Conformance Checking

Andrea Polini

Process Mining
MSc in Computer Science (LM-18)
University of Camerino

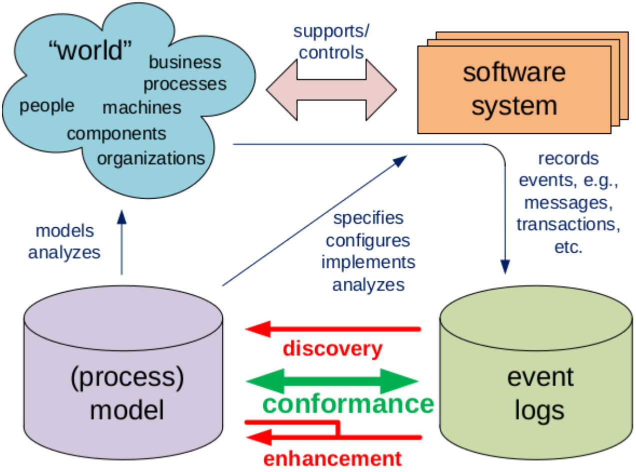
Summary

1 Introduction

Summary

1 Introduction

Conformance Checking



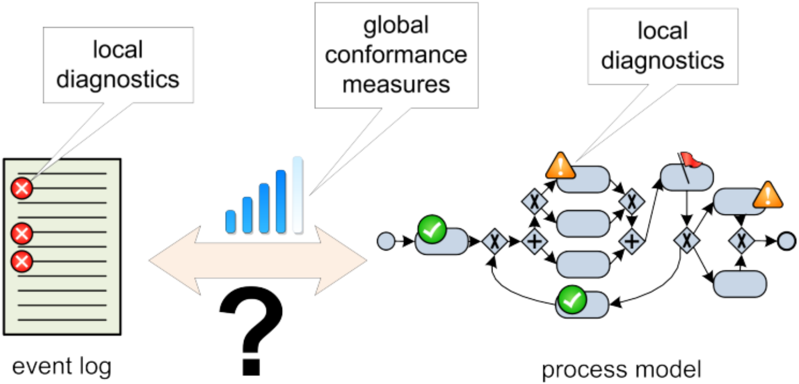
Motivations

Why?

Conformance checking relates events in the event log to activities in the process model and compares both. The goal is to find commonalities and discrepancies between the modeled behavior and the observed behavior. Conformance checking is relevant for business alignment and auditing:

- find undesirable deviations suggesting fraud or inefficiencies
- measuring the performance of process discovery algorithms
- repair models that are not aligned well with reality

Using Conformance Checking

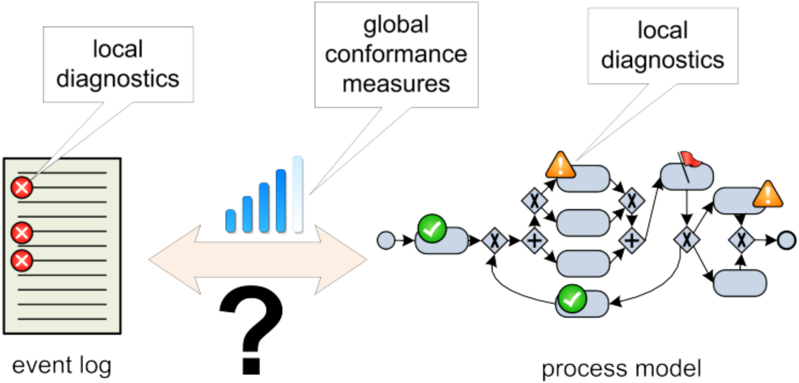


- global conformance measures – e.g. 85% of the cases in the event log can be replayed by the model
- local diagnostics – e.g. activity x was executed 15 times although this was not allowed according to the model

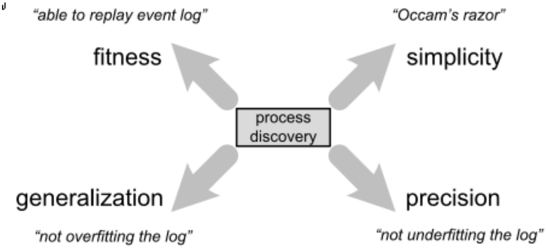
Results Interpretation

The interpretation of non-conformance depends on the purpose of the model:

- descriptive
- normative



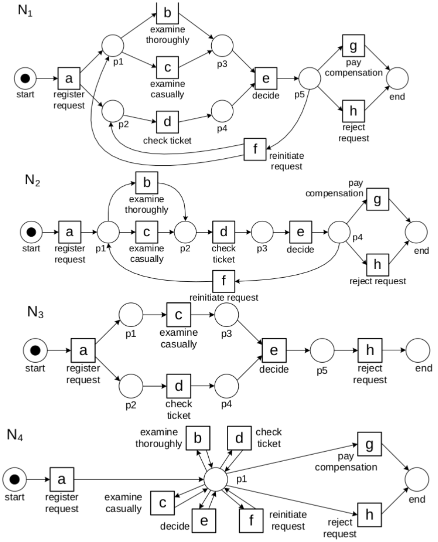
Quality criteria



Fitness function

- A naïve approach towards conformance checking would be to simply count the fraction of cases that can be "parsed completely"
  - N_1 : 1, N_2 : 0.6815, N_3 : 0.4543, N_4 : 1

Four models and one log



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

Token Based Metrics

- The fitness metric is generally defined at the level of events
  - Let's continue to replay a trace adding (and counting) tokens to enable blocked transitions, and also counting the remaining tokens at the end of the execution

Token Based Metrics

- The fitness metric is generally defined at the level of events
  - Let's continue to replay a trace adding (and counting) tokens to enable blocked transitions, and also counting the remaining tokens at the end of the execution

Let's consider model N_1, the following four counters,

- p: number of produced tokens
- c: number of consumed tokens
- m: number of added tokes
- r: number of remaining tokens,

and let's replay trace $\sigma_3 = \langle a, d, c, e, h \rangle$

Token Based Metrics

- The fitness metric is generally defined at the level of events
  - Let's continue to replay a trace adding (and counting) tokens to enable blocked transitions, and also counting the remaining tokens at the end of the execution

Let's consider model N_1, the following four counters,

- p: number of produced tokens
- c: number of consumed tokens
- m: number of added tokes
- r: number of remaining tokens,

and let's replay trace $\sigma_3 = \langle a, d, c, e, h \rangle$

Now let's replay the trace on N_2

Computing fitness at trace level

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

- What about replaying trace $\sigma_2 = \langle a, b, d, e, g \rangle$ on N_3?
- When a trace contains labels for which there is no corresponding transition the trace has to be projected on the available transitions

Computing fitness at trace level

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

- What about replaying trace $\sigma_2 = \langle a, b, d, e, g \rangle$ on N_3?
- When a trace contains labels for which there is no corresponding transition the trace has to be projected on the available transitions

Computing fitness at trace level

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

- What about replaying trace $\sigma_2 = \langle a, b, d, e, g \rangle$ on N_3?
- When a trace contains labels for which there is no corresponding transition the trace has to be projected on the available transitions

Computing fitness at trace level

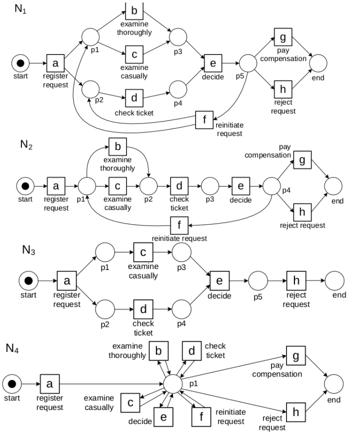
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

- What about replaying trace $\sigma_2 = \langle a, b, d, e, g \rangle$ on N_3?
- When a trace contains labels for which there is no corresponding transition the trace has to be projected on the available transitions

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

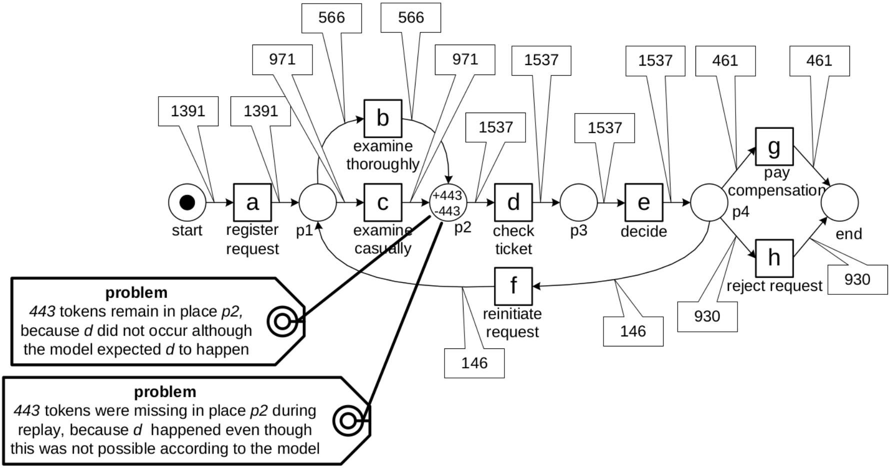
Computing fitness at the log level

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

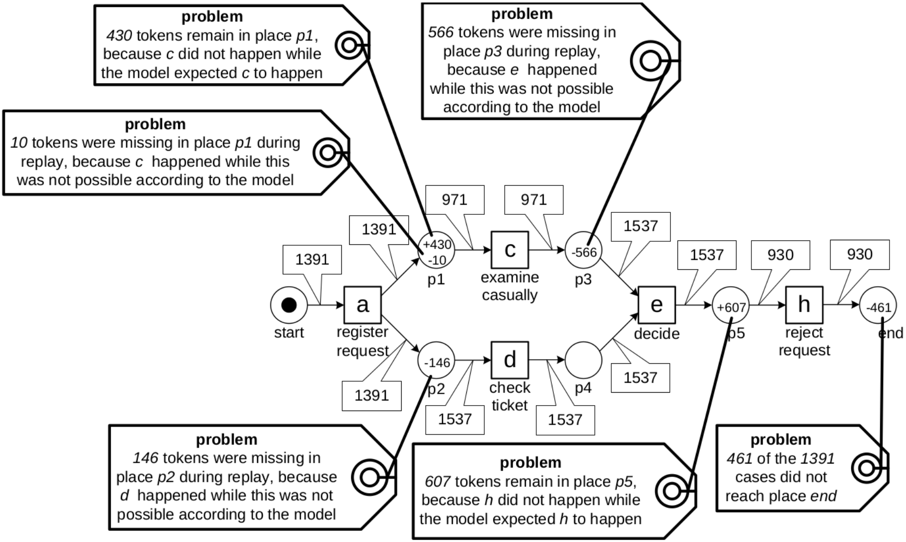


- $fitness(L_{full}, N_1) = 1$
- $fitness(L_{full}, N_2) = 0.9504$
- $fitness(L_{full}, N_3) = 0.8797$
- $fitness(L_{full}, N_4) = 1$

Diagnostics (N_2)



Diagnostics (N_3)



Further analysis