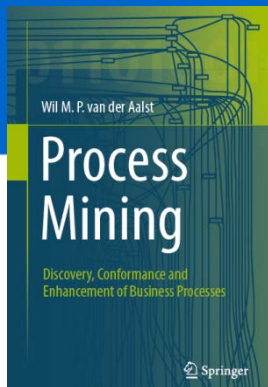


Process Mining: Data Science in Action

Conformance Checking Using Token-Based Replay

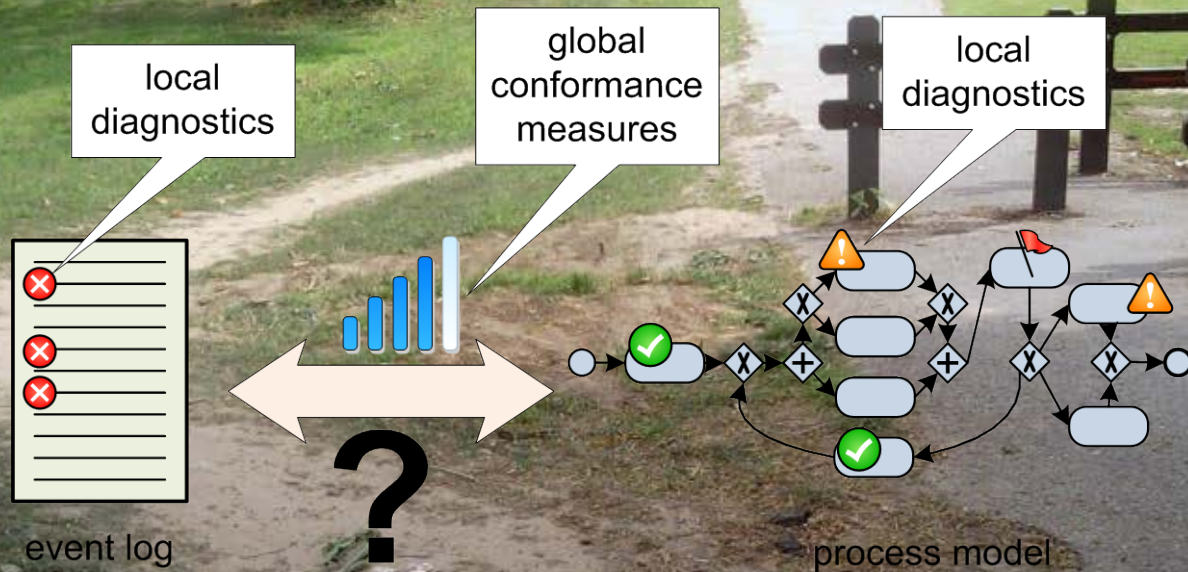
prof.dr.ir. Wil van der Aalst
www.processmining.org



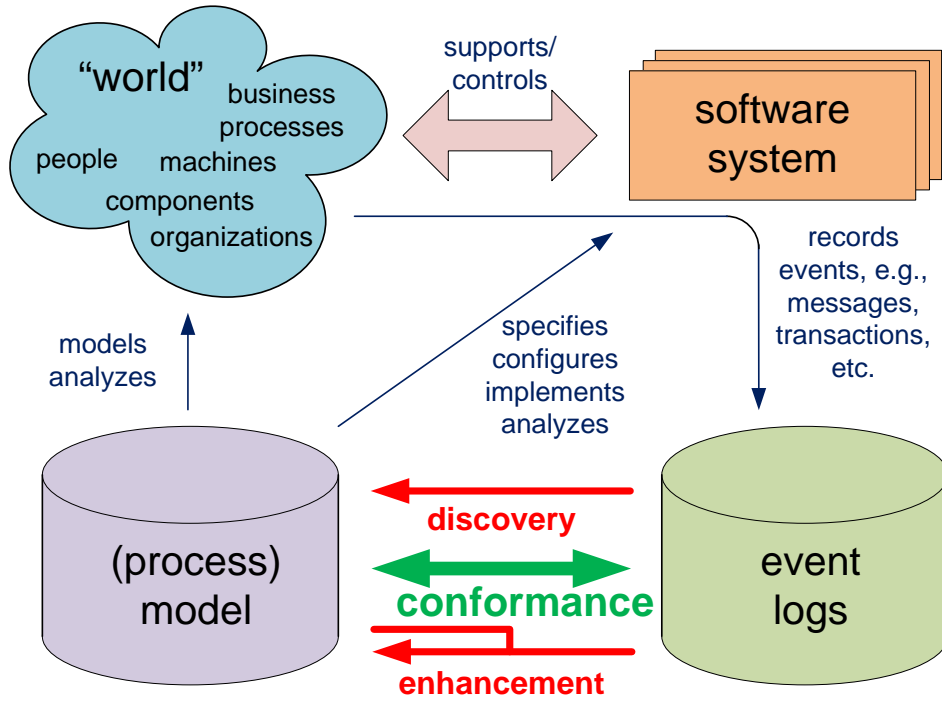
TU/e

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

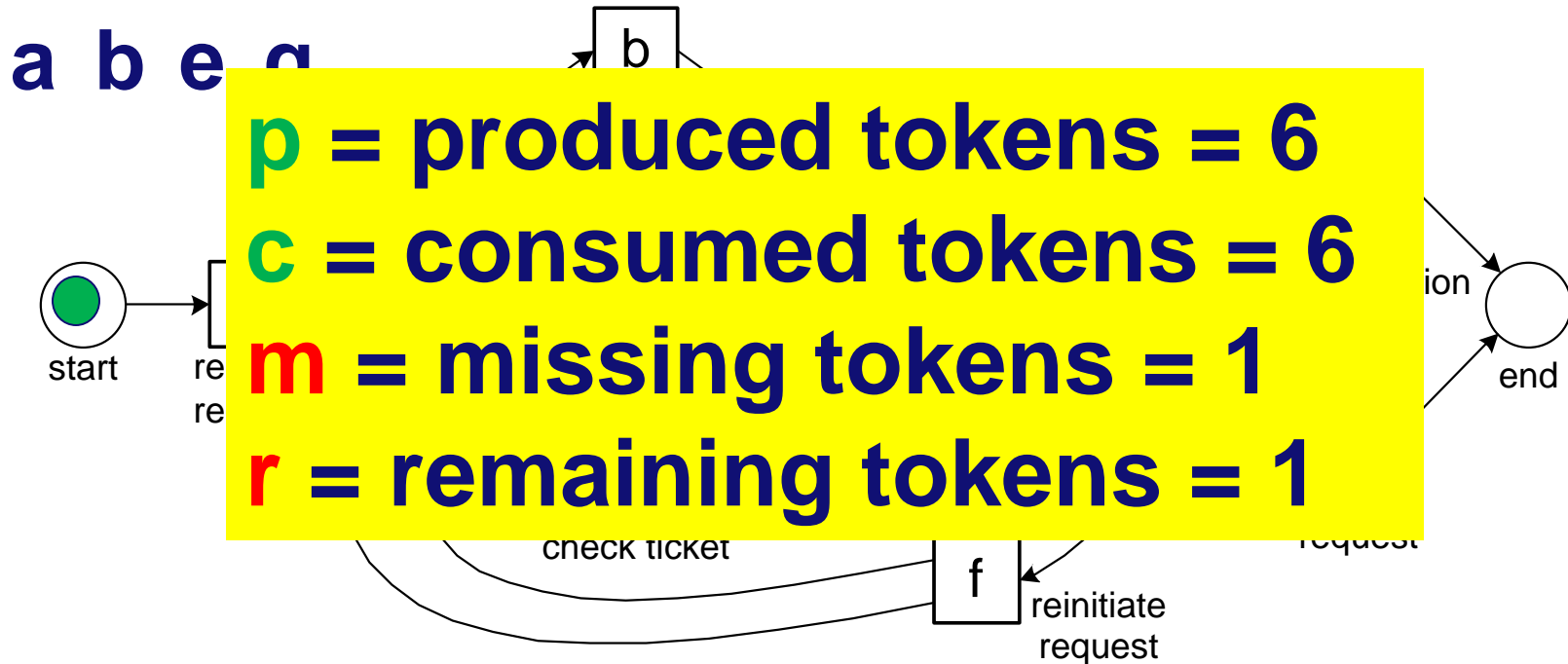


Conformance checking



1. Conformance checking using causal footprints.
2. Conformance checking based on **token-based replay**.
3. Alignment-based conformance checking.

Counting tokens while replaying



Quantifying fitness at the trace level

$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{1}{6} \right) + \frac{1}{2} \left(1 - \frac{1}{6} \right) = 0.83333$$

p = produced tokens = 6

c = consumed tokens = 6

m = missing tokens = 1

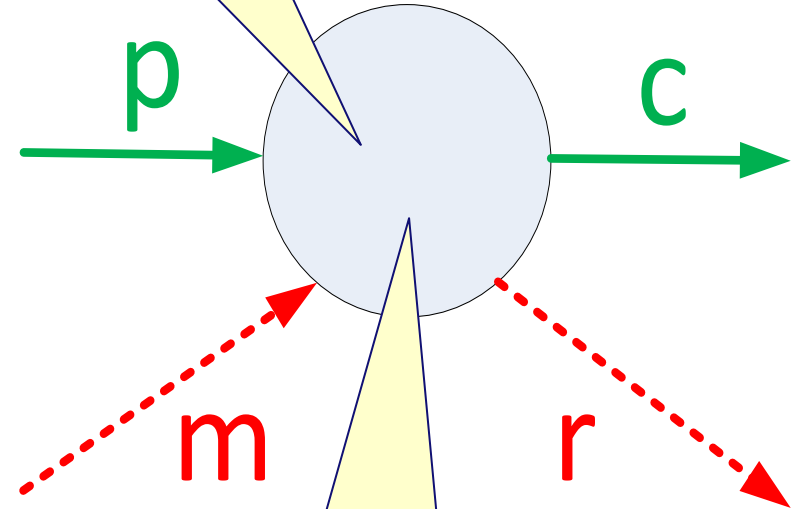
r = remaining tokens = 1

Approach (1/3)

Use four counters:

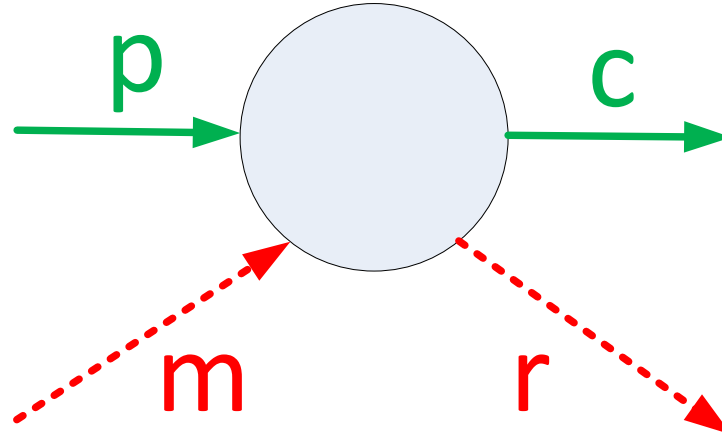
- **p** = produced tokens
- **c** = consumed tokens
- **m** = missing tokens
(consumed while not there)
- **r** = remaining tokens
(produced but not consumed)

while running
 $p+m-c$ tokens



at end no
tokens

Approach (2/3)

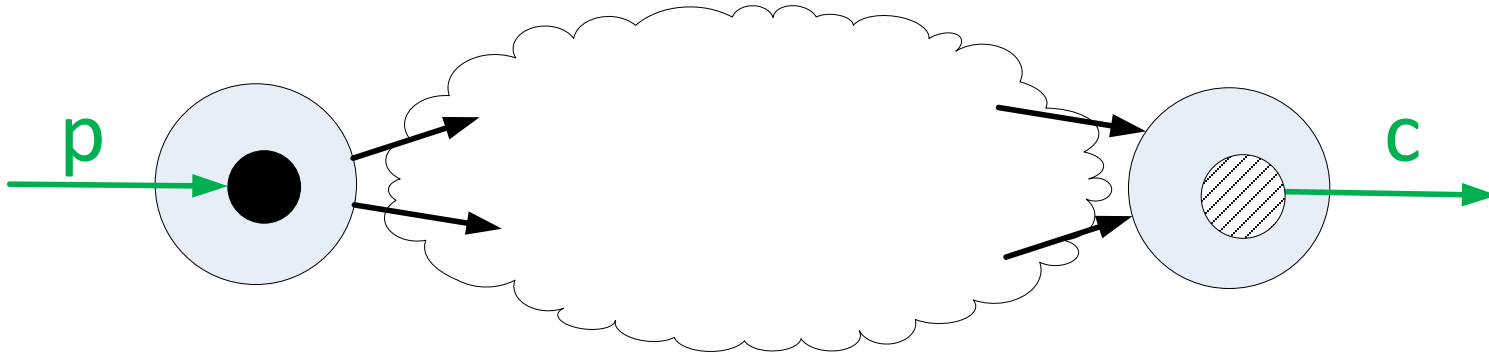


- Invariants

- At any time: $p + m \geq c \geq m$ (also per place)

- At the end: $r = p + m - c$ (also per place)

Approach (3/3)

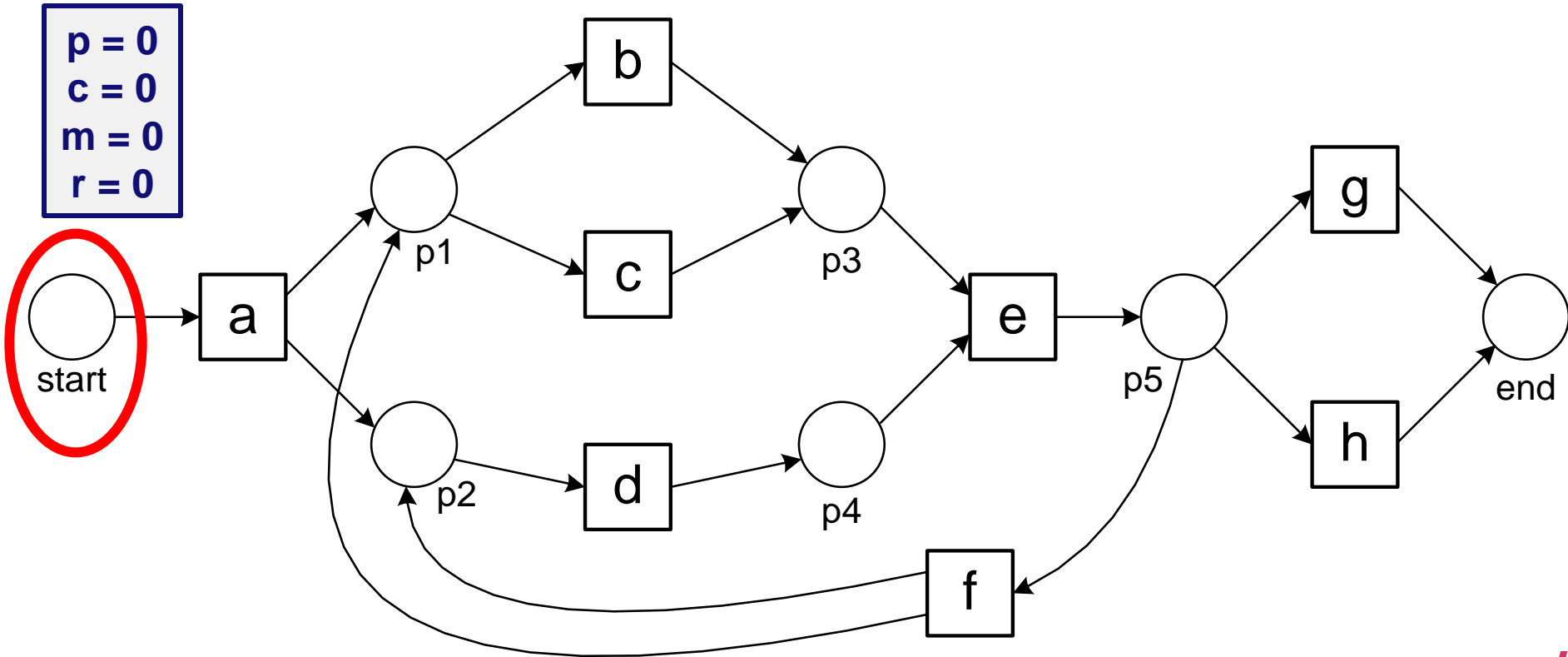


Initialization and finalization:

- In the beginning a token is **produced** for the source place: $p = 1$.
- At the end a token is **consumed** from the sink place (also if not there): $c' = c + 1$.

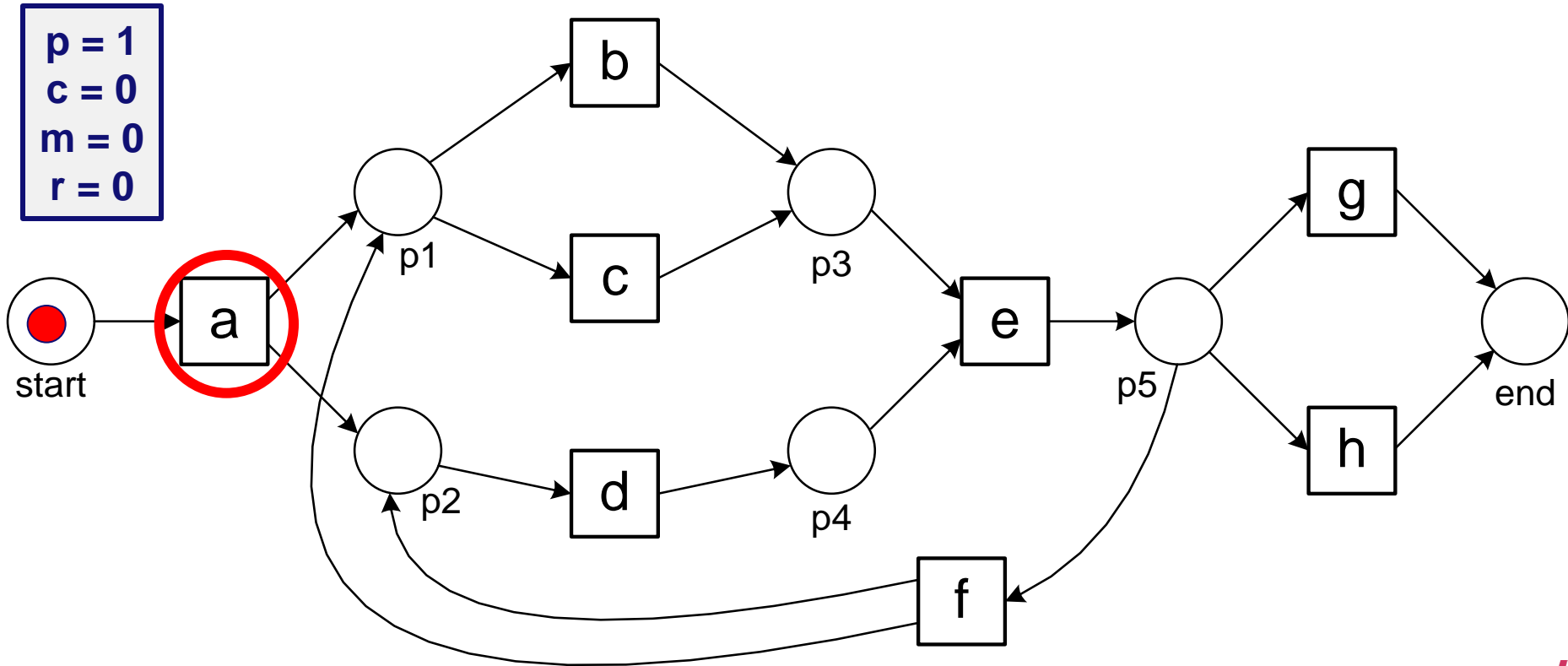
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



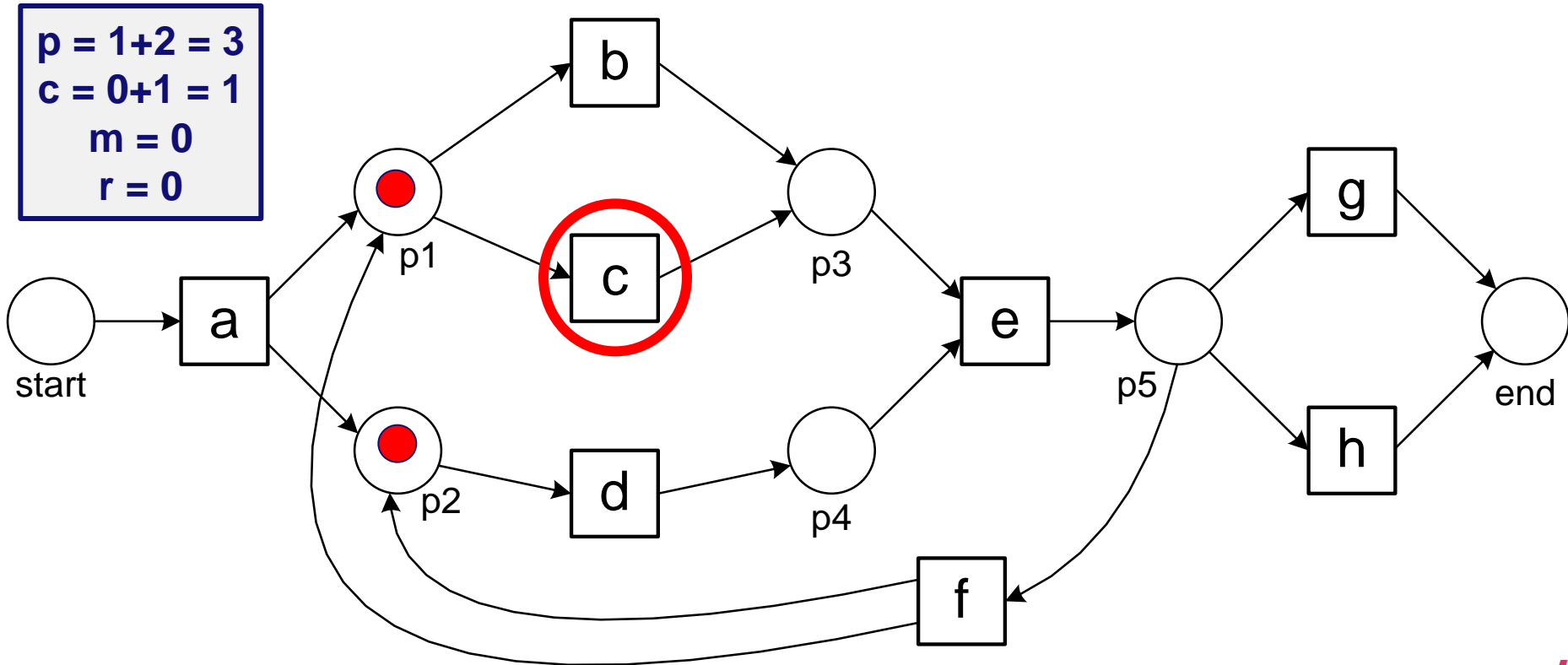
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



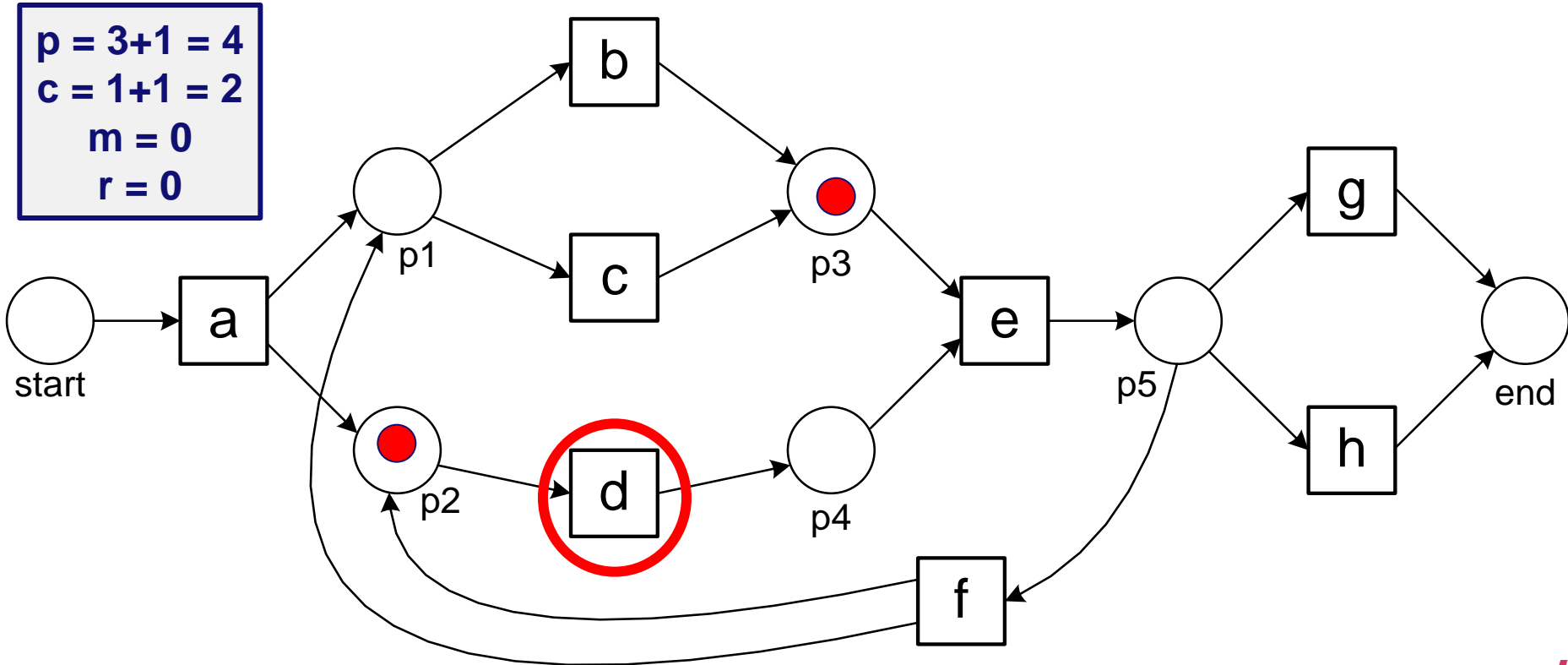
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



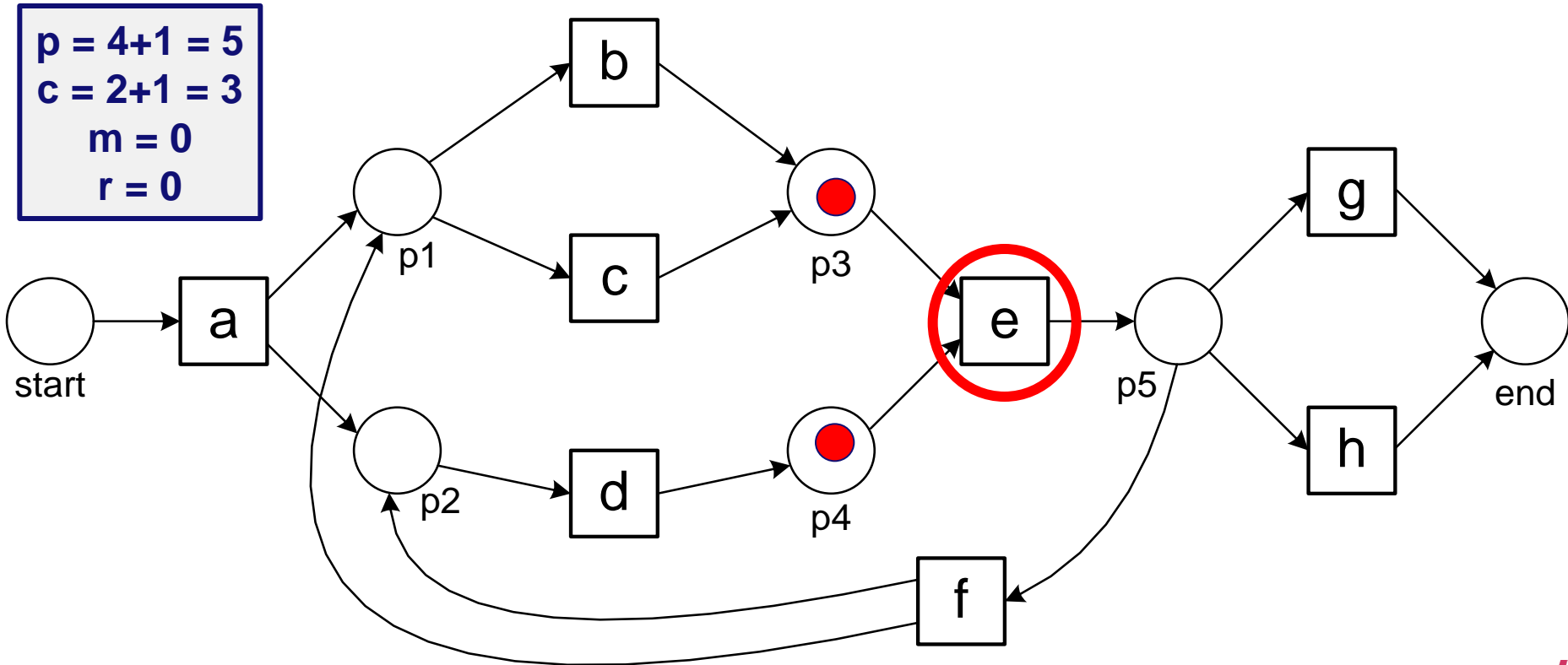
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



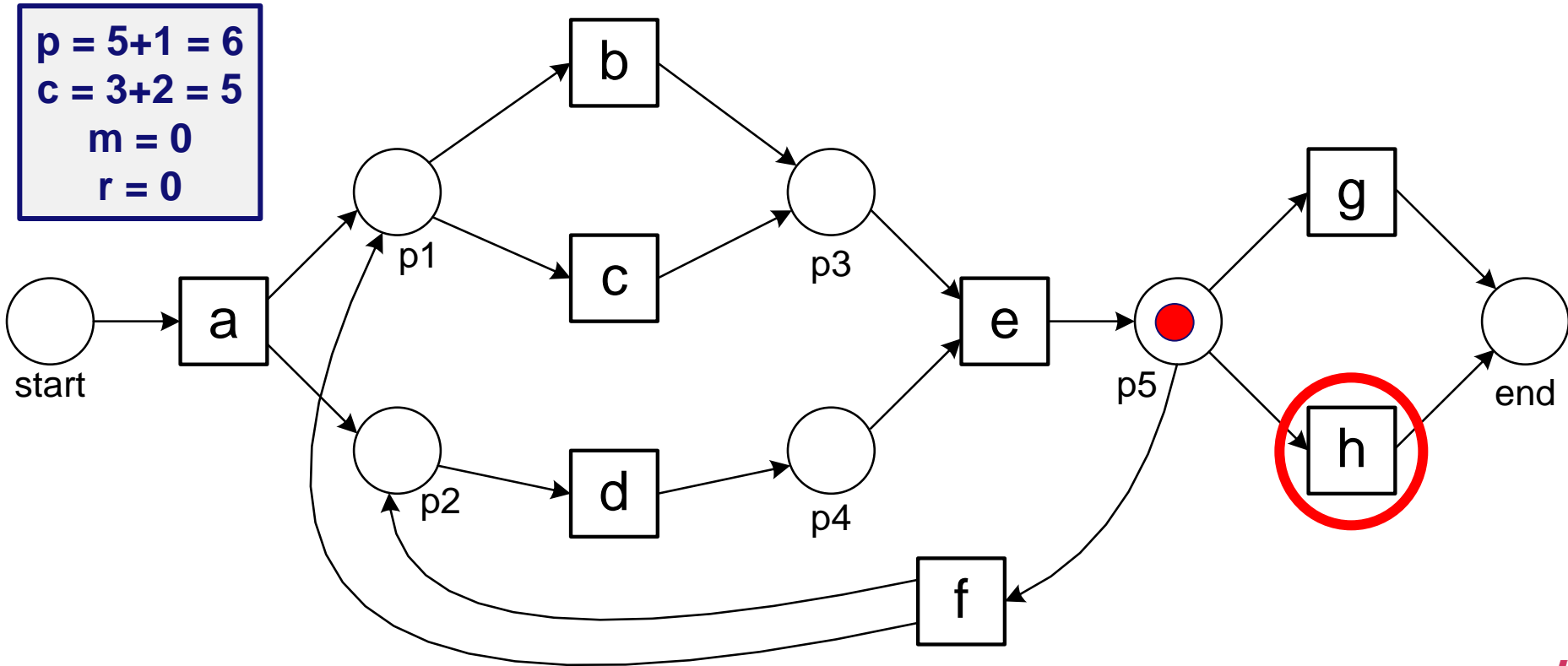
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



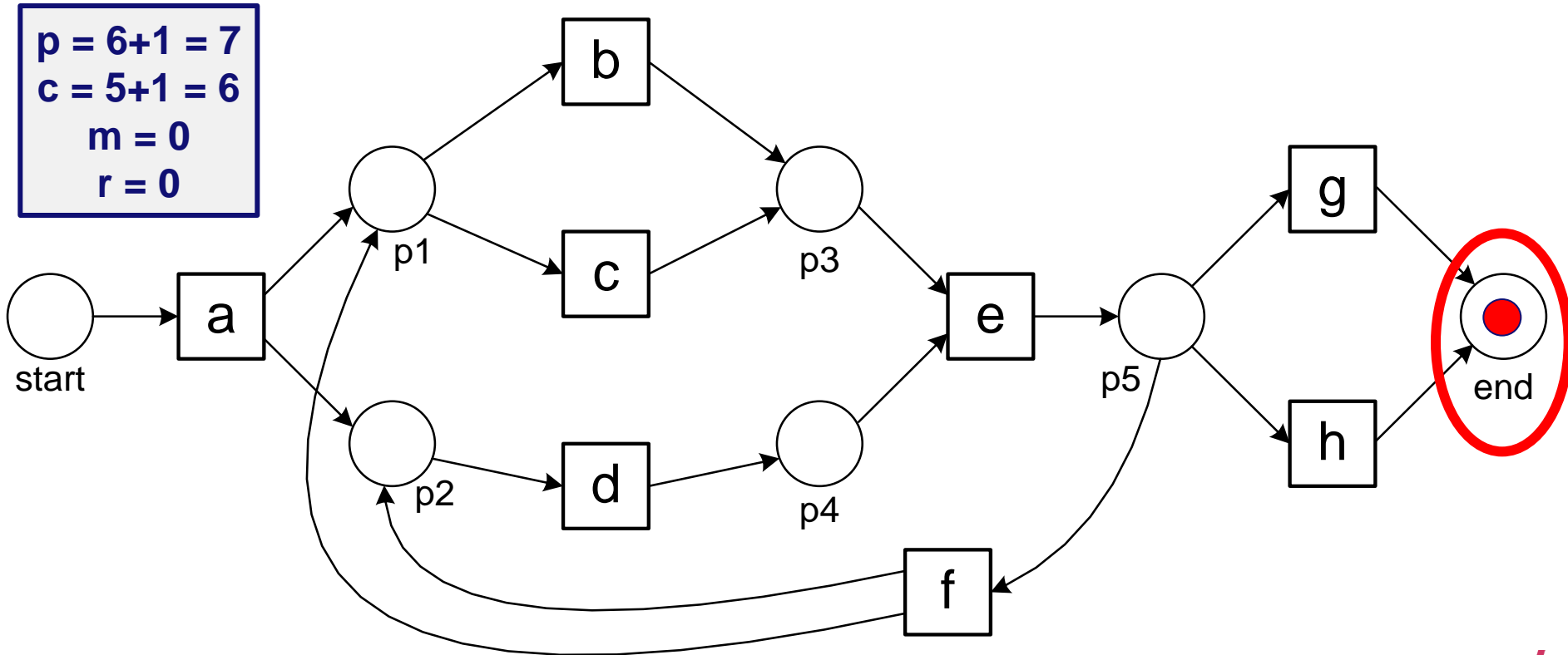
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



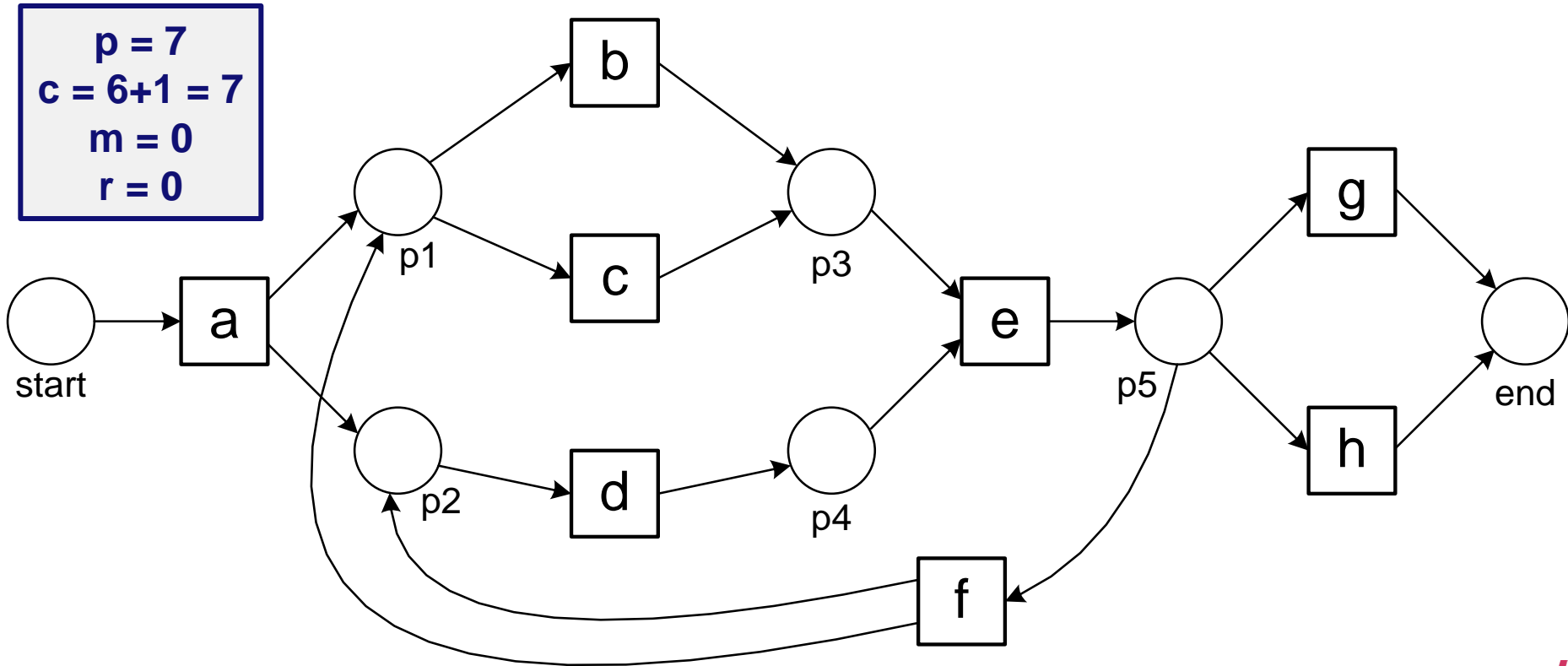
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



Replaying

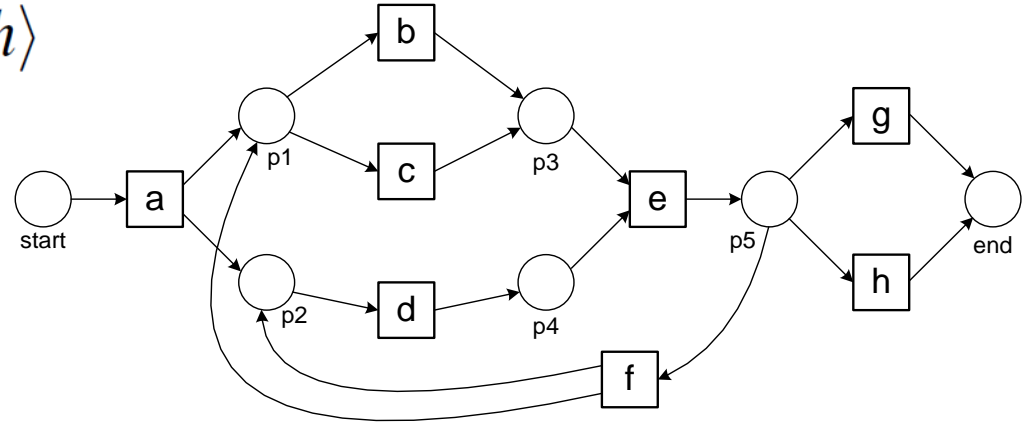
$$\sigma_1 = \langle a, c, d, e, h \rangle$$



Quantifying fitness at the trace level

p = 7
c = 7
m = 0
r = 0

$\sigma_1 = \langle a, c, d, e, h \rangle$

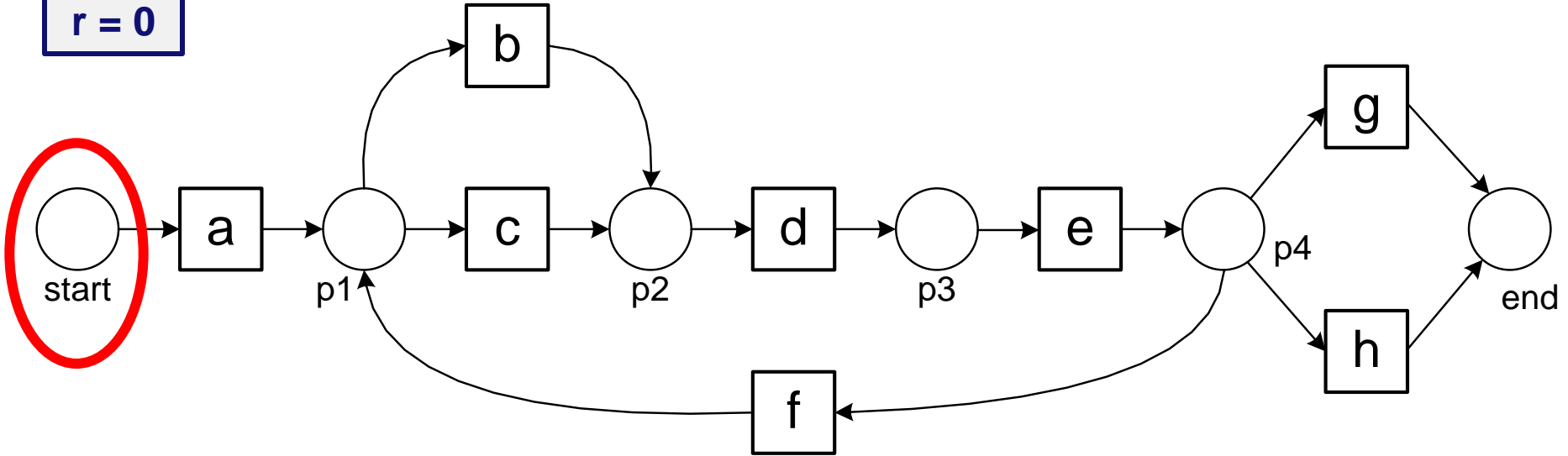


$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{\mathbf{0}}{\mathbf{7}} \right) + \frac{1}{2} \left(1 - \frac{\mathbf{0}}{\mathbf{7}} \right) = \mathbf{1}$$

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

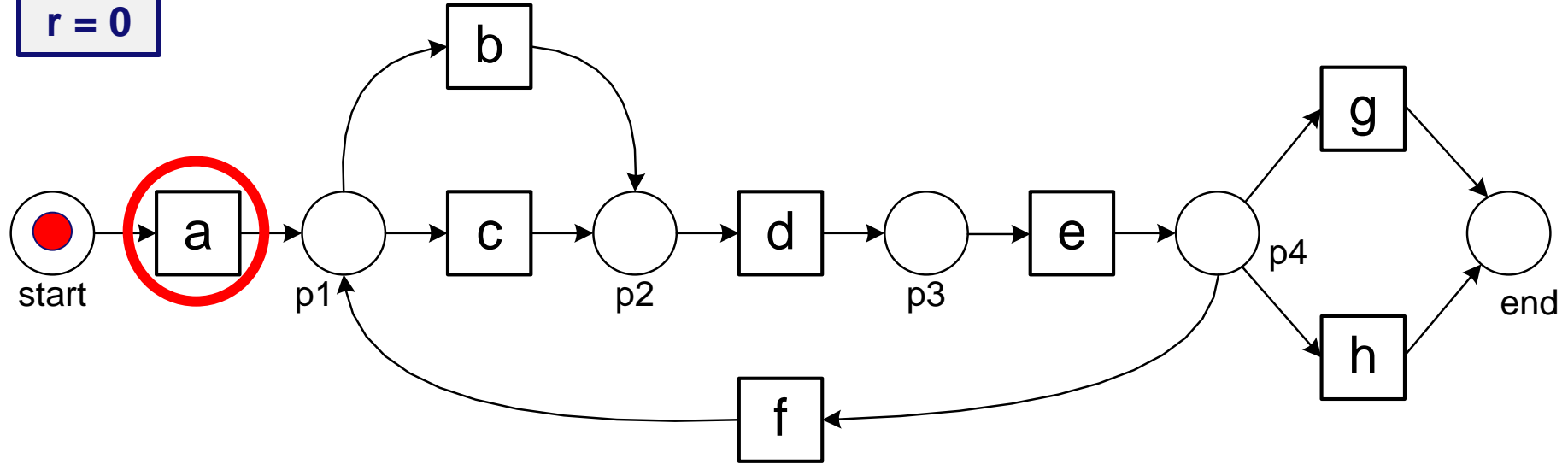
$p = 0$
 $c = 0$
 $m = 0$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

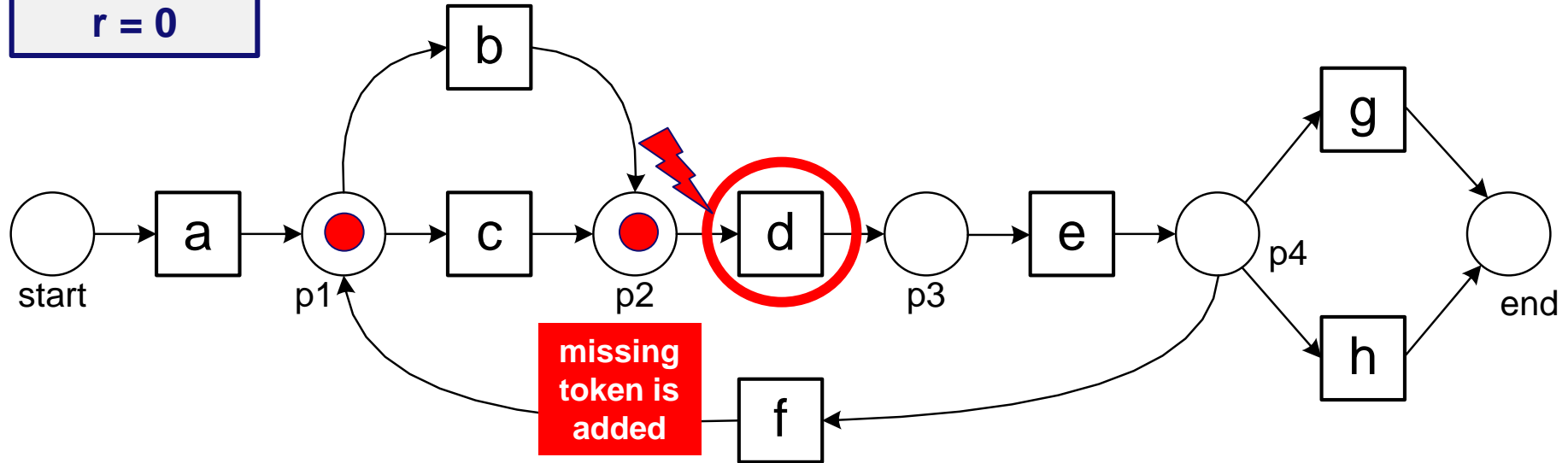
$p = 1$
 $c = 0$
 $m = 0$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

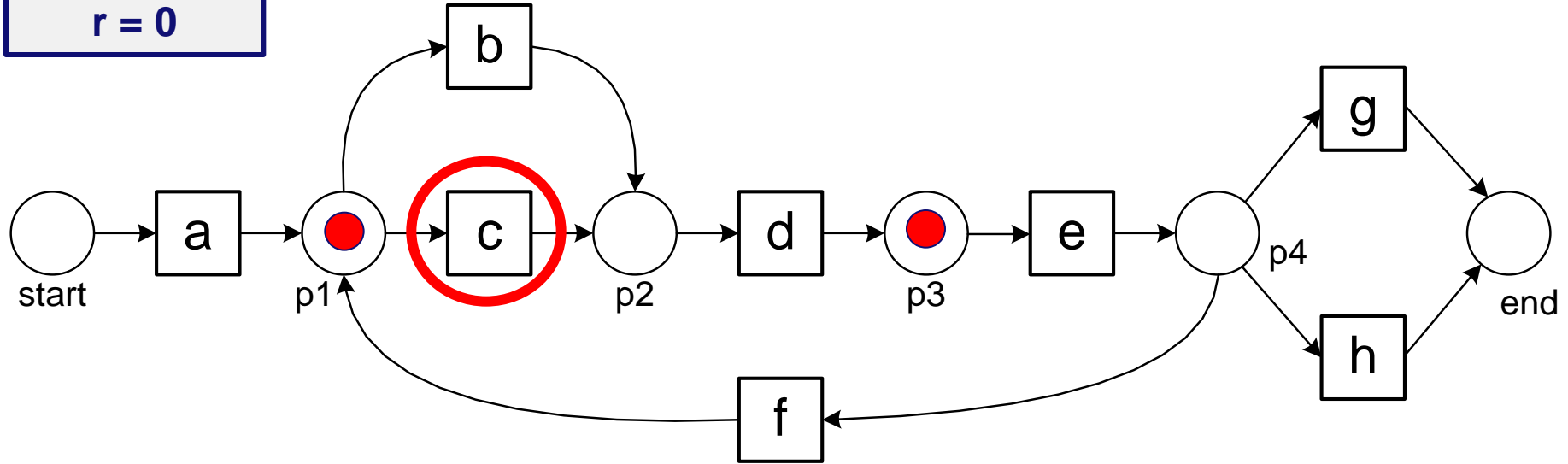
$p = 1+1 = 2$
 $c = 0+1 = 1$
 $m = 0$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

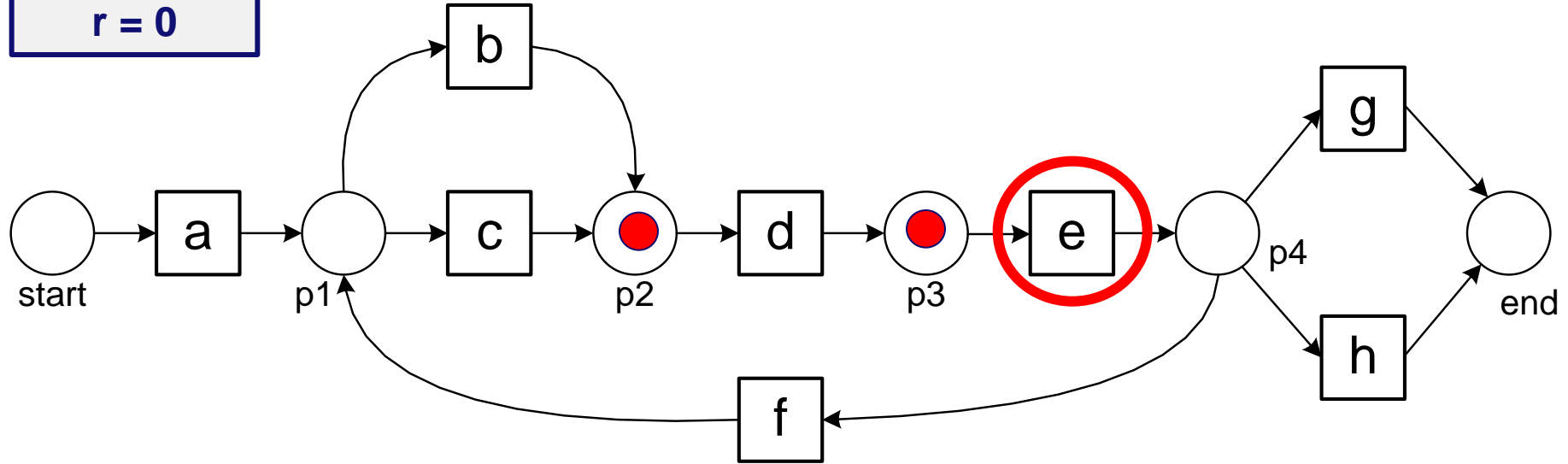
$p = 2+1 = 3$
 $c = 1+1 = 2$
 $m = 0+1 = 1$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, \textcolor{red}{e}, h \rangle$$

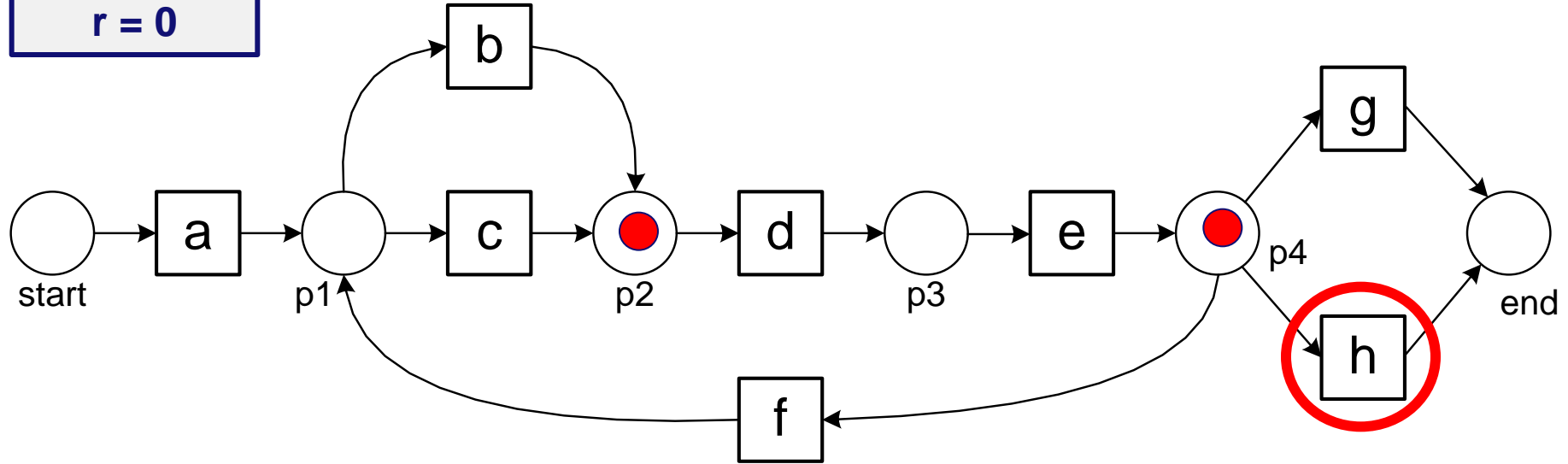
$p = 3 + 1 = 4$
 $c = 2 + 1 = 3$
 $m = 1$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

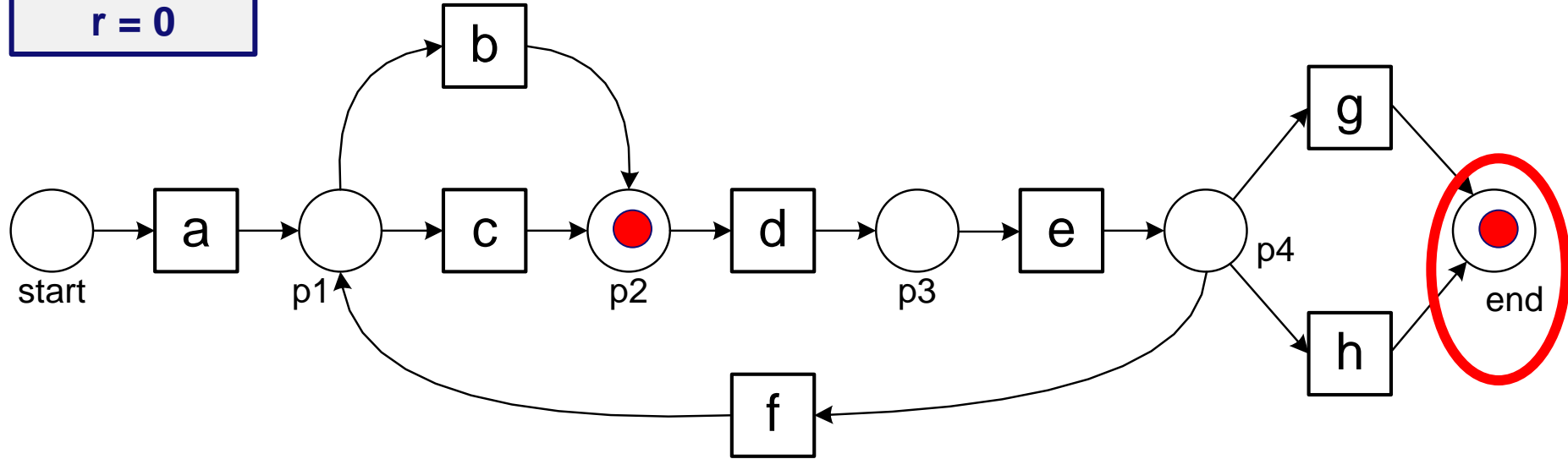
$p = 4 + 1 = 5$
 $c = 3 + 1 = 4$
 $m = 1$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

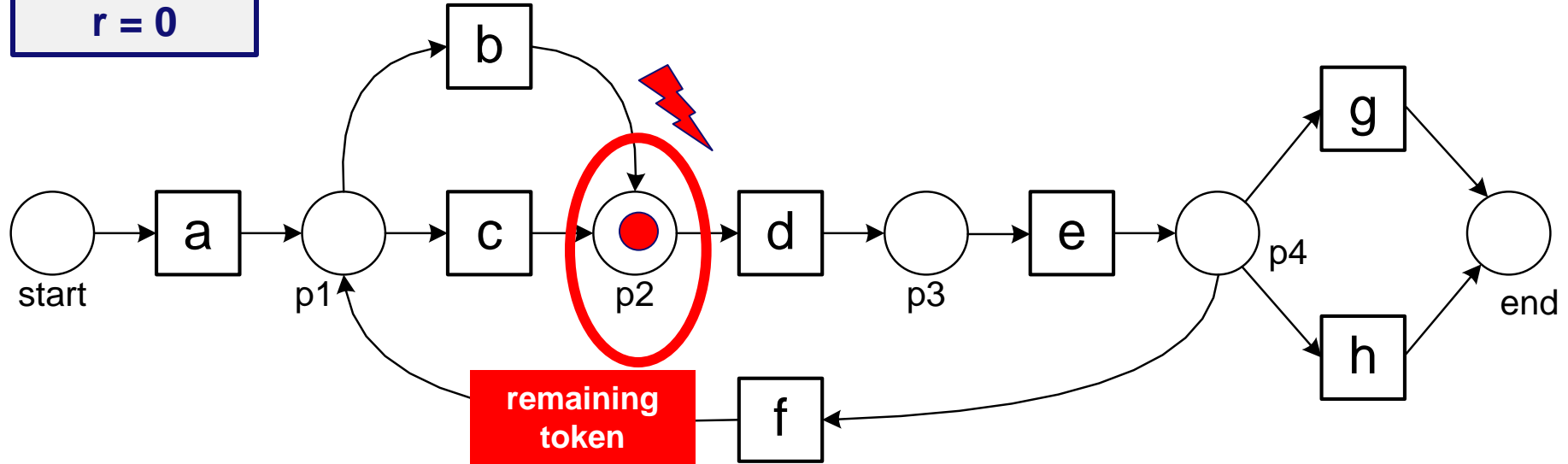
$p = 5 + 1 = 6$
 $c = 4 + 1 = 5$
 $m = 1$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

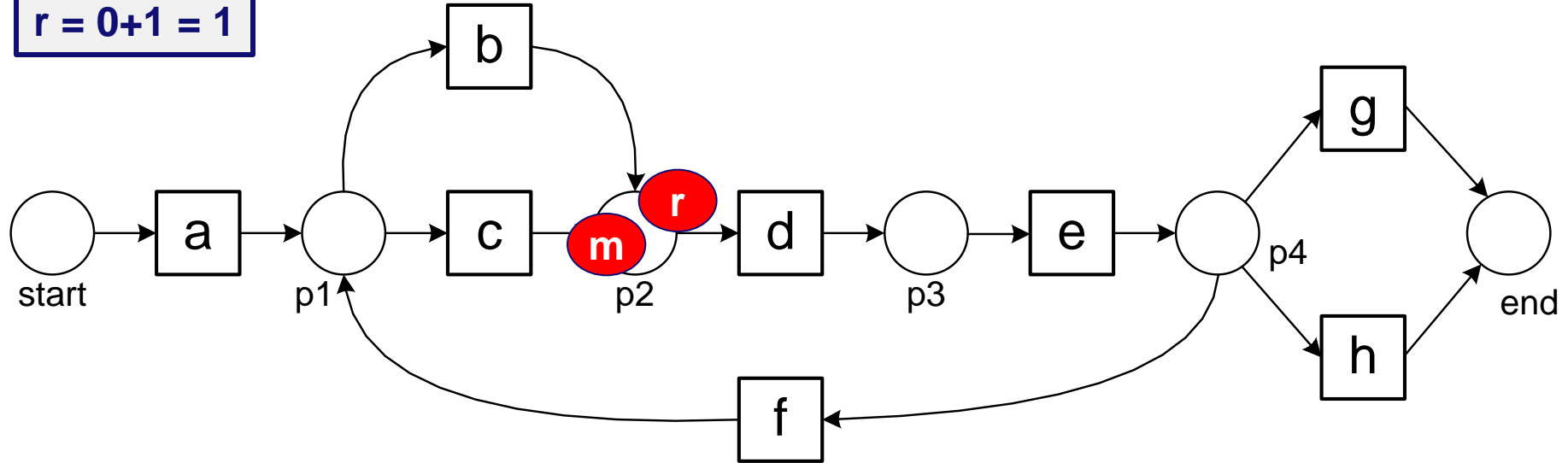
$p = 6$
 $c = 5 + 1 = 6$
 $m = 1$
 $r = 0$



Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

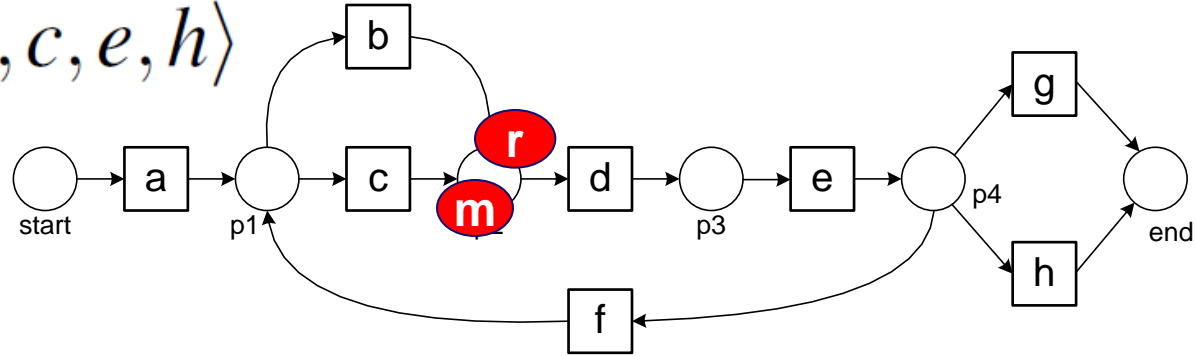
$p = 6$
 $c = 6$
 $m = 1$
 $r = 0 + 1 = 1$



Quantifying fitness at the trace level

$p = 6$
 $c = 6$
 $m = 1$
 $r = 1$

$\sigma_3 = \langle a, d, c, e, h \rangle$



$$fitness(\sigma, N) = \frac{1}{2} \left(1 - \frac{1}{6} \right) + \frac{1}{2} \left(1 - \frac{1}{6} \right) = 0.8333$$

Fitness at the log level

missing
tokens

consumed
tokens

remaining
tokens

produced
tokens

$$fitness(L, N) = \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N, \sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N, \sigma}} \right) + \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N, \sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N, \sigma}} \right)$$

Looks scary but it just needs the sums of p , c , m , and r over the multiset of traces in N .

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefdbeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefdbeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

?



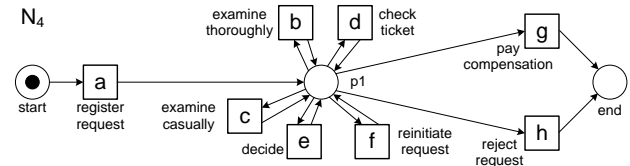
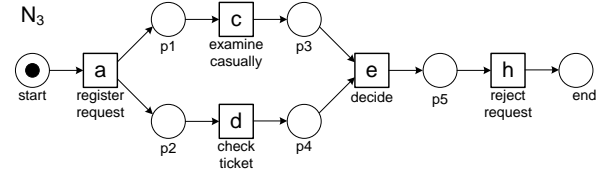
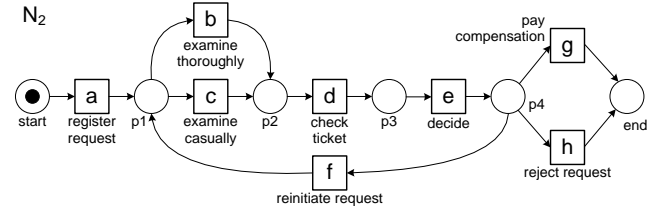
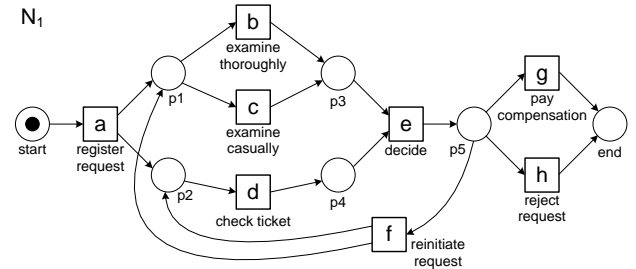
$$fitness(L, N) = \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N, \sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N, \sigma}} \right) + \frac{1}{2} \left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N, \sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N, \sigma}} \right)$$

$$fitness(L_{full}, N_1) = 1$$

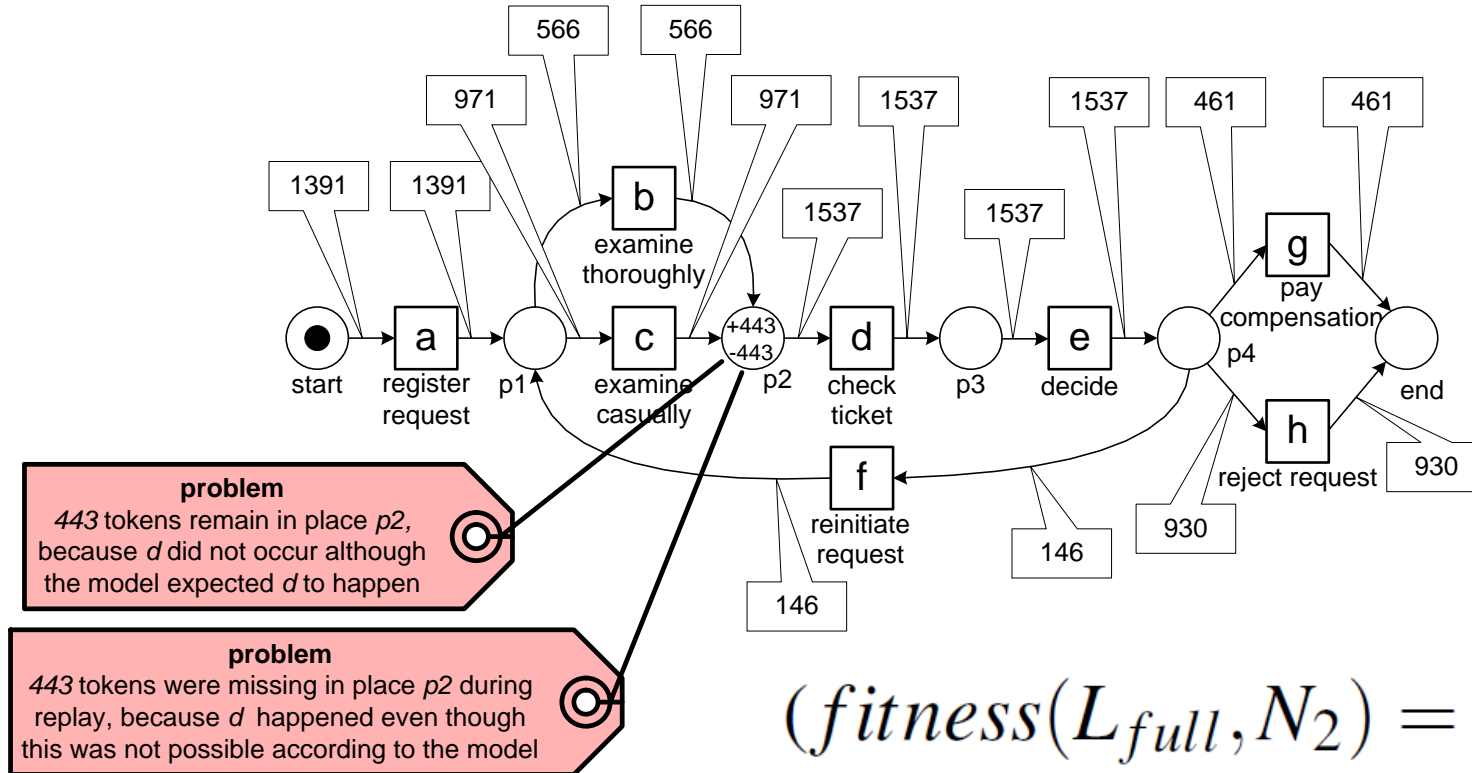
$$fitness(L_{full}, N_2) = 0.9504$$

$$fitness(L_{full}, N_3) = 0.8797$$

$$fitness(L_{full}, N_4) = 1$$

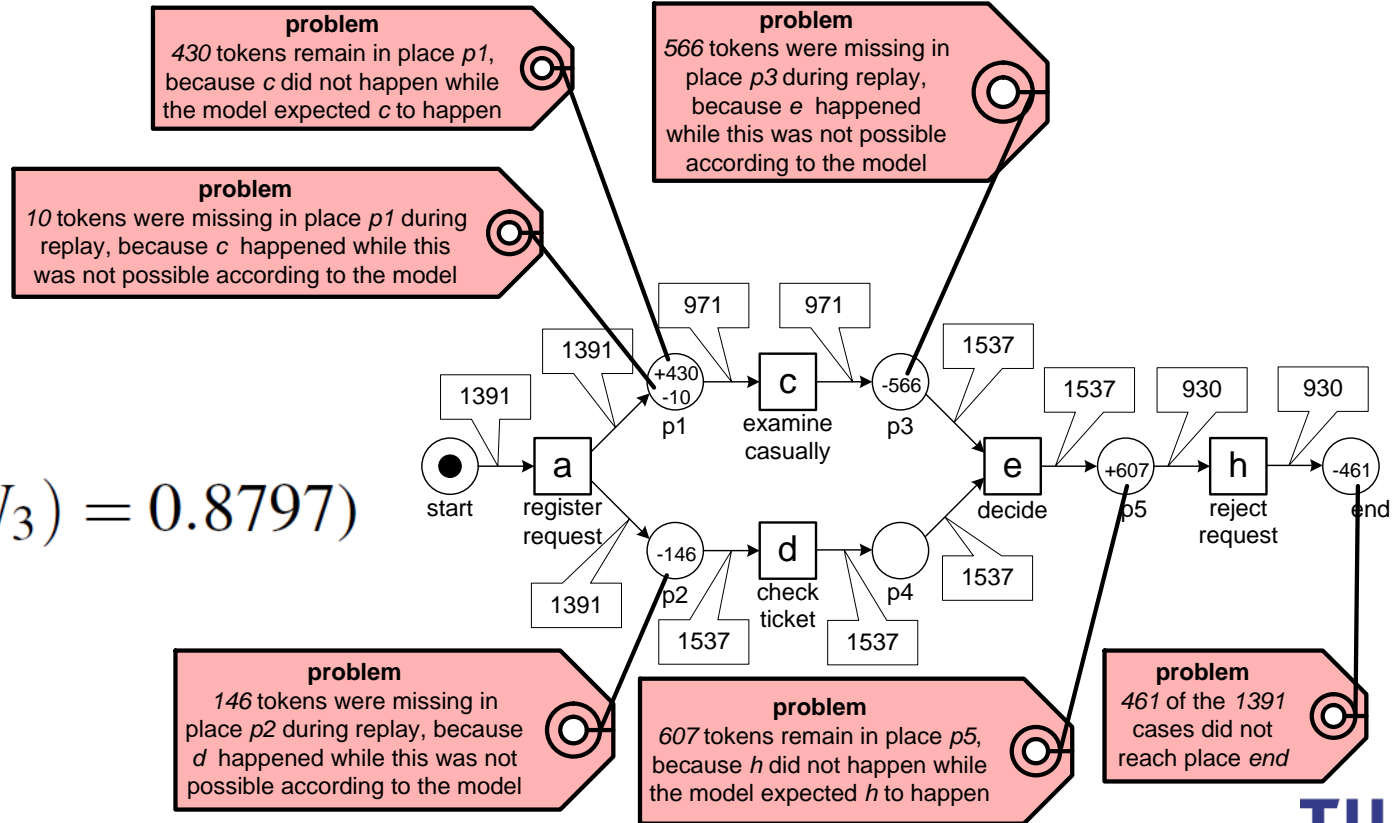


Diagnostics



Diagnostics

$$(fitness(L_{full}, N_3) = 0.8797)$$



Part I: Preliminaries

Chapter 1
Introduction

Chapter 2
Process Modeling and
Analysis

Chapter 3
Data Mining

Part II: From Event Logs to Process Models

Chapter 4
Getting the Data

Chapter 5
Process Discovery: An
Introduction

Chapter 6
Advanced Process
Discovery Techniques

Part III: Beyond Process Discovery

Chapter 7
Conformance
Checking

Chapter 8
Mining Additional
Perspectives

Chapter 9
Operational Support

Part IV: Putting Process Mining to Work

Chapter 10
Tool Support

Chapter 11
Analyzing “Lasagna
Processes”

Chapter 12
Analyzing “Spaghetti
Processes”

Part V: Reflection

Chapter 13
Cartography and
Navigation

Chapter 14
Epilogue

