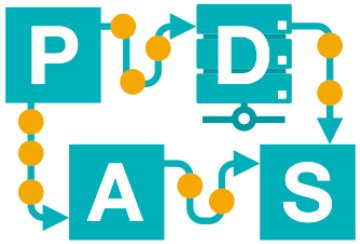
Process Mining - Instruction

Lecture 15

IDS-L15-I



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Conformance Checking

Conformance checking

Recall from the lecture:

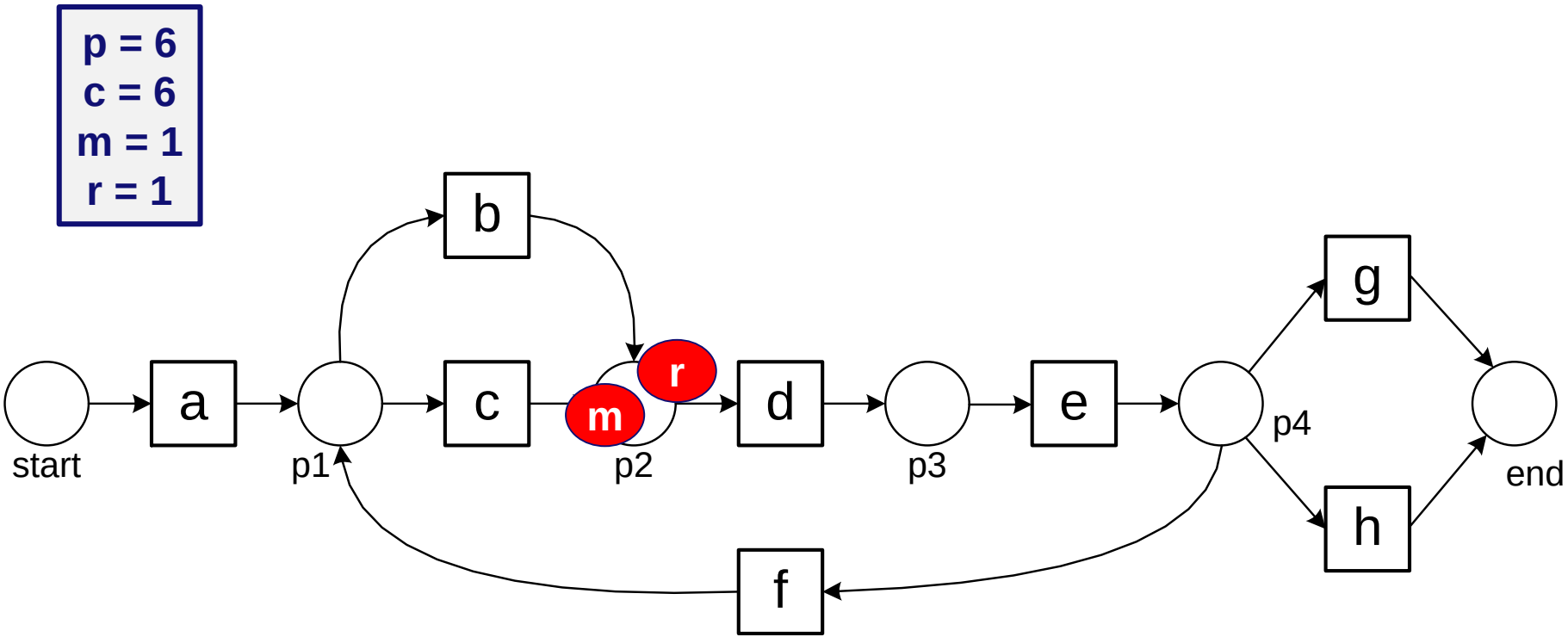
Conformance Checking compares an event log to a process model, in order to identify deviations and non-conforming behaviour.

Conformance checking

Conformance Checking can be done with a number of different methods.

One of them is through token replay: I can execute a trace on top of a model, and compute a score for conformance counting produced, consumed, missing and remaining tokens.

From lecture: replaying $\sigma_3 = \langle a, d, c, e, h \rangle$



Chair of Process and Data Science

Fitness

Once obtained the number of produced, consumed, missing and remaining token, the conformance score (fitness) of a trace is:

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

Token Replay: Exercise

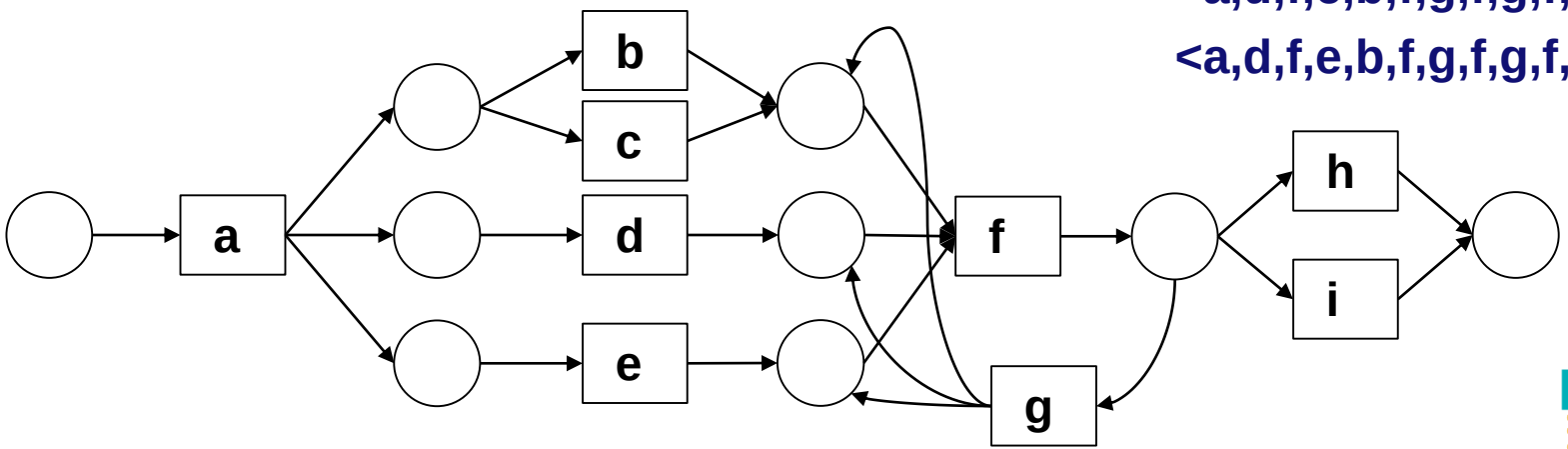
Can you obtain the counts and fitness scores for the following traces on this Petri net?

<a,d,e,b,f,g,f,g,f,h>

<a,e,b,f,g,f,g,f,>

<a,d,f,e,b,f,g,f,g,f,h>

<a,d,f,e,b,f,g,f,g,f,h,i>



Chair of Process and Data Science

Token Replay: Solution

<a,d,e,b,f,g,f,h>

p = 13

c = 13

m = 0

r = 0

<a,e,b,f,g,f,>

p = 11

c = 11

m = 2

r = 2

<a,d,f,e,b,f,g,f,h>

p = 14

c = 16

m = 3

r = 1

<a,d,f,e,b,f,g,f,h,i>

p = 15

c = 17

m = 3

r = 1

Token Replay: Solution

<a,d,e,b,f,g,f,h>

p = 13 fitness = 1

c = 13

m = 0

r = 0

<a,e,b,f,g,f,>

p = 11 fitness = 0.82

c = 11

m = 2

r = 2

<a,d,f,e,b,f,g,f,h>

p = 14 fitness = 0.87

c = 16

m = 3

r = 1

<a,d,f,e,b,f,g,f,h,i>

p = 15 fitness = 0.88

c = 17

m = 3

r = 1

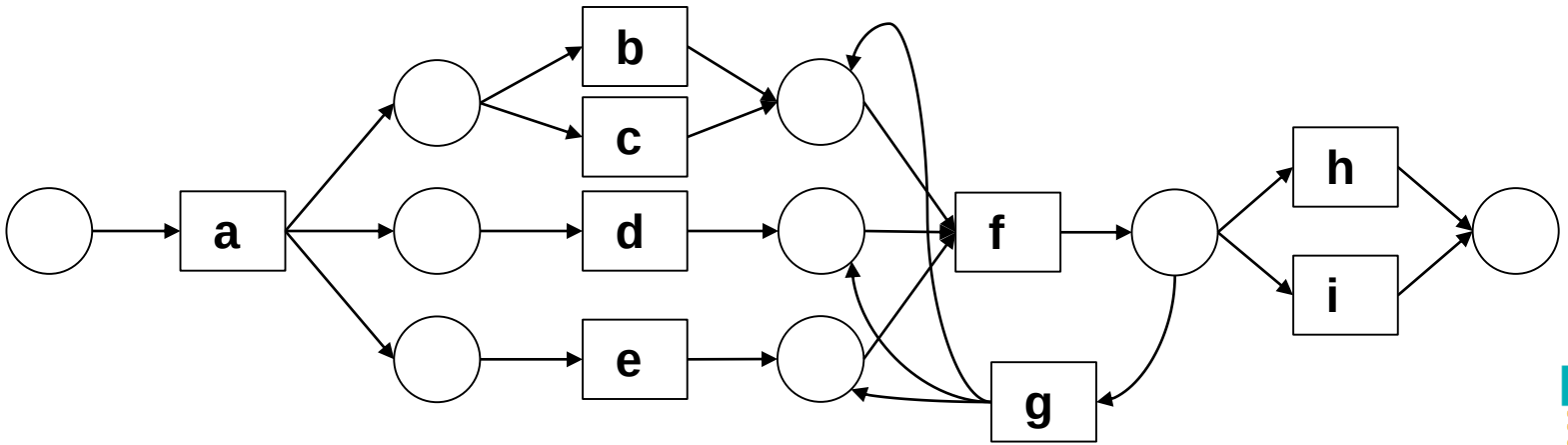
Token Replay

Anybody noticed something weird?

Token Replay

The trace <a,d,f,e,b,f,g,f,h> contains an additional f event.

The trace <a,d,f,e,b,f,g,f,h> contains an additional f event and an additional i event.



Token Replay

The trace <a,d,f,e,b,f,g,f,h> contains an additional f event. Fitness = 0.87.

The trace <a,d,f,e,b,f,g,f,h> contains an additional f event and an additional i event. Fitness = 0.88.

Why does the fitness increase with more violations?

Token Replay

This is a known limitation of the conformance based on token replay: in some cases, errors can "cancel each other out".

To avoid this there are more sophisticated techniques.

pm4py

pm4py

- pm4py: Python library for Process Mining
- Entirely developed within the PADS team
- Open source
- Source on Github: https://github.com/pm4py/
- Recently published on pip a 0.1 version

pm4py: the concept

- ProM is a general tool, aimed both at researchers, analysts and companies
  - We wanted something made by researchers for researchers
- ProM trades off simplicity for power: steep learning curve
  - Focus on simplicity
- ProM suffers from JAR hell (dependency problems)
  - Focus on ease of installation and use
- The ProM project started off around 15 years ago
  - We wanted to update the technology stack (Python)

pm4py: contents

- The published version is a “lite” edition of pm4py
  - We have a lot of other algorithms, but they need testing and stabilization
  - This lite version is more than enough for the IDS course
  - The more complete “dev” version is available on Github
- Import and export of logs and models
  - CSV, XES and Petri nets
- Process Discovery: Alpha Miner and Inductive Miner
- Conformance Checking: Token Replay

pm4py: contents

- **Still experimental!**
- **Please report bugs on Github!**

More advertisement!

- In the last lecture Wil told you about opportunities for thesis and HiWi contracts with the PADS team
- There's more!

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

- Twitter: @pads_rwth

P D
A S
Chair of Process
and Data Science

More advertisement!

- **Seminars: in the next semester, there will be a seminar on Prediction in Process Mining**

- **We will talk about how to perform prediction on process data with advanced ML techniques (e.g. Recurrent Neural Networks)**