CS 4072 - Topics in CS
Process Mining

Lecture # 25

May 30, 2022

Spring 2022

FAST - NUCES, CFD Campus

Dr. Rabia Maqsood

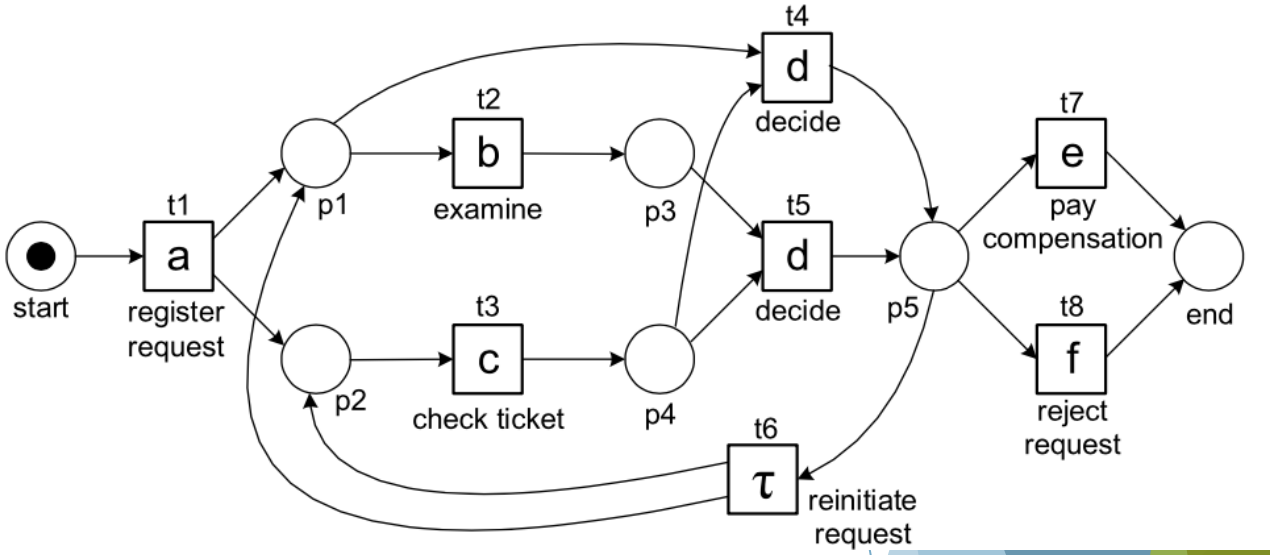
rabia.maqsood@nu.edu.pk

Today's Topics

- Conformance Checking
  - Sequence Alignment (continued)

NOTE: silent transition leaves no trail in the event log

Alignments



- Consider

$\sigma_4 = \langle a,c,d,b,c,d,c,d,c,b,d,f \rangle$ and N_5

- Following are the possible alignments:

$$\gamma_{5,4} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|} \hline a & c & d & \gg & b & c & d & \gg & c & d & \gg & c & b & d & f \\ \hline a & c & d & \tau & b & c & d & \tau & c & d & \tau & c & b & d & f \\ t1 & t3 & t4 & t6 & t2 & t3 & t5 & t6 & t3 & t4 & t6 & t3 & t2 & t5 & t8 \\ \hline \end{array}$$

Alignments

- (x,(y,t)) is a *legal move* if one of the following four cases holds:
  - x = y and y is the visible label of transition t (*synchronous move*)
  - x = ≫ and y is the visible label of transition t (*visible model move*)
  - x = ≫, y = τ and transition t is silent (*invisible model move*)
  - x ≠ ≫ and (y,t) = ≫ (*log move*)
- Other moves such as (≫, ≫) and (x , (y , t)) with x ≠ y are illegal moves.

Alignments

- To select the most appropriate alignment, we associate ***costs*** to undesirable moves and select an alignment with the lowest total costs.
- Generic cost function:
  - Cost function δ assigns costs to **legal moves**.
  - Moves where log and model agree have no costs, i.e., $\delta(x,(y,t)) = 0$ for *synchronous moves* (with $x = y$).
  - Moves in model only have no costs if the transition is invisible, i.e., $\delta(\gg,(\tau,t)) = 0$ for *invisible model moves*.
  - $\delta(\gg,(y,t)) > 0$ is the cost when the model makes a "y move" without a corresponding move of the log (*visible model move*).
  - $\delta(x, \gg) > 0$ is the cost for an "x move" in just the log (*log move*).

Computing Fitness

- The cost function can be converted into a fitness value (between 0 and 1).
- Compare the cost of an optimal alignment with a “worst case scenario” = move in log only for observed events and shortest path with only moves in model.

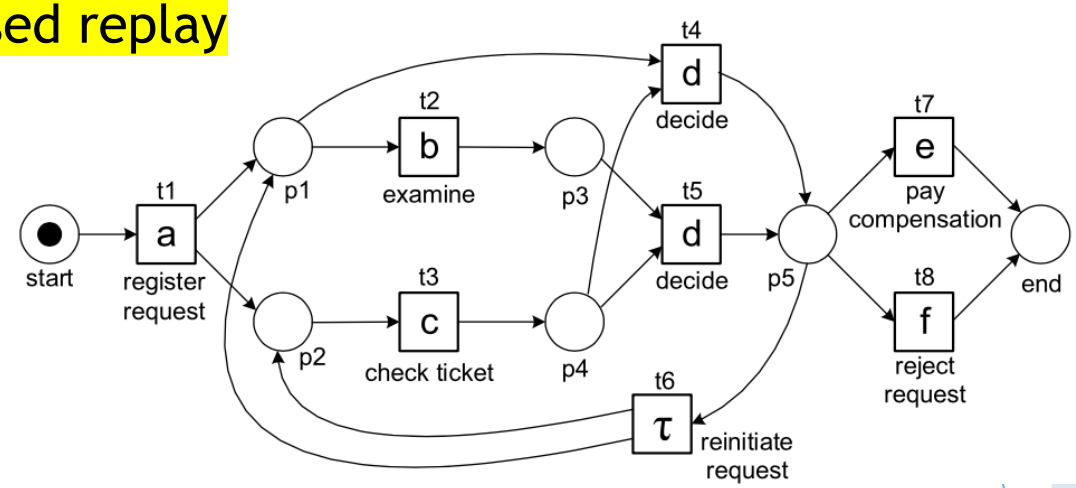
Computing Fitness

- The cost function can be converted into a fitness value (between 0 and 1).
- Let the worst-case trace alignment be represented as: $\lambda^N_{worst}(\sigma)$
- And, the optimal-case trace alignment as: $\lambda^N_{opt}(\sigma)$
- Now, the fitness of a trace can be defined as:

$$fitness(\sigma, N) = 1 - \frac{\delta(\lambda^N_{opt}(\sigma))}{\delta(\lambda^N_{worst}(\sigma))}$$

Compute the fitness using token-based replay

Computing Fitness



- For $\sigma_2 = \langle a,b,d,f \rangle$ and N_5

- Optimal cost

$$\delta(\lambda_{opt}^{N_5}(\sigma_2)) = 1$$

$$\gamma_{5,2a} = \begin{array}{|c|c|c|c|c|} a & b & \gg & d & f \\ a & b & c & d & f \\ t1 & t2 & t3 & t5 & t8 \end{array} \qquad \gamma_{5,2b} = \begin{array}{|c|c|c|c|c|} a & \gg & b & d & f \\ a & c & b & d & f \\ t1 & t3 & t2 & t5 & t8 \end{array}$$

- What is the worst-case cost?

$$\delta(\lambda_{worst}^{N_5}(\sigma_2)) = 8$$

$$\gamma_{5,2w} = \begin{array}{|c|c|c|c|c|c|c|c|} a & b & d & f & \gg & \gg & \gg & \gg \\ \gg & \gg & \gg & \gg & a & c & d & f \\ & & & & t1 & t3 & t4 & t8 \end{array}$$

- Fitness of the trace is:

$$fitness(\sigma_2, N_5) = 1 - \frac{1}{8} = 0.875$$

Computing Fitness of Event log

- As before, the fitness notion can be extended to event logs.

$$fitness(L, N) = 1 - \frac{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda^N_{opt}(\sigma))}{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda^N_{worst}(\sigma))}$$

Sum of all costs when replaying the event log using optimal alignments

Sum of all worst-case alignment costs

Token-based replay vs. Alignments

- Alignments provide more *detailed* but *easy to understand diagnostics*.
- Alignments provide more *accurate diagnostics*.
- Alignments are *configurable* through the cost function.
- Alignments can be used to *map each case onto a feasible path in model*.
- Alignments are *model independent*.
- Token-based replay provides *deterministic diagnostics* whereas multiple optimal alignments may exist for a trace. This can be addressed by deterministically picking one of possibly many optimal alignments. This does not influence the overall fitness value, but influences diagnostics based on alignments.

Diagnostics (1) token-based replay

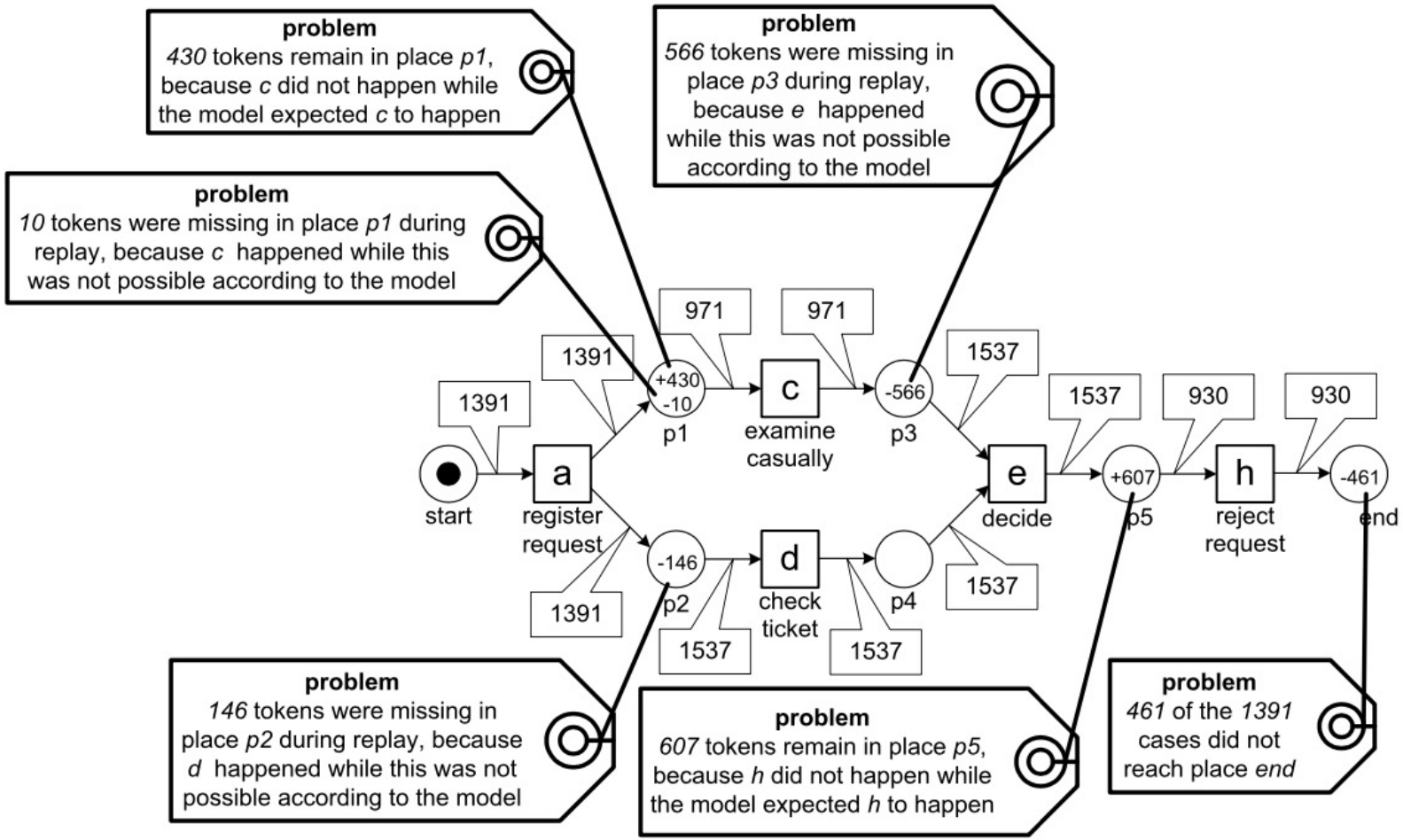


Fig. 8.7 Diagnostic information showing the deviations ($fitness(L_{full}, N_3) = 0.8797$)

Diagnostics (2) alignments

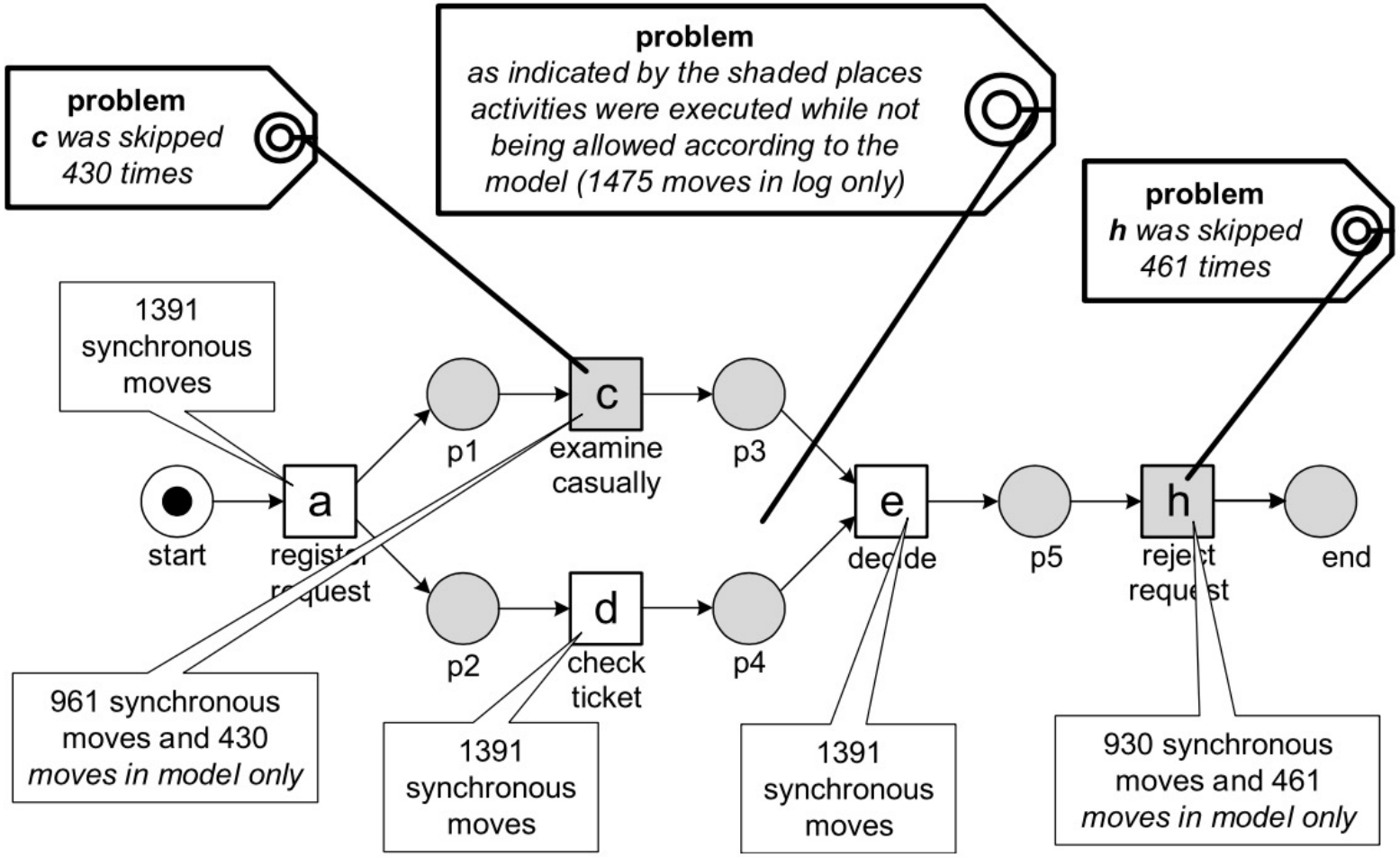


Fig. 8.11 Diagnostic information showing the deviations ($fitness(L_{full}, N_3) = 0.83676$)

Other applications of Conformance checking

Self study: Read Section 8.5

- Repairing models
- Evaluating process discovery algorithms

Reading Material

- Chapter 8: Aalst