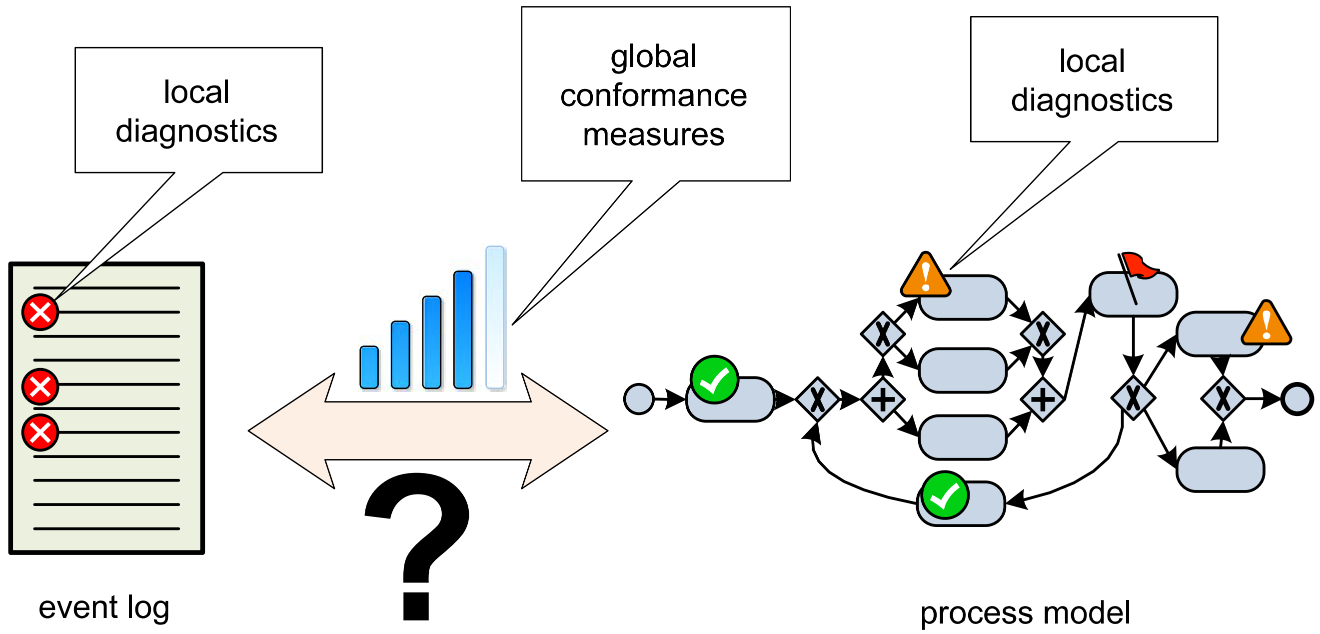
Business Processes Modelling

MPB (6 cfu, 295AA)

Roberto Bruni
http://www.di.unipi.it/~bruni

19 - Conformance checking

Object



We overview the key principles of process mining

Chapters 1, 6. Process Mining. W. van der Aalst

Conformance Checking: fitness measures

Measures and Diagnostic

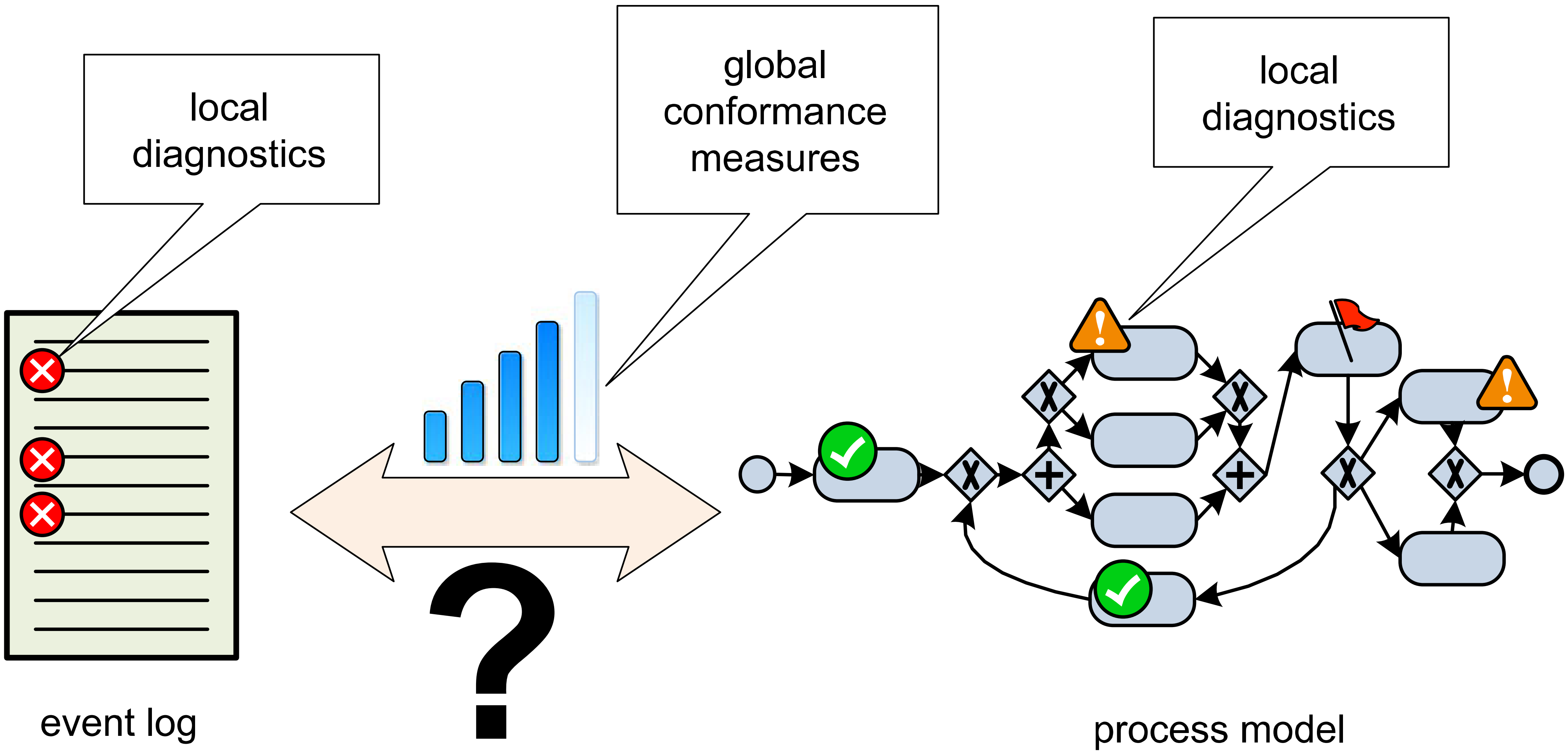


Fig. 7.1 Conformance checking: comparing observed behavior with modeled behavior. Global conformance measures quantify the overall conformance of the model and log. Local diagnostics are given by highlighting the nodes in the model where model and log disagree. Cases that do not fit are highlighted in the visualization of the log

Measuring Fitness

Fitness measures “the proportion of behaviour in the event log possible according to the model”.

Of the four quality criteria,
fitness is the closest to conformance.

A naïve approach toward conformance checking would be to count the fraction of cases that can be “**replayed**” (i.e., the proportion of cases corresponding to firing sequences leading from [start] to [end]).

Ability to replay

Can the net N replay the trace σ ?

is equivalent to ask if

does $\sigma \in L(N)$?
(is σ in the language of N ?)

when $\sigma \notin L(N)$ we say that
σ is **non-fitting** for N

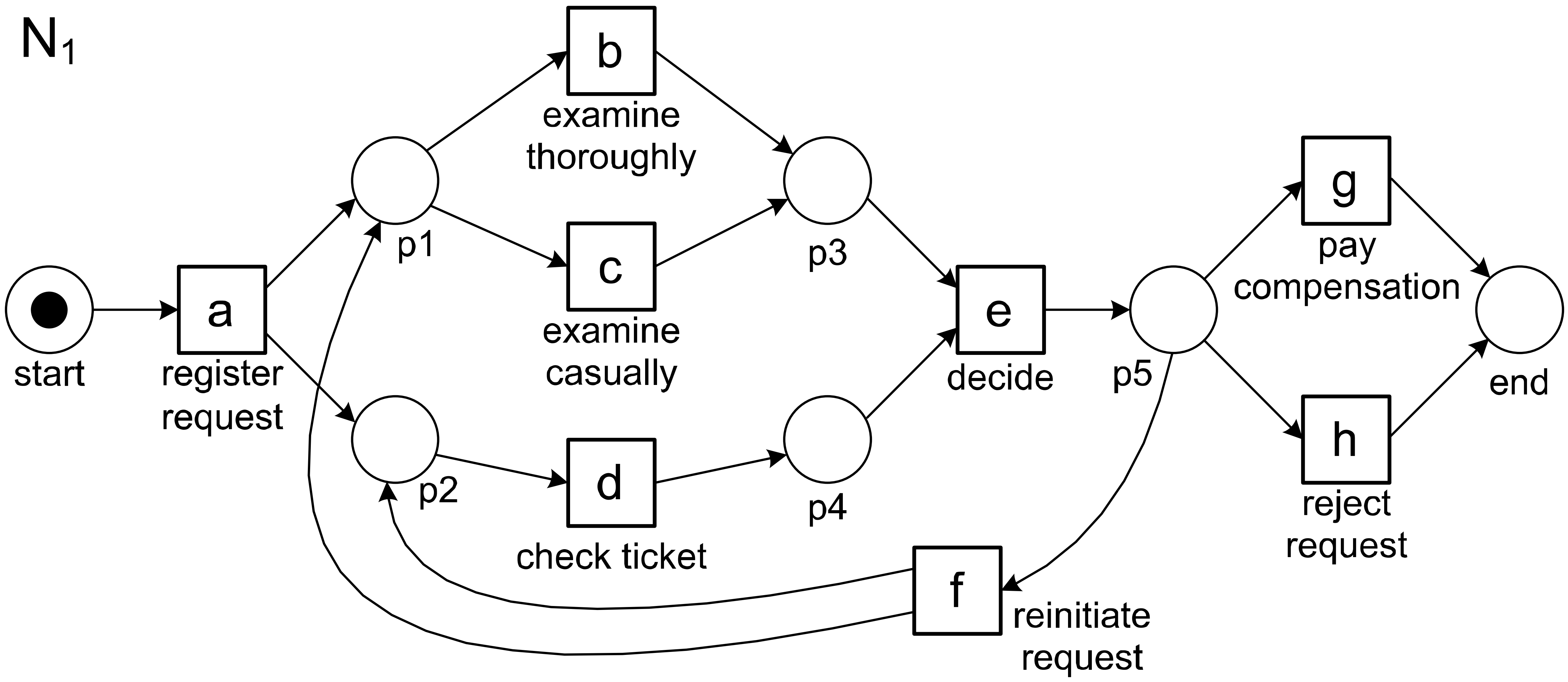
Table 7.1 Event log L_{full}: a = *register request*, b = *examine thoroughly*, c = *examine casually*, d = *check ticket*, e = *decide*, f = *reinitiate request*, g = *pay compensation*, and h = *reject request*

1391 cases

Example

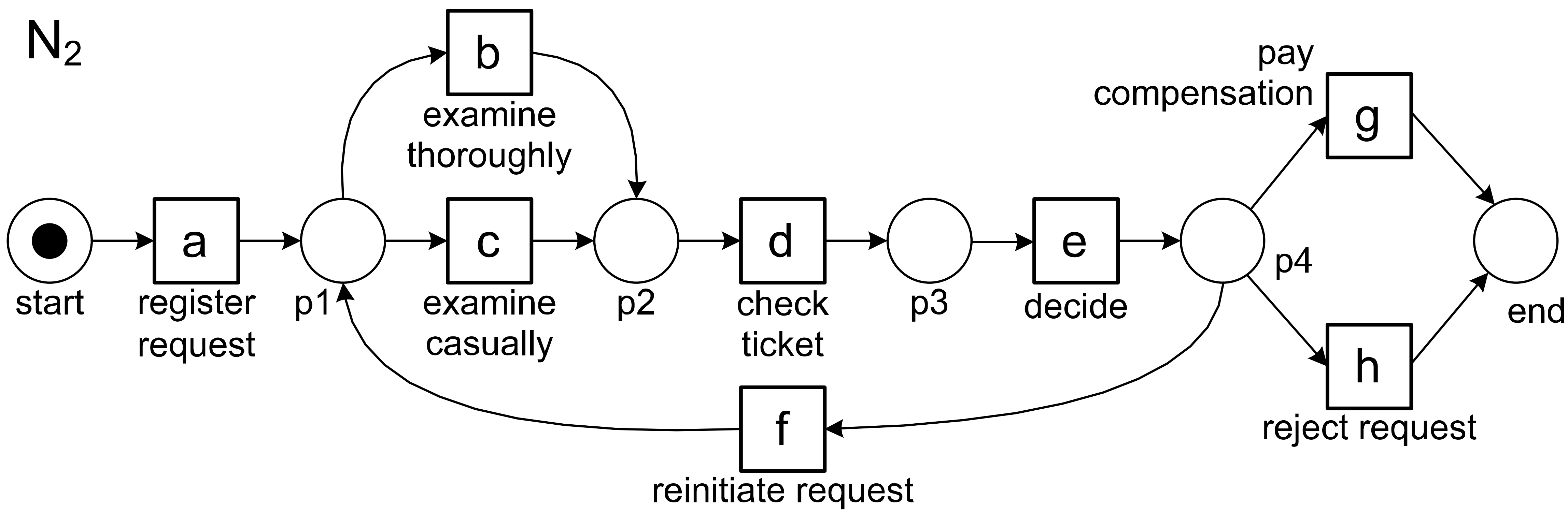
Frequency	Reference	Trace
455	σ_1	$\langle a, c, d, e, h \rangle$
191	σ_2	$\langle a, b, d, e, g \rangle$
177	σ_3	$\langle a, d, c, e, h \rangle$
144	σ_4	$\langle a, b, d, e, h \rangle$
111	σ_5	$\langle a, c, d, e, g \rangle$
82	σ_6	$\langle a, d, c, e, g \rangle$
56	σ_7	$\langle a, d, b, e, h \rangle$
47	σ_8	$\langle a, c, d, e, f, d, b, e, h \rangle$
38	σ_9	$\langle a, d, b, e, g \rangle$
33	σ_{10}	$\langle a, c, d, e, f, b, d, e, h \rangle$
14	σ_{11}	$\langle a, c, d, e, f, b, d, e, g \rangle$
11	σ_{12}	$\langle a, c, d, e, f, d, b, e, g \rangle$
9	σ_{13}	$\langle a, d, c, e, f, c, d, e, h \rangle$
8	σ_{14}	$\langle a, d, c, e, f, d, b, e, h \rangle$
5	σ_{15}	$\langle a, d, c, e, f, b, d, e, g \rangle$
3	σ_{16}	$\langle a, c, d, e, f, b, d, e, f, d, b, e, g \rangle$
2	σ_{17}	$\langle a, d, c, e, f, d, b, e, g \rangle$
2	σ_{18}	$\langle a, d, c, e, f, b, d, e, f, b, d, e, g \rangle$
1	σ_{19}	$\langle a, d, c, e, f, d, b, e, f, b, d, e, h \rangle$
1	σ_{20}	$\langle a, d, b, e, f, b, d, e, f, d, b, e, g \rangle$
1	σ_{21}	$\langle a, d, c, e, f, d, b, e, f, c, d, e, f, d, b, e, g \rangle$

Example N1



naïve fitness $\frac{1391}{1391} = 1$ The net can ``replay'' any trace

Example N2

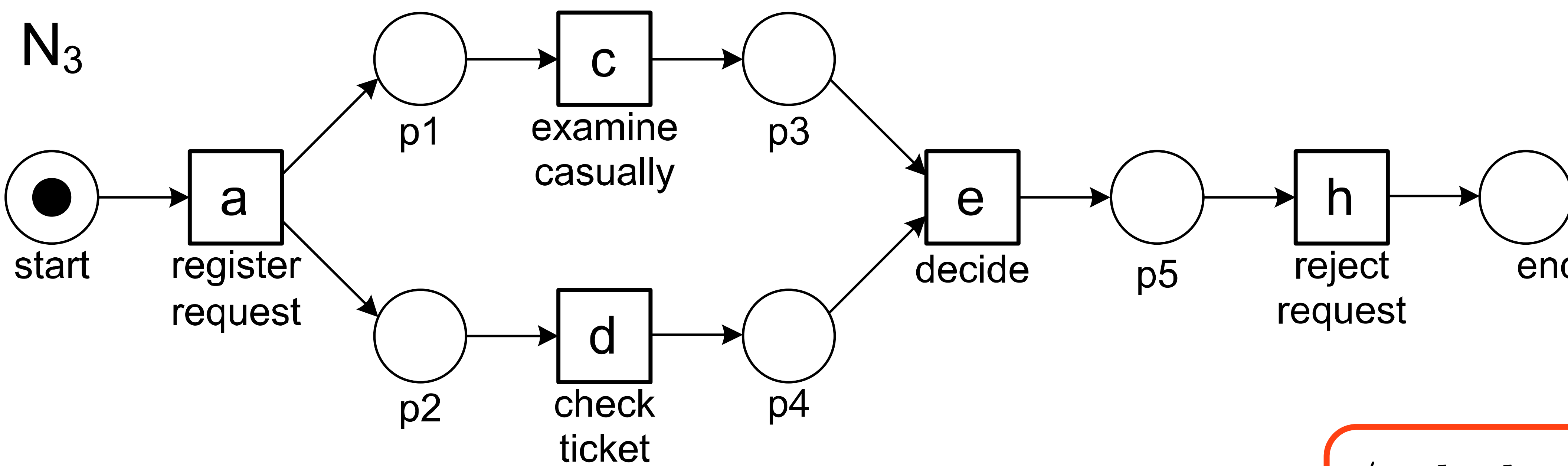


443 cases do not correspond to a firing sequence

$\langle a, d, c, e, h \rangle^{177}$
$\langle a, d, c, e, g \rangle^{82}$
$\langle a, d, b, e, h \rangle^{56}$
$\cdots$

naïve fitness $\frac{948}{1391} = 0.6815$

Example N3

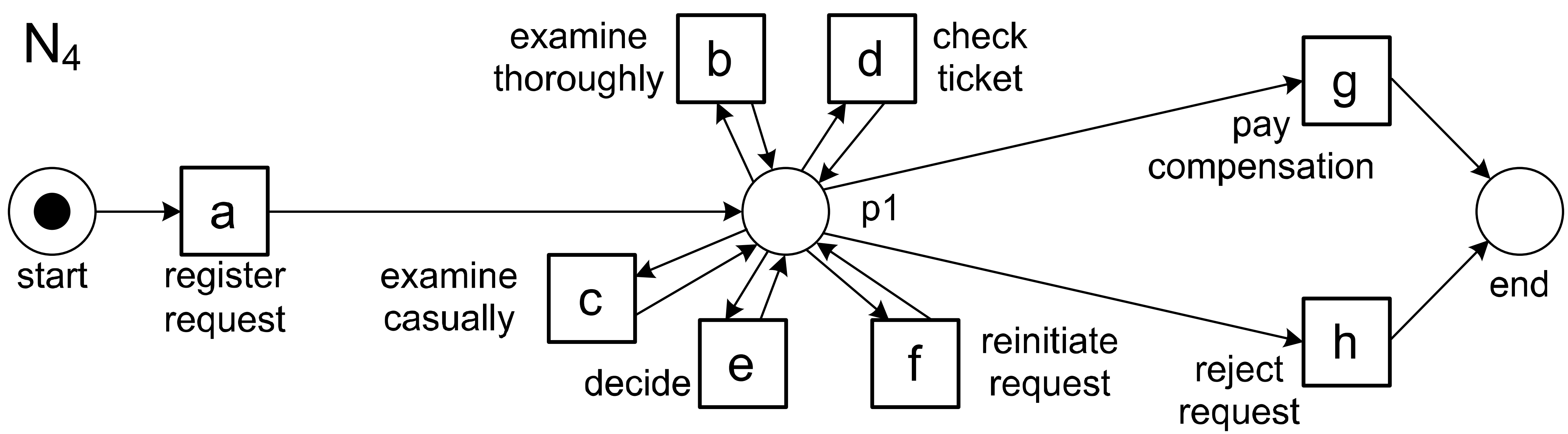


759 cases do not correspond to a firing sequence

$\langle a, b, d, e, g \rangle^{191}$
$\langle a, b, d, e, h \rangle^{144}$
$\langle a, c, d, e, g \rangle^{111}$
$\ldots$

naïve fitness $\frac{632}{1391} = 0.4543$

Example N4



"flower model" (poorly structured)

naïve fitness $\frac{1391}{1391} = 1$ The net can ``replay'' any trace

Almost Fitting Traces

This naïve fitness notion seems to be too strict as traces can differ only slightly and not be counted at all.

$$\sigma = \langle a_1, a_2, \ldots, a_{100} \rangle$$

Consider a model N1 that cannot replay σ ,
but that can replay 99 of the 100 events in σ.
Then, consider another model N2 that can only replay 10 of the 100 events in σ.

Using the naïve fitness metric, the trace would simply be classified as non-fitting for both models without acknowledging that σ was almost fitting in N1 and in complete disagreement with N2.

Missing and Remaining Tokens

We next introduce a more accurate fitness notion.

When computing the naïve fitness,
we stop replaying a trace as soon as we find a problem
(and tag that trace as non-fitting).

Let us instead just continue replaying the trace on the model
but record all situations where a transition is
forced to fire without being enabled,
i.e., we count all **missing** tokens.
Moreover, we record the tokens that **remain** at the end.

Four Counters

p (produced tokens)

r (remaining tokens)

c (consumed tokens)

m (missing tokens)

equally weighted

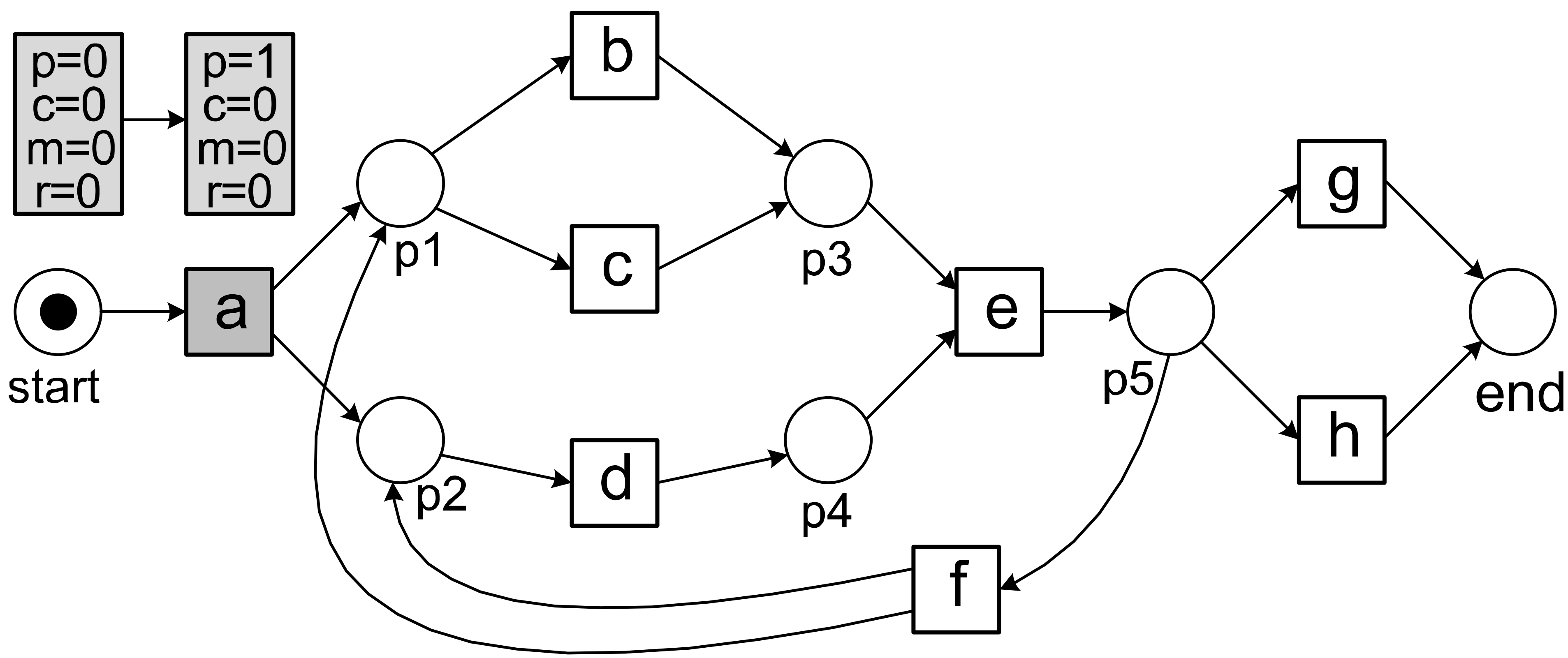
ideally m=r=0

$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

proportions of misplacement

Example: none missing, none remaining

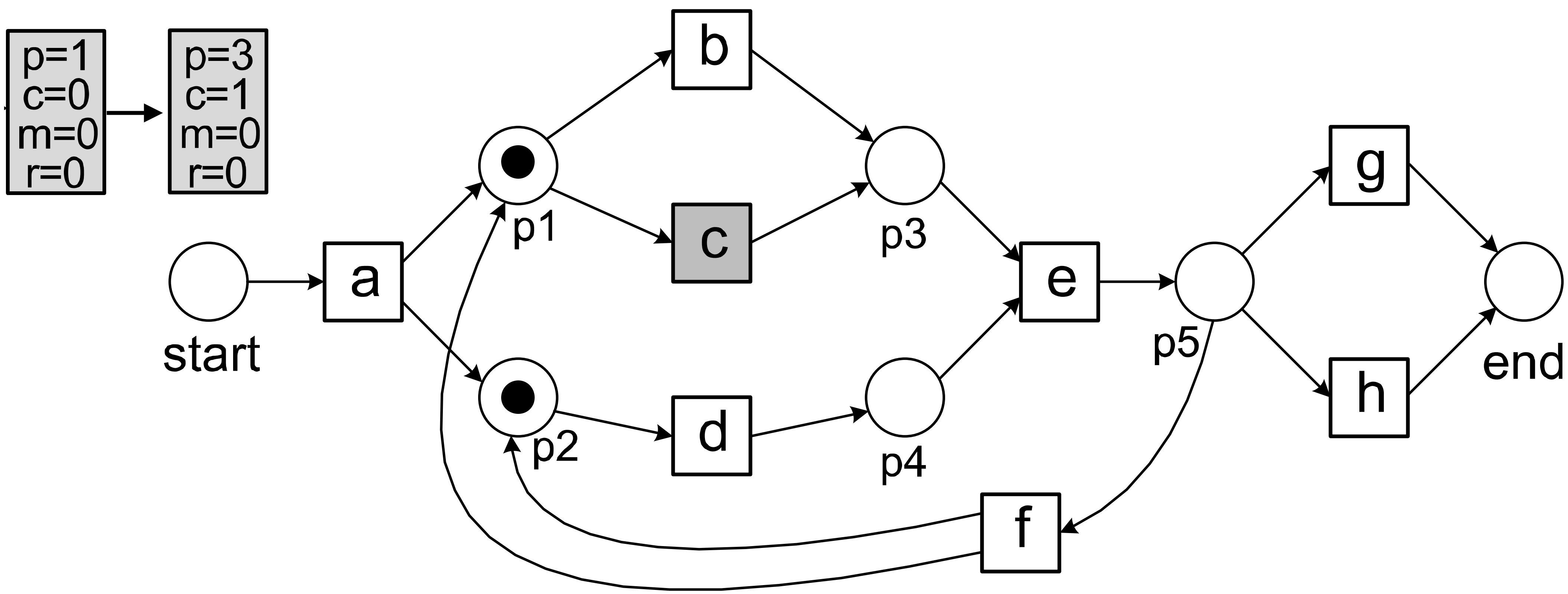
the environment produces a token for place start



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

Example: none missing, none remaining

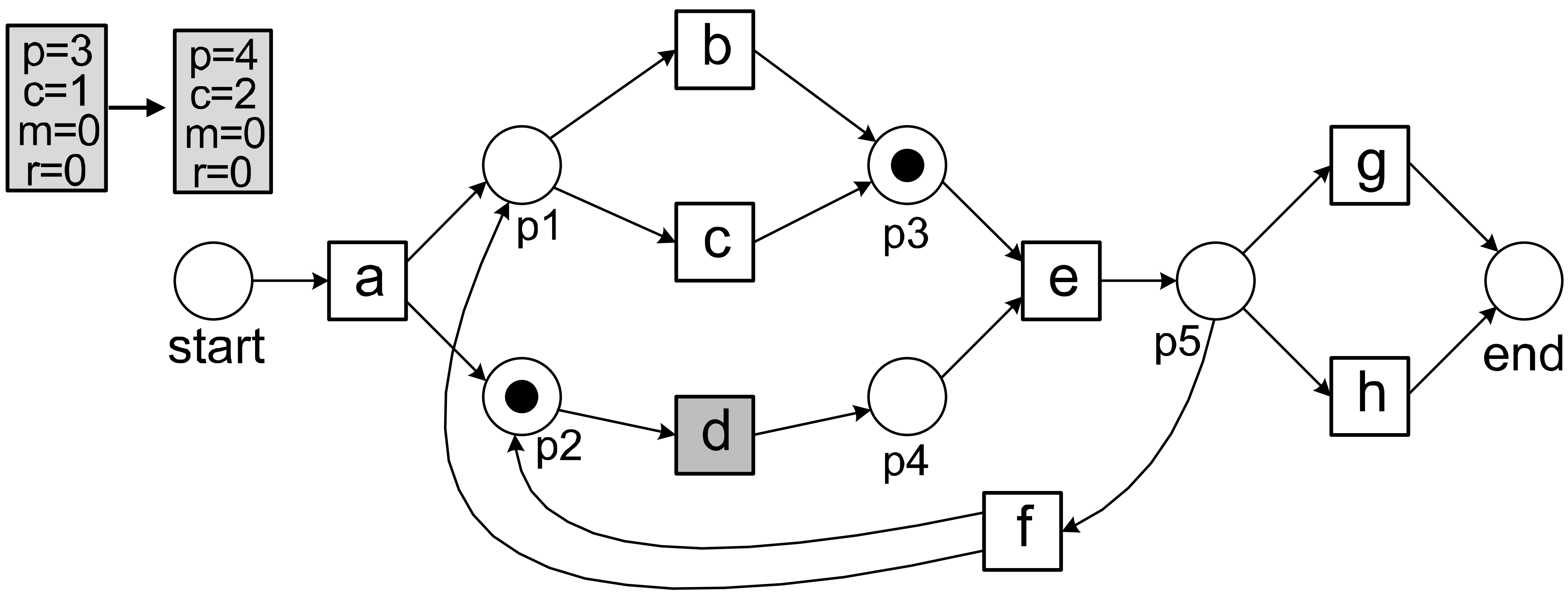
replaying a is possible
one token is consumed, two produced



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

Example: none missing, none remaining

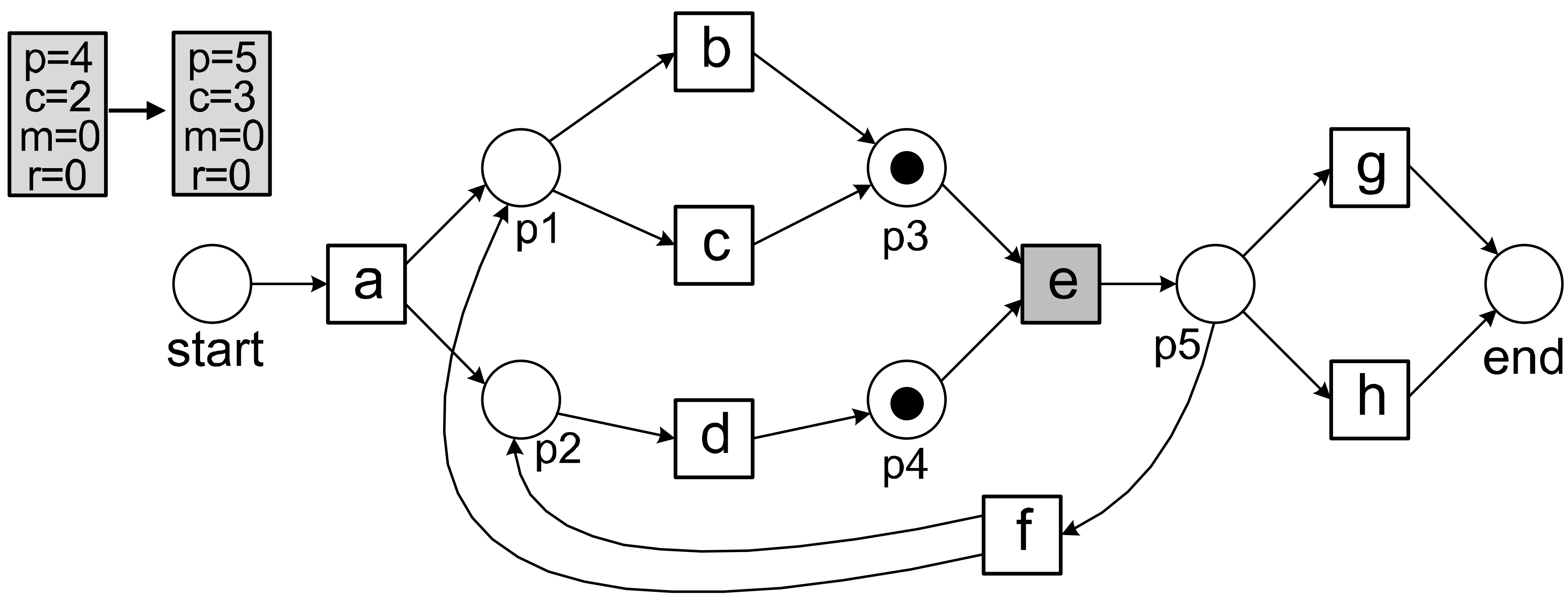
replaying c is possible
one token is consumed, one produced



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

Example: none missing, none remaining

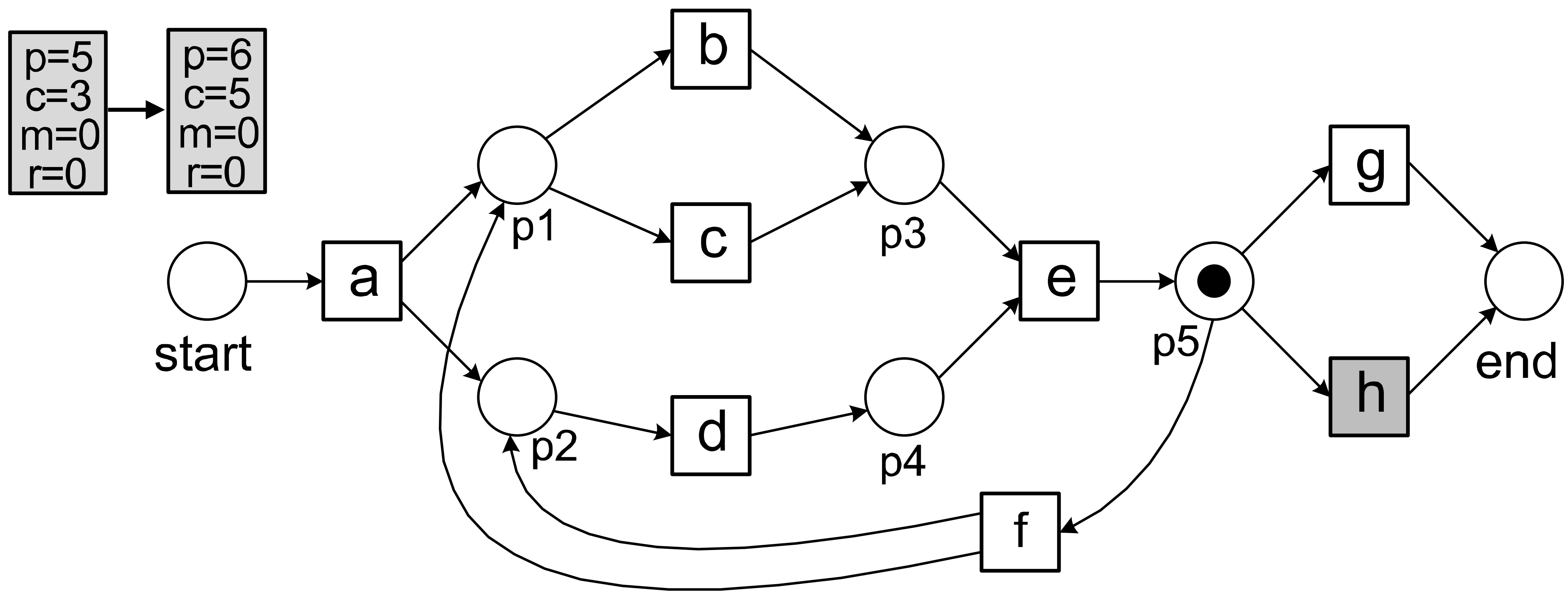
replaying d is possible
one token is consumed, one produced



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

Example: none missing, none remaining

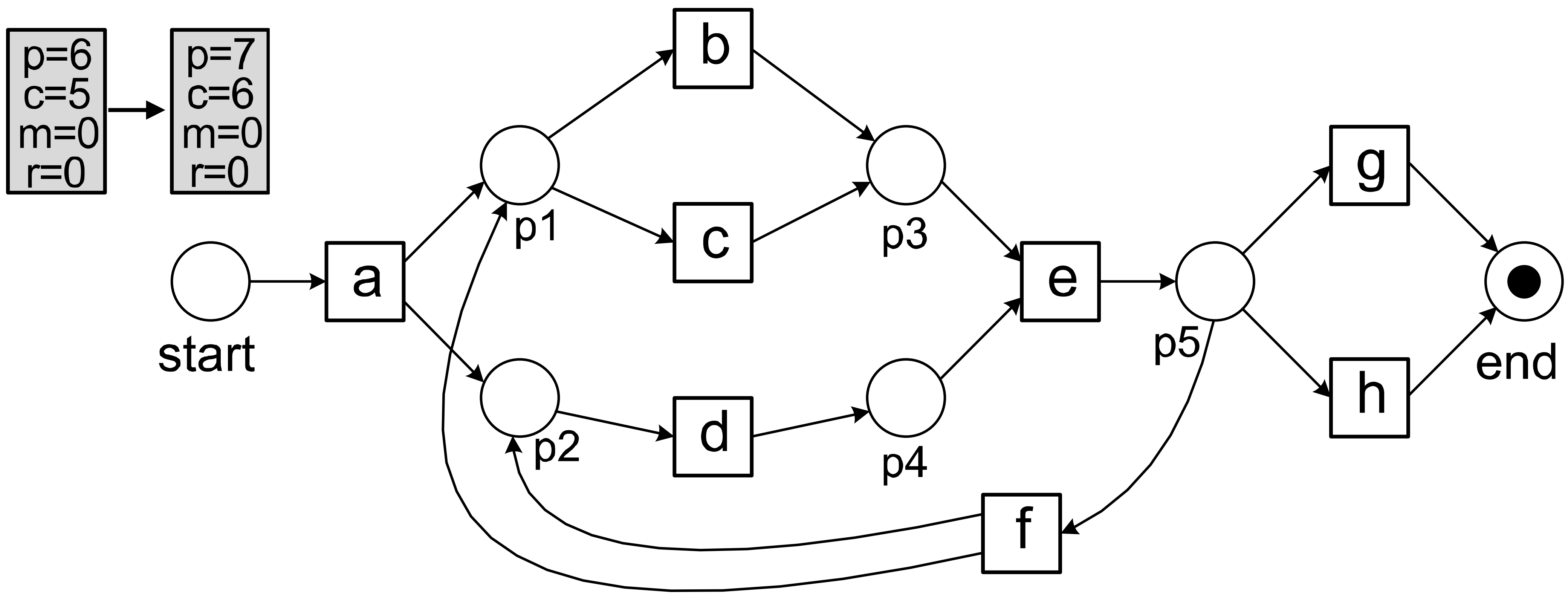
replaying e is possible
two tokens are consumed, one produced



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

Example: none missing, none remaining

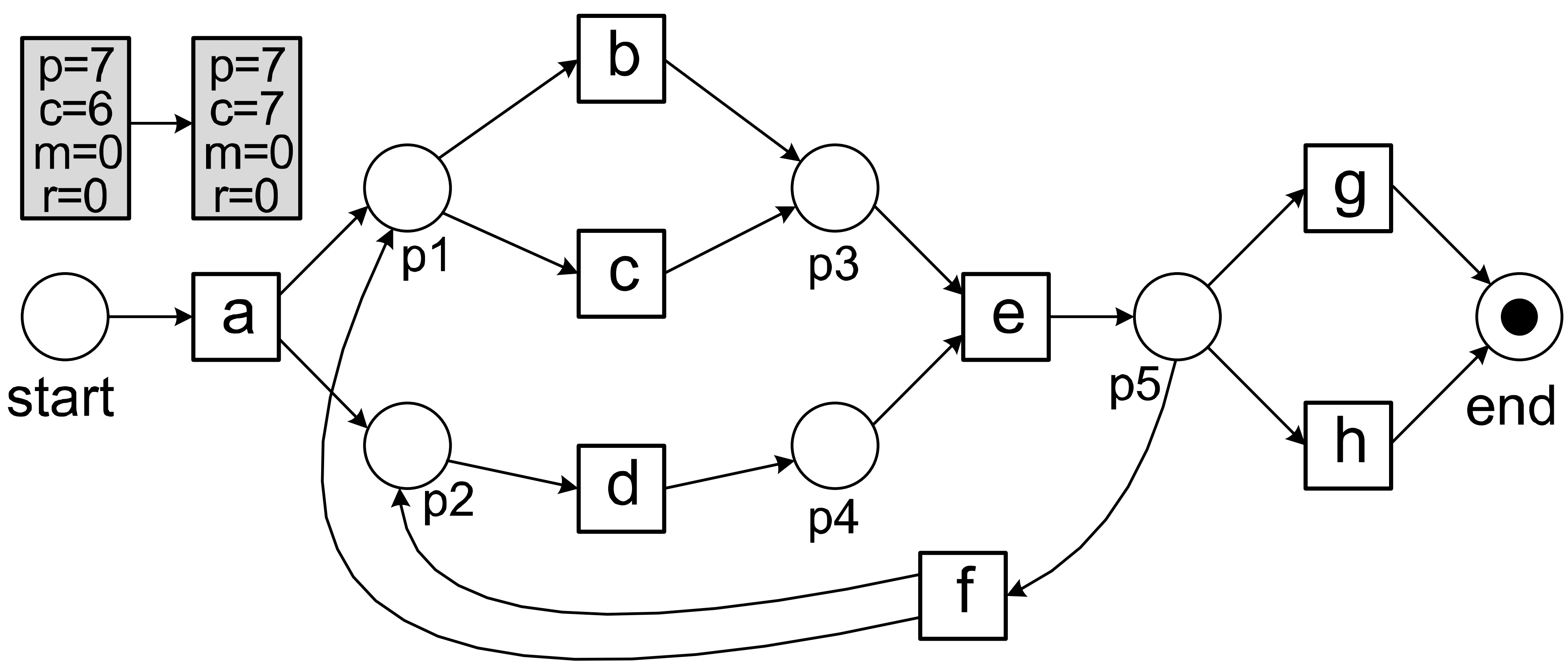
replaying h is possible
one token is consumed, one produced



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

Example: none missing, none remaining

At the end,
the environment consumes
a token from place end.

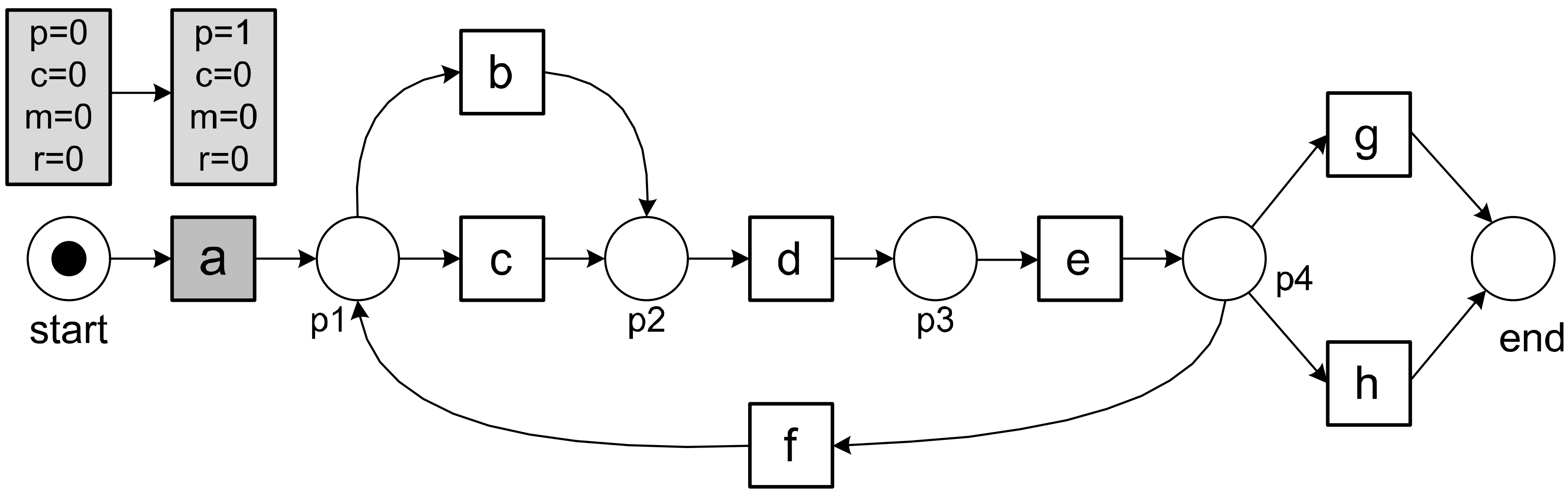


$$fitness(\sigma_1, N_1) = \frac{1}{2}(1 - \frac{0}{7}) + \frac{1}{2}(1 - \frac{0}{7}) = 1$$

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

Example: Missing Token

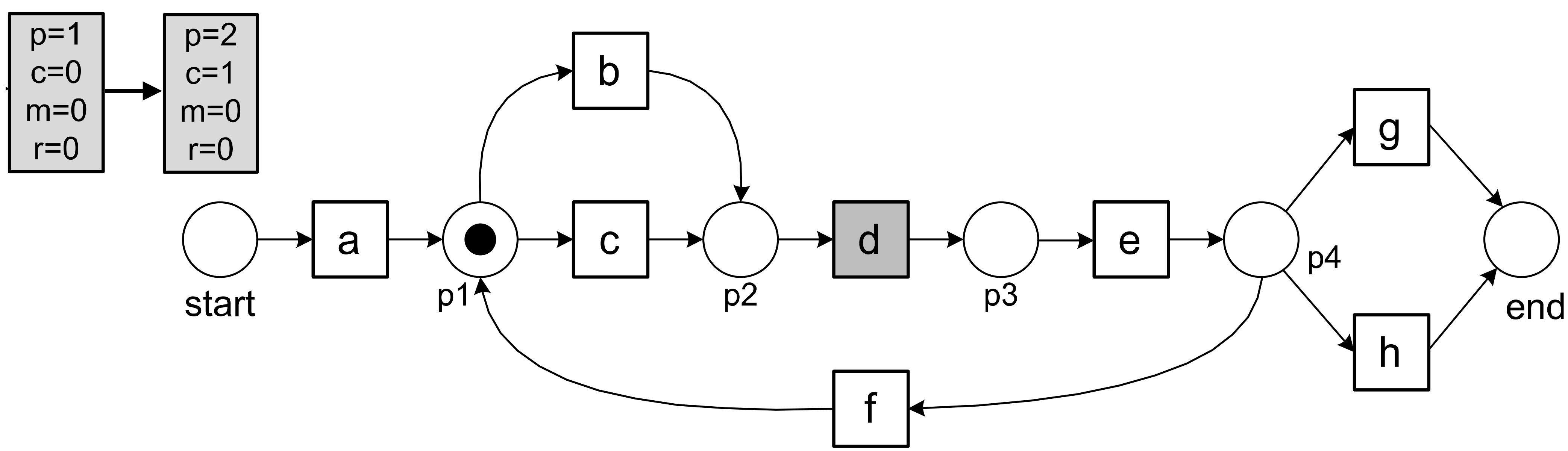
the environment produces a token for place start



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

Example: Missing Token

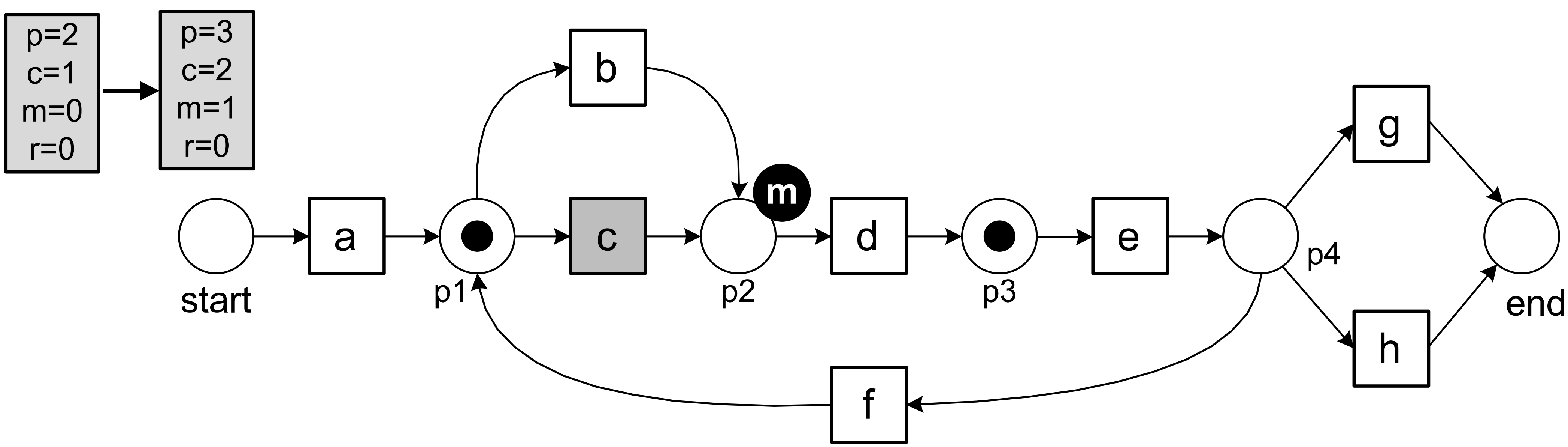
replaying a is possible
one token is consumed, one produced



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

Example: Missing Token

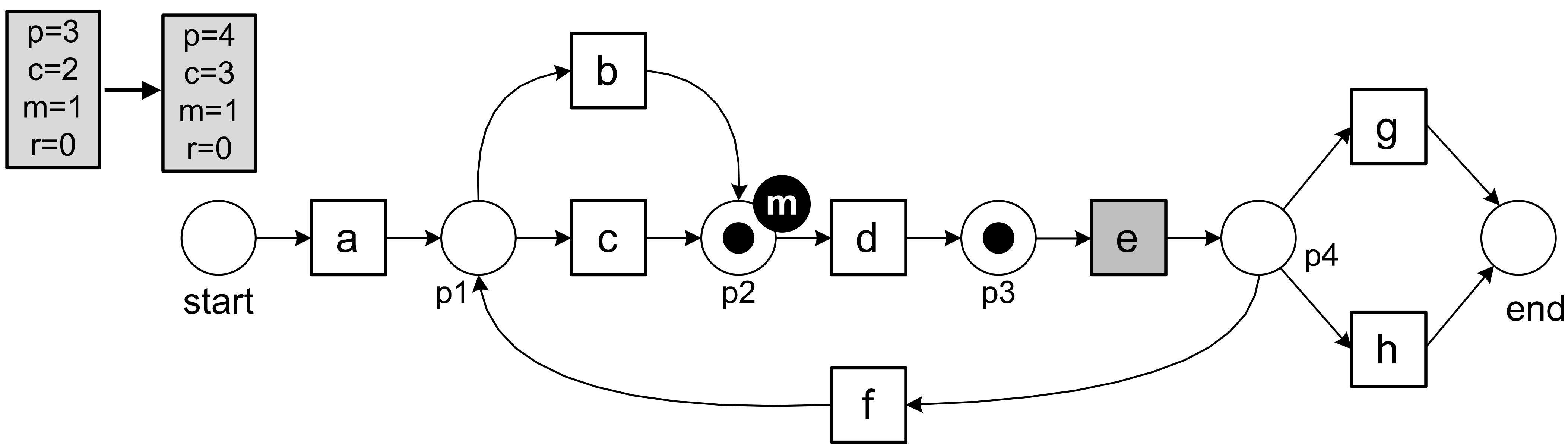
replaying d is NOT possible
one token is missing,
one produced, one consumed



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

Example: Missing Token

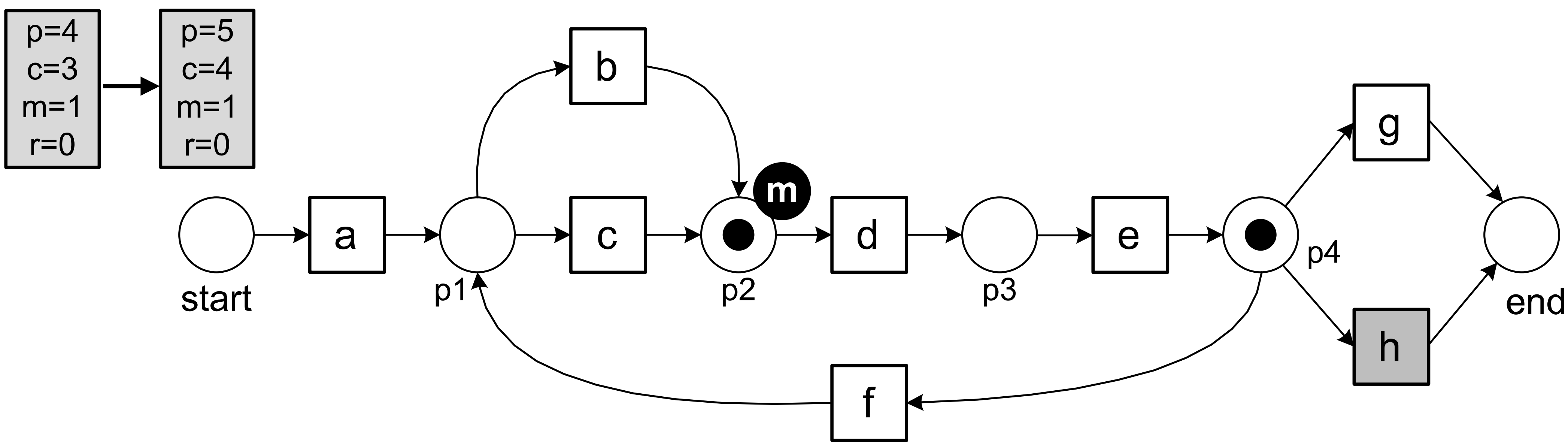
replaying c is possible
one token is produced, one consumed



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

Example: Missing Token

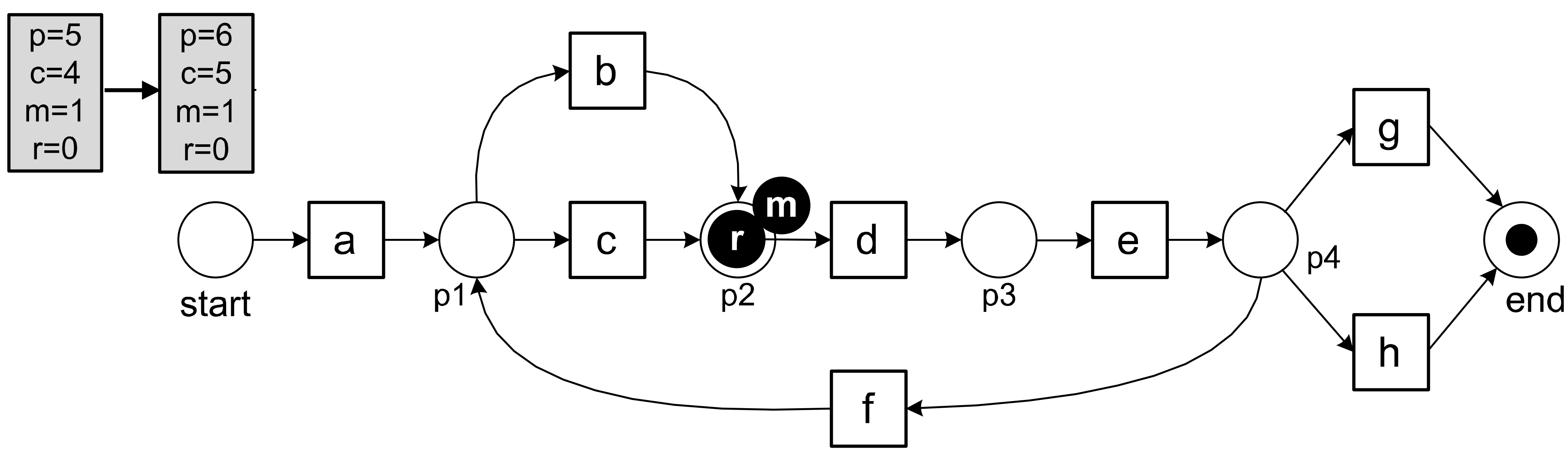
replaying e is possible
one token is produced, one consumed



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

Example: Missing Token

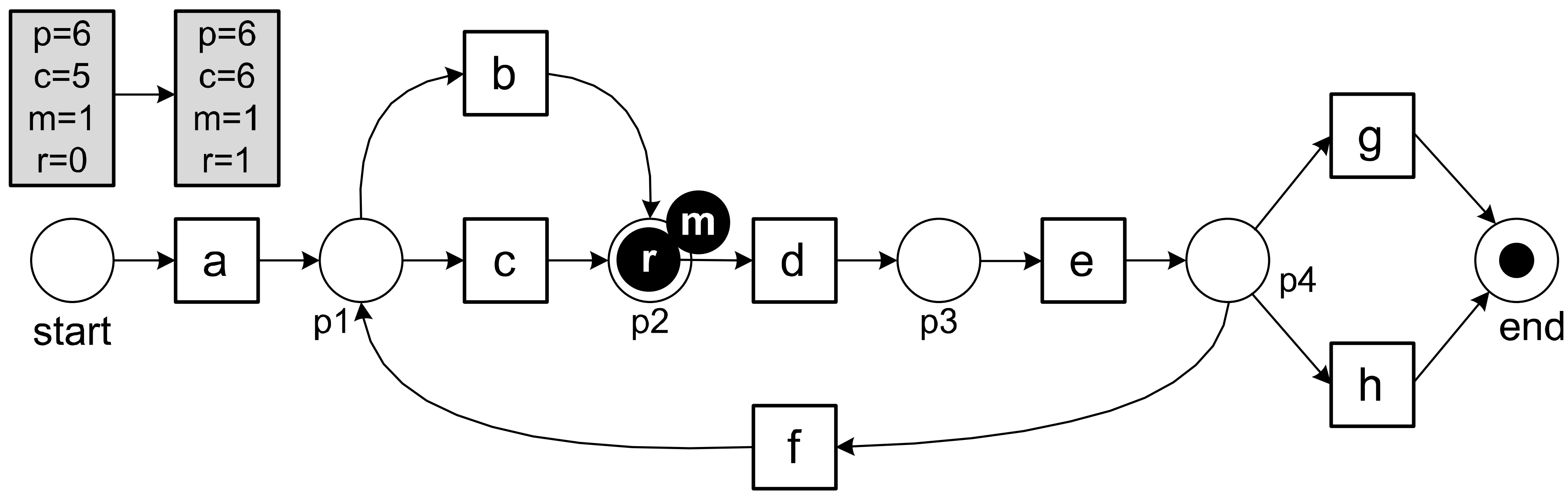
replaying h is possible
one token is produced, one consumed



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

Example: Missing Token

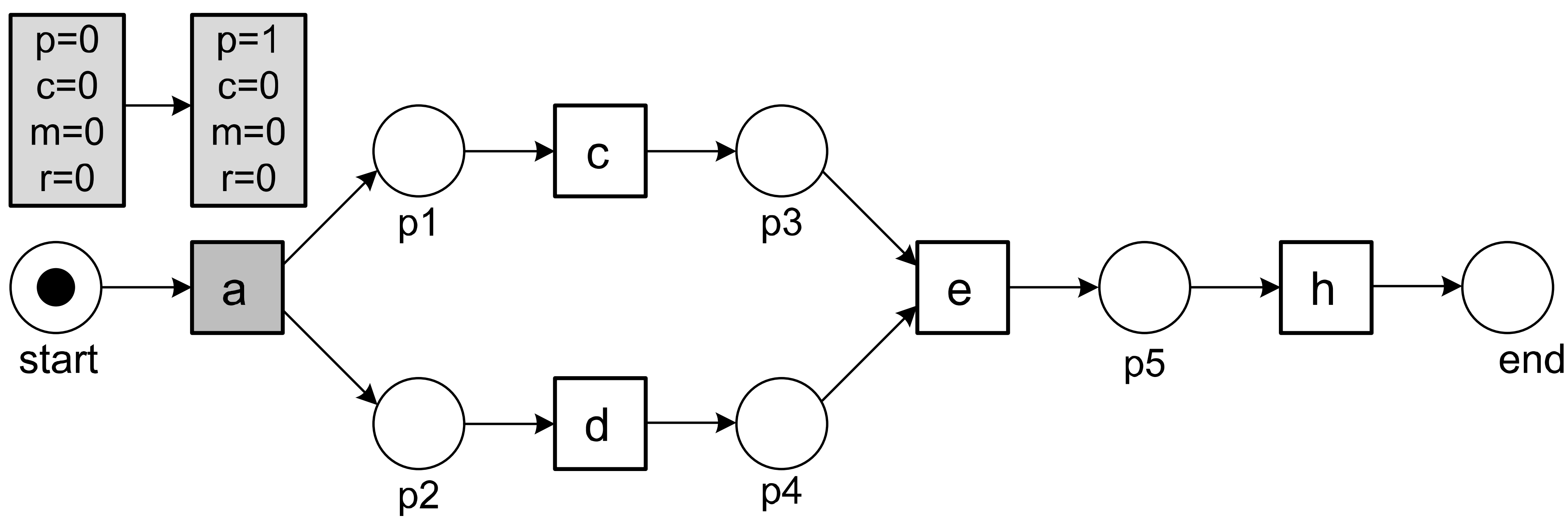
At the end,
the environment consumes
a token from place end.



$$fitness(\sigma_3, N_2) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.8333$$

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

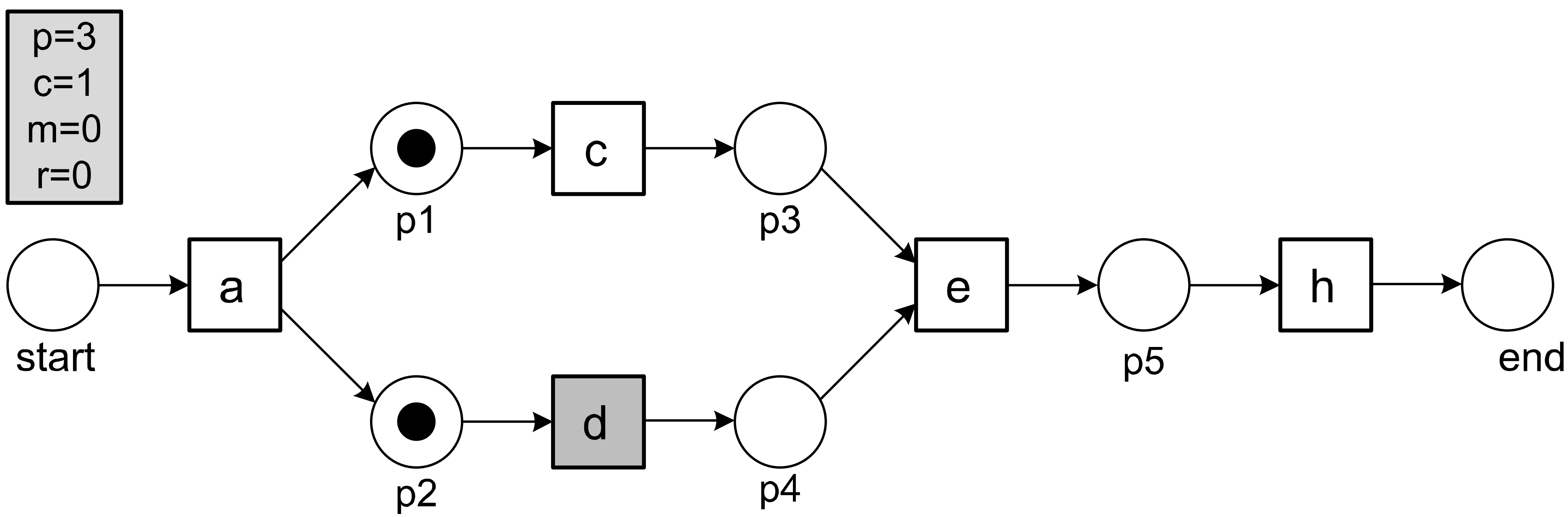
Example: Event Removal



events b and g are not present in the net
therefore we remove them from the trace

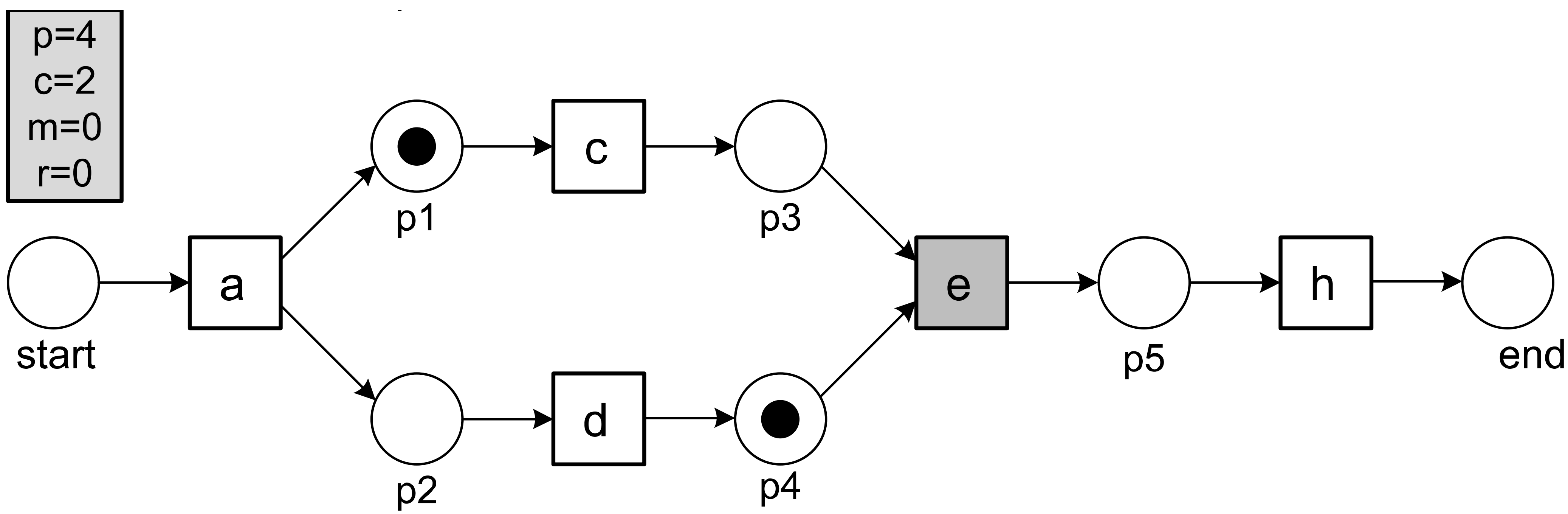
$$\sigma_2 = \langle a, b, d, e, g \rangle \qquad \sigma_2' = \langle a, d, e \rangle$$

Example: Event Removal



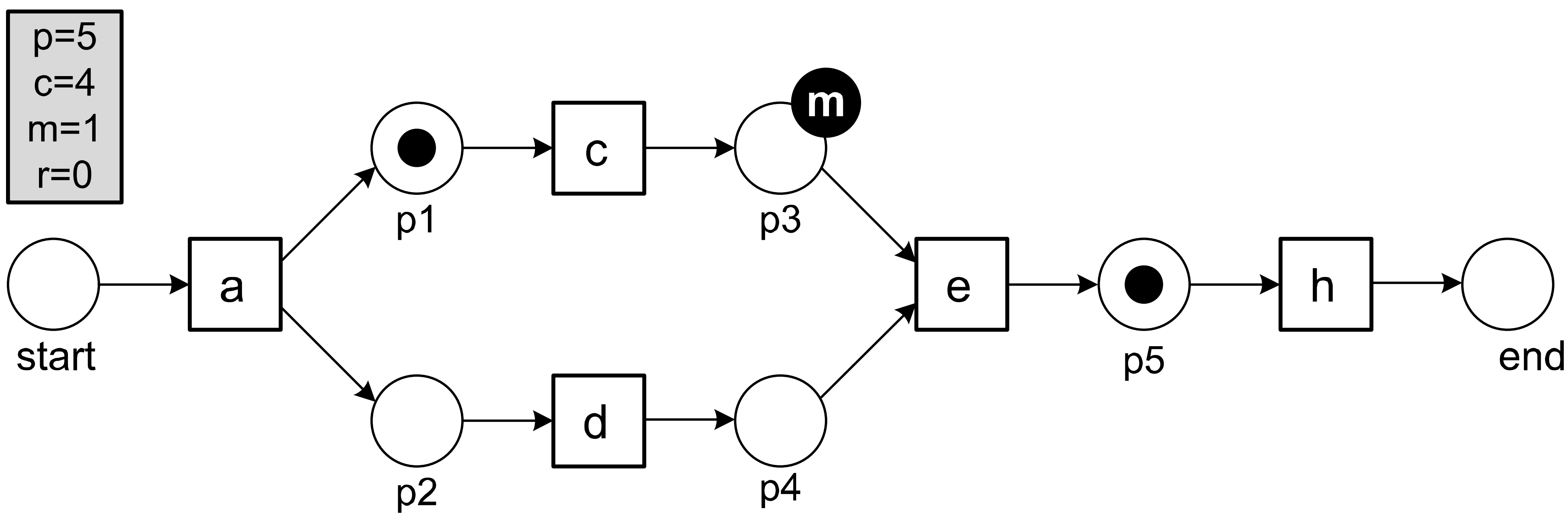
$$\sigma_2' = \langle a, d, e \rangle$$

Example: Event Removal



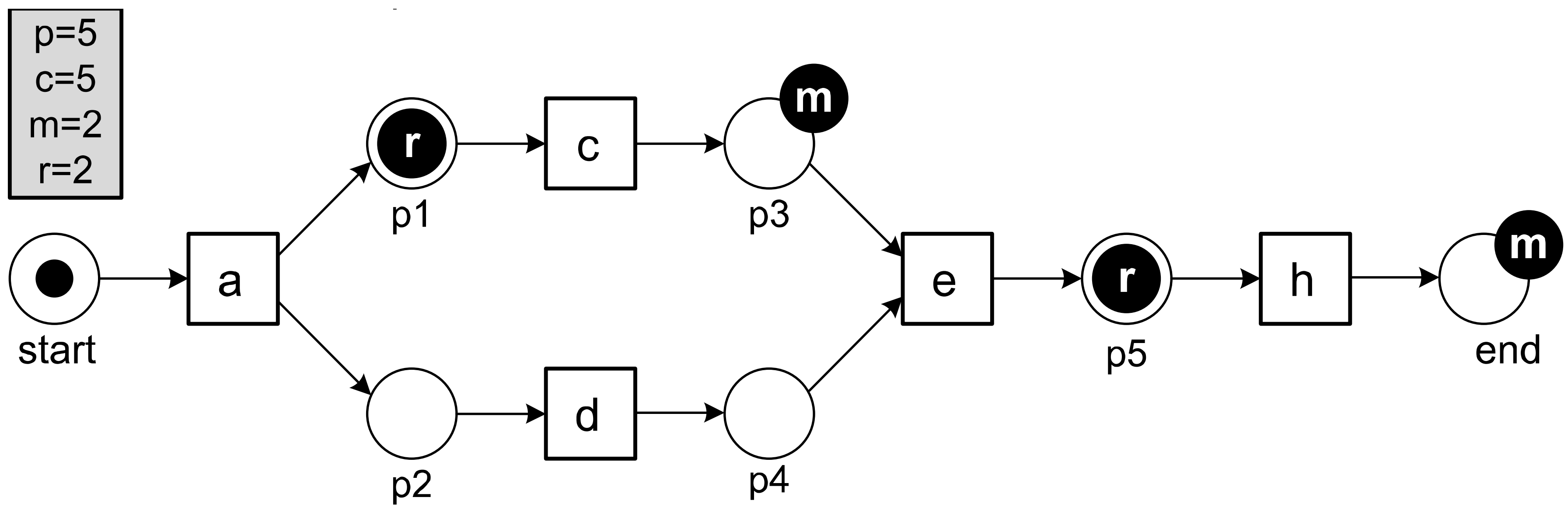
$$\sigma_2' = \langle a, d, e \rangle$$

Example: Event Removal



$$\sigma_2' = \langle a, d, e \rangle$$

Example: Event Removal



$$fitness(\sigma_2, N_3) = \frac{1}{2}\left(1 - \frac{2}{5}\right) + \frac{1}{2}\left(1 - \frac{2}{5}\right) = 0.6$$

$$\sigma_2' = \langle a, d, e \rangle$$

Fitness of a Log

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

$L(\sigma)$ is just the multiplicity of the trace σ in the log L

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

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

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

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

Diagnostic Information

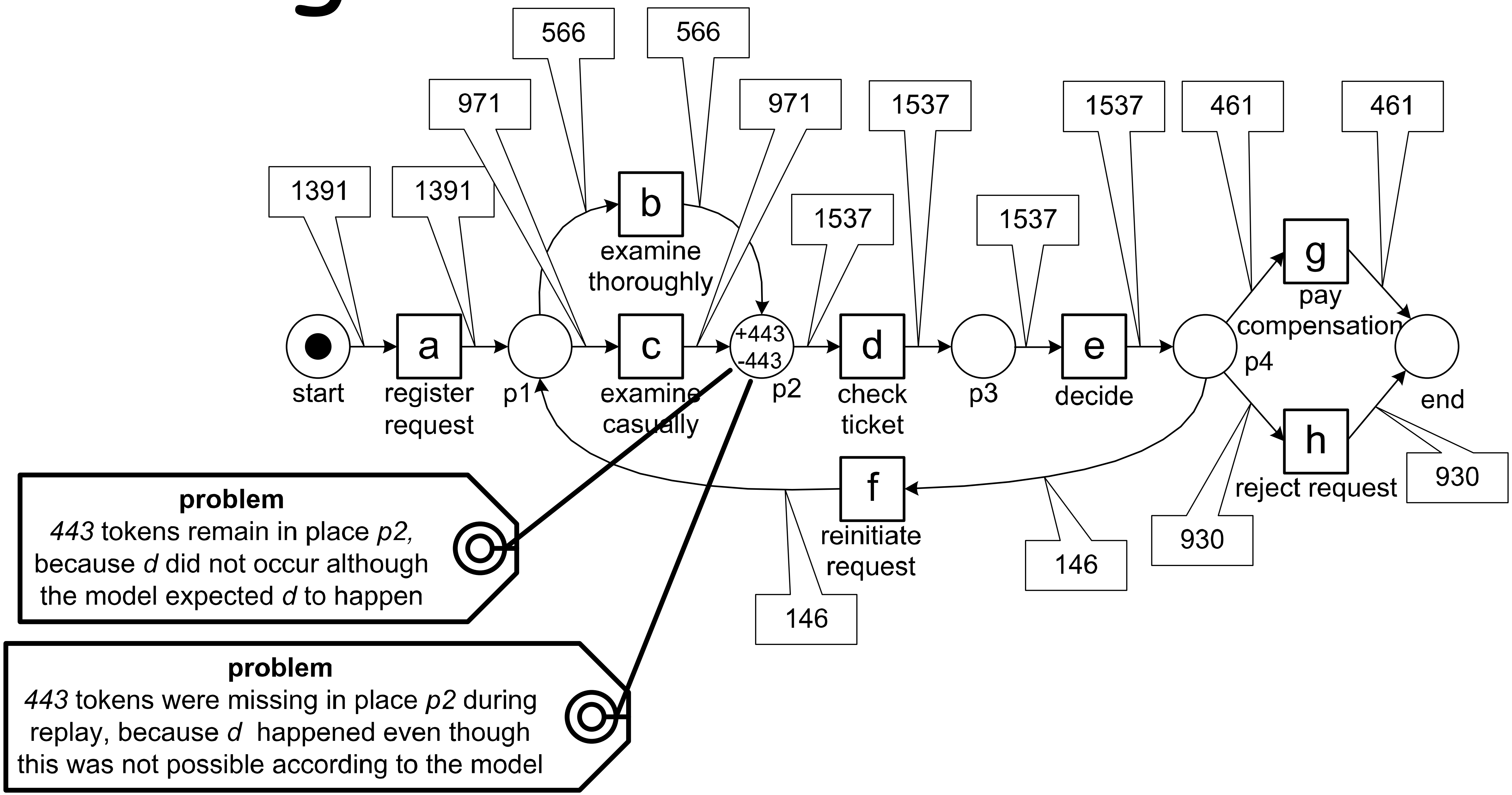


Fig. 7.6 Diagnostic information showing the deviations ($fitness(L_{full}, N_2) = 0.9504$)

Diagnostic Information

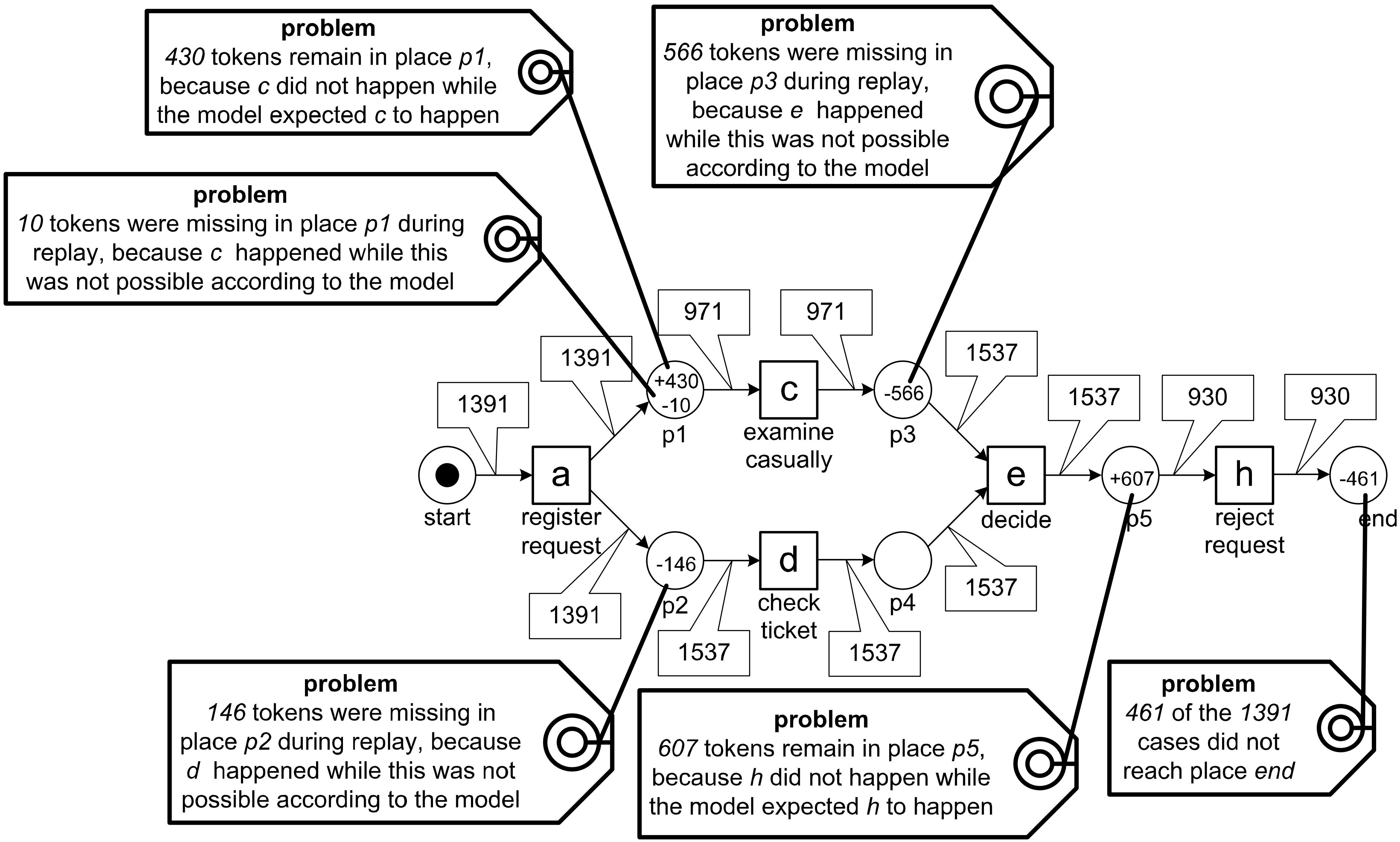


Fig. 7.7 Diagnostic information showing the deviations ($fitness(L_{full}, N_3) = 0.8797$)

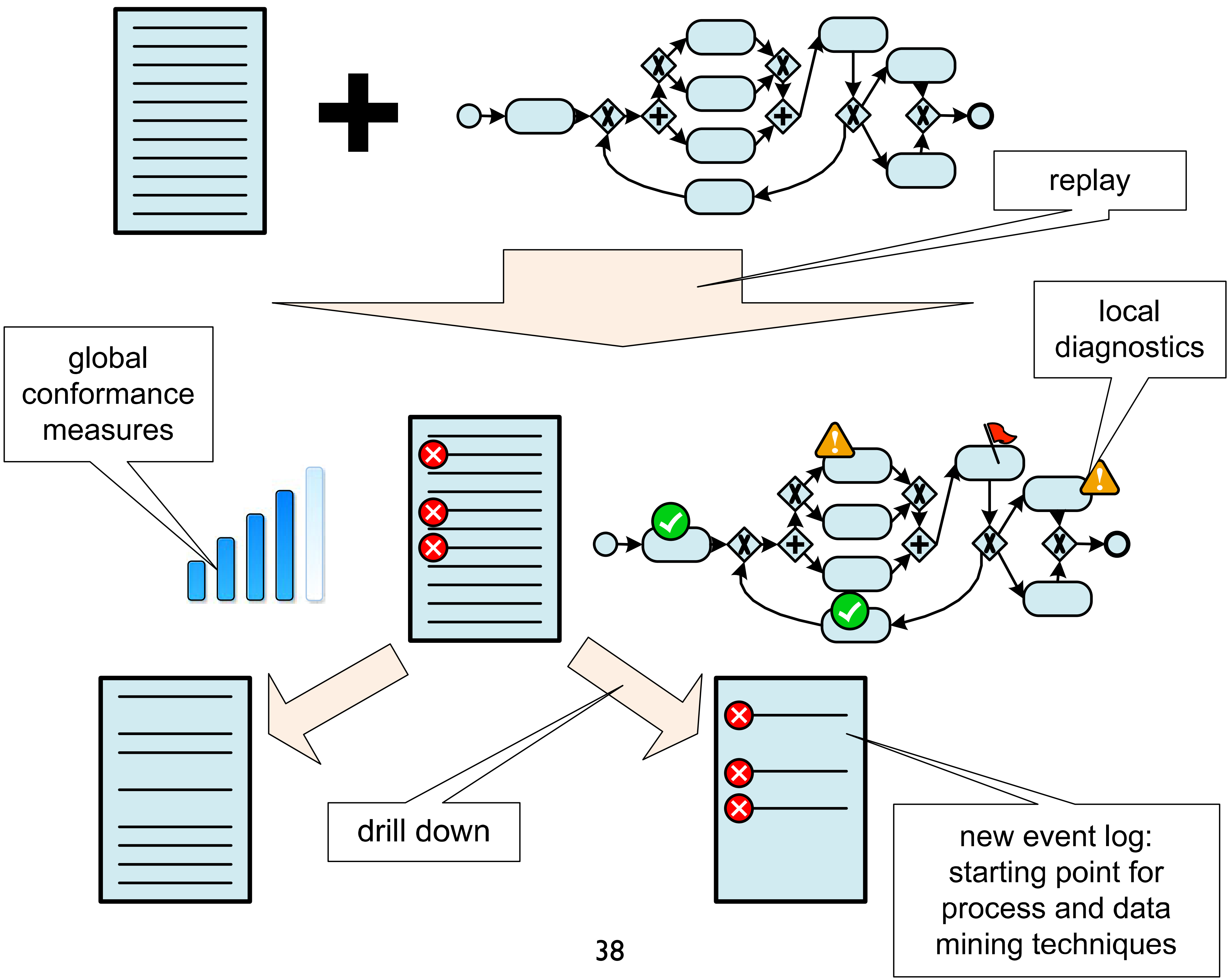
Drill Down

An event log can be split into two sublogs:
one event log containing only fitting cases and
one event log containing only non-fitting cases.

The second event log can be used to discover a different process model.

Also other data and process mining techniques can be used. For instance, it is interesting to know which people handled the deviating cases and whether these cases took longer or were more costly.
In case fraud is suspected, one may create a social network based on the event log with deviating cases.

Drill Down


replay
global conformance measures
local diagnostics
drill down
new event log: starting point for process and data mining techniques

Comparing Footprints (optional reading)

Footprint from Play-out

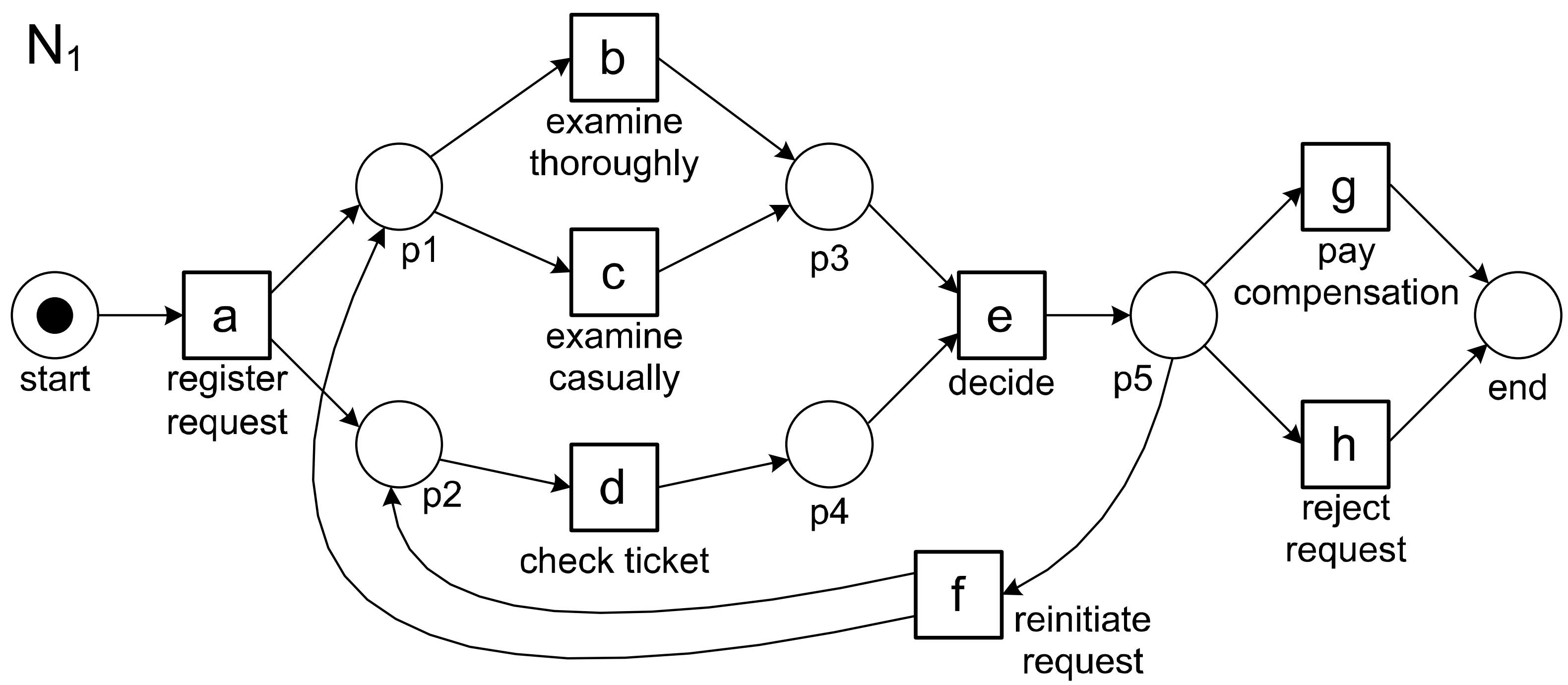
Given a workflow net, the **play-out** technique can be used to extract a **local complete** set of traces.

If we see the set of traces as an event log (without multiplicities), then we can derive the relation >.

Then, we can construct the footprint (i.e. a matrix showing causal dependencies between events) of the net model based on such relation >.

(From the viewpoint of a footprint matrix, an event log is complete if and only if all activities that can follow one another do so at least once in the log.)

Example: complete set



$\langle a\ b\ d\ e\ g \rangle$ $\langle a\ c\ d\ e\ f\ b\ d\ e\ g \rangle$

$\langle a\ d\ b\ e\ f\ d\ c\ e\ h \rangle$ $\langle a\ d\ b\ e\ f\ c\ d\ e\ h \rangle$

Footprint-based Conformance

Footprints are available for logs and models (nets).
This allows for:

log vs model conformance
(do the log and the model agree?)

model vs model conformance
(quantification of their similarities)

log vs log comparison
(*concept drift*: how does the work changes in sub-logs?)

Conformance based on footprints

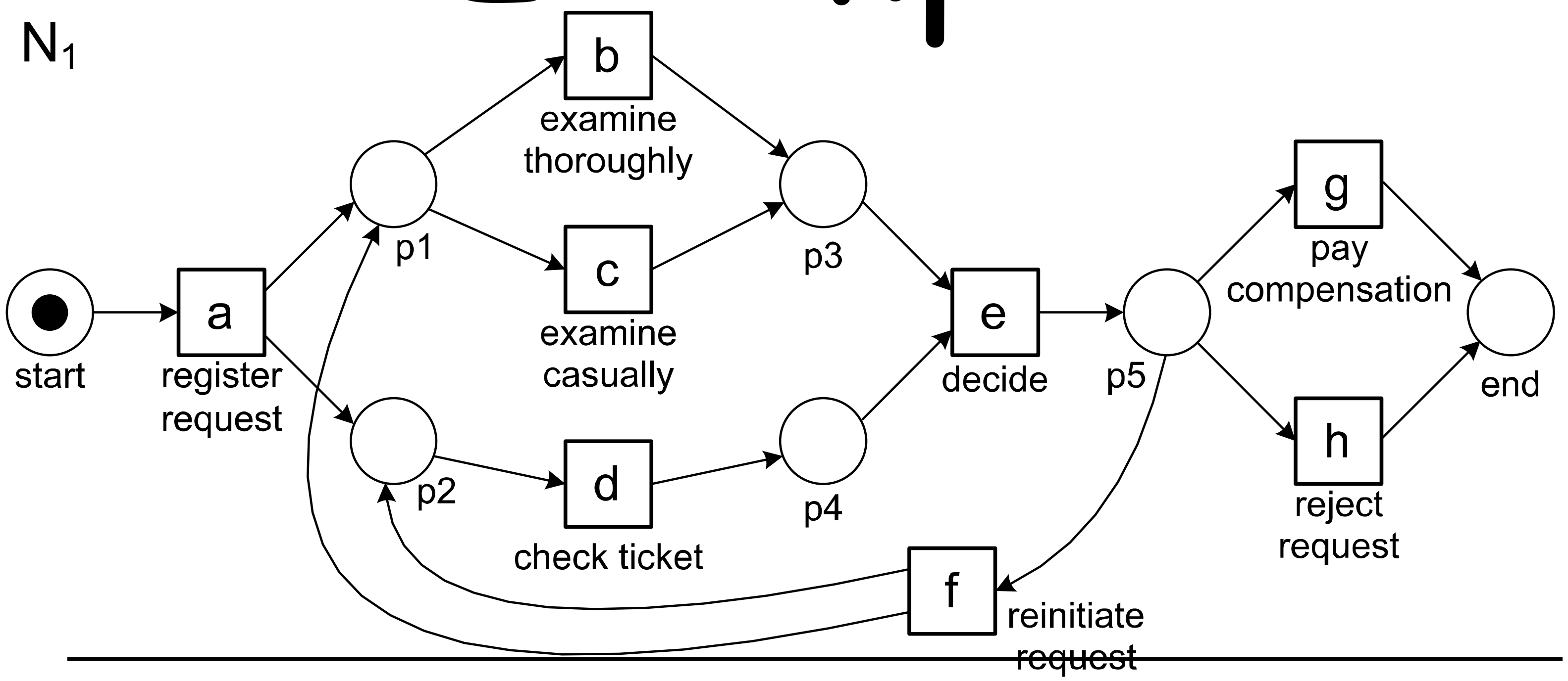
The conformance based on footprints can be computed by taking:

n: total number of cells in the footprint matrix

d: number of cells in the same positions but with different content between the two matrices

$$1 - \frac{d}{n}$$

Example



$\langle a\ b\ d\ e\ g \rangle$

$\langle a\ d\ b\ e\ f\ d\ c\ e\ h \rangle$

$\langle a\ c\ d\ e\ f\ b\ d\ e\ g \rangle$

$\langle a\ d\ b\ e\ f\ c\ d\ e\ h \rangle$

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
a	#	→	→	→	#	#	#	#
b	←	#	#	‖	→	←	#	#
c	←	#	#	‖	→	←	#	#
d	←	‖	‖	#	→	←	#	#
e	#	←	←	←	#	→	→	→
f	#	→	→	→	←	#	#	#
g	#	#	#	#	←	#	#	#
h	#	#	#	#	←	#	#	#

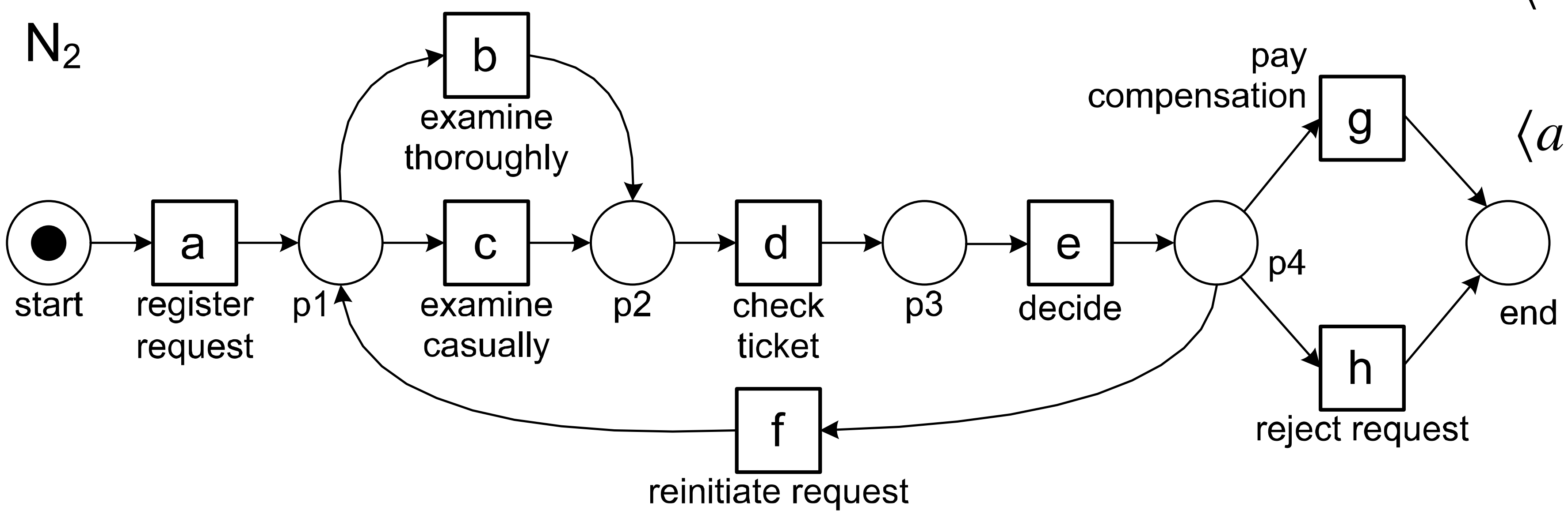
Also

Footprint of L_{full}

Example

$\langle a\ b\ d\ e\ f\ b\ d\ e\ g\rangle$

$\langle a\ c\ d\ e\ f\ c\ d\ e\ h\rangle$



	a	b	c	d	e	f	g	h
a	#	$\rightarrow$	$\rightarrow$	#	#	#	#	#
b	$\leftarrow$	#	#	$\rightarrow$	#	$\leftarrow$	#	#
c	$\leftarrow$	#	#	$\rightarrow$	#	$\leftarrow$	#	#
d	#	$\leftarrow$	$\leftarrow$	#	$\rightarrow$	#	#	#
e	#	#	#	$\leftarrow$	#	$\rightarrow$	$\rightarrow$	$\rightarrow$
f	#	$\rightarrow$	$\rightarrow$	#	$\leftarrow$	#	#	#
g	#	#	#	#	$\leftarrow$	#	#	#
h	#	#	#	#	$\leftarrow$	#	#	#

Example

	a a	b b	c c	d d	e e	f f	g g	h h
a a	# #	$\rightarrow\rightarrow$	$\rightarrow\rightarrow$	$\rightarrow$#	# #	# #	# #	# #
b b	$\leftarrow\leftarrow$	# #	# #	$\parallel$ $\rightarrow$	$\rightarrow$#	$\leftarrow\leftarrow$	# #	# #
c c	$\leftarrow\leftarrow$	# #	# #	$\parallel$ $\rightarrow$	$\rightarrow$#	$\leftarrow\leftarrow$	# #	# #
d d	$\leftarrow$#	$\parallel$ $\leftarrow$	$\parallel$ $\leftarrow$	# #	$\rightarrow\rightarrow$	$\leftarrow$#	# #	# #
e e	# #	$\leftarrow$#	$\leftarrow$#	$\leftarrow\leftarrow$	# #	$\rightarrow\rightarrow$	$\rightarrow\rightarrow$	$\rightarrow\rightarrow$
f f	# #	$\rightarrow\rightarrow$	$\rightarrow\rightarrow$	$\rightarrow$#	$\leftarrow\leftarrow$	# #	# #	# #
g g	# #	# #	# #	# #	$\leftarrow\leftarrow$	# #	# #	# #
h h	# #	# #	# #	# #	$\leftarrow\leftarrow$	# #	# #	# #

Example

$$1 - \frac{12}{64} = 0.8125$$

	a	b	c	d	e	f	g	h
a				$\rightarrow : \#$				
b				$\parallel : \rightarrow$	$\rightarrow : \#$			
c				$\parallel : \rightarrow$	$\rightarrow : \#$			
d	$\leftarrow : \#$	$\parallel : \leftarrow$	$\parallel : \leftarrow$			$\leftarrow : \#$		
e		$\leftarrow : \#$	$\leftarrow : \#$					
f				$\rightarrow : \#$				
g								
h								