

# **Analysing differences between business process similarity measures**

**Michael Becker, Ralf Laue**

**1st Int. Workshop on Process Model Collections (PMC 2011)**

**29.08.11, Clermont-Ferrand**

# Outline

- Introduction
  - Steps to measure similarity
  - Requirements for similarity measures
- Analysis
  - Model changes
  - Types of similarity measures
- Conclusion
  - Results and discussion
  - Future research

# Introduction

## Step to measure similarity

- Map: identify correspondences between activities
  - Sounds easy but may be non-trivial
- **Measure: calculate similarity / distances**
  - Focal point of the paper
- Evaluate: rank models according to intended use

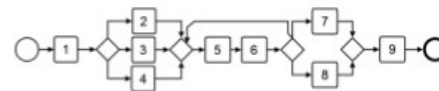
## Requirements for measures

- Following metric spaces
  - 1) Non-negativity
  - 2) Symmetry
  - 3) Equivalence
    - 1) Trace equivalence
  - 4) Triangle inequality
- Furthermore
  - 5) Respect commonalities and differences
  - 6) Overall similarity based on similarity between elements
  - 7) Applicable on arbitrary models
  - 8) Efficient calculation

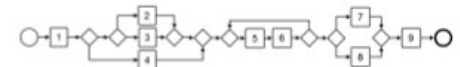
# Analysis

# Model changes

- Initial point: moderately sized BPMN model
- Apply different change operations (Dijkman2007, Weber2008, Weidlich2009)



(a)  $V_0$ : original BPMN model



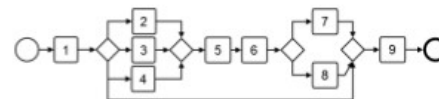
(b)  $V_1$ : model with same set of traces as  $V_0$



(c)  $V_2$ : model with modified connector types



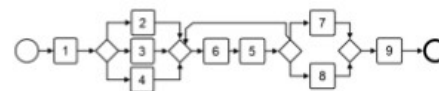
(d)  $V_3$ : model with additional activities



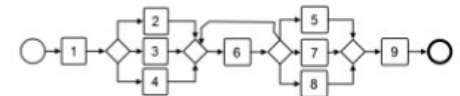
(e)  $V_4$ : model with modified control flow arcs



(f)  $V_5$ : model with modified control flow



(g)  $V_6$ : model with modified order of activities 5 and 6



(h)  $V_7$ : model with modified activity 5

## **Types of measures**

- Correspondences of nodes and edges
- Graph edit distance
- Causal dependencies between activities
- Comparison of traces / logs

## **Correspondences of nodes and edges**

- Foundation: amount of common nodes and arcs in models
- Rather simple measures
- Efficient calculation possible

## Graph edit distance

- Foundation: amount of operations necessary to transform one process into the other
  - Insert/delete vertices/edges
  - Substitute/move vertices

## **Causal dependencies between activities**

- Foundation: possible execution order of activities
  - Sequential
  - Parallel
  - Exclusive
  - Arbitrary

## Comparison of traces/logs

- Foundation: abstract from the actual process (graph) structure
- Commonalities in traces and logs (derived by simulation etc.)

## Tool support: ProM plug-in

- Plug-in to calculate various similarity measures
- API can be used standalone in other projects, too
- Source available at SourceForge: [prom-similarity](#)

The screenshot displays the ProM plug-in interface. It features two windows showing activity diagrams for 'Imported - v0.xml - AML file' and 'Imported - v3.xml - AML file'. The diagrams consist of nodes (green rectangles) and edges (black lines) connected by arrows, representing process flows. The top window shows a diagram with a zoom level of 50%. The bottom window shows a similar diagram, also at 50% zoom. To the right of the diagrams is a 'Similarity Measures' panel. This panel lists various similarity metrics and their values:

- Correspondences between nodes and edges in the BPM
- CommonActivityName Similarity: 0.8461538461538461
- NodeMatching Similarity: 0.8461538461538461
- FeatureBased Similarity: 0.8181818181818182
- CommonNodesEdges Similarity: 0.4
- Edit distance between graphs
- GraphEditDistance Similarity: 0.6298701298701299
- LaRosa Similarity: 0.8278213507625273
- Causal Dependencies Between Activities
- DependencyGraph Similarity: 0.037037037037035
- TAR Similarity: 0.037037037037035
- CausalFootprint Similarity: 0.16977320573294194
- State-Space Based Approaches

# Conclusion

## Results

- Similarity values of 23 different measures applied on variants
  - Very different results

# Results: absolute and relative

	Similarity between $V_0$ and ...						
	$V_1$	$V_2$	$V_3$	$V_4$	$V_5$	$V_6$	$V_7$
Measures based on the correspondence of nodes and edges (not taking into account the control flow)							
Percentage of Common Activity Names [15]	1.00	1.00	0.82	1.00	1.00	1.00	1.00
Label Matching Similarity [4]	1.00	1.00	0.82	1.00	1.00	1.00	1.00
Similarity of Activity Labels [6]	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Feature-Based Activity Similarity [18]	1.00	1.00	0.82	1.00	1.00	1.00	1.00
Percentage of Common Nodes and Edges [19]	1.00	1.00	0.40	0.95	0.58	0.76	0.79
Node- and Link-Based Similarity [20]	1.00	1.00	0.64	0.96	0.72	0.90	0.90
Measures based on graph edit distances							
Graph Edit Distance [4]	1.00	1.00	0.63	0.97	0.73	0.86	0.12
Graph Edit Distance [22]	0.05	0.04	0.20	0.33	0.03	0.33	0.17
Label Similarity and Graph Edit Distance [23]	0.81	1.00	0.60	0.96	0.61	0.79	0.84
Label Similarity and Graph Edit Distance [10]	0.05	0.03	0.06	0.33	0.03	0.14	0.20
Number of High-Level Change Operations [24]	1.00	0.17	0.20	0.33	0.14	0.50	0.50
Comparing BPM Represented as Trees [26]	1.00	1.00	0.07	0.12	0.06	0.14	0.14
Edit Distance Between Reduced Models [26]	1.00	0.07	1.00	1.00	0.00	0.80	0.81
Measures that analyse causal dependencies between activities							
Comparing Dependency Graphs [32, 33]	1.00	1.00	0.04	0.33	0.06	0.09	0.10
Comparing Dependency Graphs [34]	1.00	0.93	0.54	0.90	0.51	0.98	0.83
Reference Similarity [35]	not defined ( $V_0$ has a loop!)						
TAR-Relationship [35]	1.00	0.57	0.04	0.85	0.11	0.41	0.47
Causal Behavioural Profiles [36]	1.00	0.93	0.63	0.93	0.22	0.98	0.89
Causal Footprints [4]	1.00	1.00	0.45	0.80	0.59	0.97	0.84
State-Space as n-grams [40]	1.00	0.10	0.04	0.33	0.05	0.09	0.10
Measures that compare state spaces or logs							
Longest Common Subsequence of Traces [42]	1.00	0.86	0.79	1.00	0.43	0.93	0.90
Similarity Based on Principal Transition Sequences [43]	1.00	0.83	0.61	0.84	0.20	0.85	0.83
Similarity Based on Traces [44]	1.00	0.90	0.33	0.83	0.22	0.72	0.65

	Similarity between $V_0$ and ...						
	$V_1$	$V_2$	$V_3$	$V_4$	$V_5$	$V_6$	$V_7$
Measures based on the correspondence of nodes and edges (not taking into account the control flow)							
Percentage of Common Activity Names [15]							
Label Matching Similarity [4]							
Similarity of Activity Labels [6]							
Feature-Based Activity Similarity [18]							
Percentage of Common Nodes and Edges [19]							
Node- and Link-Based Similarity [20]							
Measures based on graph edit distances							
Graph Edit Distance [4]							
Graph Edit Distance [22]							
Label Similarity and Graph Edit Distance [23]							
Label Similarity and Graph Edit Distance [10]							
Number of High-Level Change Operations [24]							
Comparing BPM Represented as Trees [26]							
Edit Distance Between Reduced Models [26]							
Measures that analyse causal dependencies between activities							
Comparing Dependency Graphs [32, 33]							
Comparing Dependency Graphs [34]							
TAR-Relationship [35]							
Causal Behavioural Profiles [36]							
Causal Footprints [4]							
State-Space as n-grams [40]							
Measures that compare state spaces or logs							
Longest Common Subsequence of Traces [42]							
Similarity Based on Principal Transition Sequences [43]							
Similarity Based on Traces [44]							

## Discussion: Requirements

- Adherence of measures to identified requirements
  - No measure fulfils all requirements
  - Is this necessary?

	1	2	3	3a	5	6	7	8
Measures based on the correspondence of nodes and edges (not taking into account the control flow)								
Percentage of Common Activity Names [6]	yes	yes	no	no	yes	yes	yes	yes
Label Matching Similarity [3]	yes	no	no	no	yes	yes	yes	yes
Similarity of Activity Labels [5]	yes	no	no	no	yes	yes	yes	yes
Feature-Based Activity Similarity [11]	yes	yes	no	no	yes	yes	yes	yes
Percentage of Common Nodes and Edges [12]	yes	yes	no	no	yes	yes	yes	yes
Node- and Link-Based Similarity [10]	yes	yes	no	no	yes	yes	yes	yes
Measures based on graph edit distances								
Graph Edit Distance [3]	yes	yes	yes	no	yes	yes	yes	yes
Graph Edit Distance [30]	yes	yes	yes	no	no	no	yes	yes
Label Similarity and Graph Edit Distance [13]	yes	yes	no	no	yes	yes	yes	yes
Label Similarity and Graph Edit Distance [26]	yes	yes	yes	no	no	yes	yes	yes
Number of High-Level Change Operations [16]	yes	yes	no	yes	yes	no	n/a	yes
Comparing BPMs Represented as Trees [17]	yes	yes	no	no	yes	no	yes	yes
Distance Between Reduced Models [7]	yes	no	no	no	yes	no	yes	yes
Measures that analyse causal dependencies between activities								
Comparing Dependency Graphs [8, 9]	yes	yes	no	no	yes	no	yes	yes
Comparing Dependency Graphs [22]	yes	yes	no	no	yes	no	n/a	yes
Reference Similarity [21]	yes	yes	no	yes	yes	no	yes	no
TAR-Relationship [21]	yes	yes	no	no	yes	no	yes	no
Causal Behavioural Profiles [20]	yes	yes	no	no	yes	no	no	yes
Causal Footprints [3]	yes	yes	no	no	yes	no	yes	no
Set of Traces as n-grams [14]	yes	no	no	no	no	no	yes	no
Measures that compare sets of traces or logs								
Longest Common Subsequence of Traces [18]	yes	yes	no	no	yes	no	yes	no
Similarity Based on Principal Transition Sequences [31]	yes	yes	no	no	yes	no	yes	yes
Similarity Based on Traces [19]	yes	yes	no	no	yes	no	yes	yes

## Discussion: Results

- Measures are developed for different purposes, e.g.
  - Find related/similar models
  - Measure conformance to reference models
  - Discover services
- Different types of measures show different application areas
  - However, specific advantages and disadvantages must be considered

## Future research

- Similarity measure ontology to structure the area
  - Based on established types
- Life cycle based recommendations
  - Design, implementation, execution need different measures with different properties
- Quantitative comparison of models: SAP R/3 reference model
- Identify discrepancies between human perception of similarity and results of similarity measures
  - Questionnaire

**Backup**

**Thanks for the attention!**

## References

- Dijkman2007: A classification of differences between similar business processes, EDOC 2007
- Weber2008: Change patterns and change support features - enhancing flexibility in process-aware information systems, Data Knowl. Eng. 66 (2008)
- Weidlich2009: Vertical alignment of process models - how can we get there?, Enterprise, Business-Process and Information Systems Modeling 2009

## Requirements in detail

- Non-negativity:  $\text{dist}(M_0, M_1) \geq 0$
- Symmetry:  $\text{dist}(M_0, M_1) = \text{dist}(M_1, M_0)$
- Equivalence:  $\text{dist}(M_0, M_1) = 0 \leftrightarrow M_0 = M_1$ 
  - Trace equiv.:  $\text{dist}(M_0, M_1) = 0 \leftrightarrow \Sigma(M_0) = \Sigma(M_1)$
- Triangle inequality:  $\text{dist}(M_0, M_2) \leq \text{dist}(M_0, M_1) + \text{dist}(M_1, M_2)$