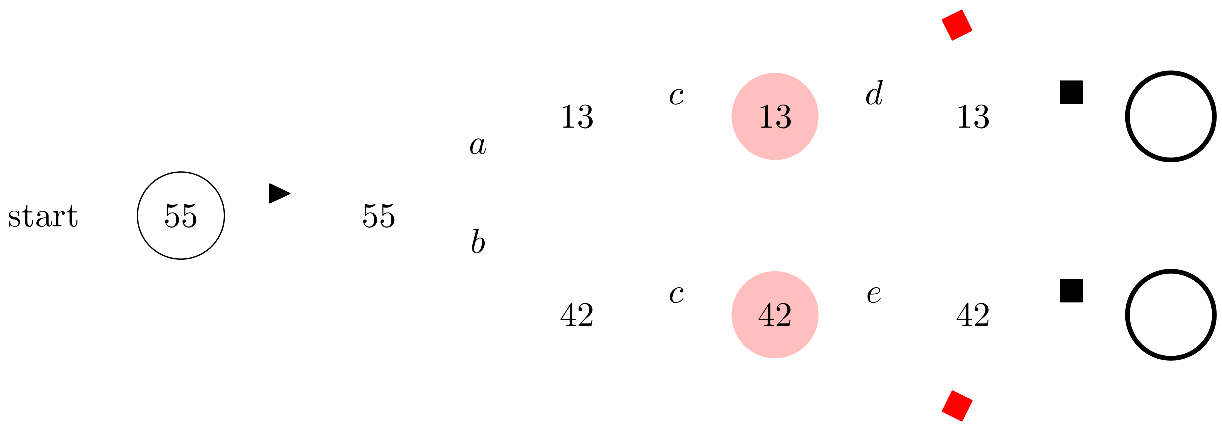


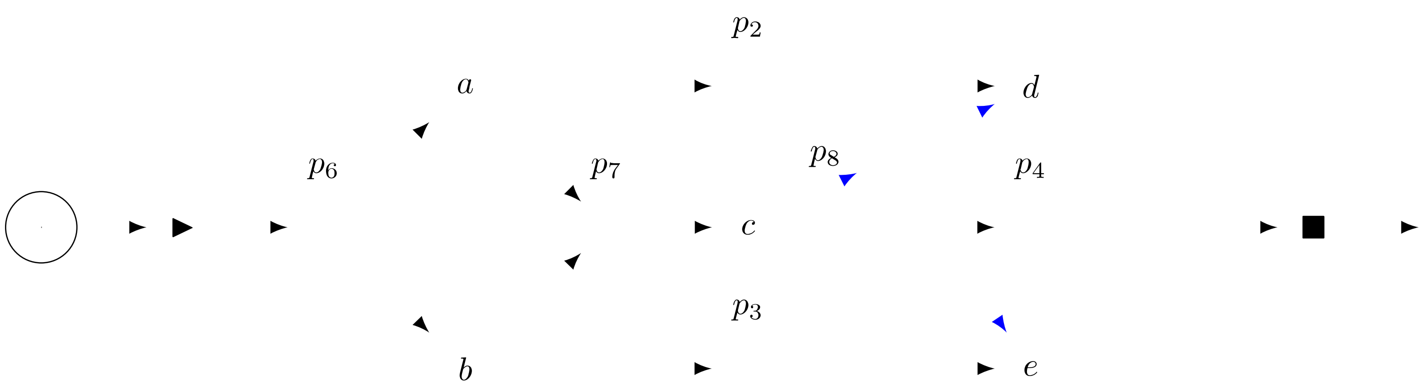
Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_8 = (c|d, e)$





$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	~~26~~ 13	~~13~~ 0
e	~~84~~ 42	~~42~~ 0
$\blacksquare$	110	55

$$precision(P) = 1 - \frac{55}{330+55} \approx 0.857\ (> 0.75)$$

$\Rightarrow add(p_8)$

$$P_{PotImpl}(p_8) = \{(a|d), (b|e)\}$$

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Example

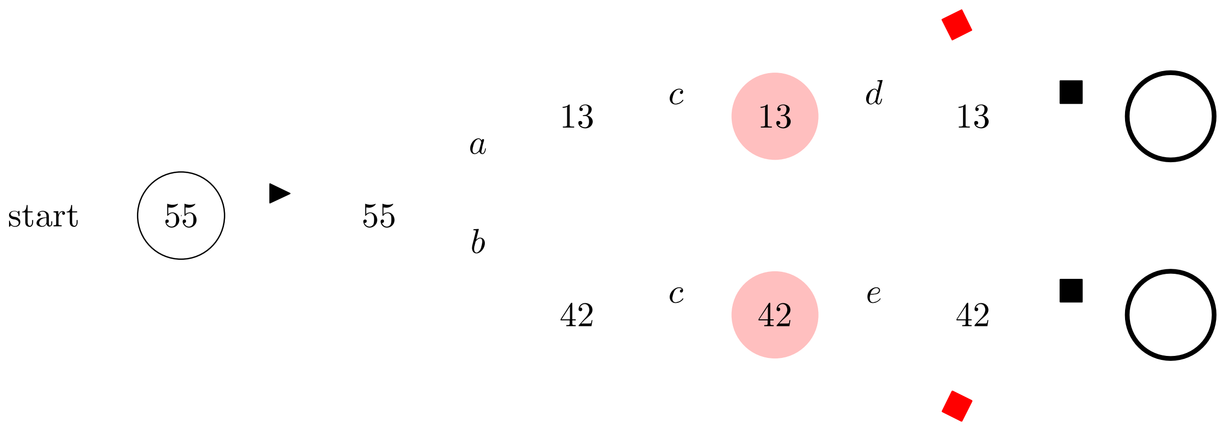
$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

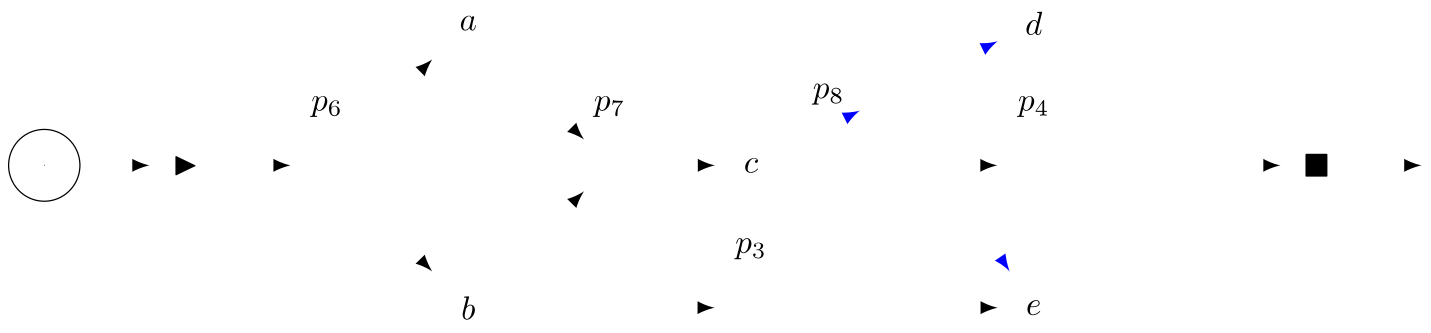
$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	110	55

Candidate place: $p_8 = (c|d,e)$

Pot. implicit place:

$p_2 = (a|d)$



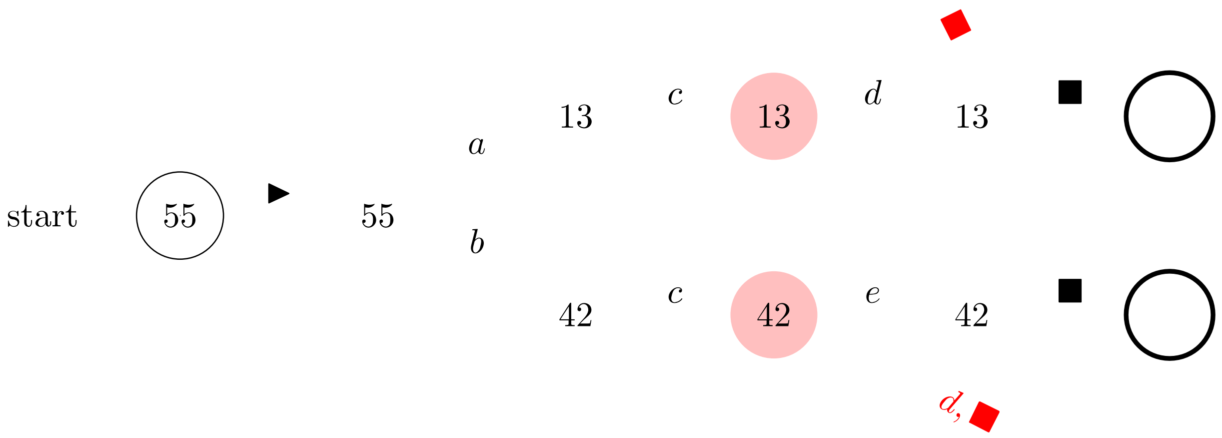


Example

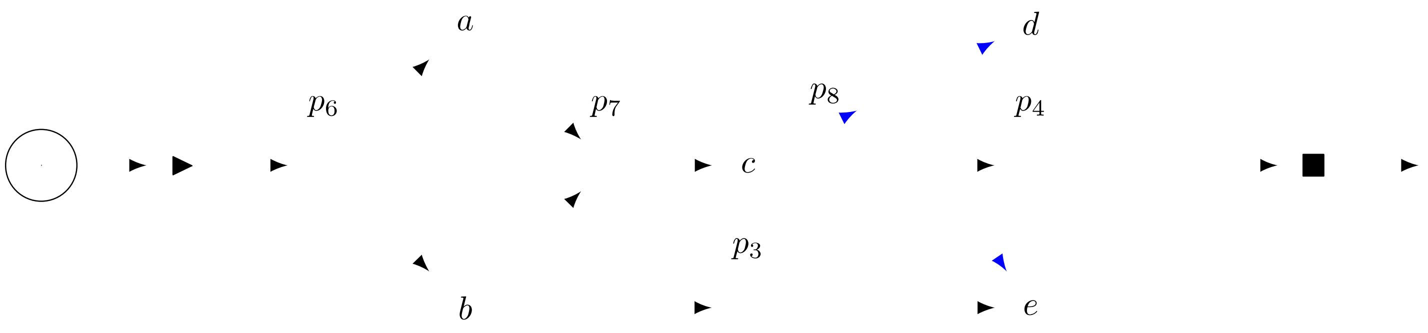
$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

Candidate place: $p_8 = (c|d, e)$

Pot. implicit place: $p_2 = (a|d)$



$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	~~13~~ 55	~~0~~ 42
e	42	0
$\blacksquare$	110	55



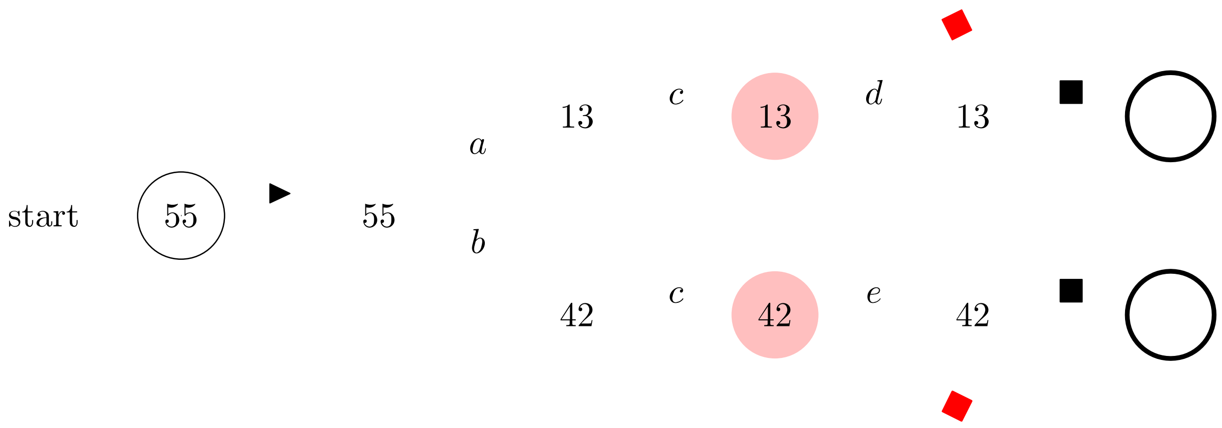
$precision(P) = 1 - \frac{97}{372+55}$
$\approx 0.787\ (< 0.857)$

$\Rightarrow keep(p_2)$

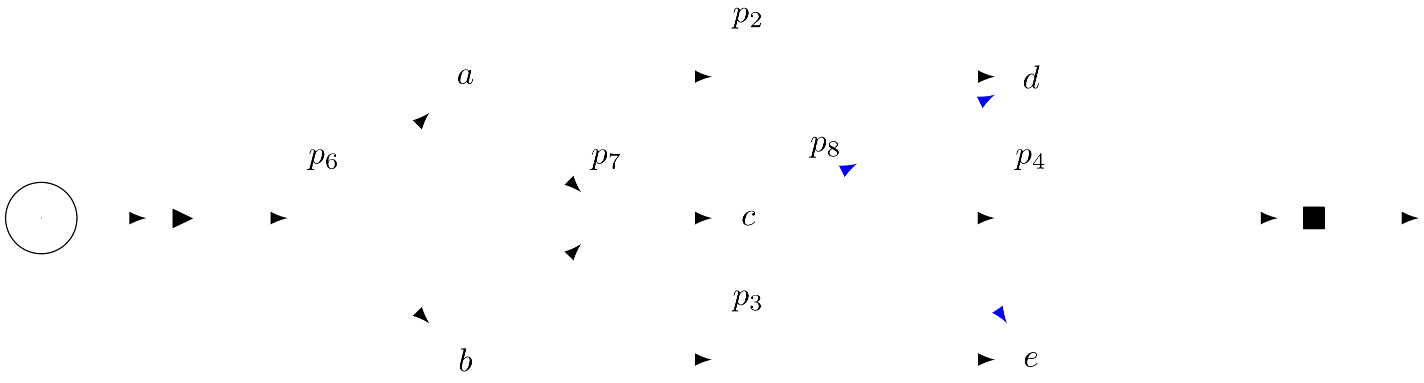
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_8 = (c|d, e)$



$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	110	55

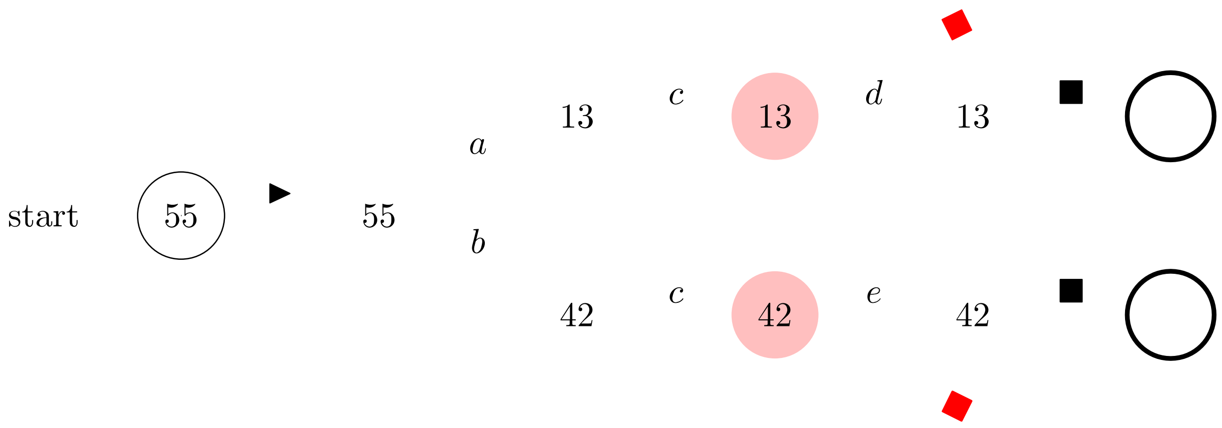


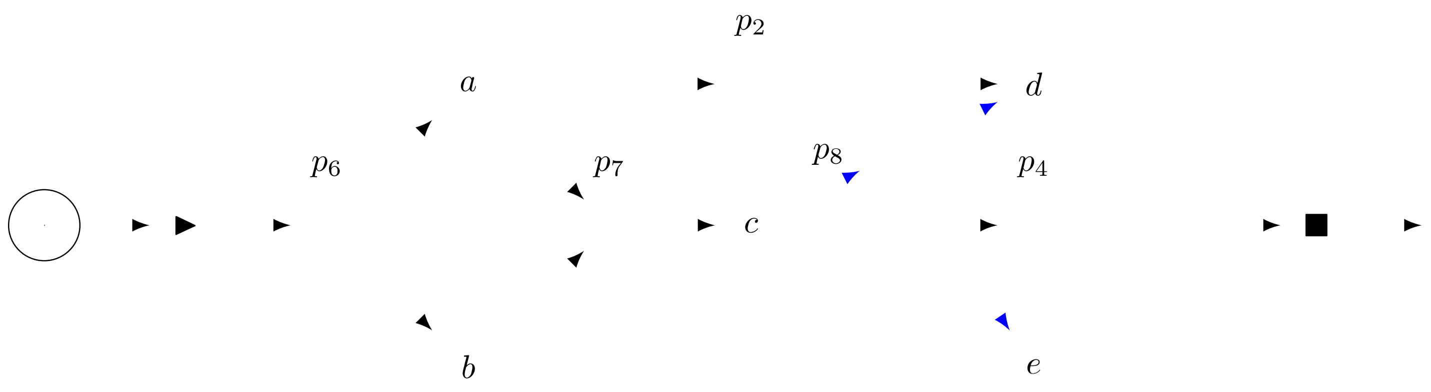
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_8 = (c|d, e)$

Pot. implicit place: $p_3 = (b|e)$





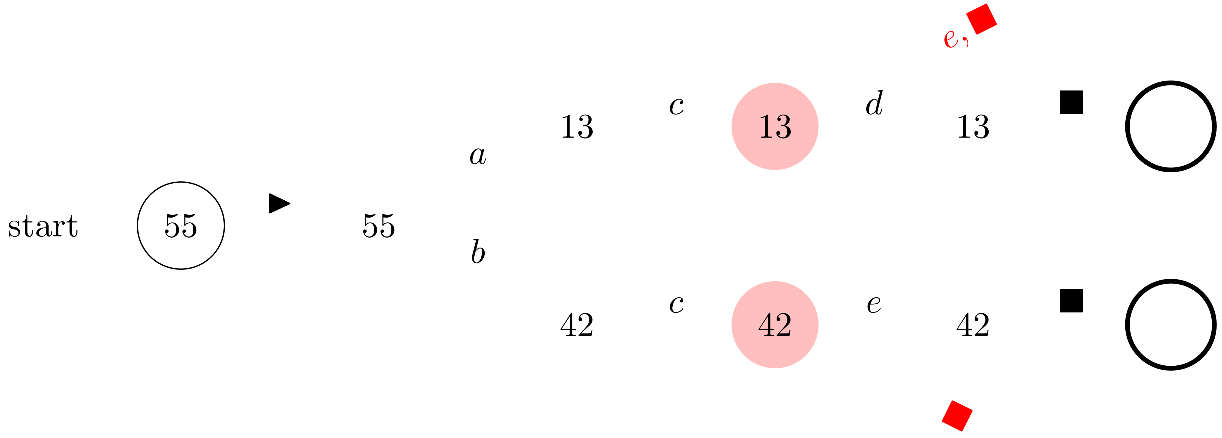
$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	110	55

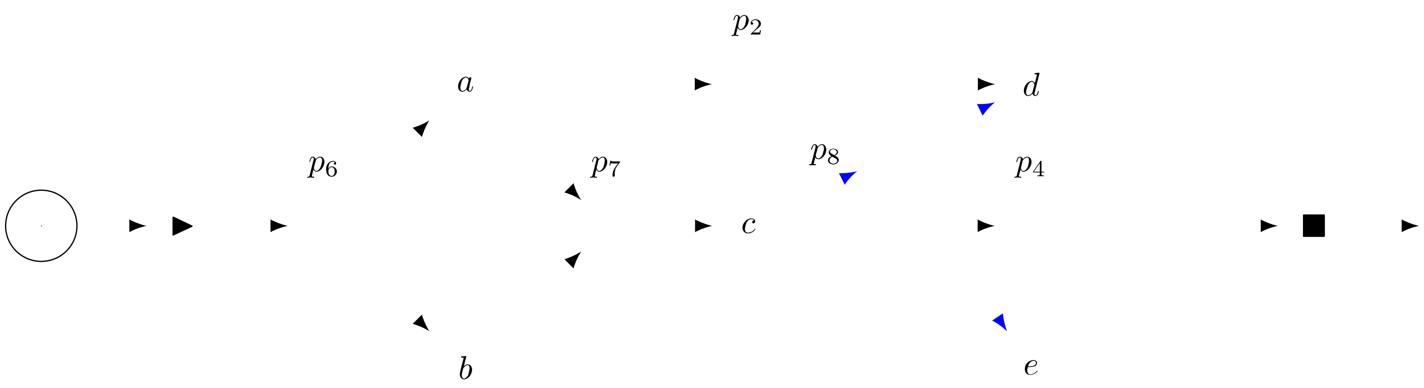
Example

$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

Candidate place: $p_8 = (c|d, e)$

Pot. implicit place: $p_3 = (b|e)$





·	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	~~42~~ 55	~~0~~ 13
■	110	55

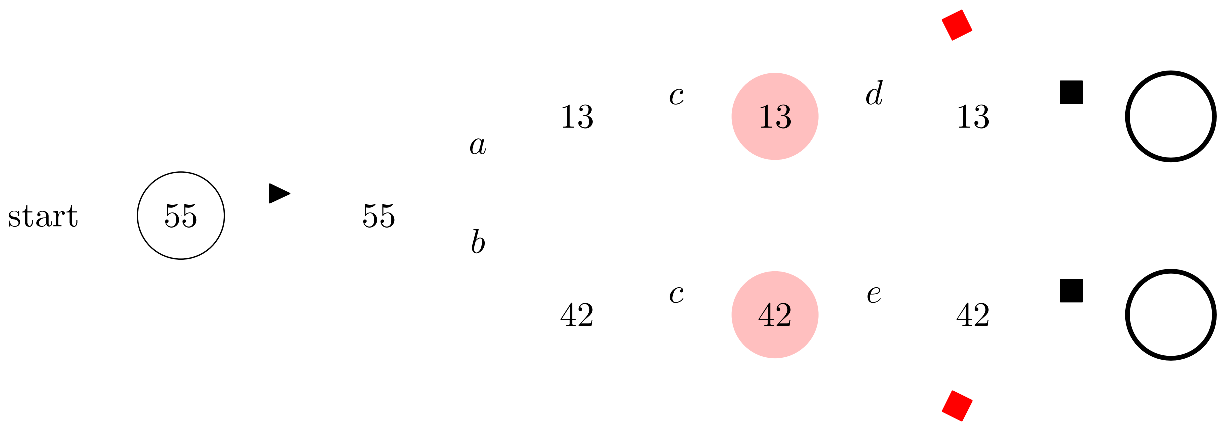
$$precision(P) = 1 - \frac{68}{343+55} \approx 0.829\ (< 0.857)$$

$\Rightarrow keep(p_3)$

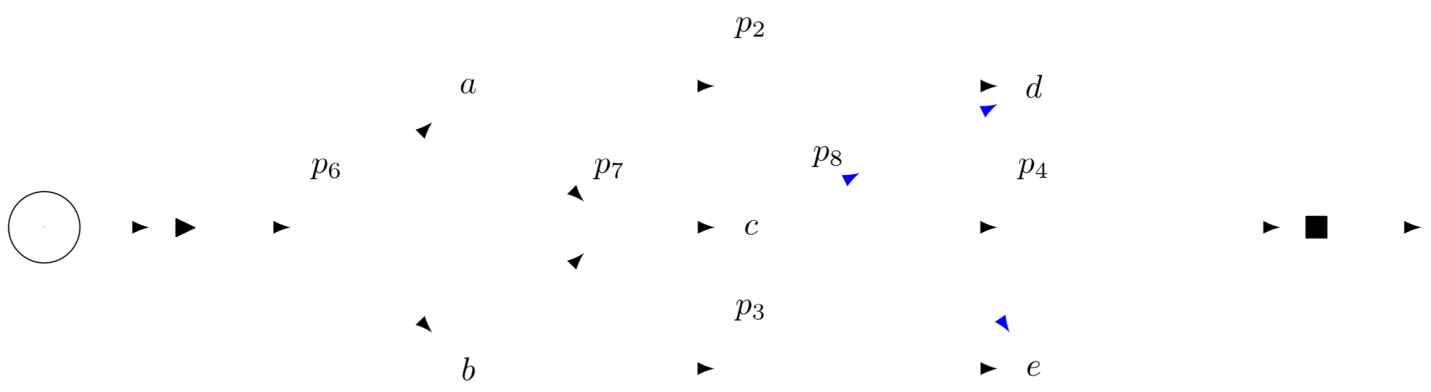
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_8 = (c|d, e)$



$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	110	55

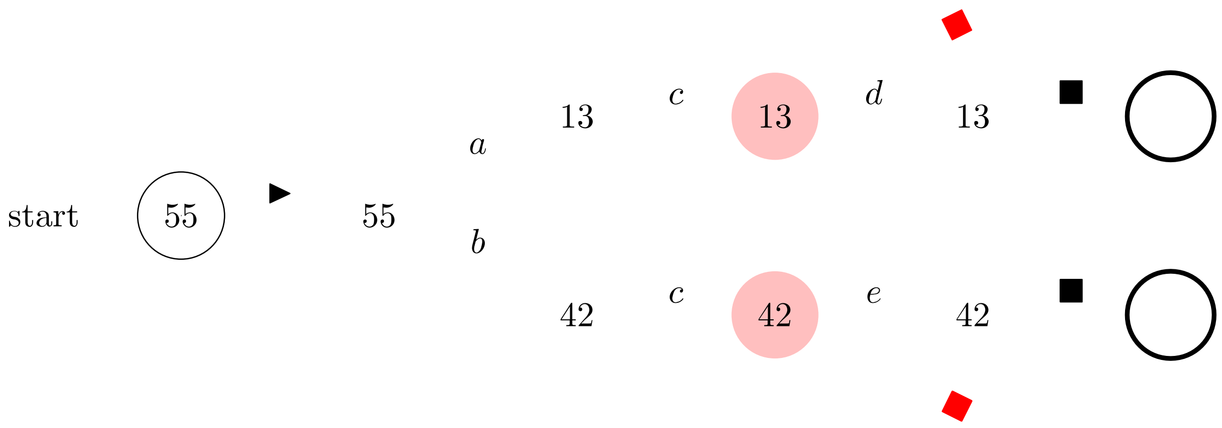


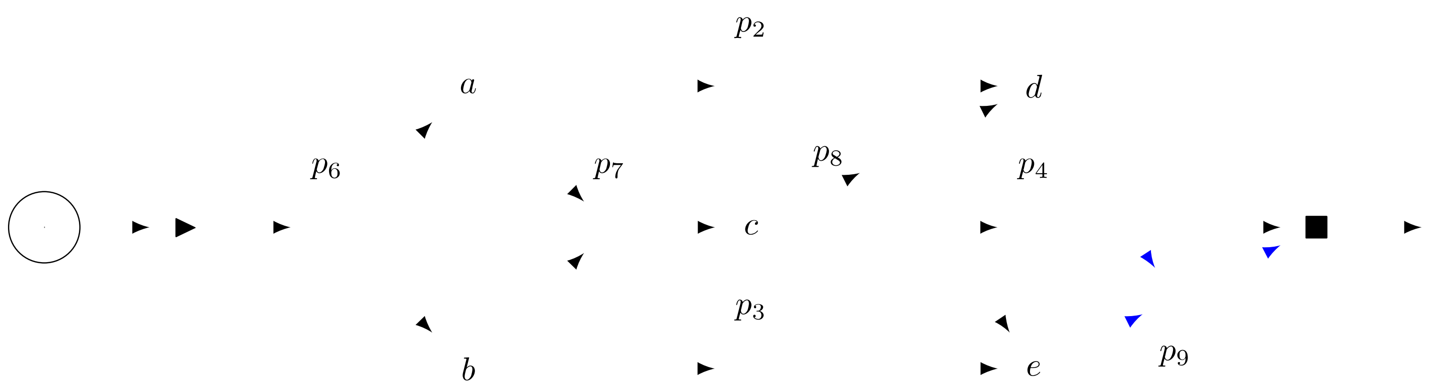
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	110	55

Candidate place: $p_9 = (d, e | \blacksquare)$

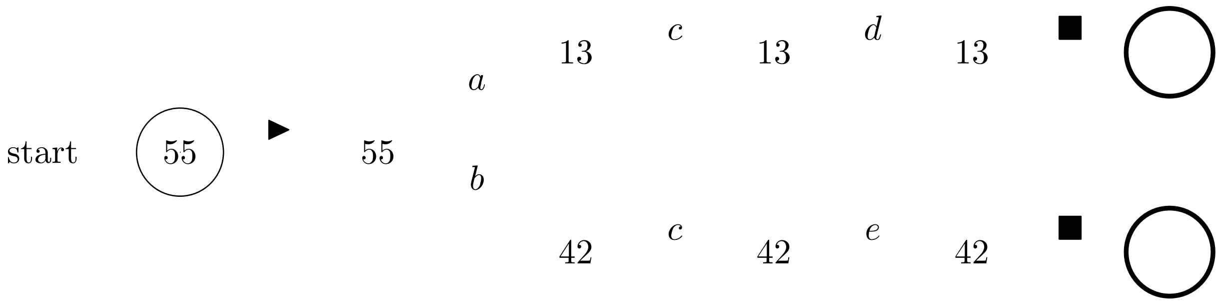




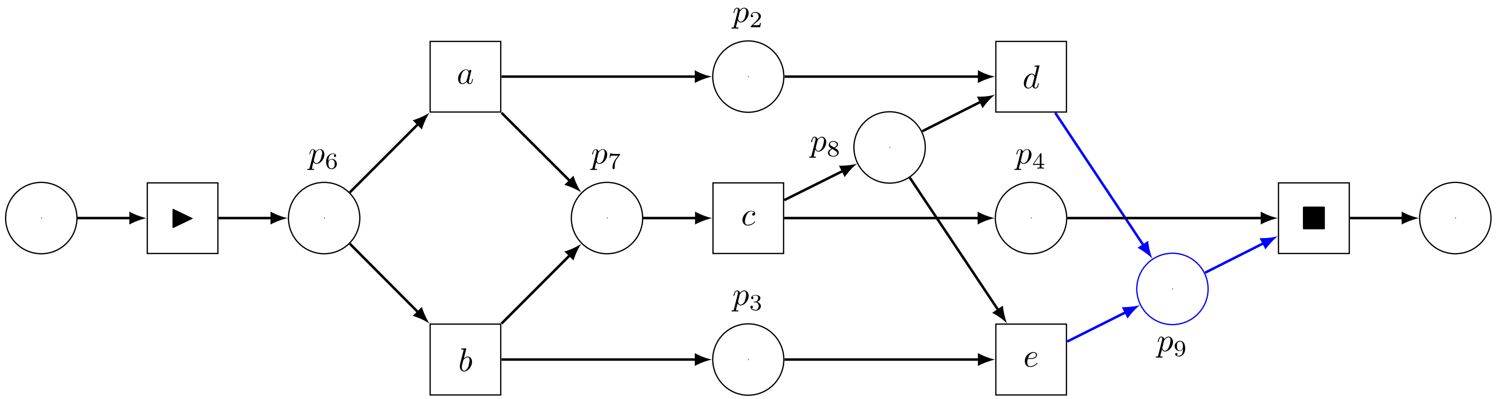
Example

$$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$$

Candidate place: $p_9 = (d, e | \blacksquare)$



$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	~~110~~ 55	~~55~~ 0



$$precision(P) = 1 - \frac{0}{275+55} \approx 1\ (> 0.857)$$

$\Rightarrow add(p_9)$

$$P_{PotImpl}(p_9) = \{(c|\blacksquare)\}$$

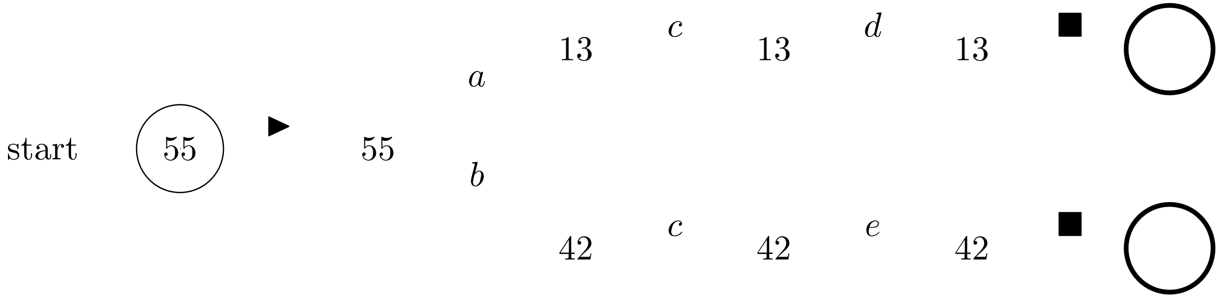
Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Example

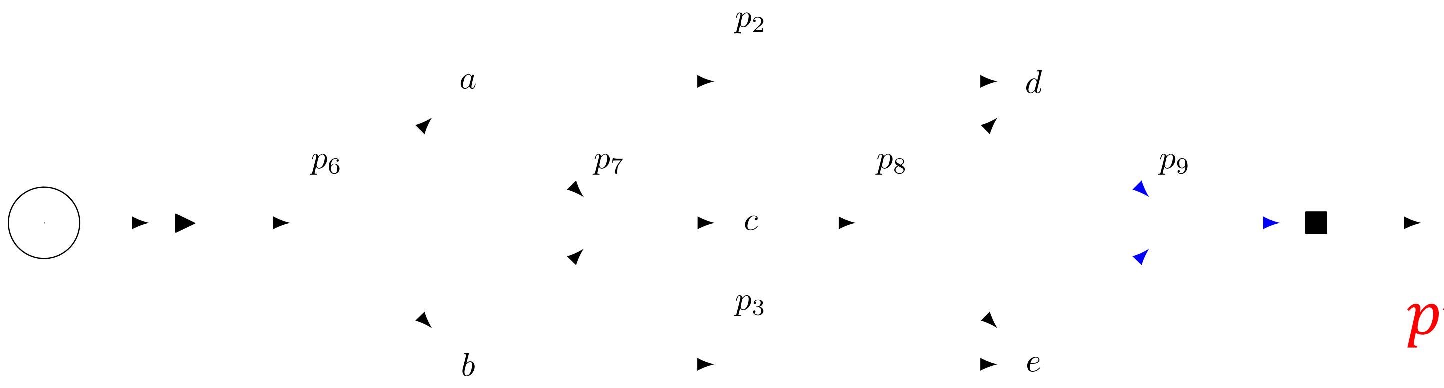
$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$

Candidate place: $p_9 = (d, e | \blacksquare)$

Pot. implicit place: $p_4 = (c | \blacksquare)$



$\cdot$	$A(\cdot)$	$E(\cdot)$
a	55	0
b	55	0
c	55	0
d	13	0
e	42	0
$\blacksquare$	55	0



$precision(P) = 1$

$\Rightarrow revoke(p_4)$

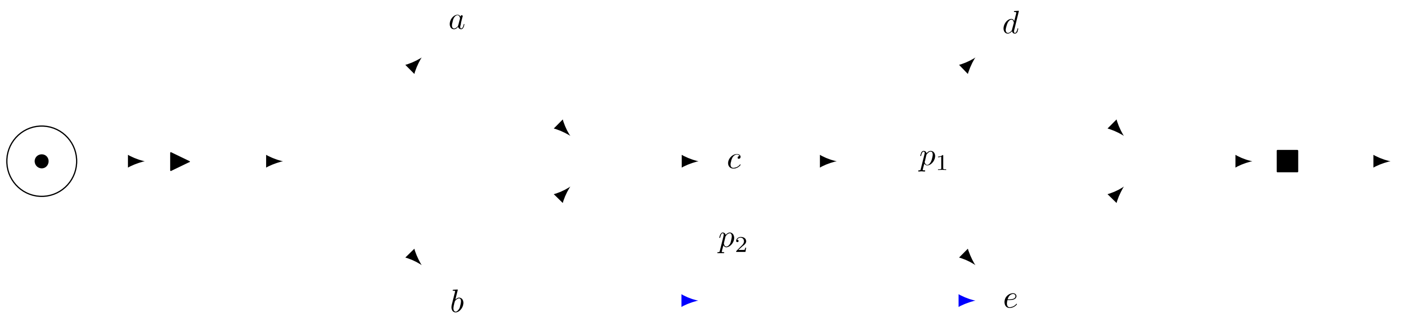
$precision \geq \rho \Rightarrow$ prematurely abort!

How do we do this efficiently?

Approximating ETC-Precision efficiently

Observation:

If $p = (I|O)$ is added to / removed from the intermediate result, we only have to reevaluate $E(o)$ and $A(o)$, for all $o \in O$.

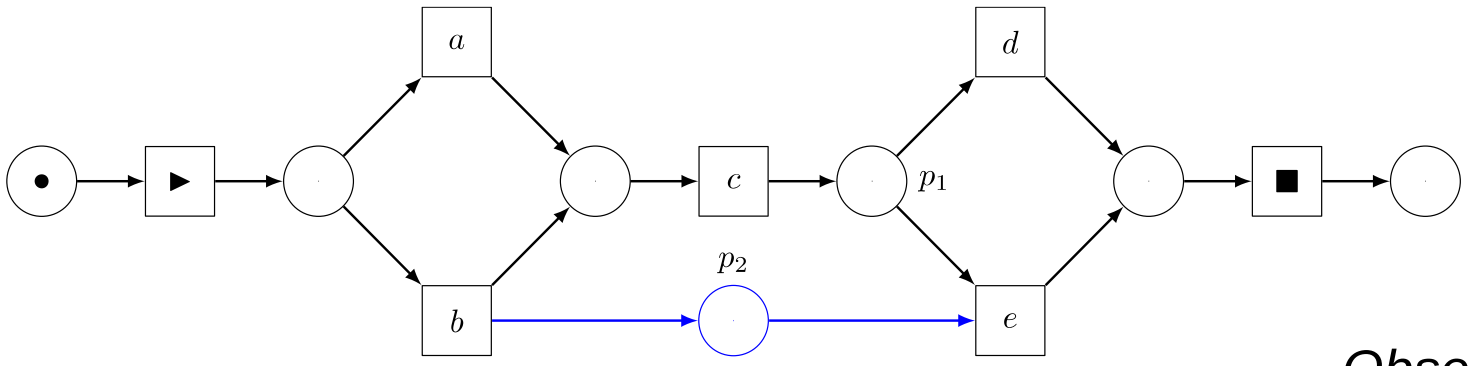


Approximating ETC-Precision efficiently

"When is an activity allowed / reflected / escaping during replay?"

⇒ Prefix Automaton

Calculate the marking histories of its input places



$H_L(\cdot)$	▶	a	c	d	■	▶	b	c	e	■
p_1	0	0	1	0	0	0	0	1	0	0
p_2	0	0	0	0	0	0	1	1	0	0

Observation:

$H_L(p)$ is place-dependent
⇒ only replay L once per (new) place
⇒ store $H_L(p)$ for all $p \in P$

Experimental Results

Implementation

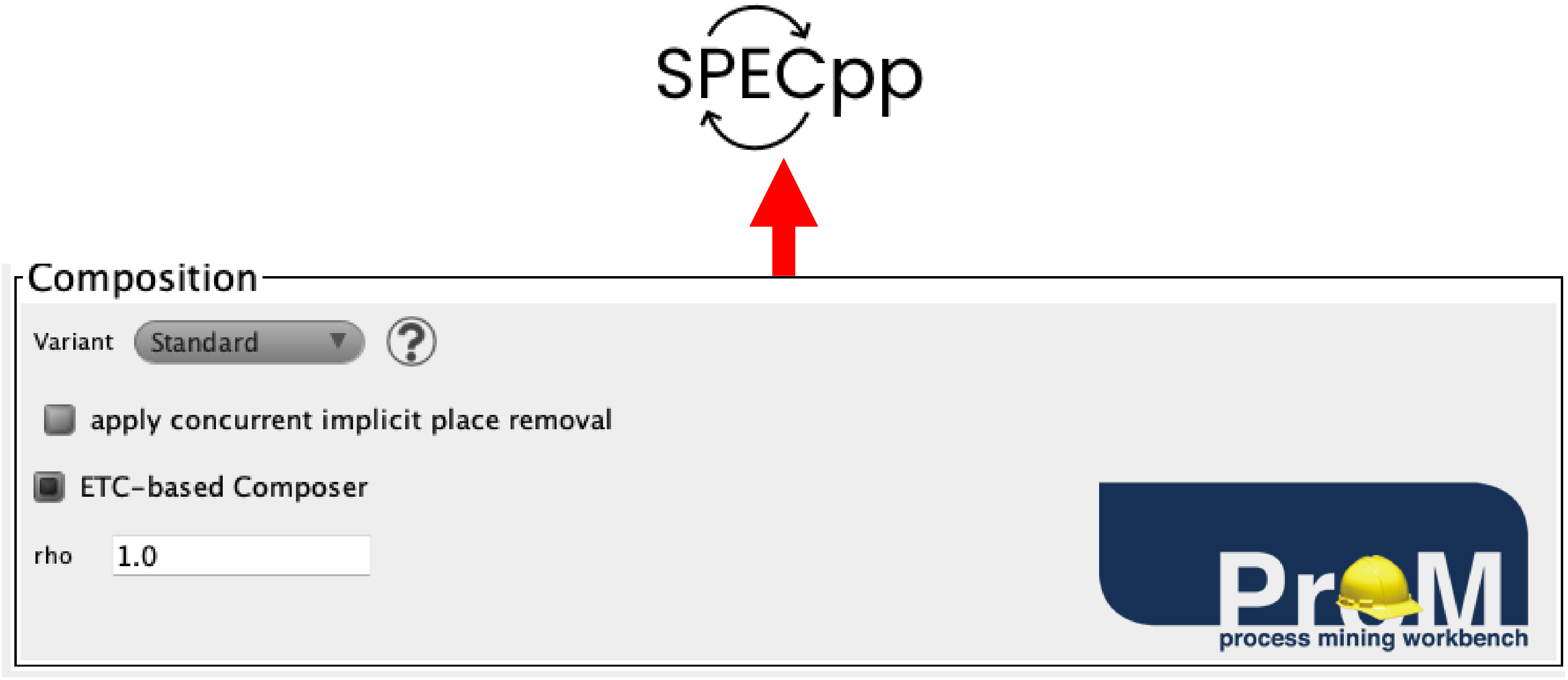


Image: promtools.org (accessed: 27.09.23)

Experimental Setup

- PADS HPC cluster (AMD Threadripper CPU: 12x3,5 GHz, 128 GB RAM)
- Six (real-life and artificial event logs)

type	name	#activities	#traces	#variants
real	HD2017	16	4580	226
	RTFM	13	150370	231
	Sepsis	18	1050	846
artificial	Repair	10	1104	77
	Reviewing	16	100	96
	Teleclaims	13	3512	12

Suitability for Implicit Place Removal $(d = 4)$

Remaining implicit places (missed by the ETC-based Composer) in relation to all fitting places:

No remaining implicit places for $\tau = 1$

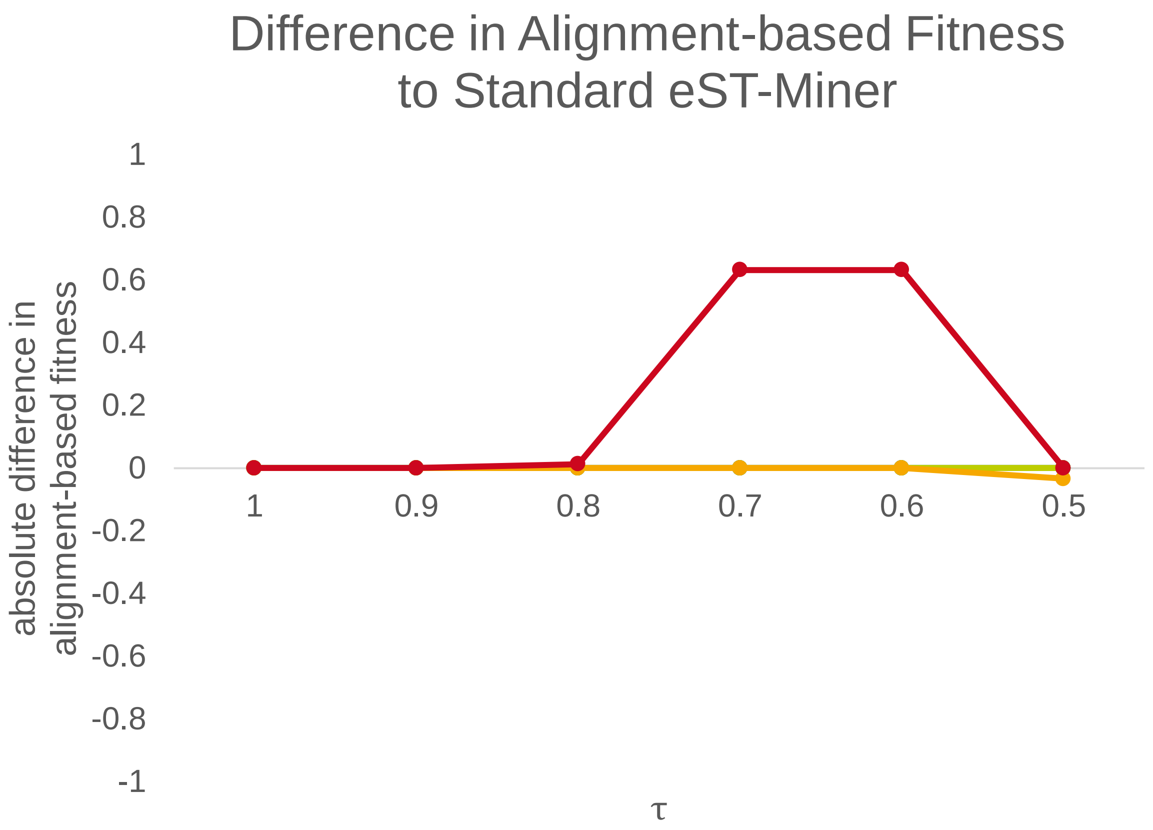
τ	HD2017	RTFM	Sepsis	Repair	Reviewing	Teleclaims
1	0/20	0/32	0/29	0/21	0/88	0/82
0.9	14/4390	3/567	15/543	5/59	0/88	3/103
0.8	32/5282	11/846	15/978	4/63	0/88	3/238
0.7	40/8125	20/989	18/1521	5/104	0/88	3/372
0.6	43/8642	10/1452	33/2987	7/127	0/88	3/414
0.5	47/8642	5/2535	42/3614	8/252	2/681	2/1035
AVG	29/5850	8/1070	21/1611	5/104	0/187	2/374

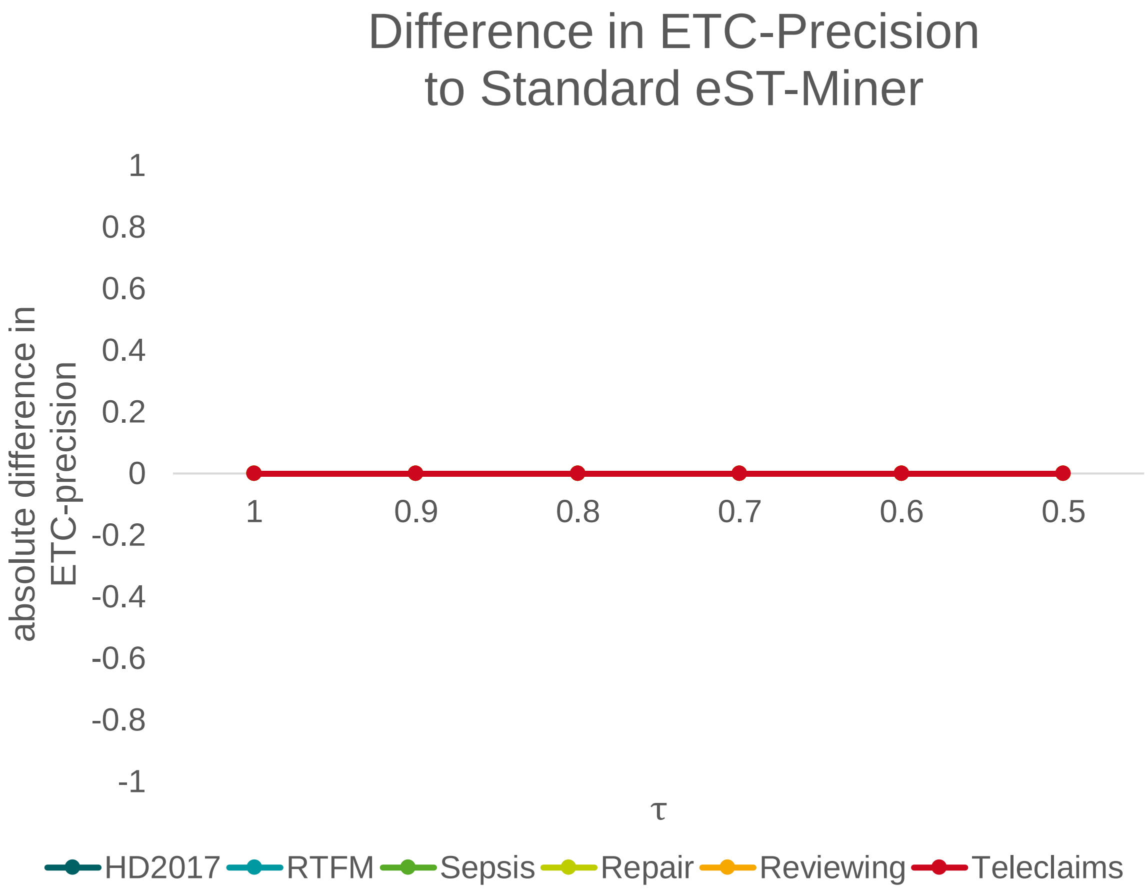
Only few implicit places left for $\tau < 1$, solve ILP to remove remaining in PP

(fast, since only called once and few places as input)

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

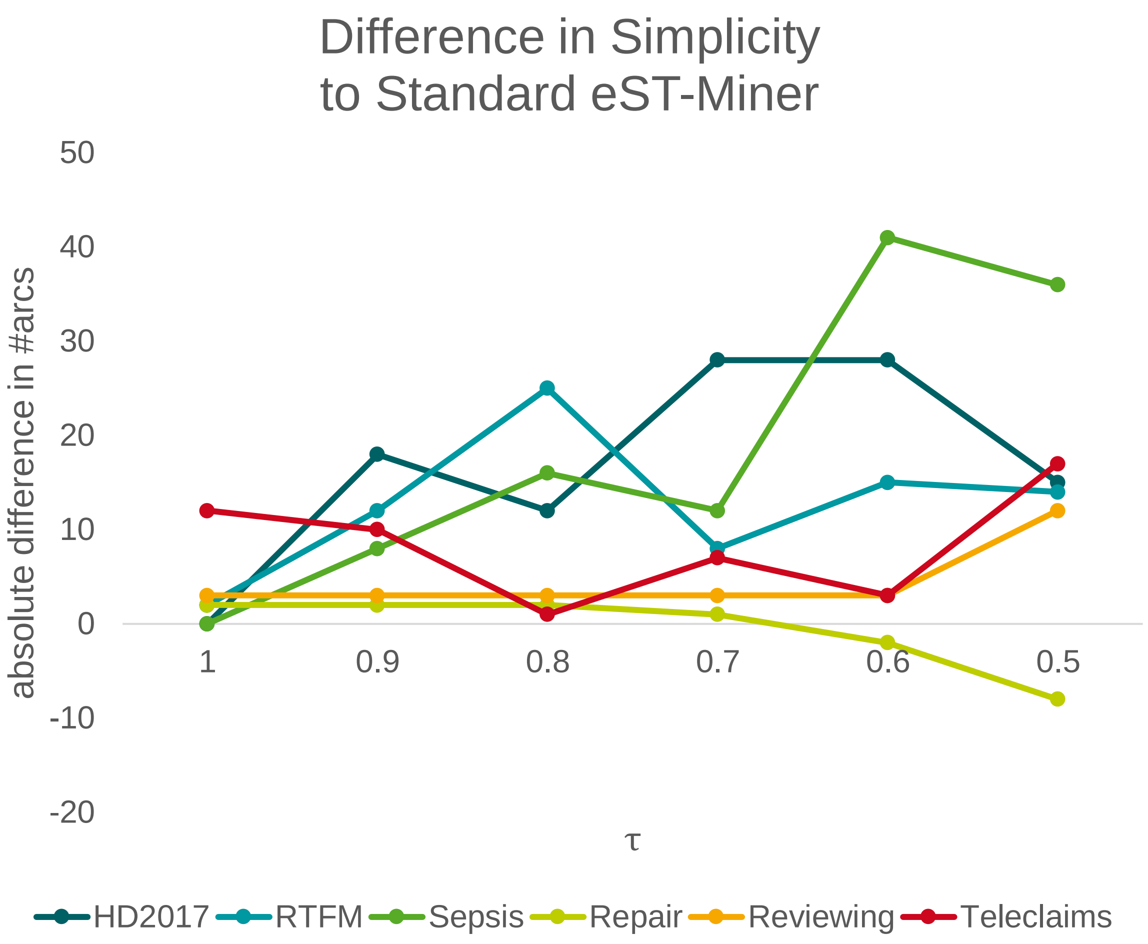
Suitability for Implicit Place Removal $(d = 4)$





⇒ (almost) no difference in quality

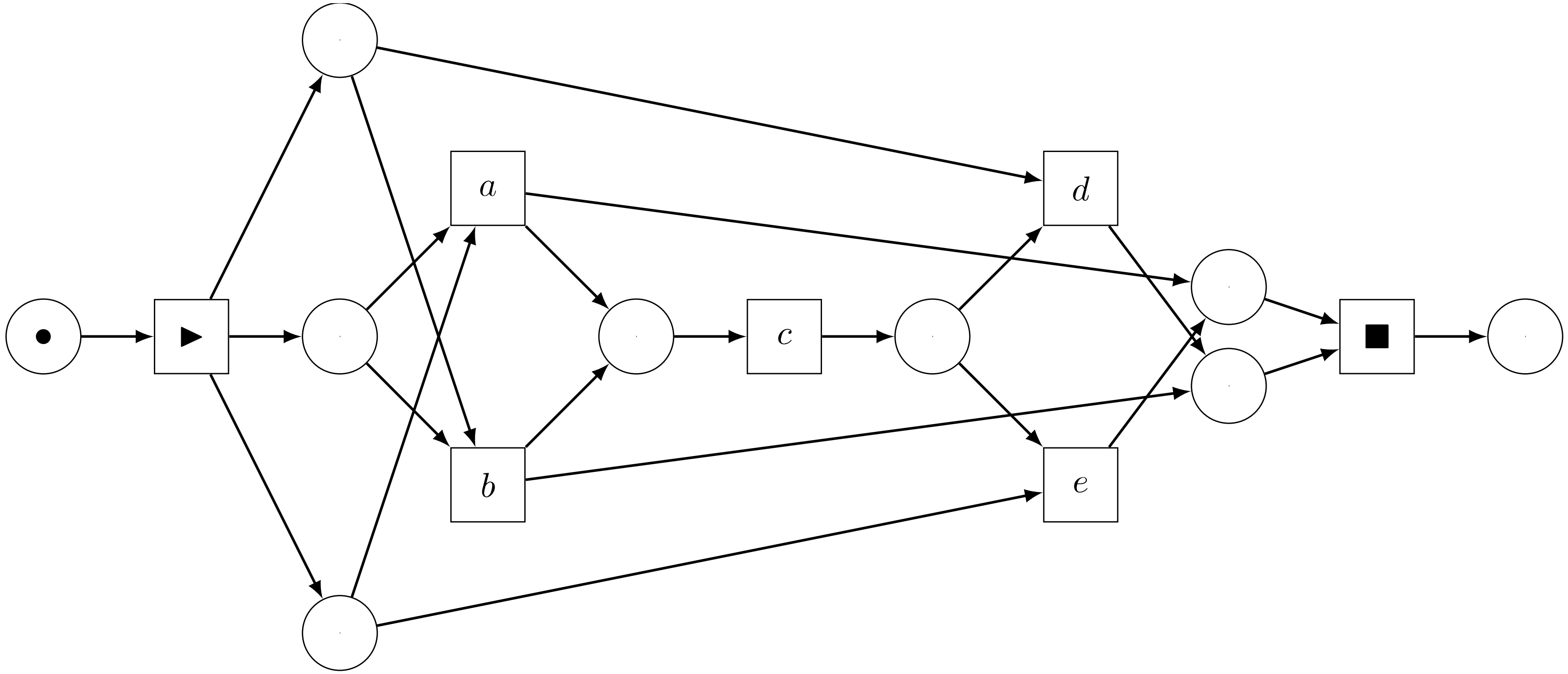
Problem: Overly Complex Models $(d = 4)$



Reason: Proposal order matters

$L = [\langle \blacktriangleright, a, c, d, \blacksquare \rangle^{13}, \langle \blacktriangleright, b, c, e, \blacksquare \rangle^{42}]$ (previous example, different proposal order)

Final Model:



Performance Comparison of IPR-Techniques

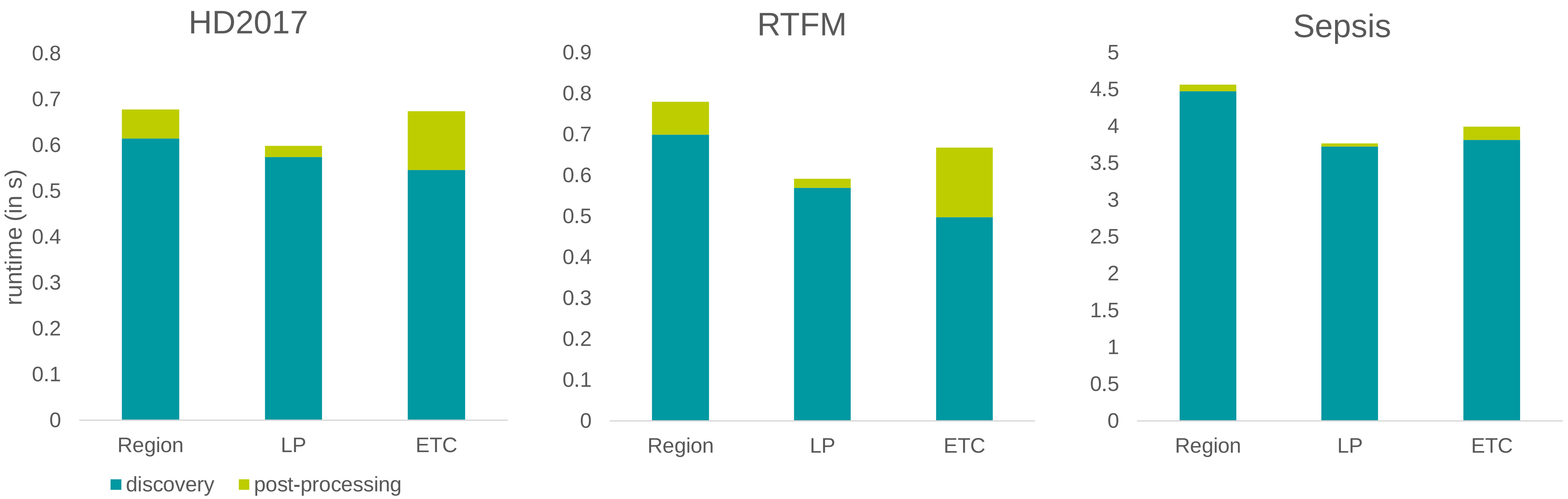
(user-definable) noise threshold parameter τ

$\tau = 1.0$

- ETC-based composer (*ETC*)
- region-based IPR (*Region*)
- LP-based IPR (*LP*)

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

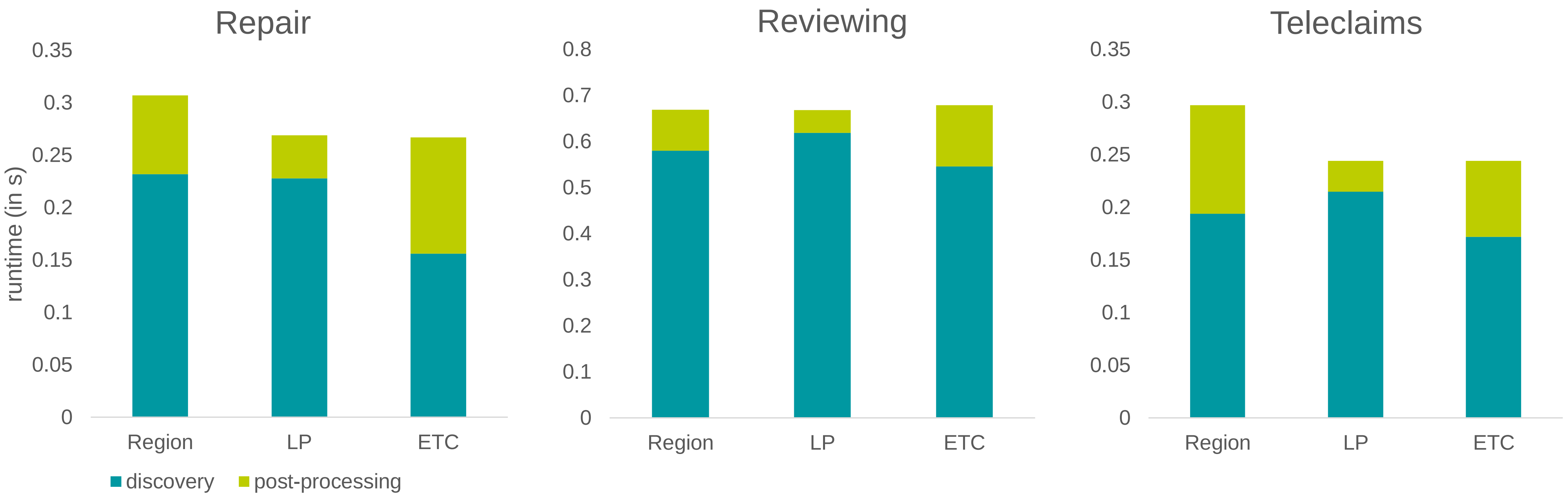
Performance Comparison of IPR-Techniques (Real-Life Event Logs, $\tau = 1, d = 4$)



For $\tau = 1$, all techniques have comparable runtime

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Performance Comparison of IPR-Techniques (Artificial Event Logs, $\tau = 1, d = 4$)



For $\tau = 1$, all techniques have comparable runtime

Performance Comparison of IPR-Techniques

(user-definable) noise threshold parameter τ

$\tau < 1.0$

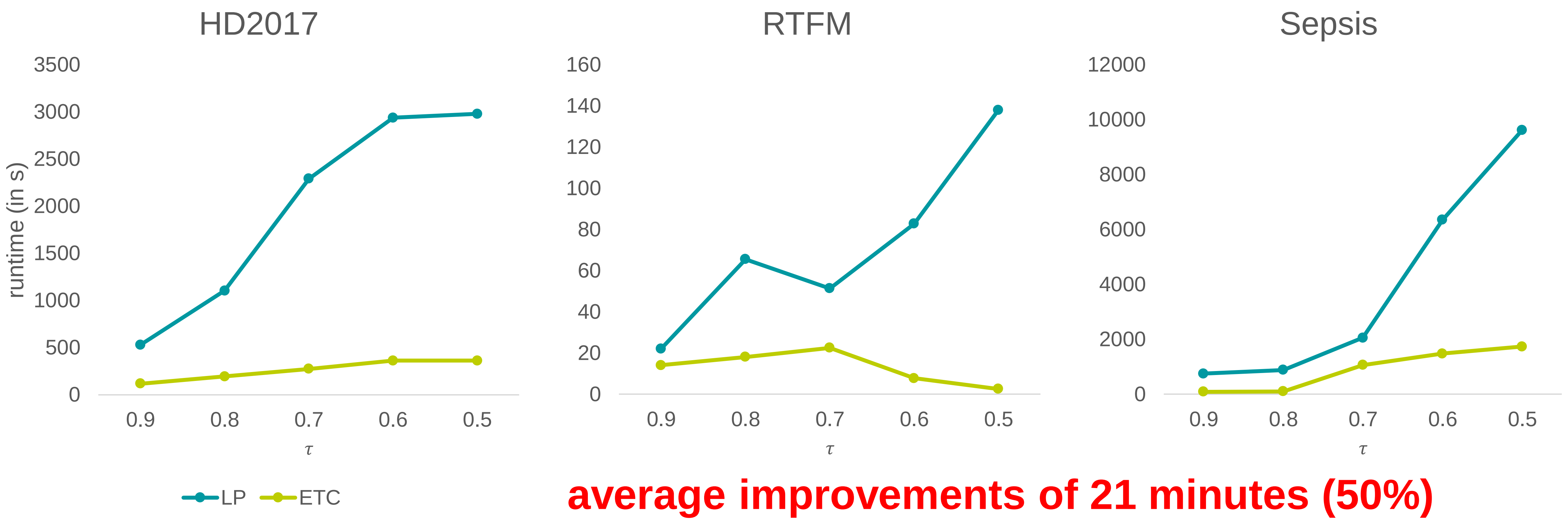
- ETC-based composer (*ETC*)
- ~~region-based IPR (*Region*)~~
- LP-based IPR (*LP*)

Performance Comparison of IPR-Techniques (Real-Life Event Logs, $\tau < 1, d = 4$)

average improvements of 10 seconds (30%)

Performance Comparison of IPR-Techniques (Real-Life Event Logs, $\tau < 1, d = 7$)

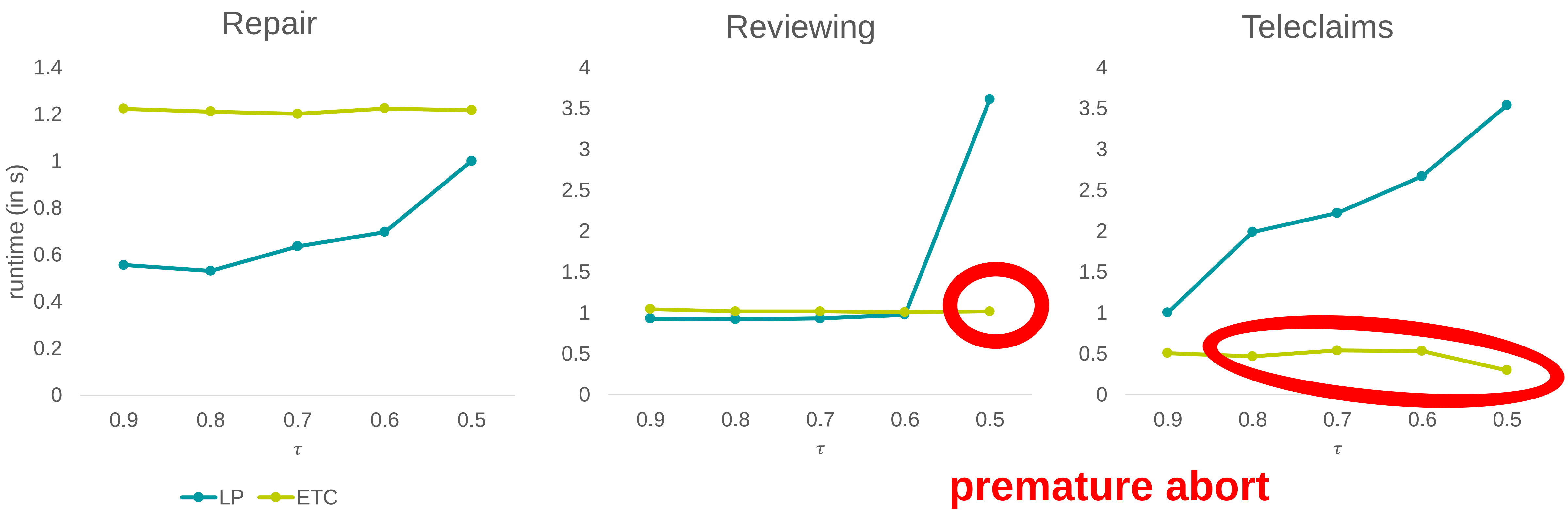
initial example



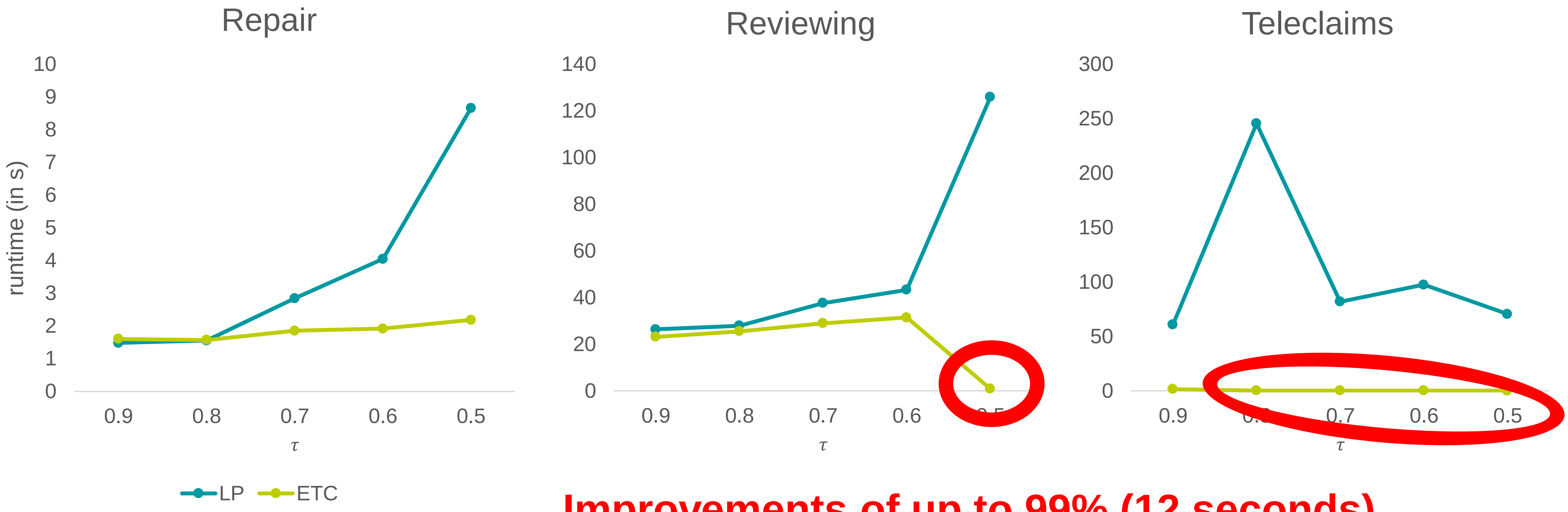
average improvements of 21 minutes (50%)

PADS Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Performance Comparison of IPR-Techniques (Artificial Event Logs, $\tau < 1, d = 4$)



Performance Comparison of IPR-Techniques (Artificial Event Logs, $\tau < 1, d = 7$)



Improvements of up to 99% (12 seconds)

Limitations

For bigger event logs, we are still unable to discover process models efficiently ($d = 4$):

name	#activities	#traces	#variants
BPI-Challenge 2019	**44**	251.734	11.973

- $\tau = 1$: 30 min
- $\tau < 1$: >2 h

eST-Miner in Process Discovery

eST-Miner in Process Discovery

Our improvements in performance enable us to compare the eST-Miner to other process discovery techniques.
Grid search for the ”best model” (considering F1-score) across all parameters of:

- eST-Miner (δ-Variant)
- Alpha Miner
- Inductive Miner
- Heuristic Miner
- Split Miner
- ~~Region-based Algorithms~~ ← too slow (not practically usable)

eST-Miner in Process Discovery

log	eST	Alpha	Heuristic	Inductive	Split
HD2017	**0.983**	0.354	0.952	0.916	0.962
RTFM	0.966	-	0.974	0.895	**0.979**
Sepsis	0.731	-	**0.829**	0.724	-
Repair	0.826	0.310	0.757	0.820	**0.848**
Reviewing	0.809	0.877	**0.999**	0.877	0.959
Teleclaims	0.962	-	0.976	0.959	**1**
% of models that have no possible firing sequence	0%	50%	44.44%	0%	39.1%

not all discovered models satisfy the requirements to compute alignment-based fitness (within 2 hours)

eST-Miner in Process Discovery

- Discovers models in competitive runtime for mid-sized, real-live event logs ($d = 4$)
  - Real-life event logs: < 13s
  - Artificial event logs: < 2s
- Discovers high-quality models (balances well between fitness and precision)
- Can handle noise and infrequent behavior
- Discovers complex control-flow structures
- Provides guarantees:
  - δ-Variant: models can replay at least $\tau \cdot |L|$ traces

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Limitations & Future Work

Limitations:

- Poor performance on logs with many different activities
- Some implicit places remain undetected ($\tau < 1$)
- Some models are overly complex

Future Work: ETC-based Composer

- Usage for conformance checking (incorporate non-fitting log traces instead of using alignments)
- Usage for other process discovery techniques
- Usage of information on precision for the eST-Miner:
  - Prune the search space
  - Guide the search

Thank You For Your Attention!

Key Takeaways: ETC-based Composer

- Efficiently (re)calculates precision of (expanding) Petri nets while incorporating non-fitting log traces
- Place classification: precision-guided implicit place avoidance of (expanding) Petri nets
- Extends the discovery with information on precision

- Exemplary use case: eST-Miner + ETC-based Composer → competitive choice for process discovery