Reality is not always what it seems!

Using process mining and conformance checking to find out what is really going on in your system

Prof.dr.ir. Wil van der Aalst

Eindhoven University of Technology,

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

w.m.p.v.d.aalst@tue.nl

Outline

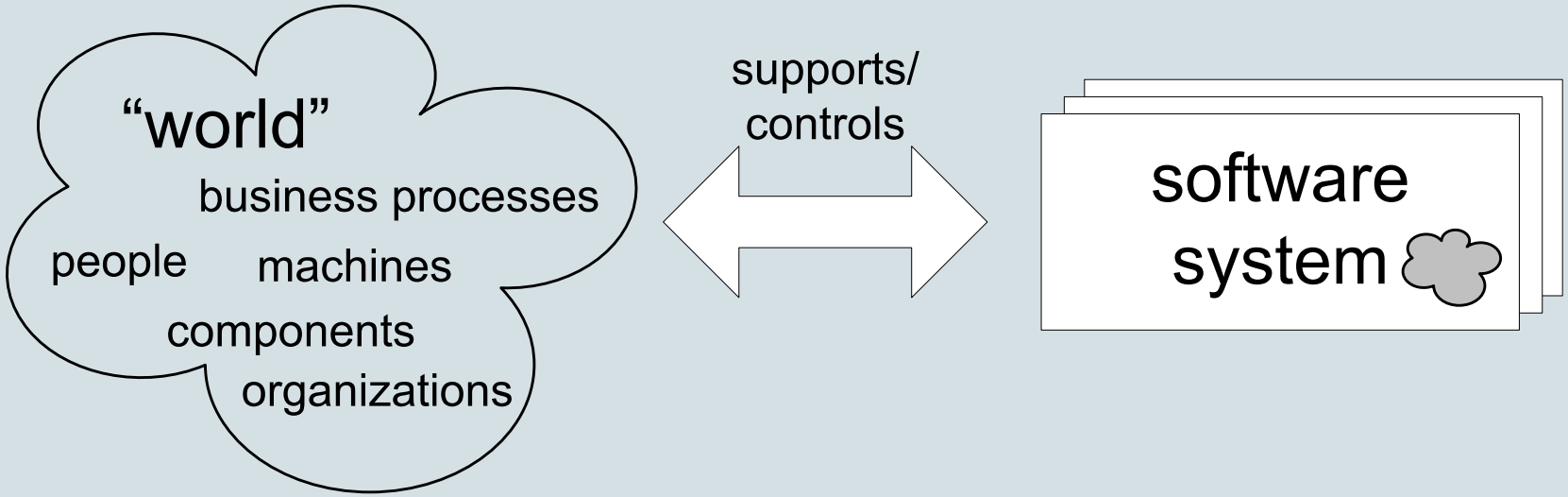
- Process Analysis: From verification to process mining
- Process Mining: Running example
- Discovery
- Conformance checking
- Reality Check
- Conclusion

The work of many people!

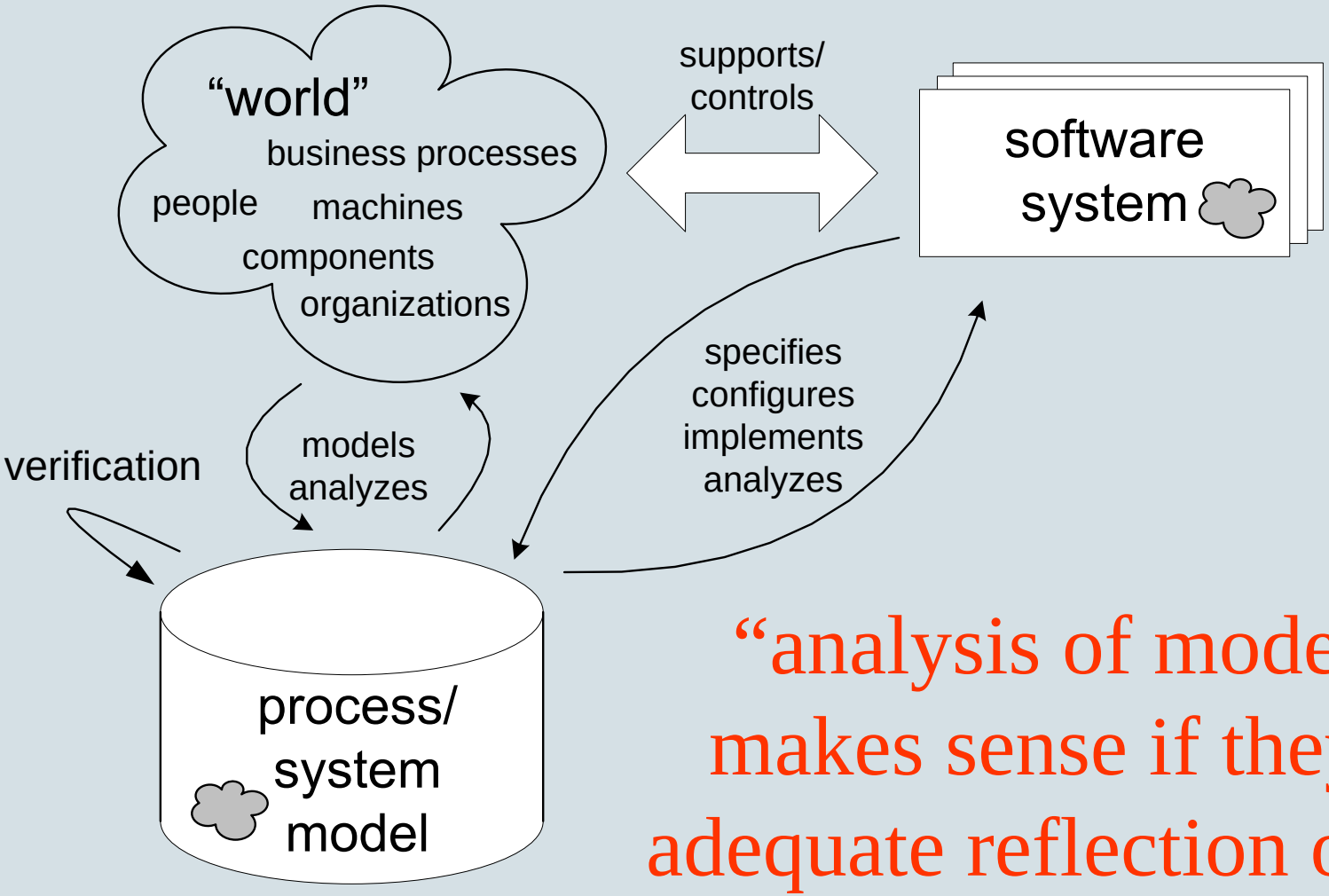
Thanks to Ton Weijters, Boudewijn van Dongen, Ana Karla Alves de Medeiros, Anne Rozinat, Christian Günter, Eric Verbeek, Ronny Mans, Minseok Song, Laura Maruster, Huub de Beer, Peter van den Brand, Jan Mendling, Andriy Nikolov, Jianmin Wang, Lijie Wen, Irene Vanderfeesten, Mariska Netjes, Steffi Rinderle, Walid Gaaloul, Gianluigi Greco, Antonella Guzzo, etc. etc.

Overview Process Analysis

Software systems are the mirror image of the “world”



Dual role of process models



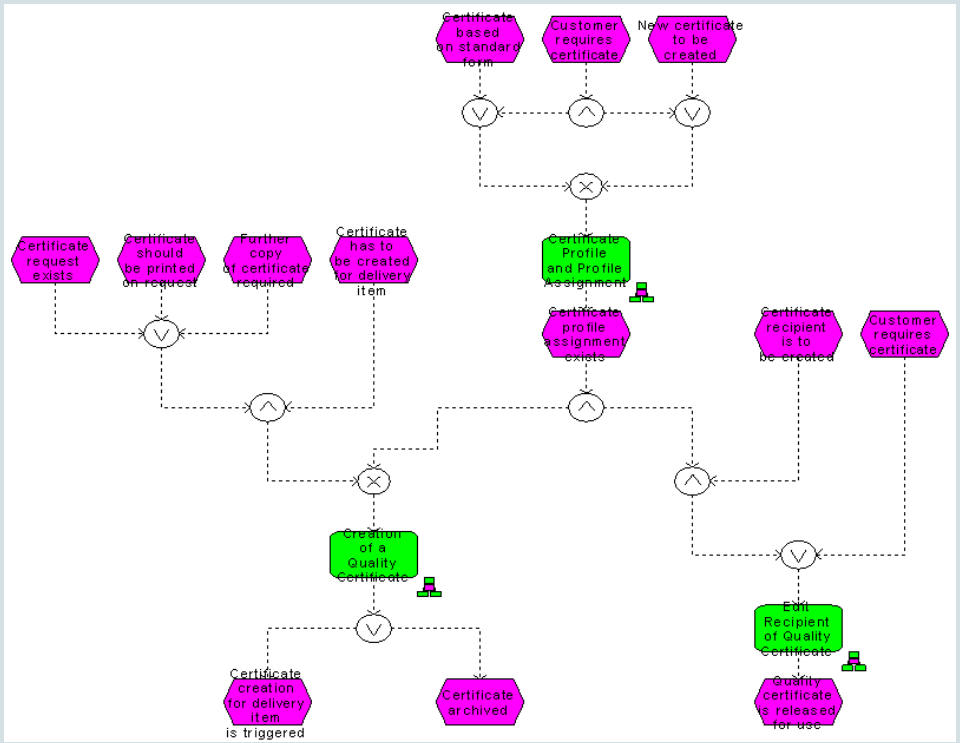
"analysis of models only makes sense if they are an adequate reflection of reality"

the disconnect …

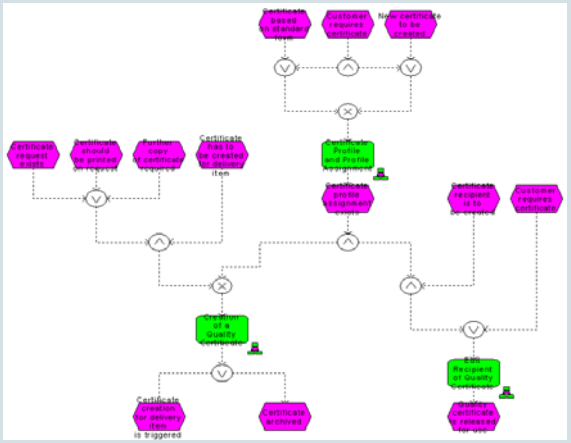
Example: Verification of the SAP Reference model (Joint work with Jan Mendling)

- The SAP reference model contains more than 600 non-trivial process models expressed in terms of Event-driven Process Chains (EPCs).

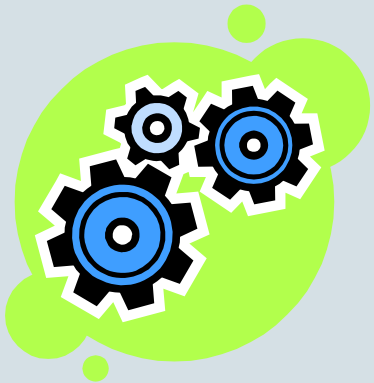
Hierarchy Level	Models	eEPC	Function Allocation Diagram	Process Selection Diagram	Role Activity Diagram	EPC	Error
1	1	1	0	0	0	0	0
2	58	29	0	29	0	0	0
3	175	73	0	0	0	102	15
4	1226	724	0	0	0	502	19
5	8384	3035	3035	0	2014	0	0
All Levels	9844	3862	3035	29	2014	604	34



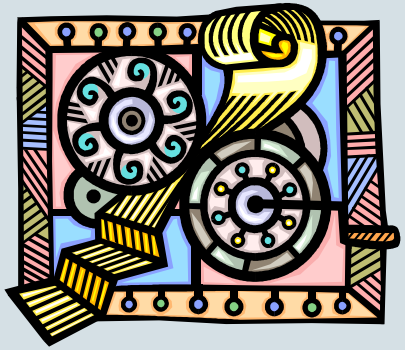
Approach



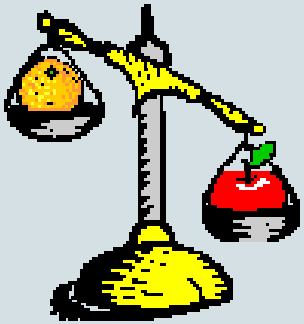
604 non-trivial process models



model analysis

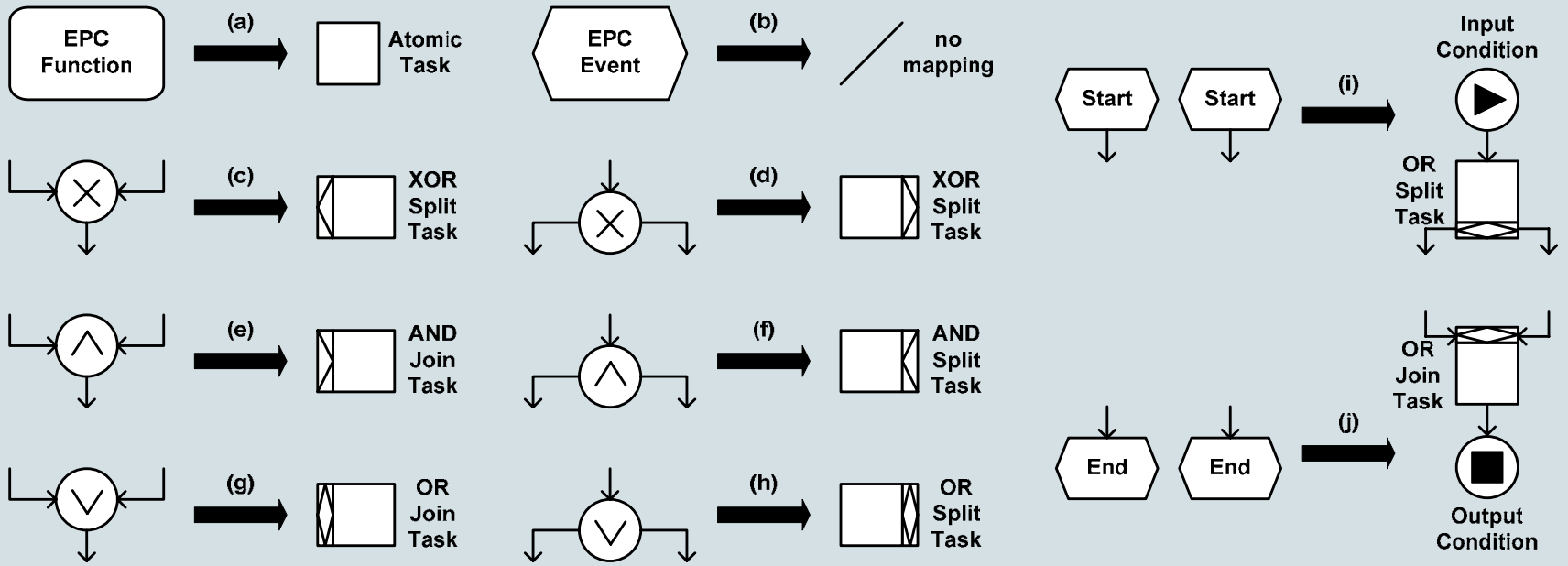


collect characteristics



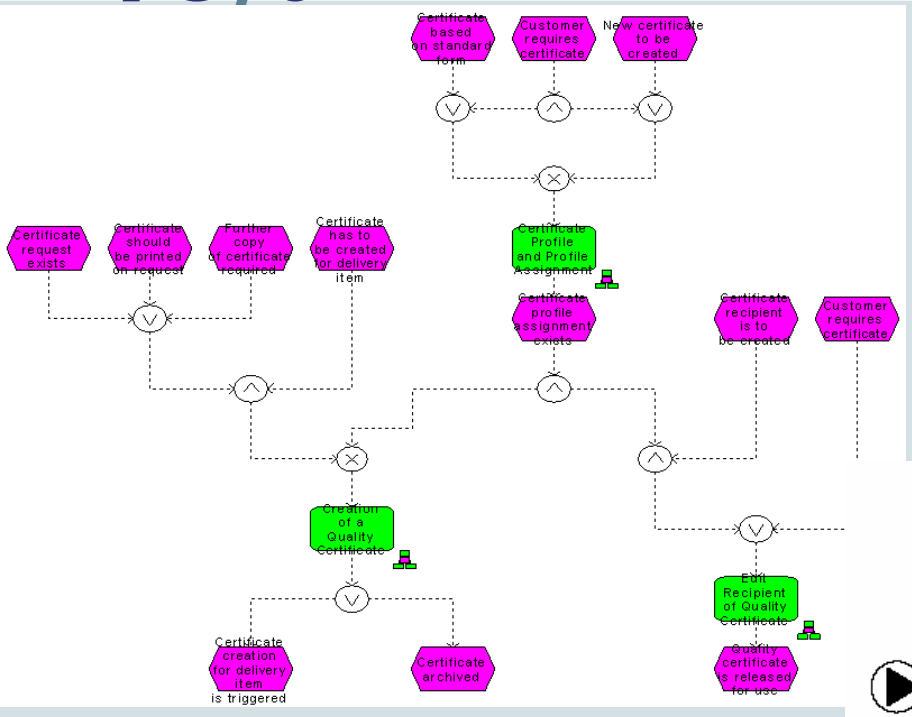
compare

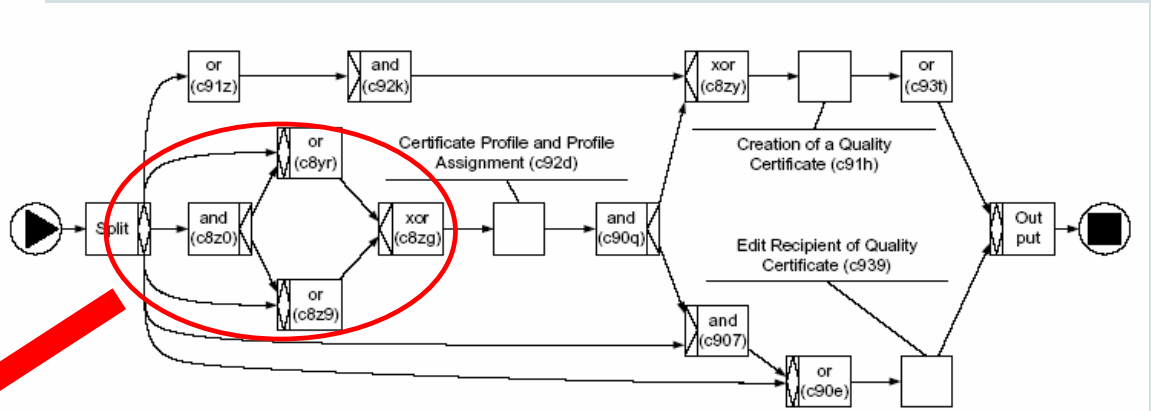
Simplistic approach: YAWL + invariants

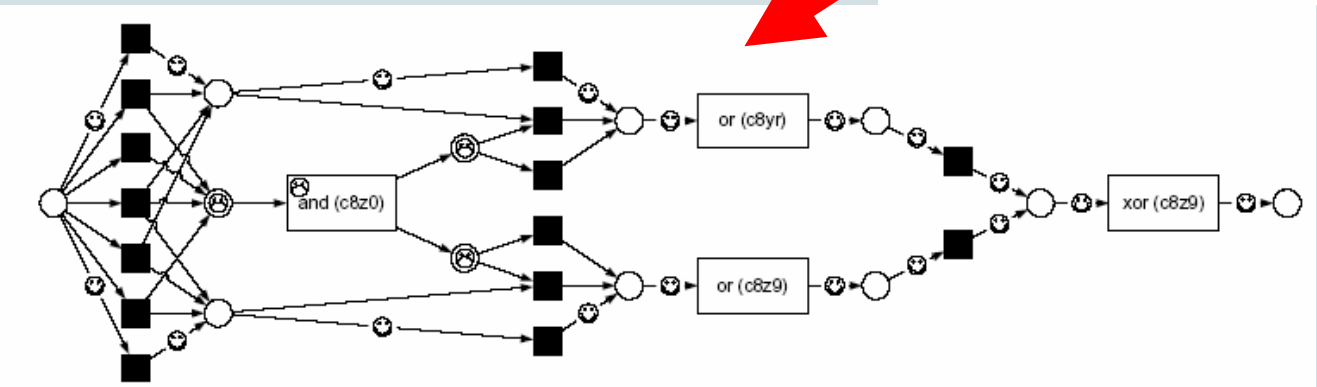


Analysis using **transition invariants**, i.e., only lower bound! ProM allows for more precise analysis

Simplistic approach: YAWL + Petri net invariants







Branch	Model	%	EPC	%	$E_{av.}$	$F_{av.}$	$C_{av.}$	$A_{av.}$	Cycle	Error	%
Asset Accounting	461	4.7%	43	7.1%	13.9	4.0	5.2	23.3	0	7	16.3%
Benefits Administration	50	0.5%	6	1.0%	9.5	3.3	5.8	19.7	3	0	0.0%
Compensation Management	122	1.2%	18	3.0%	7.6	3.4	3.3	13.7	3	1	5.6%
Customer Service	402	4.1%	41	6.8%	16.5	3.6	9.0	29.5	3	1	2.4%
Enterprise Controlling	599	6.1%	22	3.6%	14.3	10.1	6.1	32.1	0	3	13.6%
Environment, Health, Safety	102	1.0%	19	3.1%	3.5	2.7	1.2	7.0	0	0	0.0%
Financial Accounting	614	6.2%	54	8.9%	13.0	4.0	5.1	21.8	0	3	5.6%
Position Management	4	0.0%	0	0.0%	0.0	0.0	0.0	0.0	0	0	n.a.
Inventory Management	184	1.9%	3	0.5%	15.0	7.0	6.0	28.0	2	0	0.0%
Organizational Management	37	0.4%	5	0.8%	12.0	3.0	6.6	24.0	3	0	0.0%
Payroll	541	5.5%	7	1.2%	5.7	3.1	2.1	11.4	0	1	14.3%
Personnel Administration	15	0.2%	4	0.7%	7.3	1.5	4.0	12.3	0	0	0.0%
Personnel Development	60	0.6%	10	1.7%	8.7	2.5	4.4	15.6	3	1	10.0%
Personnel Time Management	87	0.9%	12	2.0%	10.8	3.0	5.3	19.5	1	2	16.7%
Plant Maintenance	399	4.1%	35	5.8%	20.5	4.2	11.4	37.8	9	1	2.9%
Procurement	444	4.5%	37	6.1%	6.7	3.5	2.7	12.4	0	2	5.4%
Product Data Management	366	3.7%	26	4.3%	4.5	5.4	2.2	13.7	0	0	0.0%
Production	296	3.0%	17	2.8%	8.8	3.0	2.9	13.7	0	1	5.9%
Production Planning	194	2.0%	17	2.8%	5.7	2.9	3.0	11.5	0	0	0.0%
Project Management	347	3.5%	36	6.0%	8.5	3.8	2.2	14.0	0	0	0.0%
Quality Management	209	2.1%	20	3.3%	20.5	3.8	11.7	37.8	1	1	5.0%
Real Estate Management	169	1.7%	6	1.0%	12.7	6.5	7.3	27.0	1	1	16.7%
Recruitment	56	0.6%	9	1.5%	7.4	2.6	4.1	13.8	3	0	0.0%
Retail	842	8.6%	1	0.2%	7.0	5.0	2.0	11.0	0	0	0.0%
Revenue & Cost Controlling	568	5.8%	19	3.1%	16.5	10.2	7.9	36.0	1	1	5.3%
Sales & Distribution	703	7.1%	76	12.6%	10.6	3.1	4.3	16.6	0	1	1.3%
Training & Event Management	95	1.0%	12	2.0%	13.0	2.7	6.2	22.2	0	1	8.3%
Travel Management	116	1.2%	1	0.2%	24.0	7.0	16.0	48.0	0	0	0.0%
Treasury	1761	17.9%	48	7.9%	10.5	3.5	4.5	18.1	0	6	12.5%
All 29 Branches	9844	100%	604	100%	11.5	4.0	5.2	20.8	33	34	5.6%

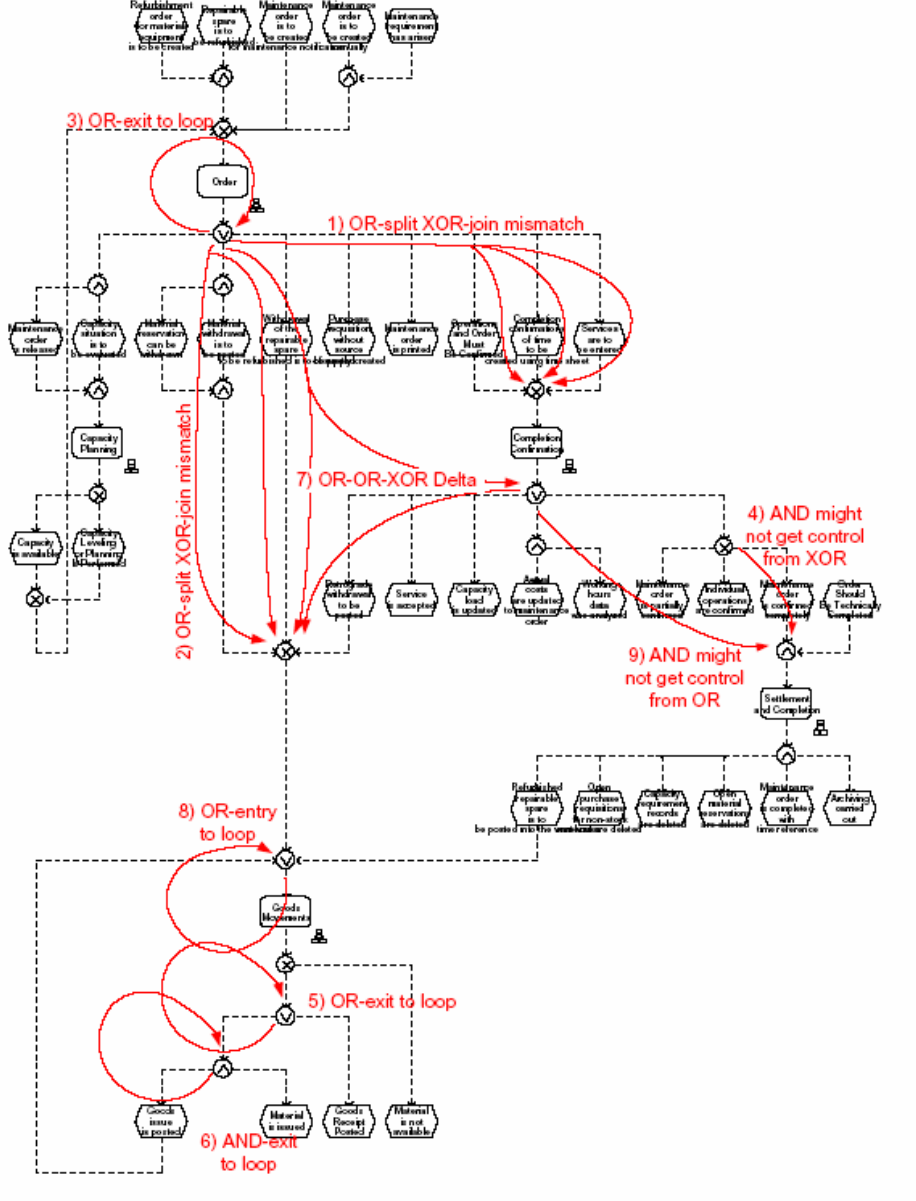
5.6%

5.6% is a lower bound!

- Using more refined techniques more errors are found, e.g., using reduction rules and state-space analysis it can be shown that 20.9% of the SAP models are incorrect (126/604).
- Other large repositories of EPC models:
  - Collection of 381 non-trivial EPCs from a German process reengineering project in the service sector
  - Collection of 935 non-trivial EPCs from the Austrian financial industry
  - Collection of 83 non-trivial EPCs from three different consulting companies
- Total: 2003 non-trivial EPCs

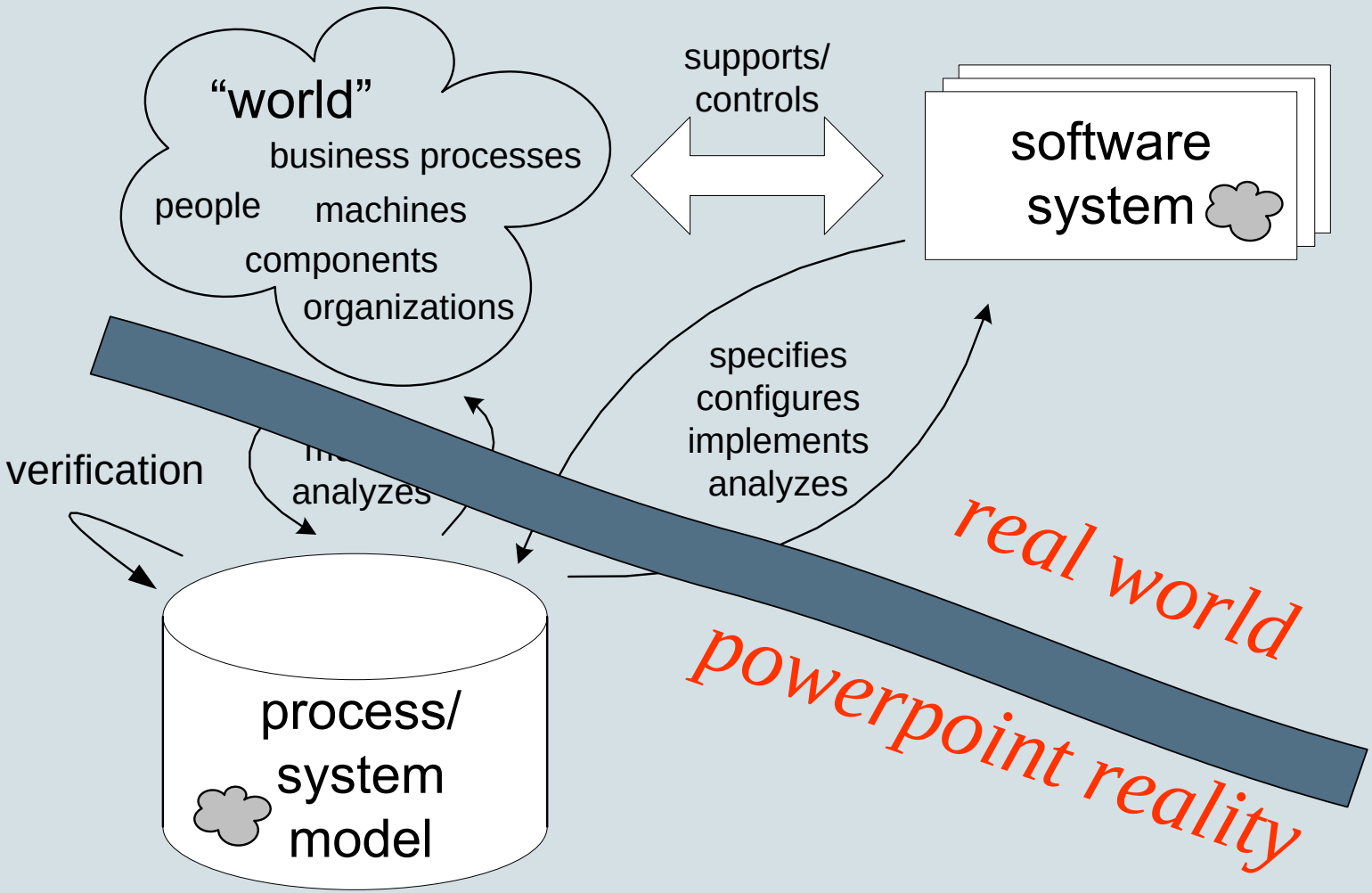
Overview results

Parameter Parameter	Complete Sample	SAP Ref. Model
xoEPC errors	154	90
Unreduced EPCs	156	103
ProM error EPCs	115	75
EPCs with errors	215	126
EPCs in total	2003	604
Error ratio	10.7%	20.9%

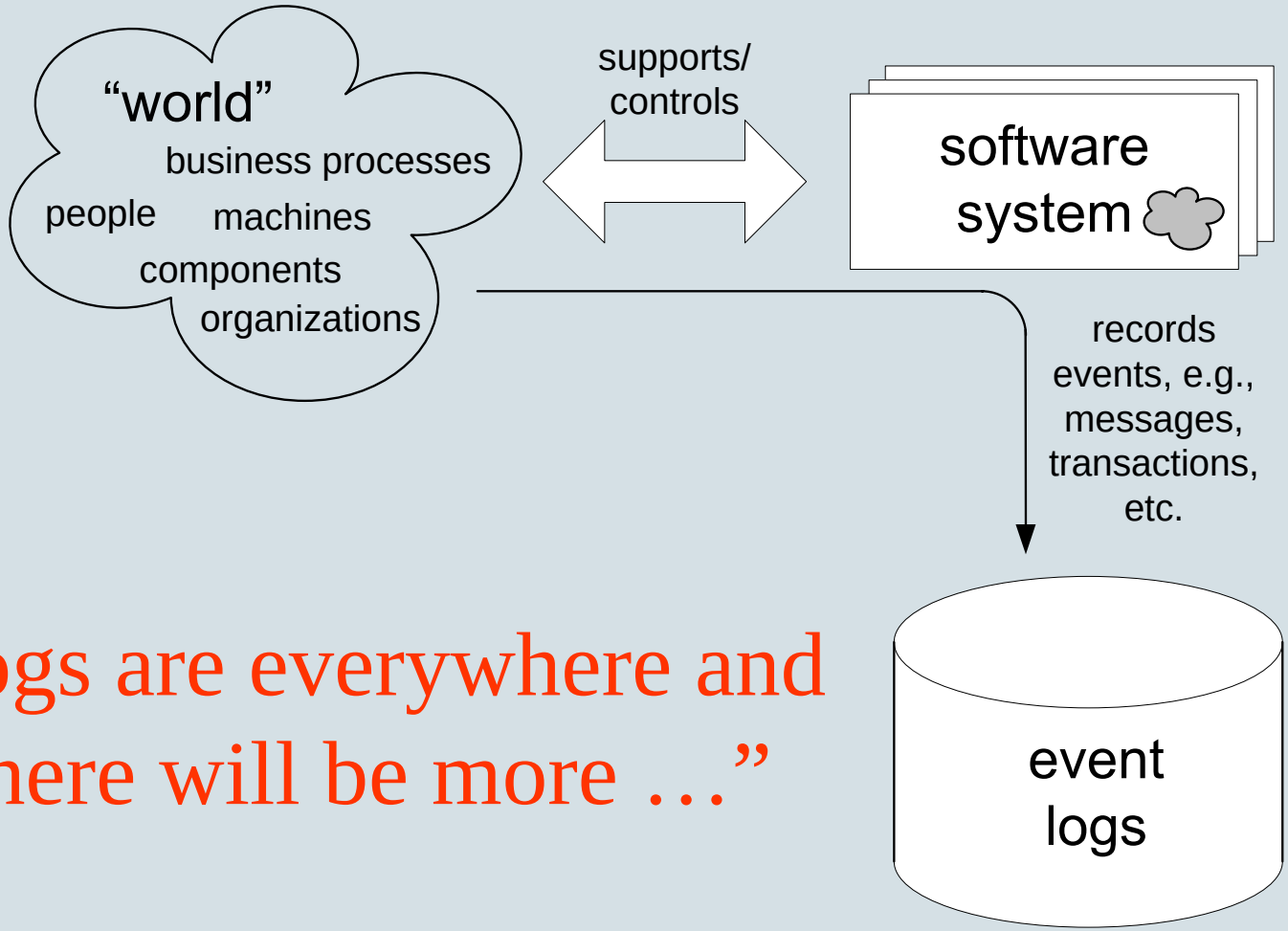


- Designers make errors (10.7
- Errors can be predicted (95.
- Process verification is mature, but models are not!
- Disconnect between ref. models and systems cf. SAP

Limitations of using models as a starting point



Event logs are a reflection of reality



"logs are everywhere and there will be more …"

Examples:

staffware. THE POWER OF PROCESS

IBM

Pallas Athena PROCESS MANAGEMENT

FileNet An IBM Company

océ

ORACLE

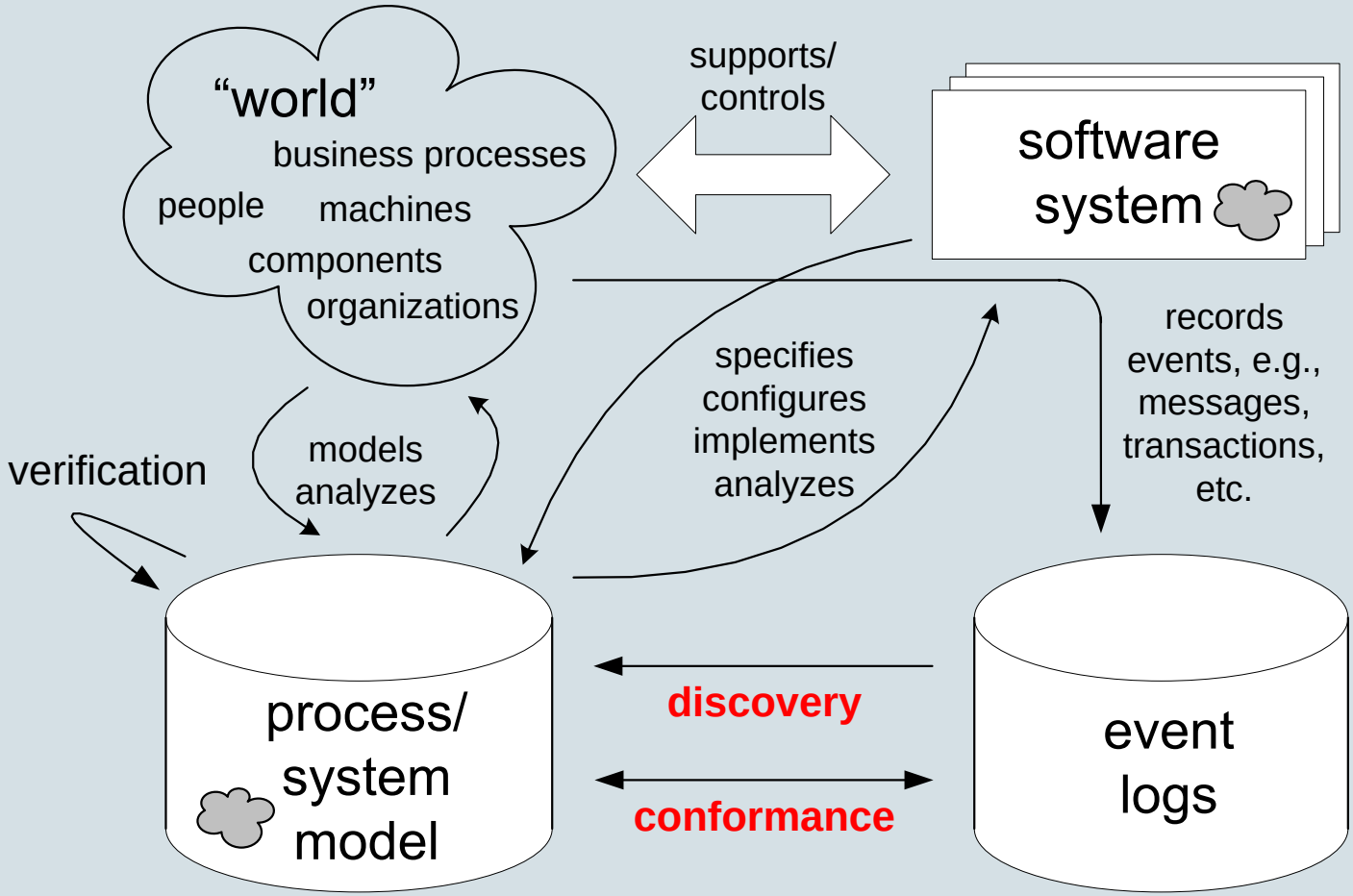
SAP

PHILIPS sense and simplicity

ASML

WebSphere. software

Process mining: Linking events to models



Toy example to explain basic idea:

Reviewing of papers for journal

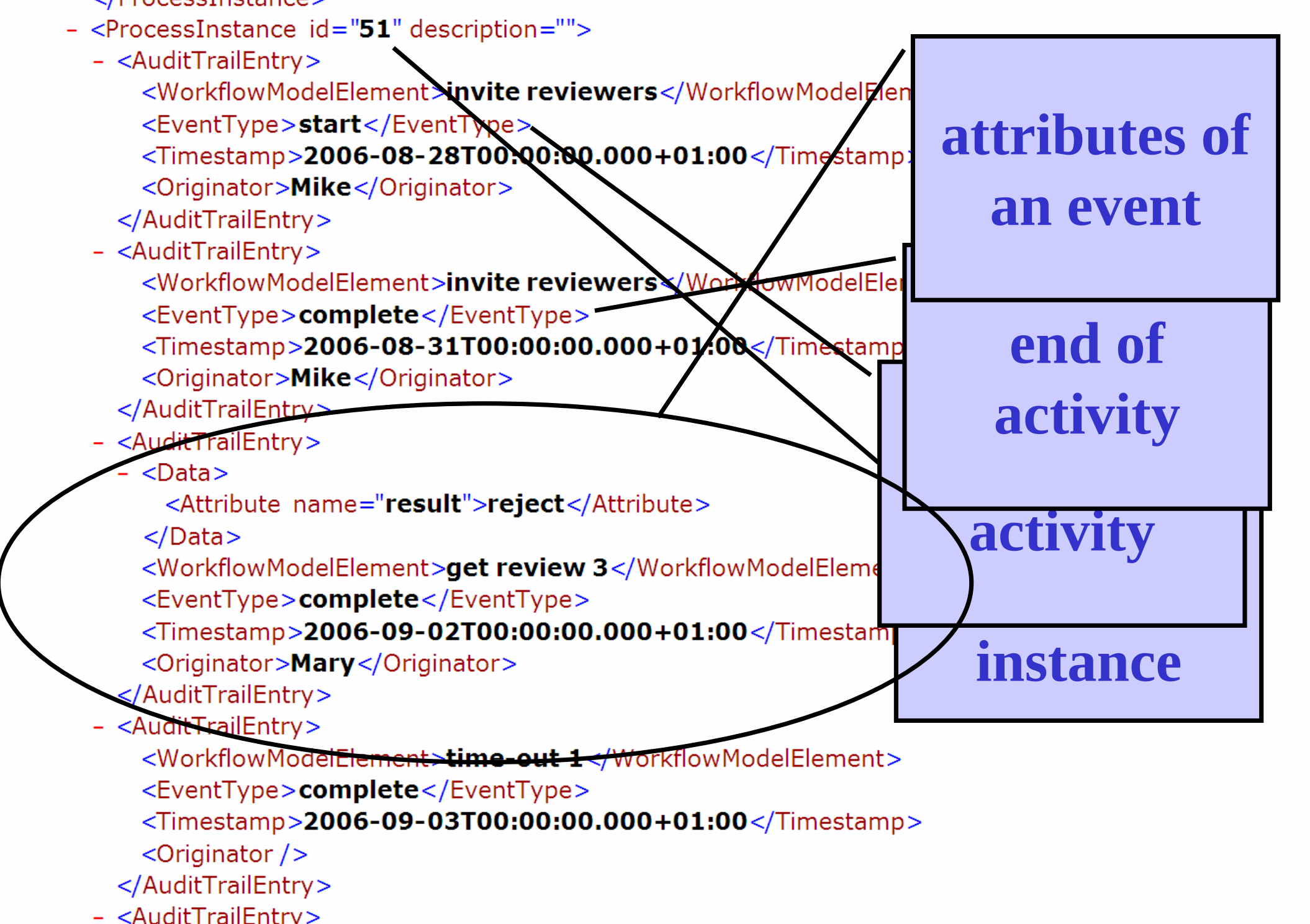
☺

Event log:

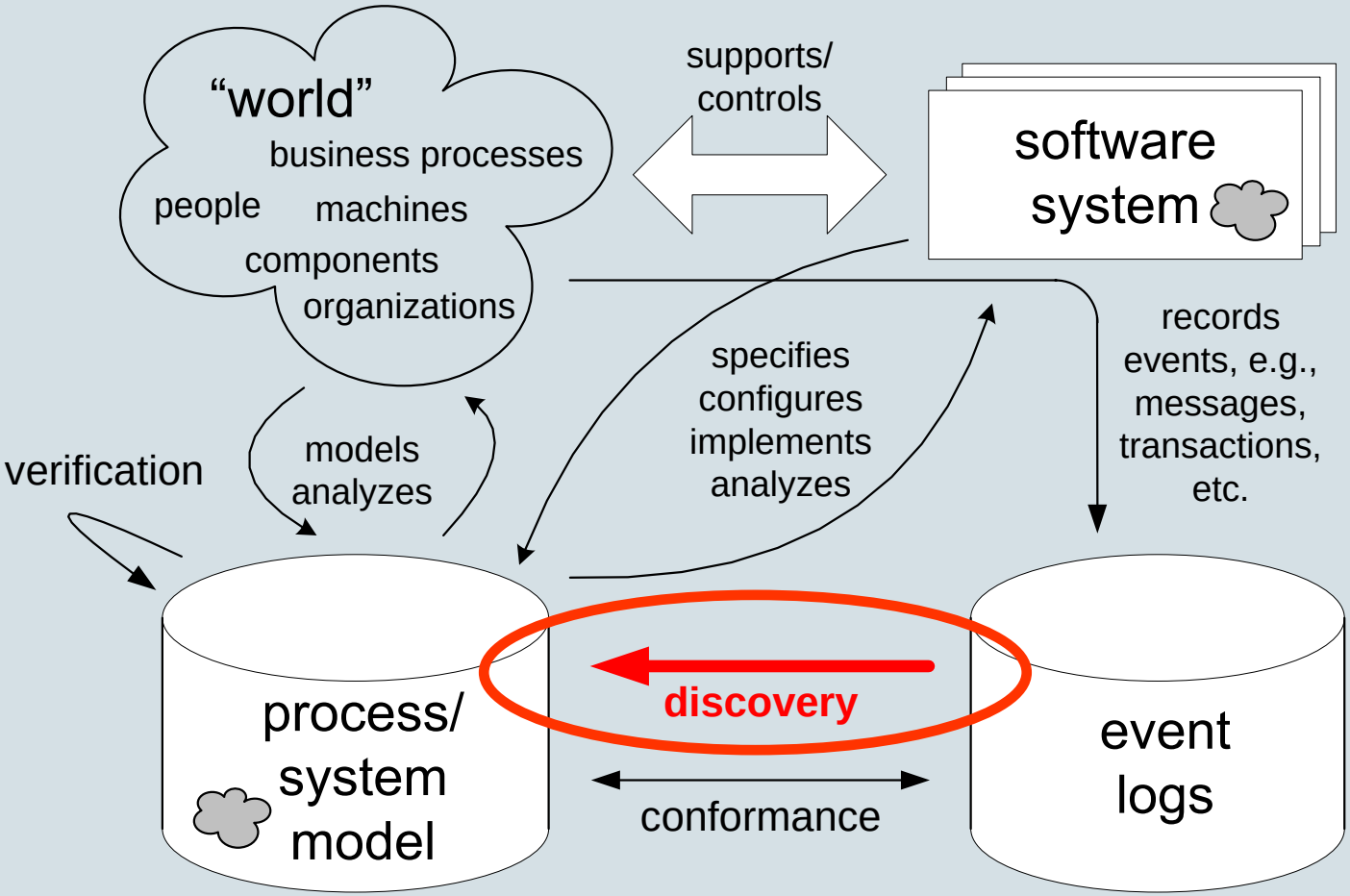
- processes
  - process instances
    - events

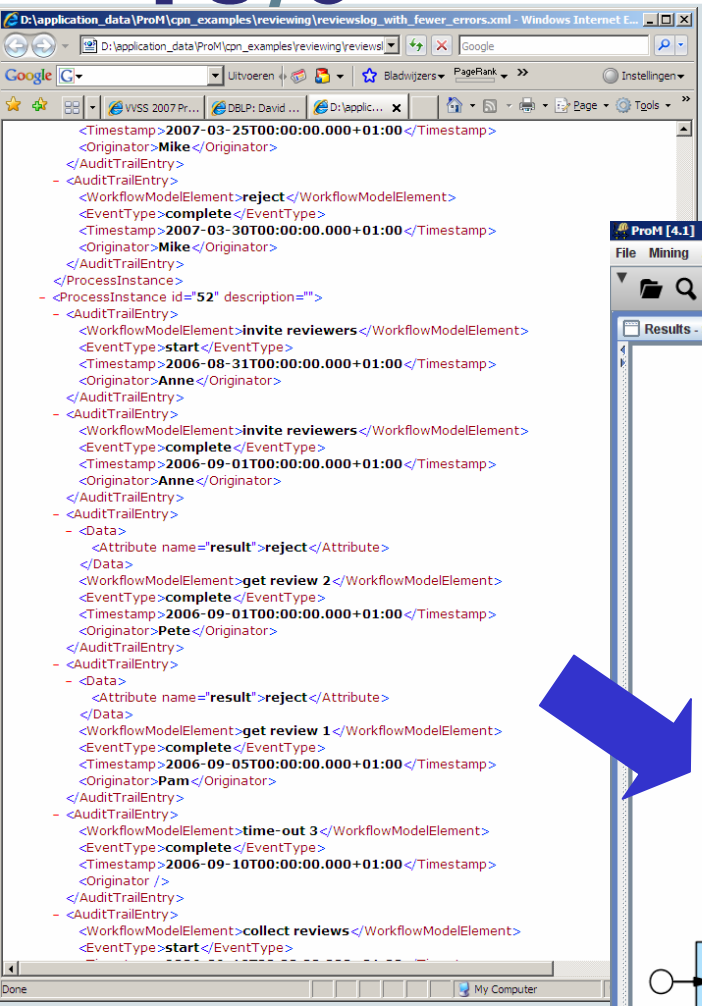
Per event:

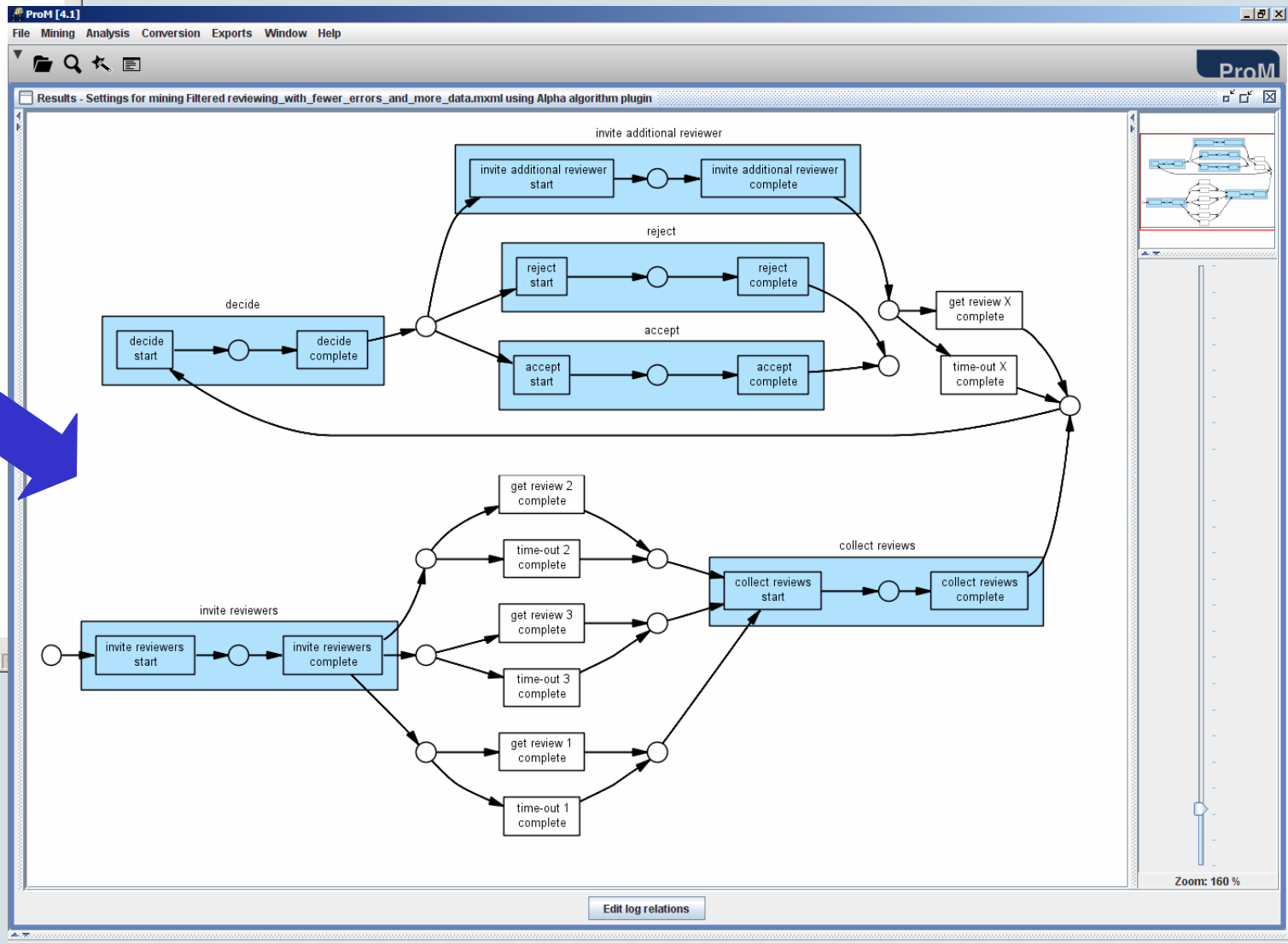
- activity name
- (event type)
- (originator)
- (timestamp)
- (data)


- <ProcessInstance id="51" description="">
- <AuditTrailEntry>
<WorkflowModelElement>invite reviewers</WorkflowModelElem
<EventType>start</EventType>
<Timestamp>2006-08-28T00:00:00.000+01:00</Timestamp>
<Originator>Mike</Originator>
</AuditTrailEntry>
- <AuditTrailEntry>
<WorkflowModelElement>invite reviewers</WorkflowModelEle
<EventType>complete</EventType>
<Timestamp>2006-08-31T00:00:00.000+01:00</Timestamp
<Originator>Mike</Originator>
</AuditTrailEntry>
- <AuditTrailEntry>
- <Data>
<Attribute name="result">reject</Attribute>
</Data>
<WorkflowModelElement>get review 3</WorkflowModelEleme
<EventType>complete</EventType>
<Timestamp>2006-09-02T00:00:00.000+01:00</Timestam
<Originator>Mary</Originator>
</AuditTrailEntry>
- <AuditTrailEntry>
<WorkflowModelElement>time-out 1</WorkflowModelElement>
<EventType>complete</EventType>
<Timestamp>2006-09-03T00:00:00.000+01:00</Timestamp>
<Originator />
</AuditTrailEntry>
- <AuditTrailEntry>
attributes of an event
end of activity
activity
instance

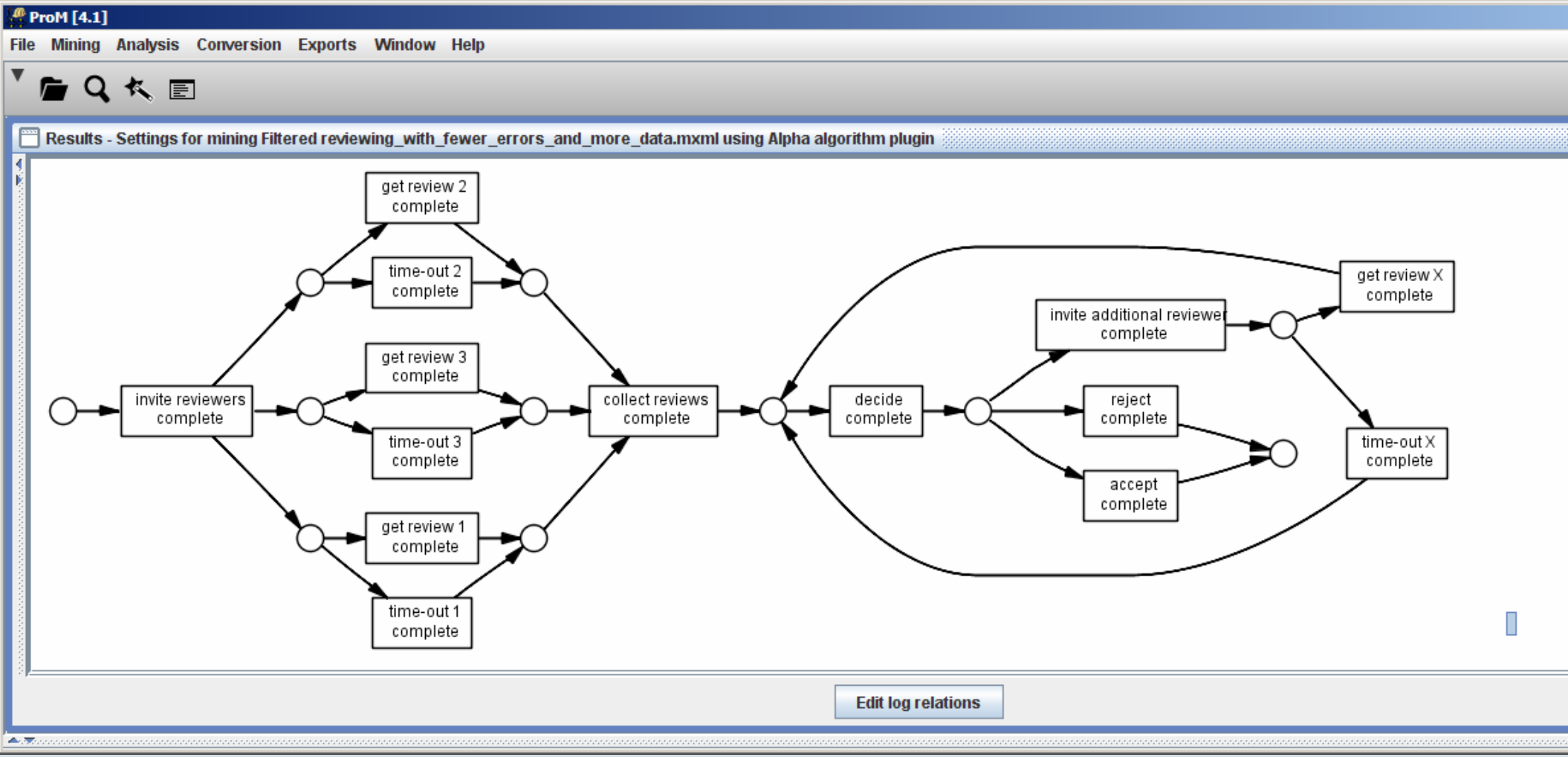
Discovery

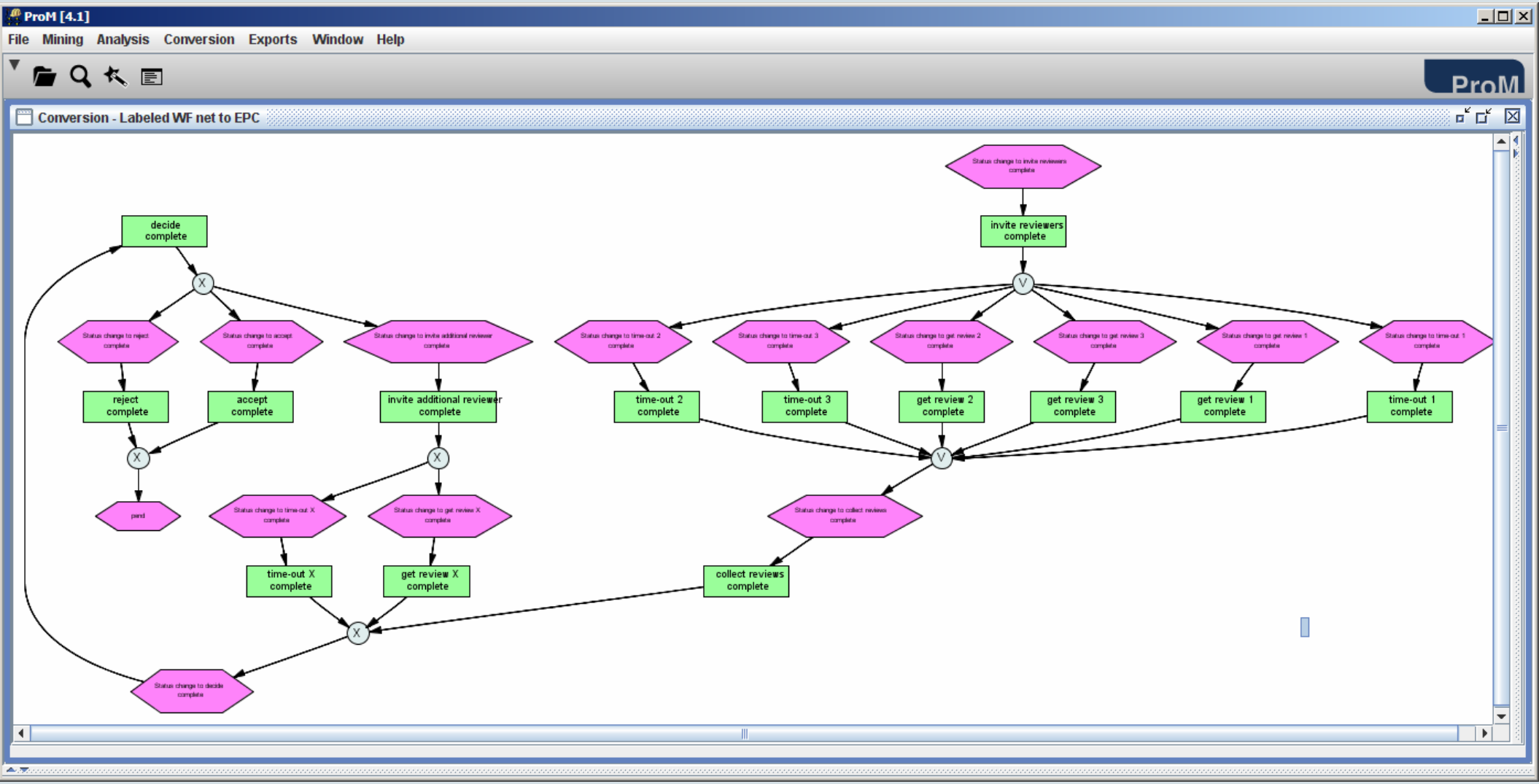

"world"
business processes
people
machines
components
organizations
supports/
controls
software
system
records
events, e.g.,
messages,
transactions,
etc.
specifies
configures
implements
analyzes
models
analyzes
verification
process/
system
model
discovery
conformance
event
logs


D:\application_data\ProM\cpn_examples\reviewing\reviewslog_with_fewer_errors.xml - Windows Internet E...
<Timestamp>2007-03-25T00:00:00.000+01:00</Timestamp>
<Originator>Mike</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>reject</WorkflowModelElement>
<EventType>complete</EventType>
<Timestamp>2007-03-30T00:00:00.000+01:00</Timestamp>
<Originator>Mike</Originator>
</AuditTrailEntry>
</ProcessInstance>
<ProcessInstance id="52" description="">
<AuditTrailEntry>
<WorkflowModelElement>invite reviewers</WorkflowModelElement>
<EventType>start</EventType>
<Timestamp>2006-08-31T00:00:00.000+01:00</Timestamp>
<Originator>Anne</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>invite reviewers</WorkflowModelElement>
<EventType>complete</EventType>
<Timestamp>2006-09-01T00:00:00.000+01:00</Timestamp>
<Originator>Anne</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<Data>
<Attribute name="result">reject</Attribute>
</Data>
<WorkflowModelElement>get review 2</WorkflowModelElement>
<EventType>complete</EventType>
<Timestamp>2006-09-01T00:00:00.000+01:00</Timestamp>
<Originator>Pete</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<Data>
<Attribute name="result">reject</Attribute>
</Data>
<WorkflowModelElement>get review 1</WorkflowModelElement>
<EventType>complete</EventType>
<Timestamp>2006-09-05T00:00:00.000+01:00</Timestamp>
<Originator>Pam</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>time-out 3</WorkflowModelElement>
<EventType>complete</EventType>
<Timestamp>2006-09-10T00:00:00.000+01:00</Timestamp>
<Originator />
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>collect reviews</WorkflowModelElement>
<EventType>start</EventType>
Done
My Computer


ProM [4.1]
File Mining Analysis Conversion Exports Window Help
ProM
Results - Settings for mining Filtered reviewing_with_fewer_errors_and_more_data.mxml using Alpha algorithm plugin
invite additional reviewer
invite additional reviewer start
invite additional reviewer complete
reject
reject start
reject complete
decide
decide start
decide complete
accept
accept start
accept complete
get review X complete
time-out X complete
get review 2 complete
time-out 2 complete
get review 3 complete
time-out 3 complete
get review 1 complete
time-out 1 complete
collect reviews
collect reviews start
collect reviews complete
invite reviewers
invite reviewers start
invite reviewers complete
Zoom: 160 %
Edit log relations

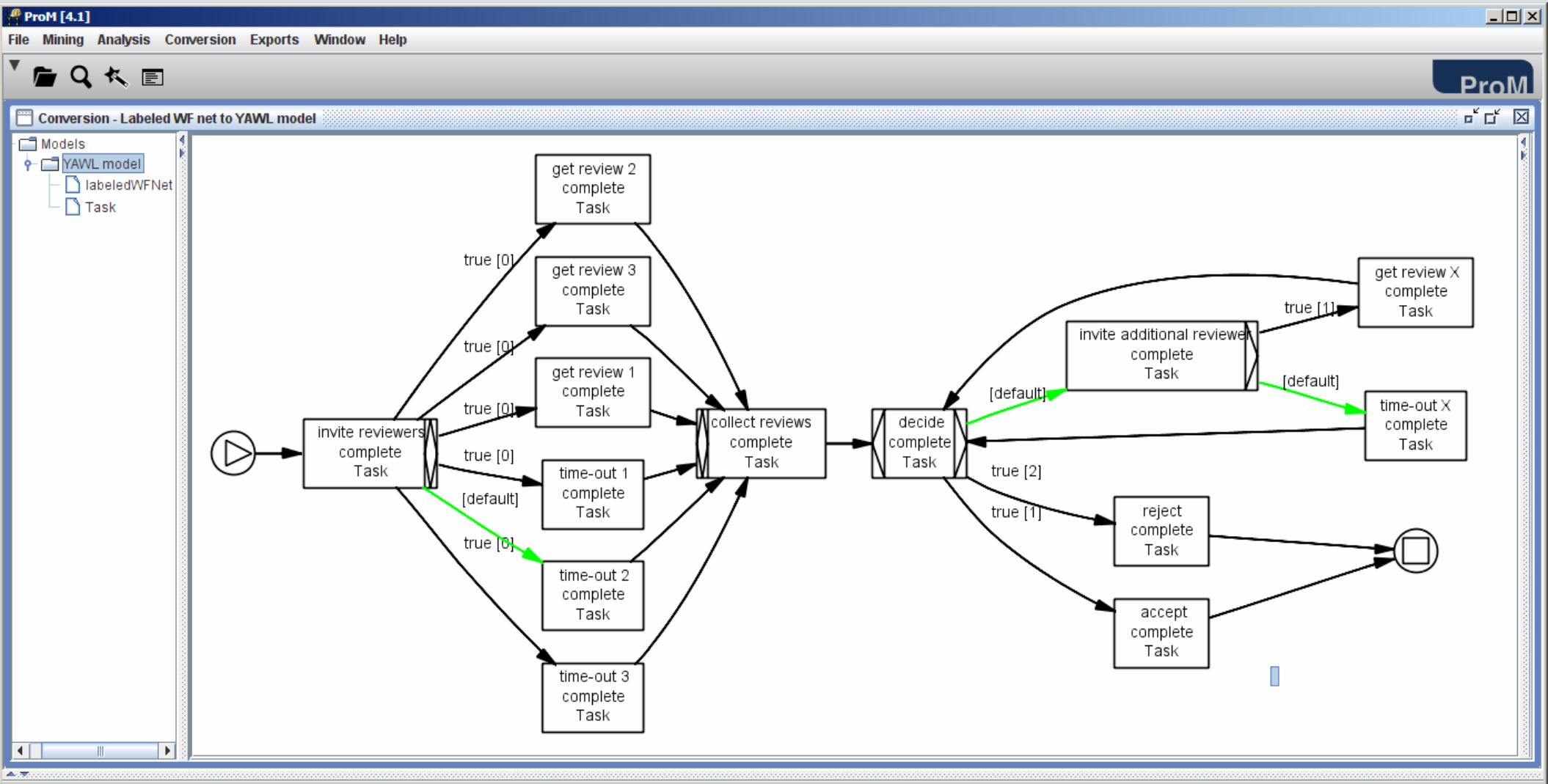
No transactional information





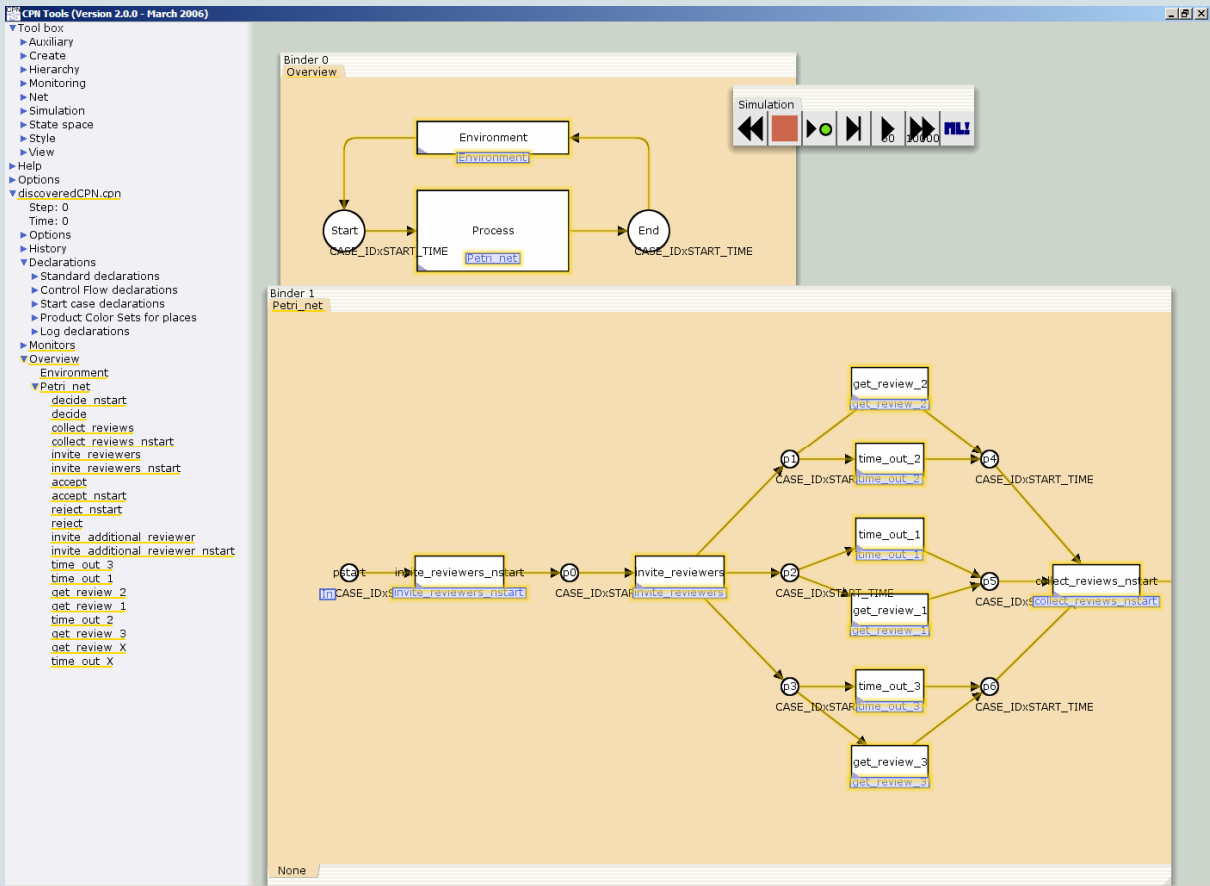
Corresponding EPC model (used by SAP,ARIS, etc)

YAWL model (executable workflow model)



Conversions/exports/imports

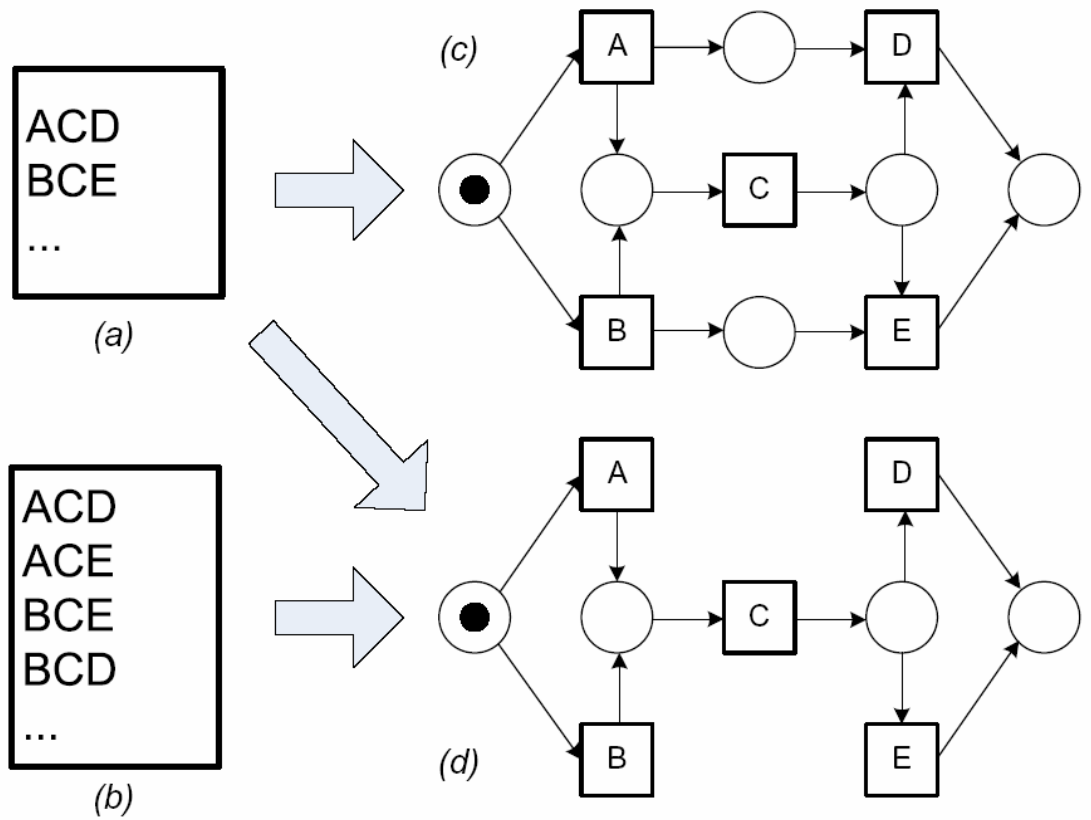
- ARIS – ARIS PPM
- BPEL 1.1 (WebSphere/Oracle)
- YAWL
- CPN Tools
- Petrify
- Woflan
- Heuristics nets
- …



Theory of Regions !?

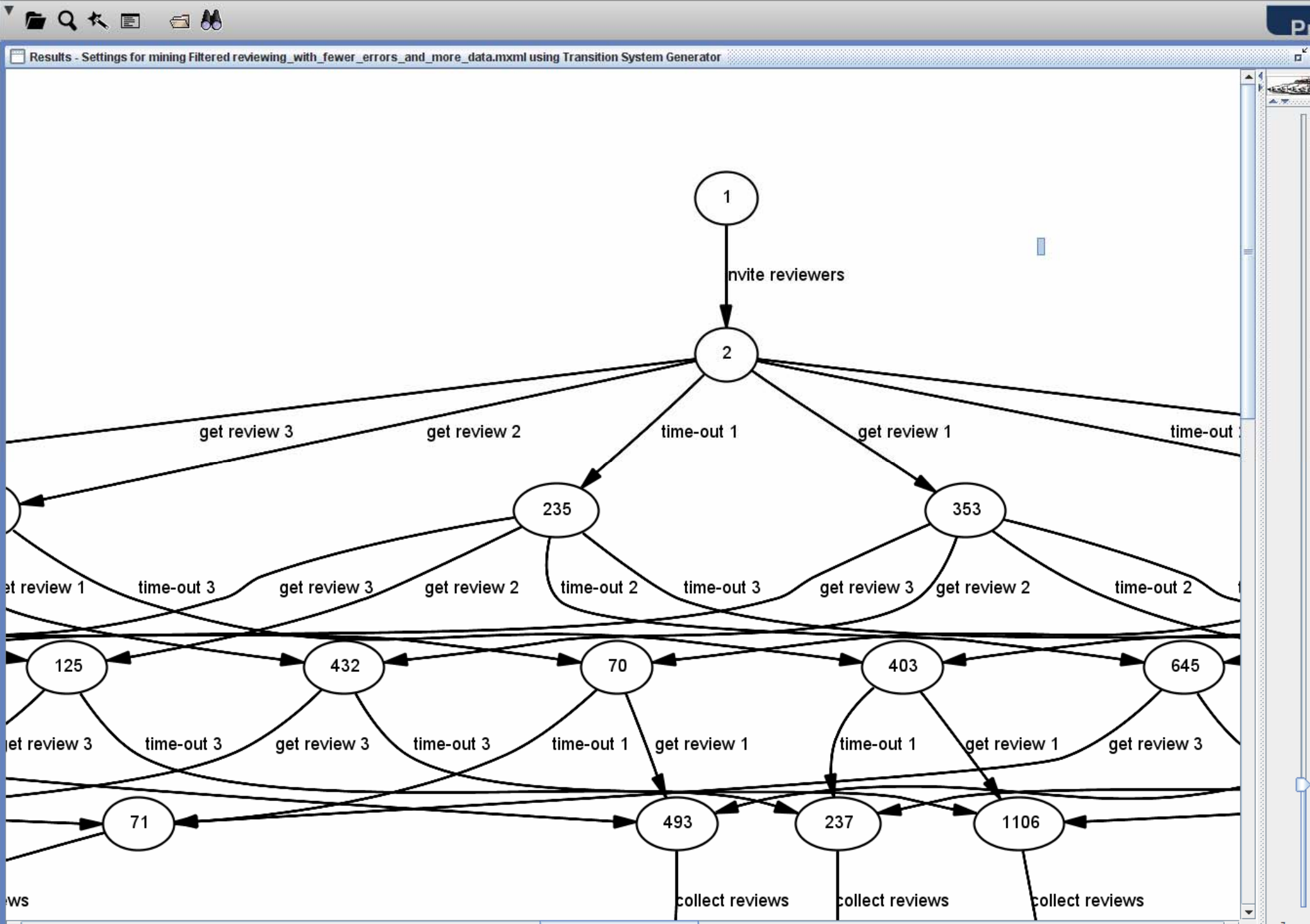
Differences with synthesis:

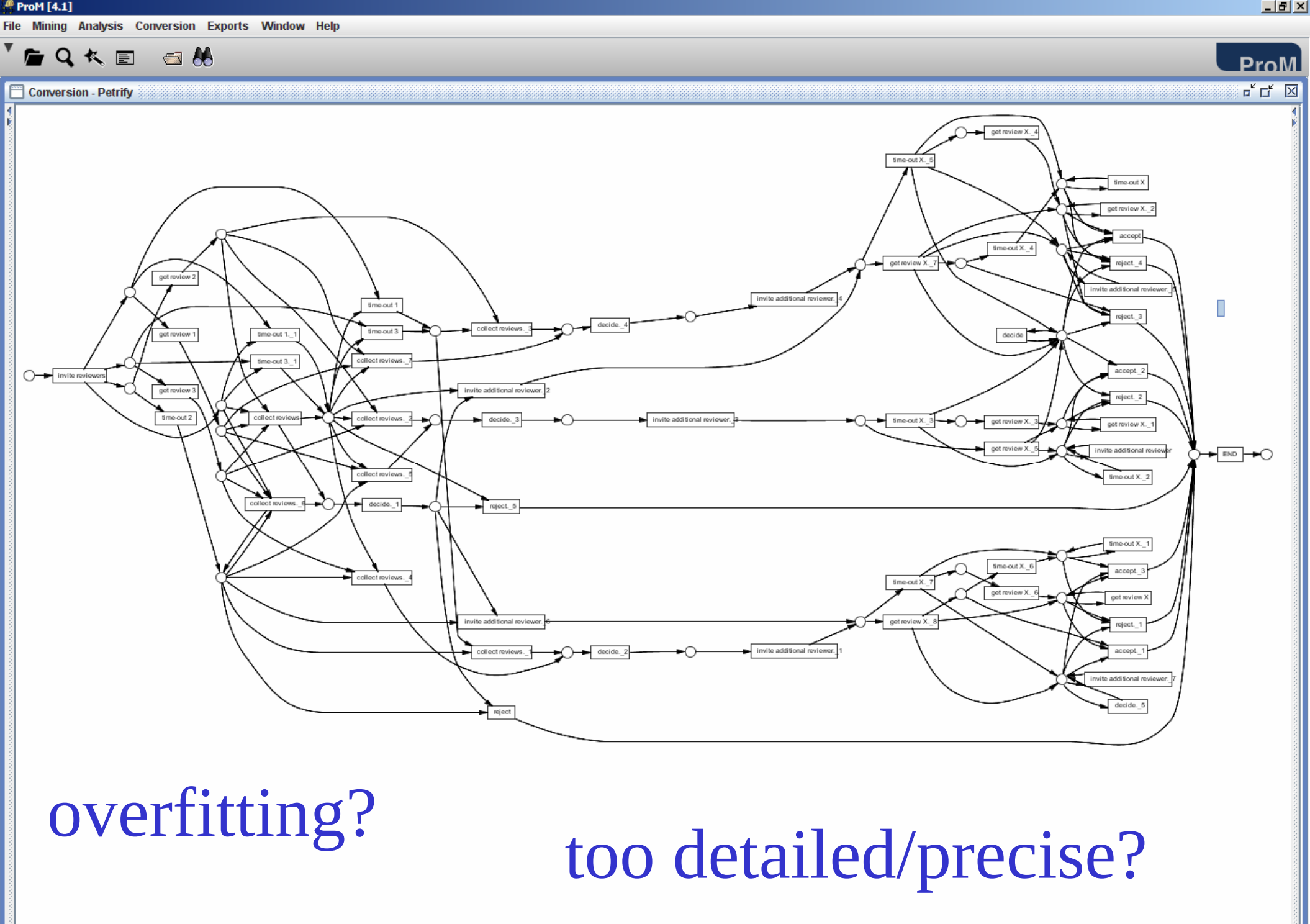
- **aim** (exact/executable model – insight)
- **input** (complete/perfect information – partial/noisy information)



Two steps:

1. log2transition-system
2. transition-system2Petri-net

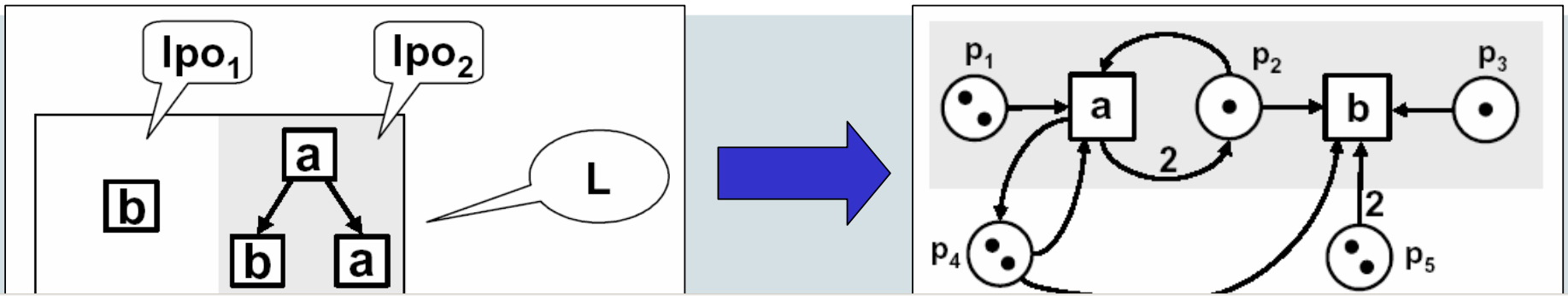

Results - Settings for mining Filtered reviewing_with_fewer_errors_and_more_data.mxml using Transition System Generator
1
nvite reviewers
2
get review 3
get review 2
time-out 1
get review 1
time-out
235
353
t review 1
time-out 3
get review 3
get review 2
time-out 2
time-out 3
get review 3
get review 2
time-out 2
125
432
70
403
645
et review 3
time-out 3
get review 3
time-out 3
time-out 1
get review 1
time-out 1
get review 1
get review 3
71
493
237
1106
ws
collect reviews
collect reviews
collect reviews

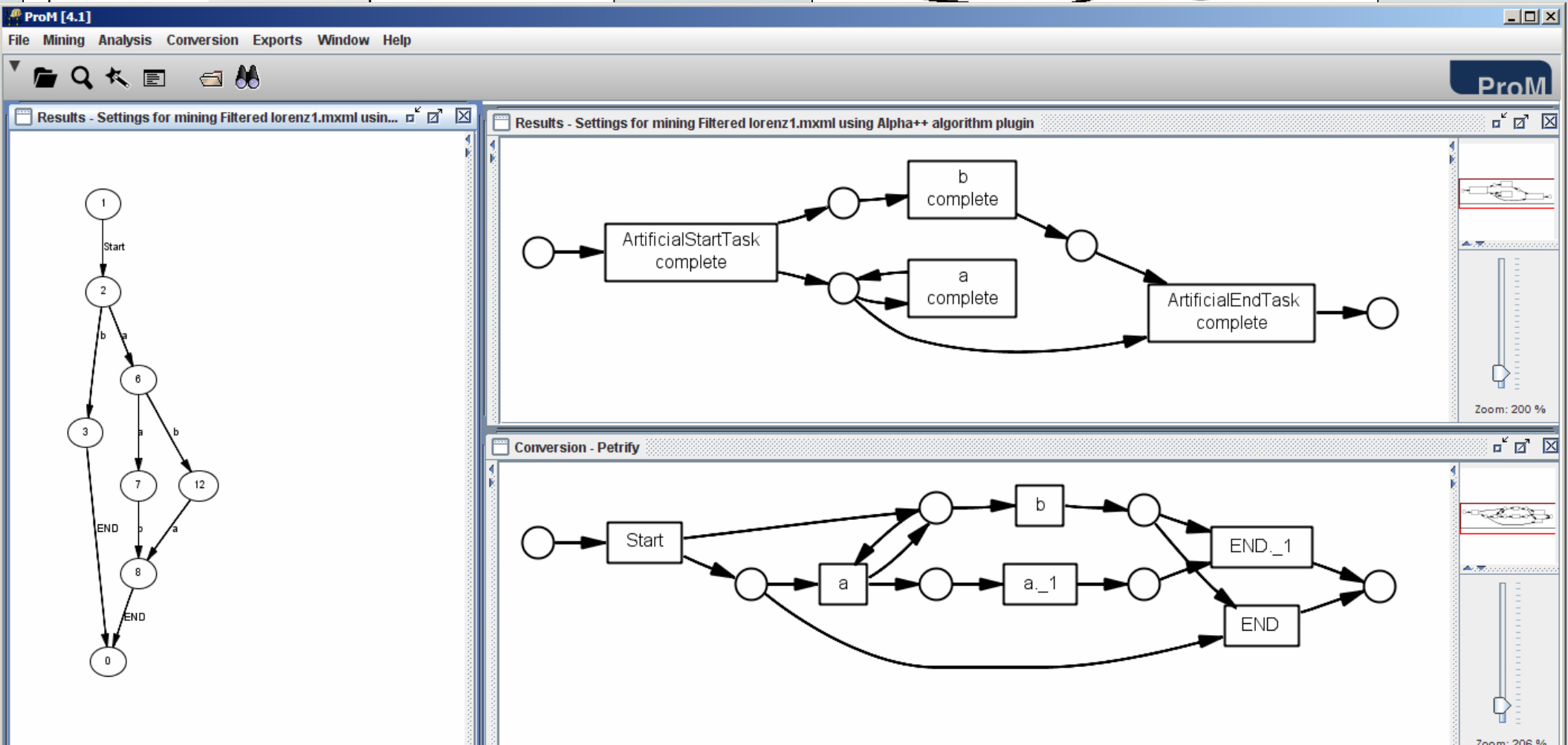


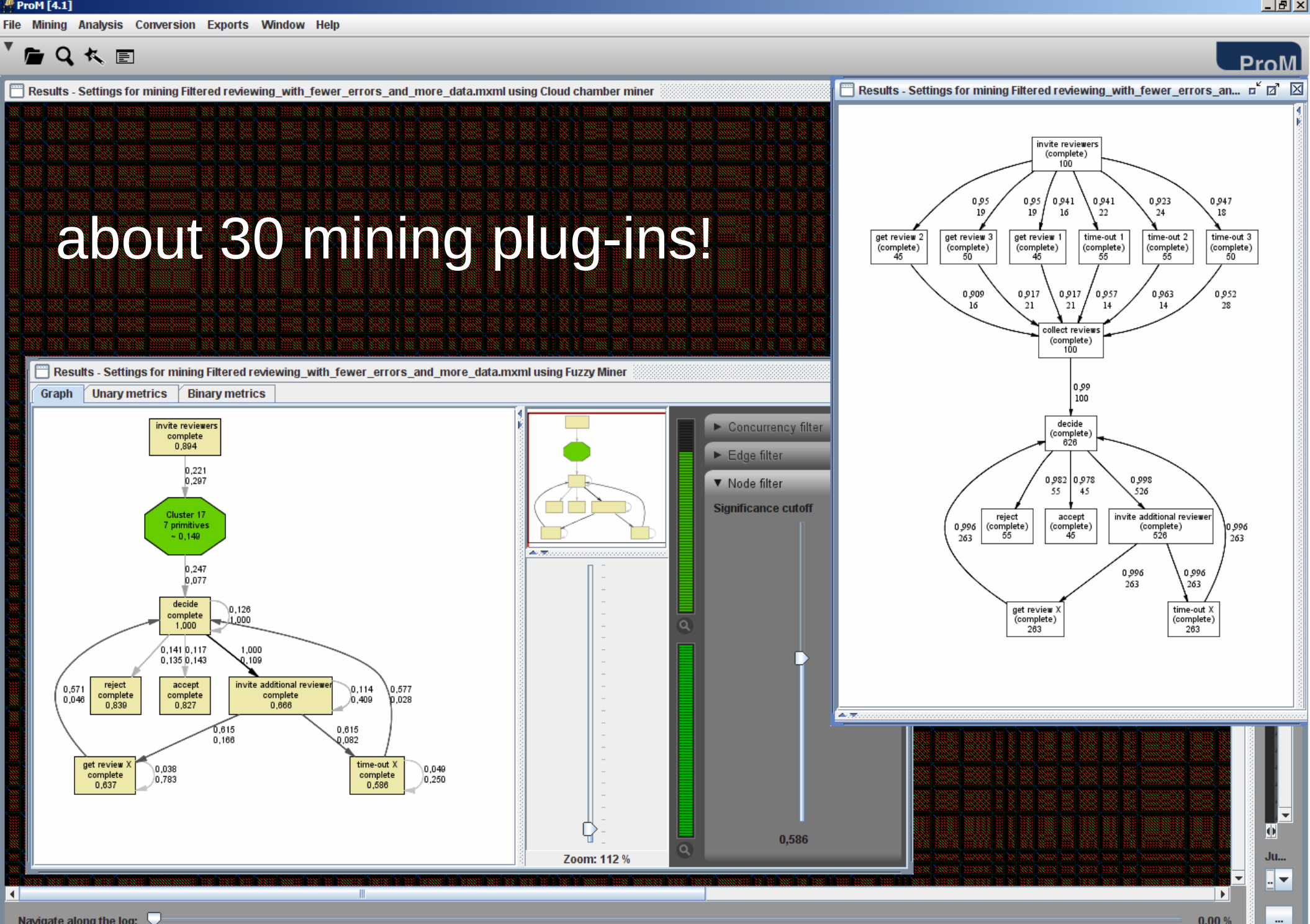
overfitting?

too detailed/precise?

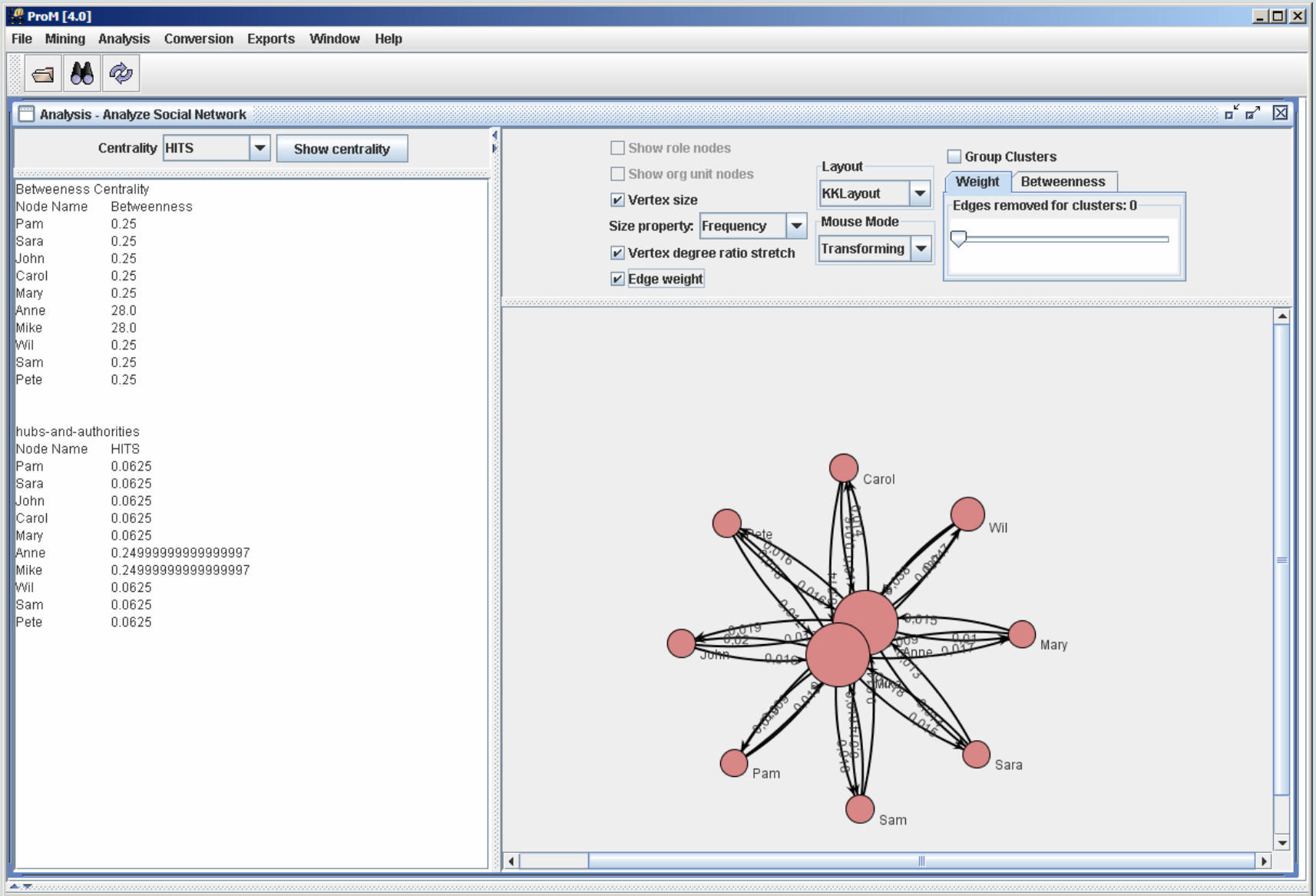
Example "Eichstätt paper"





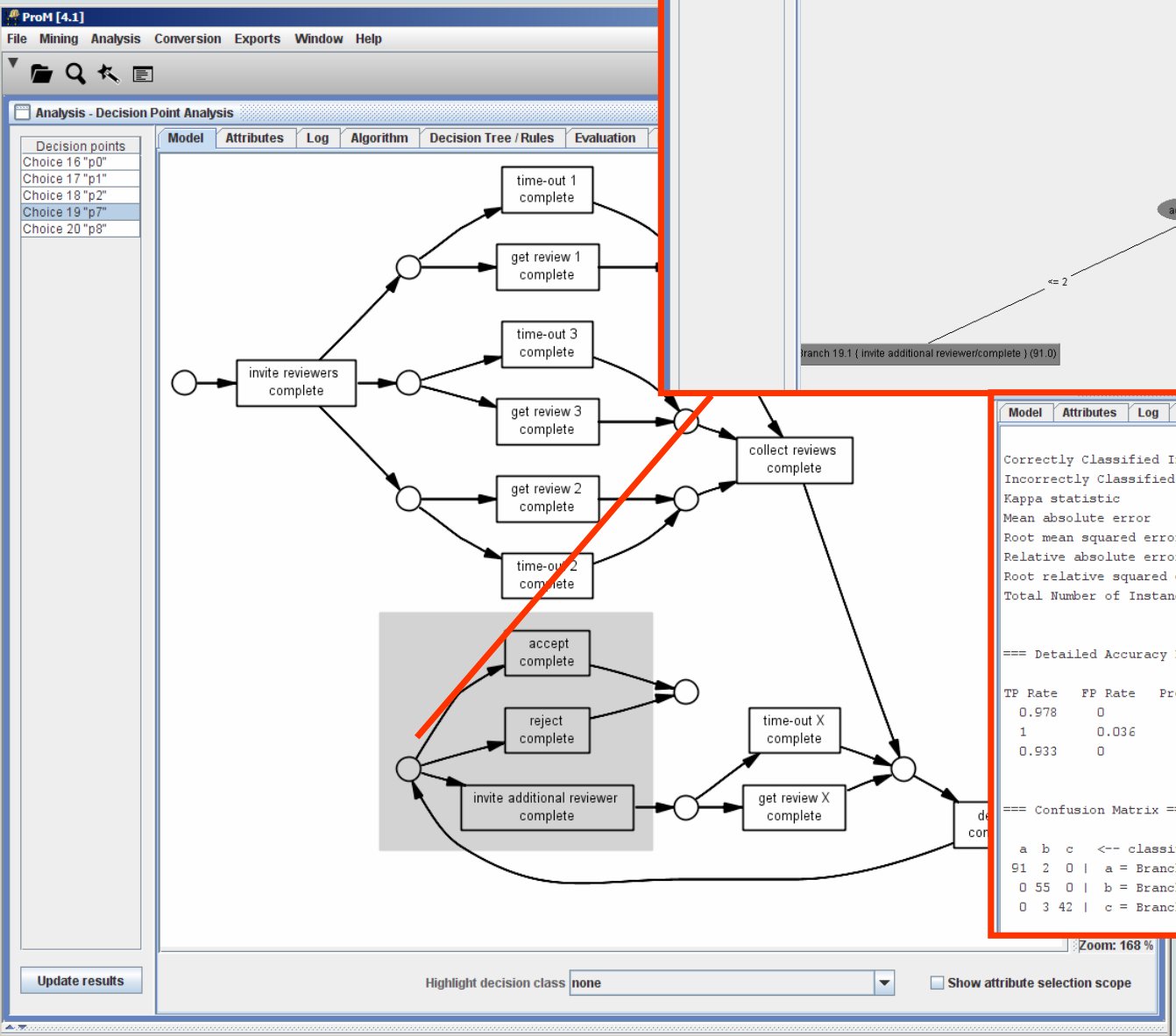

ProM [4.1]
File Mining Analysis Conversion Exports Window Help
Results - Settings for mining Filtered reviewing_with_fewer_errors_and_more_data.mxml using Cloud chamber miner
about 30 mining plug-ins!
Results - Settings for mining Filtered reviewing_with_fewer_errors_an...
invite reviewers (complete) 100
get review 2 (complete) 45
get review 3 (complete) 50
get review 1 (complete) 46
time-out 1 (complete) 55
time-out 2 (complete) 55
time-out 3 (complete) 50
collect reviews (complete) 100
decide (complete) 626
reject (complete) 55
accept (complete) 46
invite additional reviewer (complete) 526
get review X (complete) 263
time-out X (complete) 263
Results - Settings for mining Filtered reviewing_with_fewer_errors_and_more_data.mxml using Fuzzy Miner
Graph
Unary metrics
Binary metrics
invite reviewers complete 0,894
Cluster 17 7 primitives ~ 0,149
decide complete 1,000
reject complete 0,839
accept complete 0,827
invite additional reviewer complete 0,666
get review X complete 0,637
time-out X complete 0,586
Concurrency filter
Edge filter
Node filter
Significance cutoff
0,586
Zoom: 112 %
Navigate along the log:

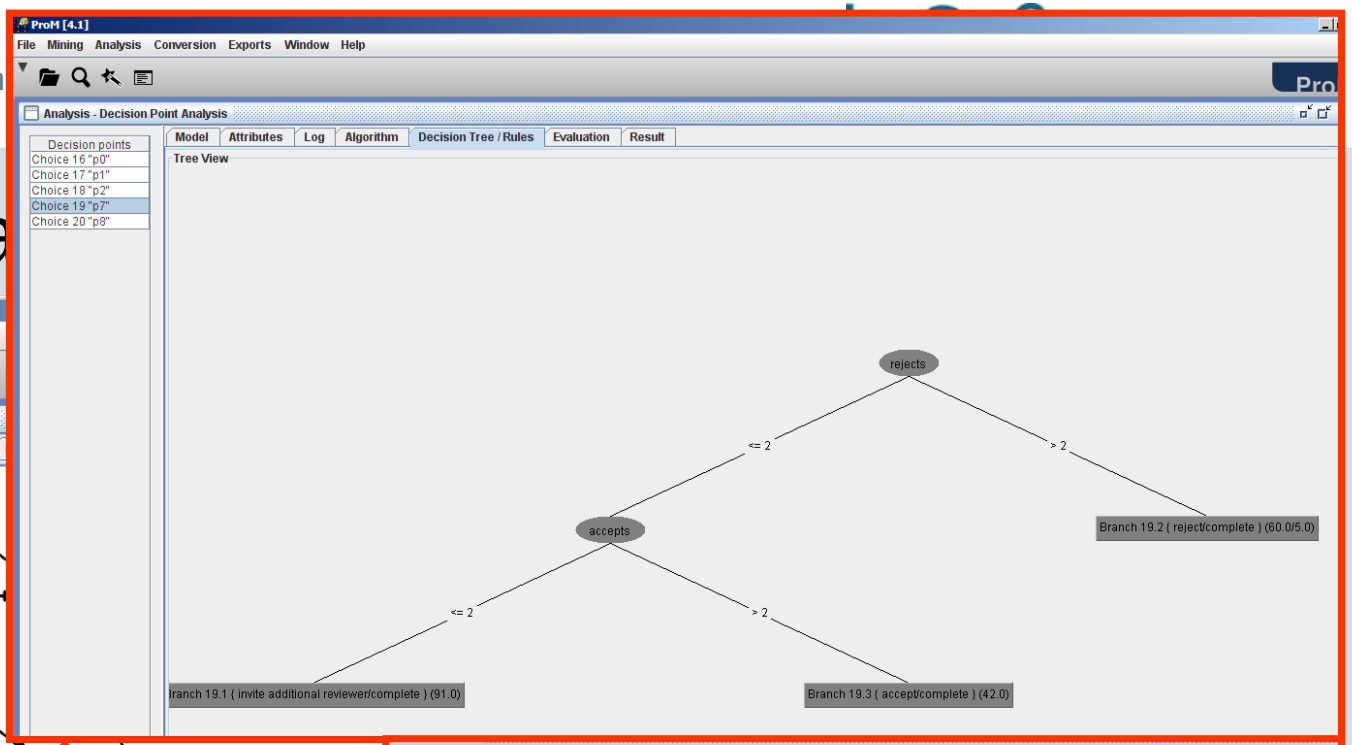
Social network analysis

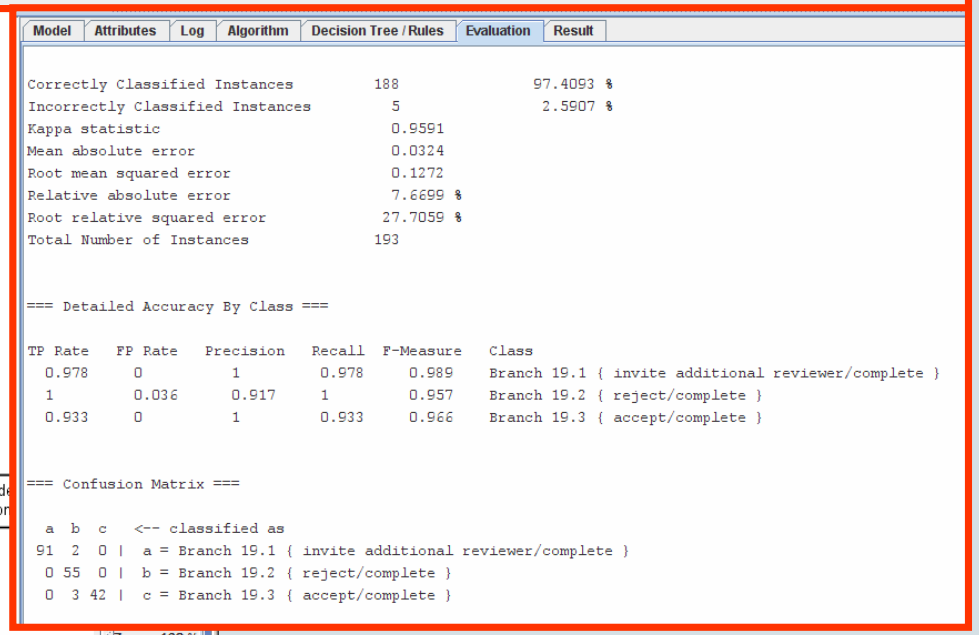


TU/e technische universiteit ein

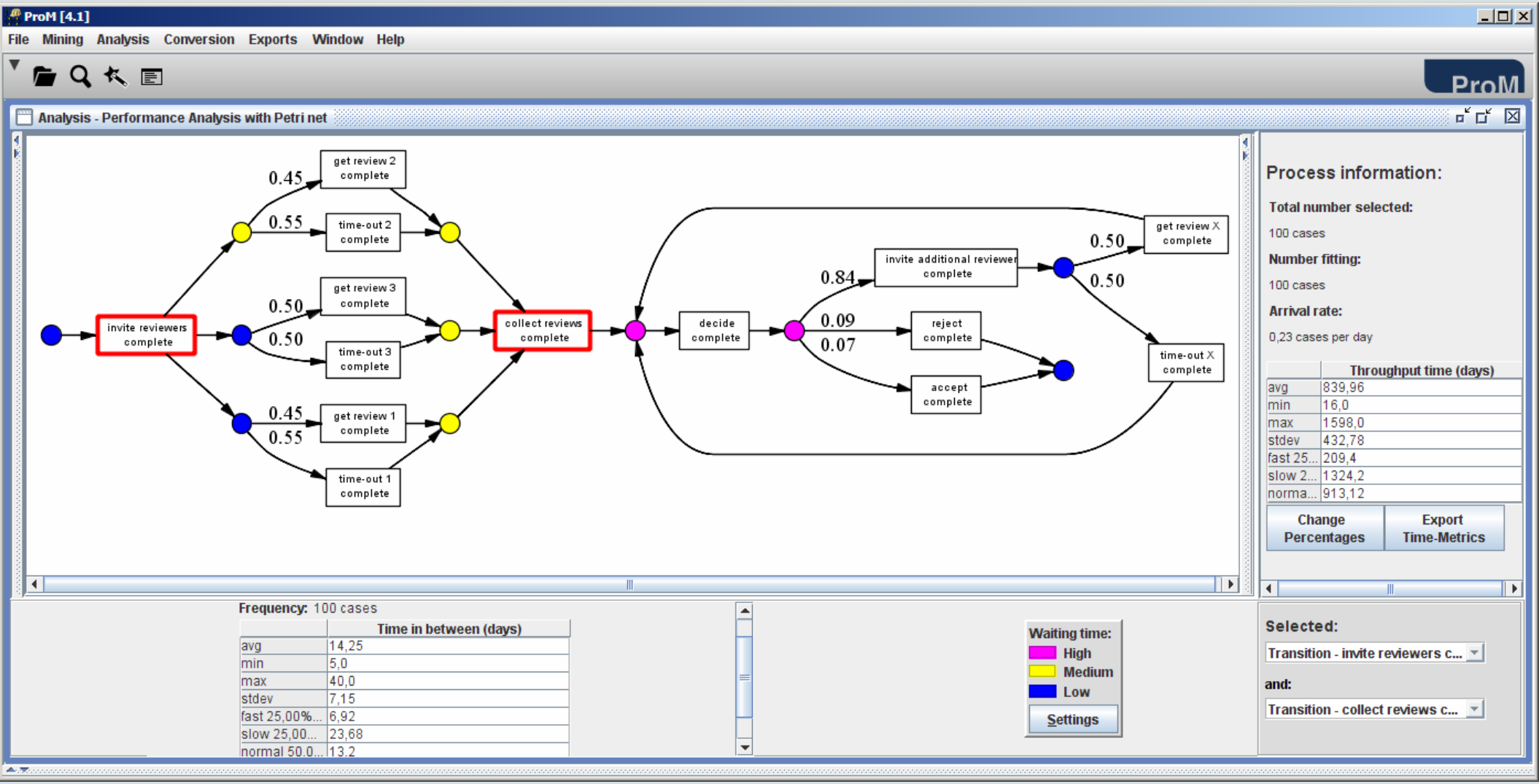
Decision point ana





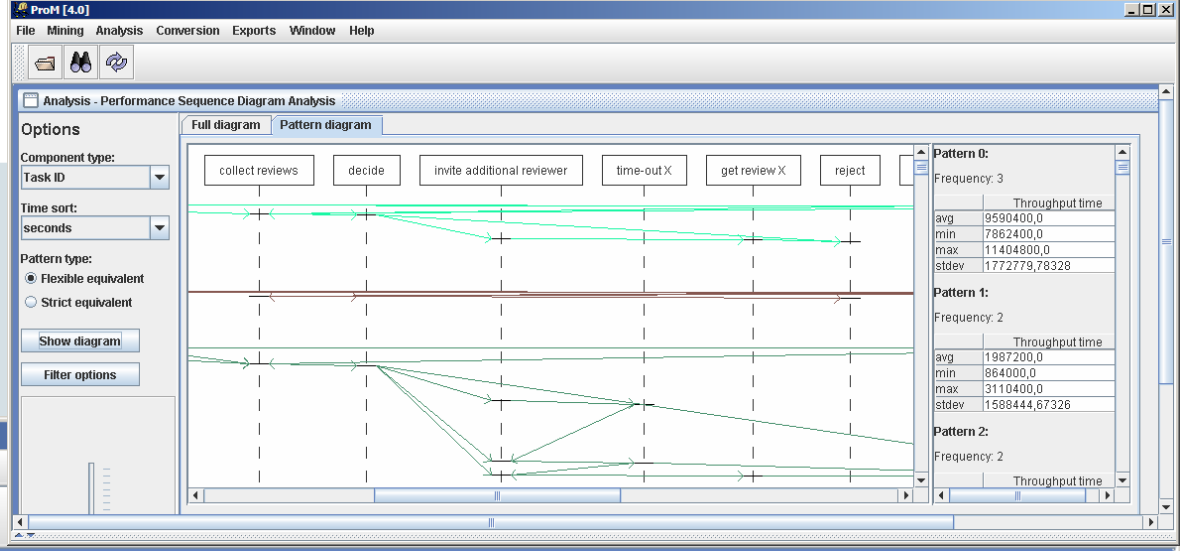


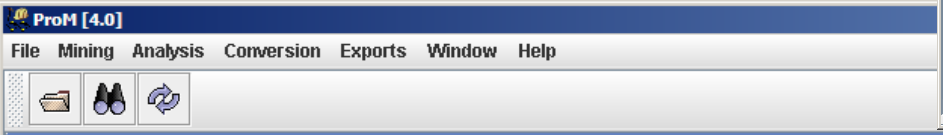
Performance analysis

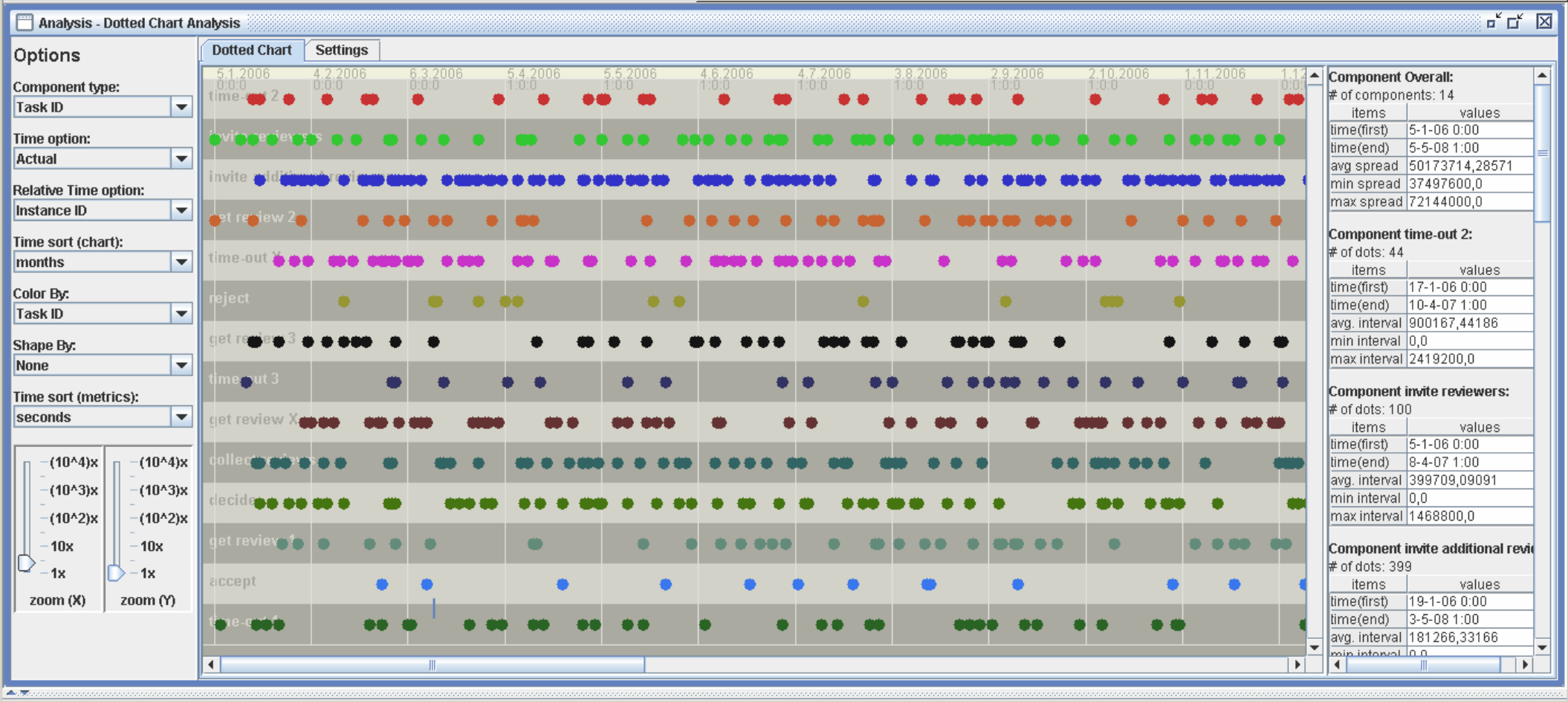


TU/e technische universiteit eindhoven

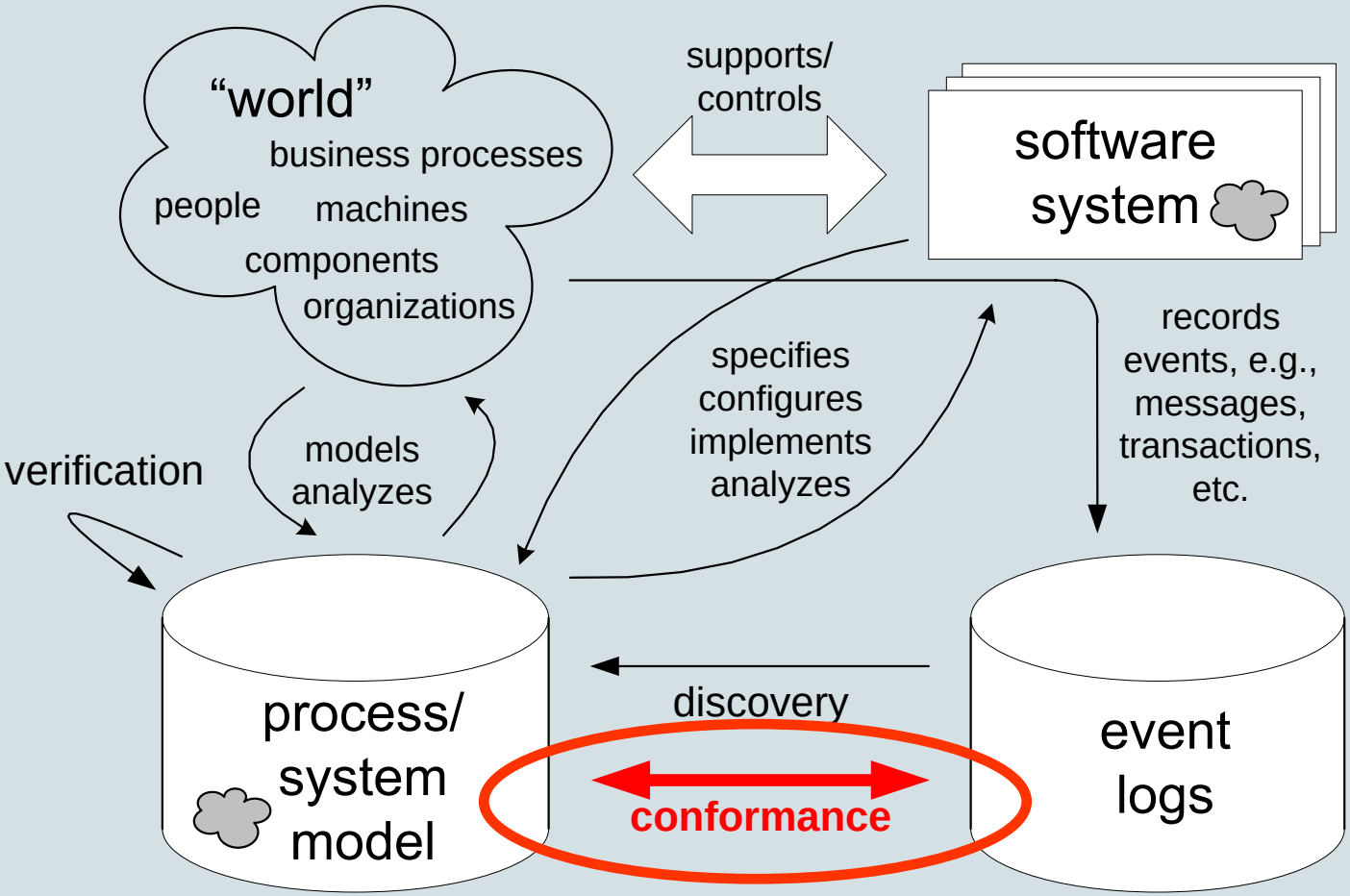
Discovering patterns



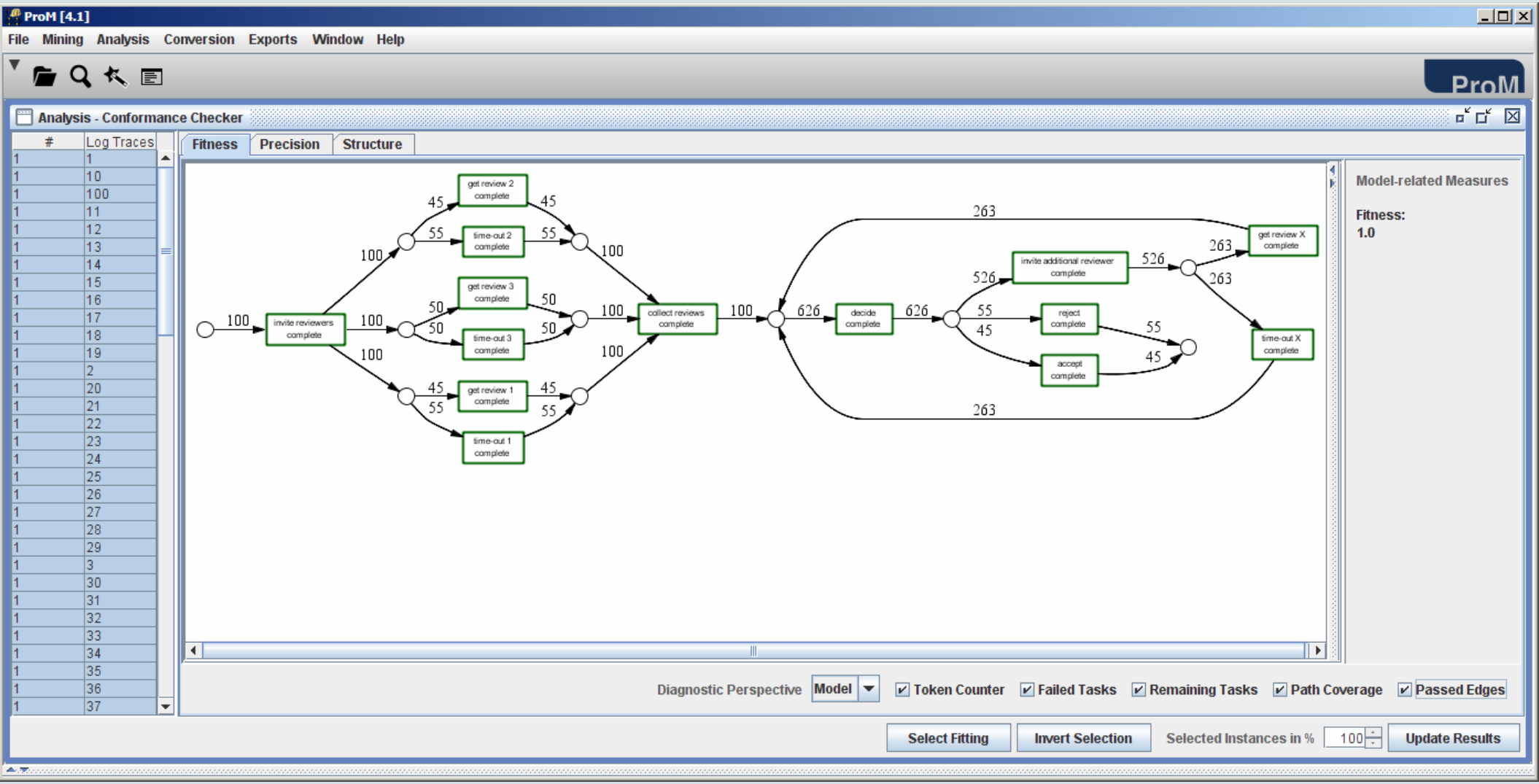




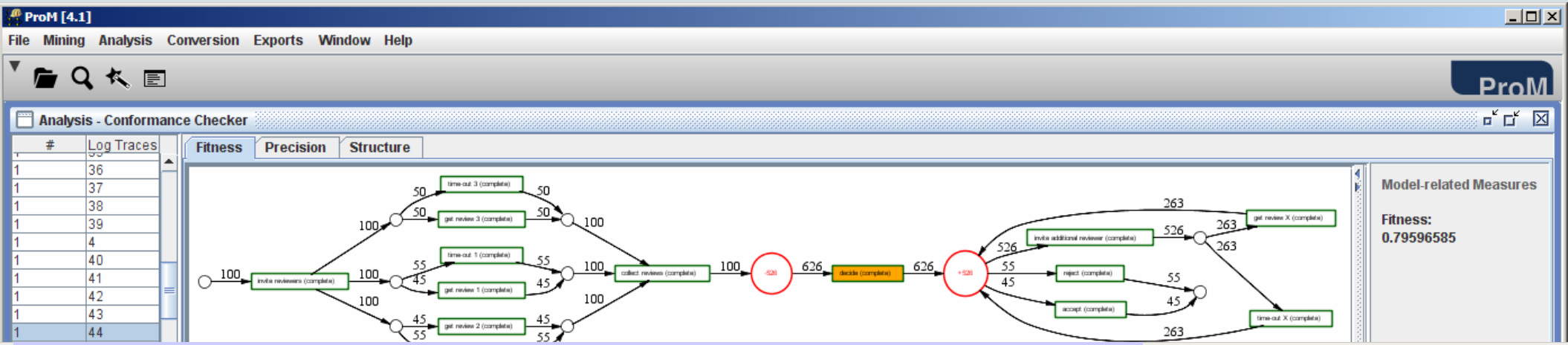
Conformance Checking



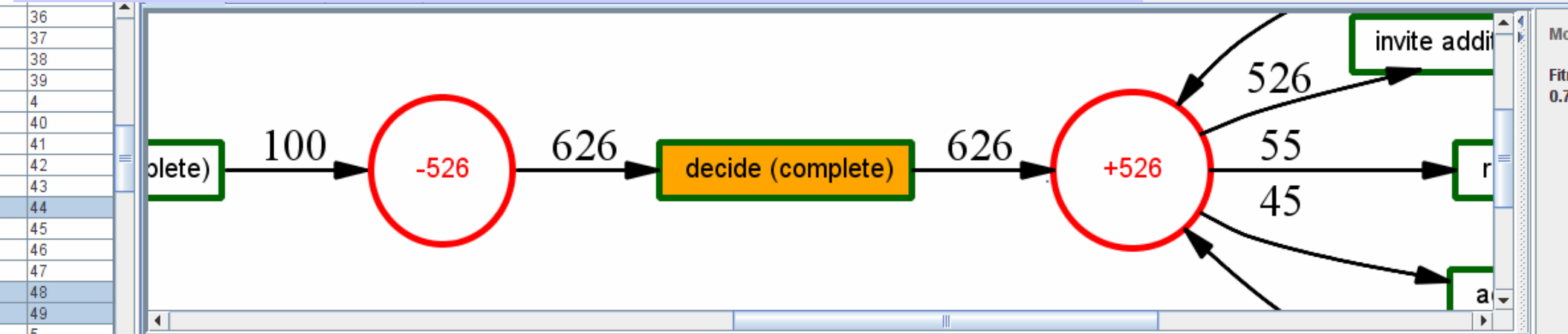
Comparing the discovered model with the log (f=1)



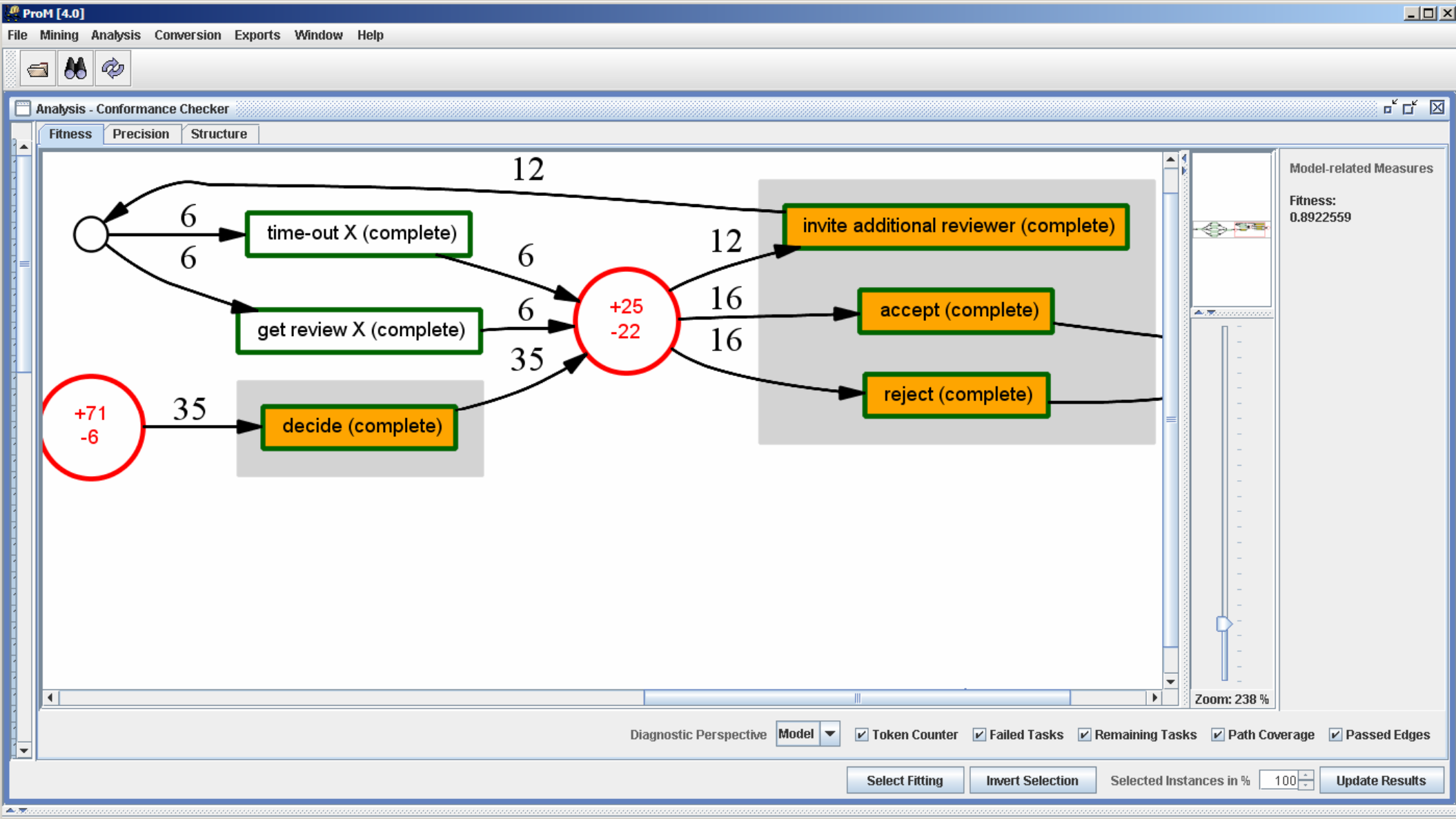
Different process model, same log (f=0.796)



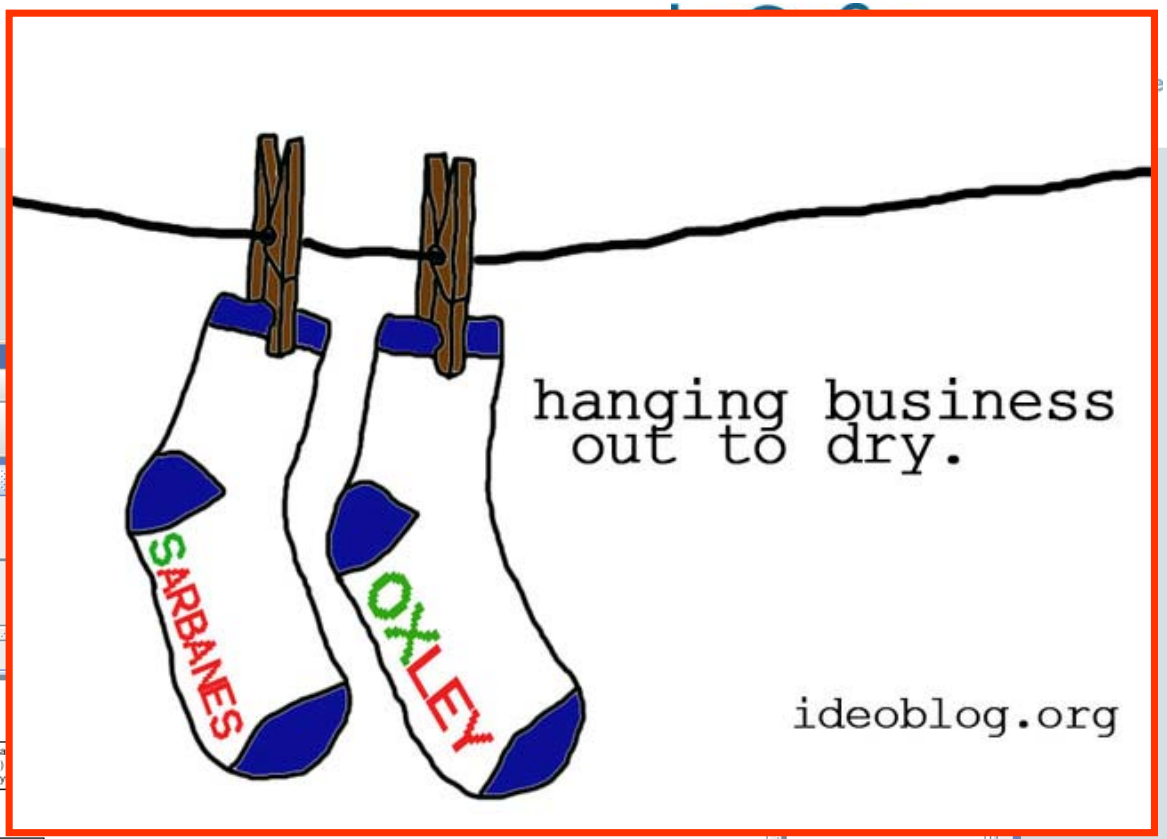
Decision cannot repeated according to model but can be repeated in reality!

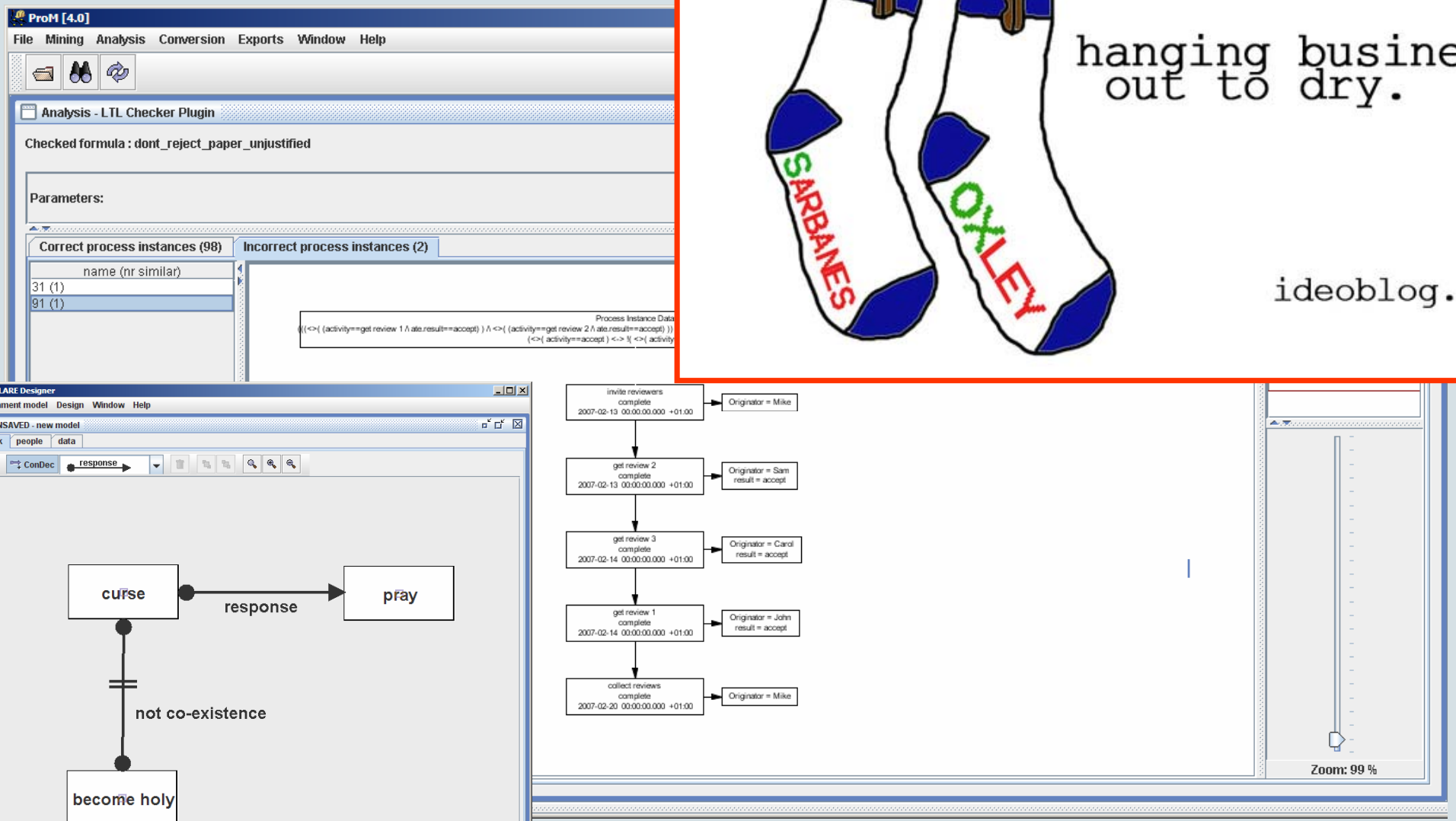


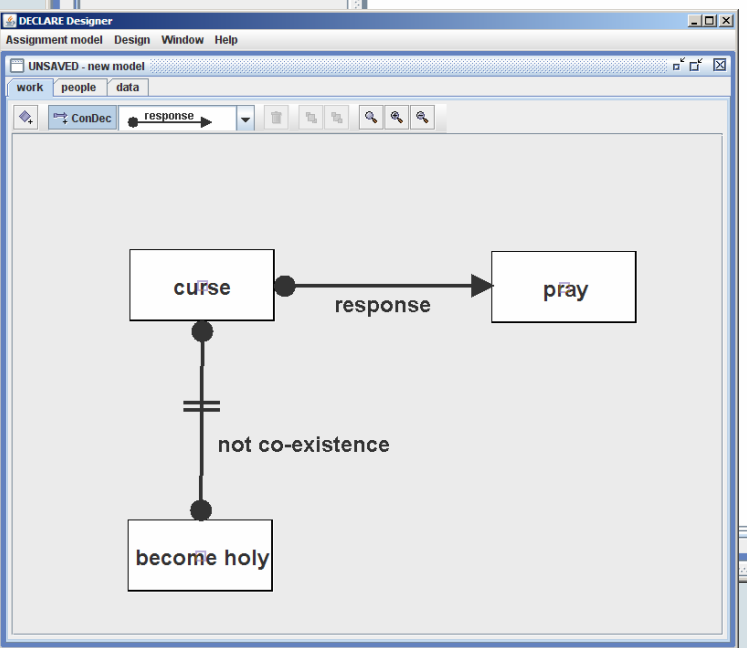
Adding deviations to the log (f=0.89)



LTL checker plug-in



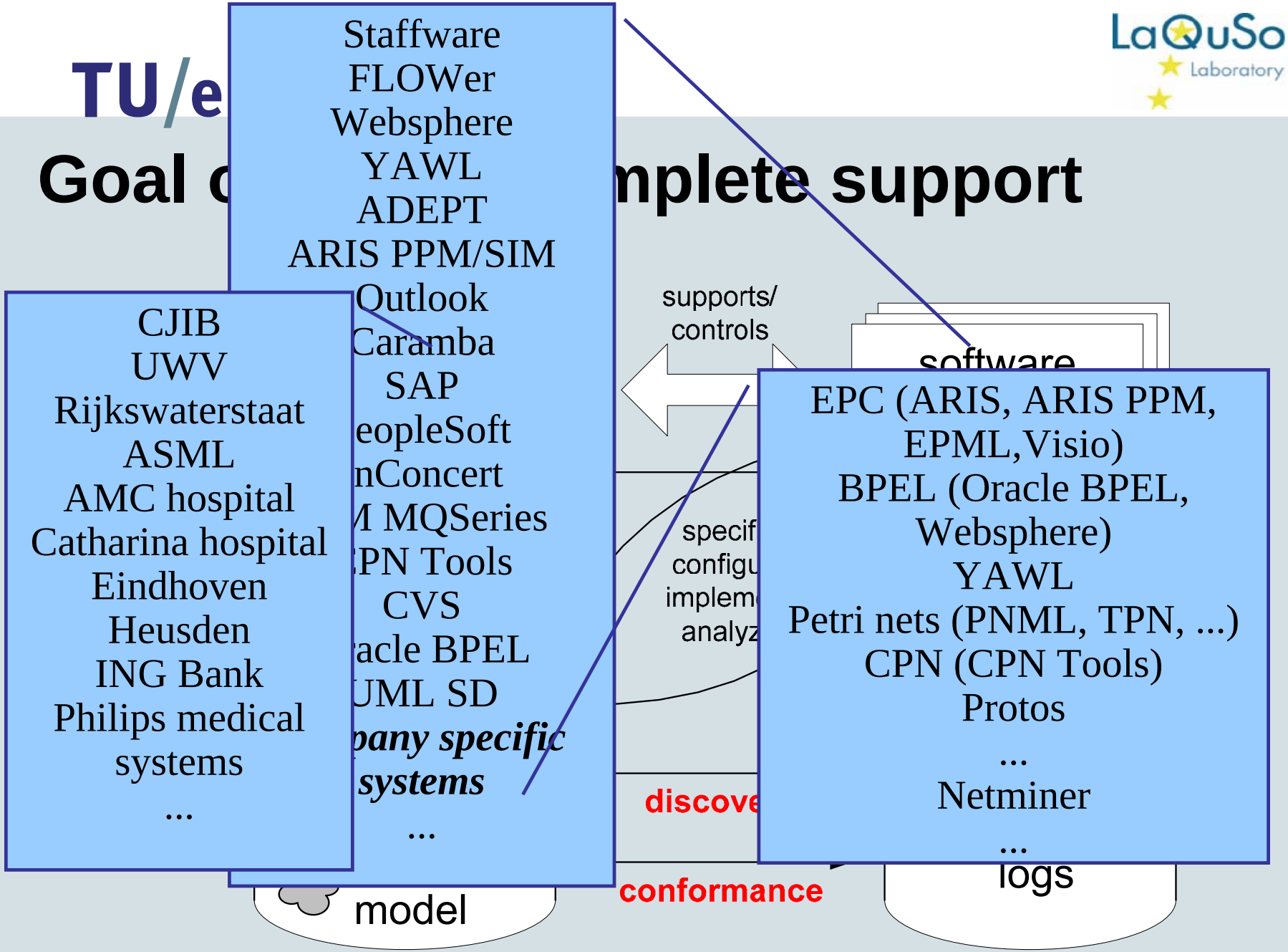




Demo?

LaQuSo
Laboratory for Quality Software

Goal c mplete support



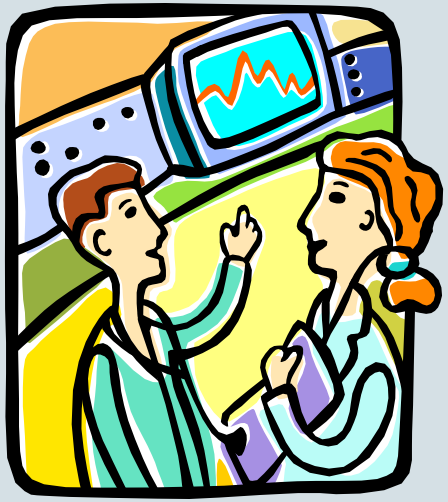

REALITY
CHECK
AHEAD

Reality Check

- Process mining on structured/administrative workflow-like logs is relatively easy.
- However, let us look at two ***extreme*** logs:
  - A log from a hospital with information on treatments, complications, and diagnoses.
  - A log from a manufacturer of high-tech system with information on system tests.

First example: Hospital data

- Information on treatment, complication, and diagnosis events.
- Data:
  - 2712 cases (all unique)
  - 29258 events
  - +/- 10.8 events per case
  - 264 different events (activities)

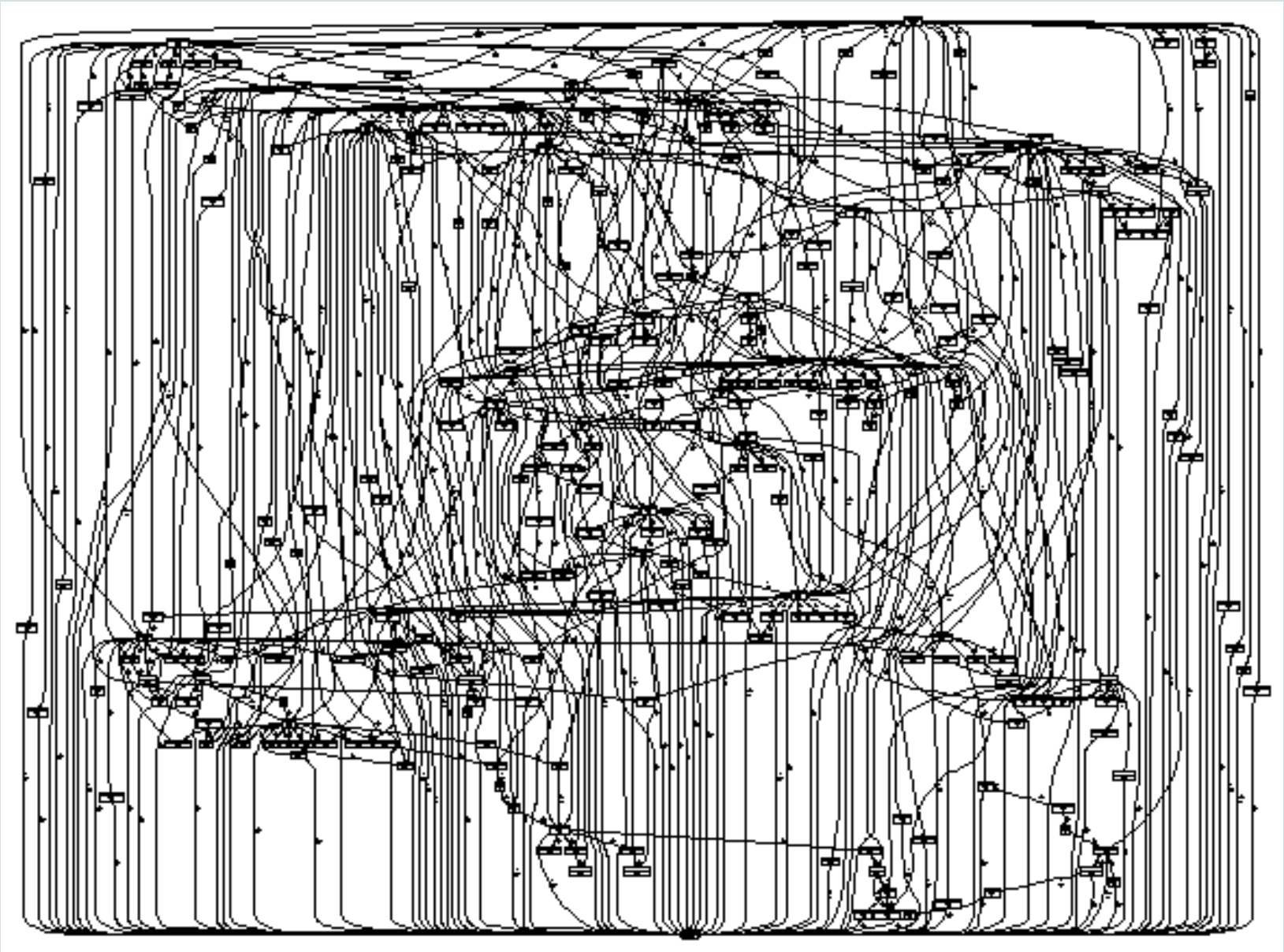


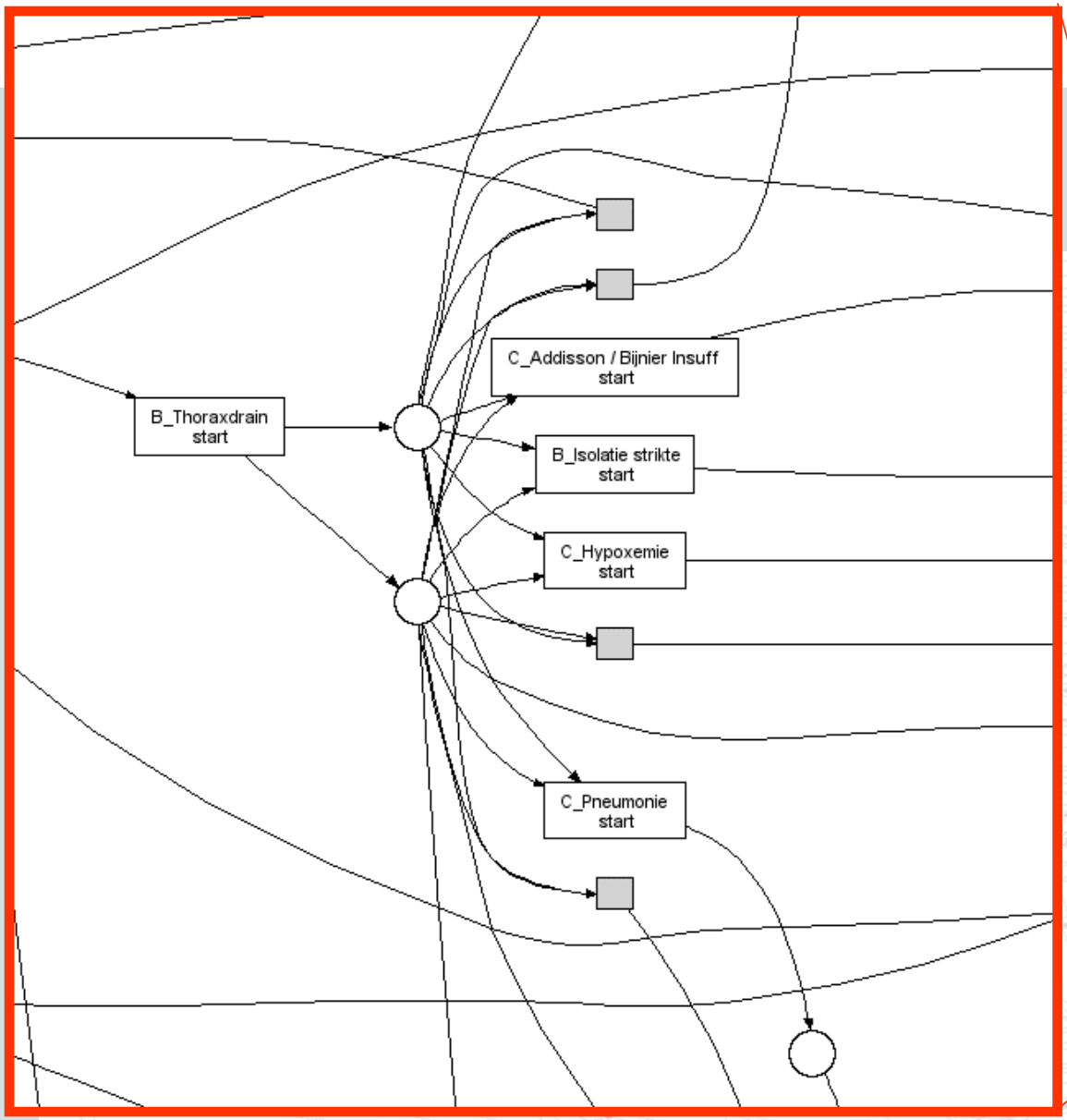
Frequency of activities

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
B_Perifeer infuus	start	2837	9,696%
B_Maagsonde	start	2430	8,305%
B_Beademing	start	2187	7,475%
B_Catheter a Demeure	start	2096	7,164%
B_Basiszorg	start	2010	6,87%
B_Arterie lijn op OK	start	2002	6,843%
B_O2 masker/slang	start	1954	6,679%
B_Thoraxdrain	start	1863	6,367%

C_N Phrenicus Paralyse	start	1	0,003%
C_TIA	start	1	0,003%
B_Horizontaal	start	1	0,003%
C_Cholecystitis, acalc	start	1	0,003%
C_Decubitus hak st. 3a	start	1	0,003%
C_Druk necrose elders	start	1	0,003%
B_Decubitus zorg stadium 3b	start	1	0,003%
C_Haemolyse	start	1	0,003%
B_Decubitus zorg stadium 4b	start	1	0,003%
B_Isolatie Beschermend	start	1	0,003%
B_Donor Weefsel	start	1	0,003%
C_Polyurie (>40ml/kg/24u)	start	1	0,003%
C_Decubitus overig st. 3a	start	1	0,003%
C_Intra-peritoneaal Abces	start	1	0,003%

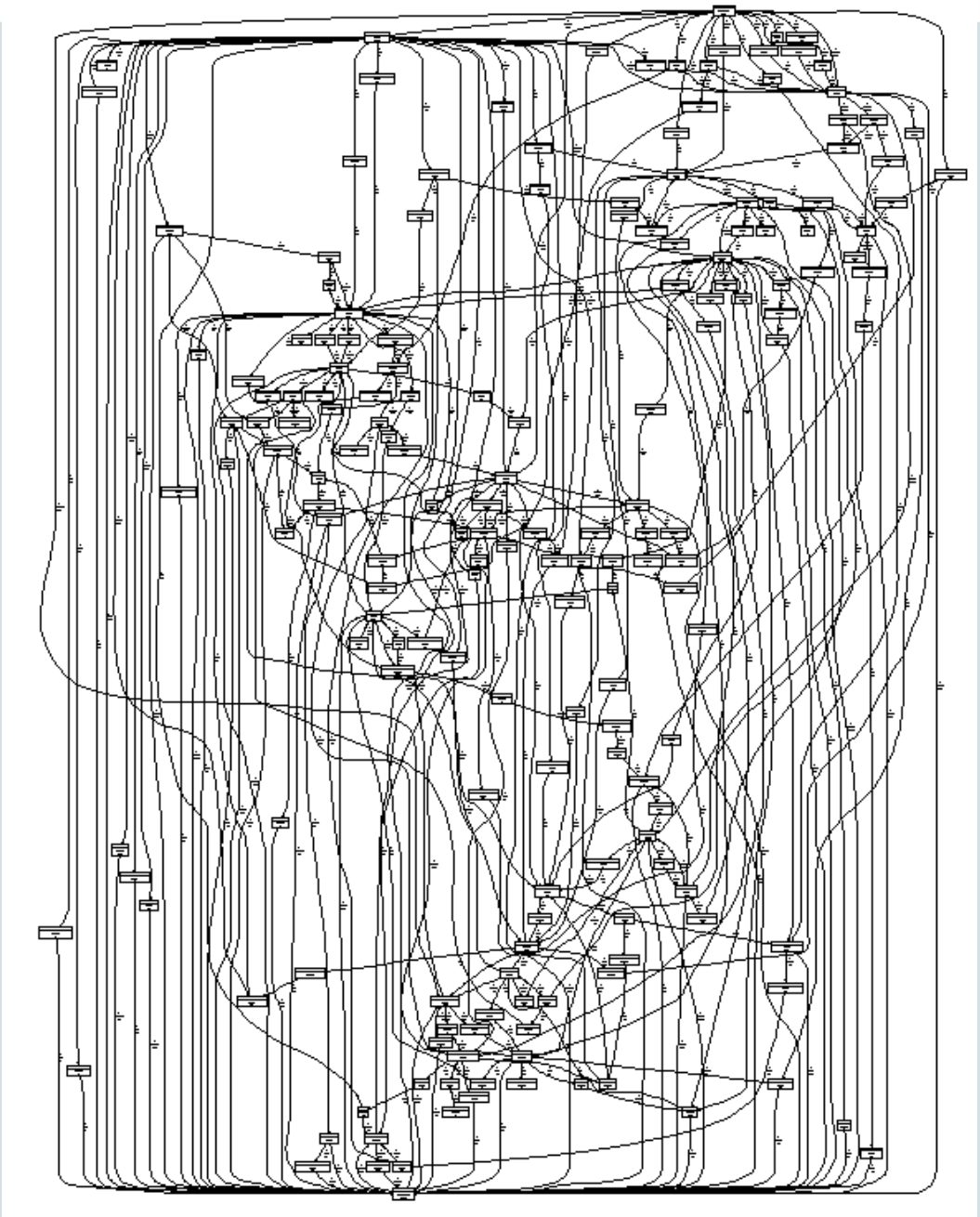
Heuristics miner




B_Thoraxdrain start
C_Addisson / Bijnier Insuff start
B_Isolatie strikte start
C_Hypoxemie start
C_Pneumonie start

Selection: Care after hart surgery

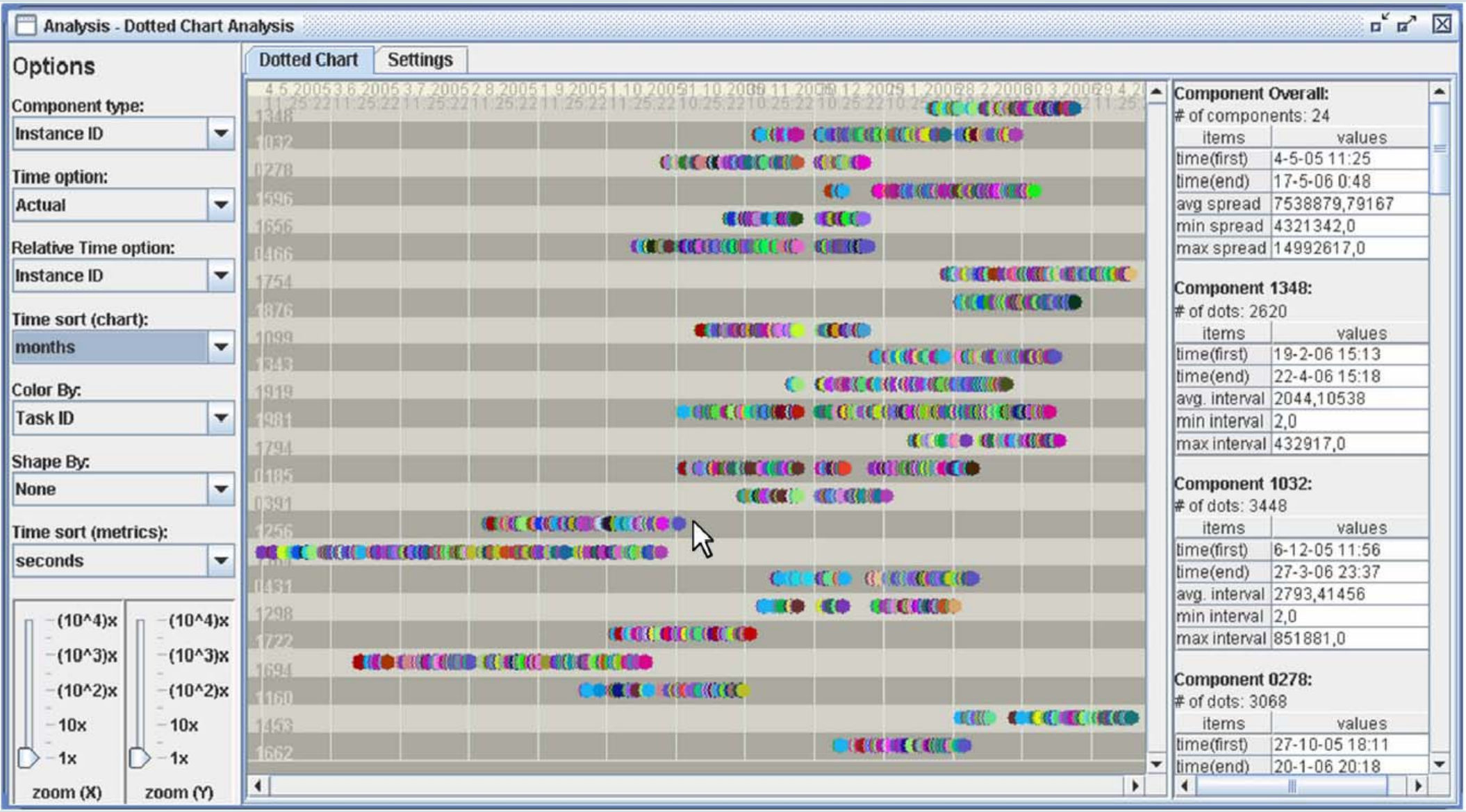
- Data
  - 874 cases (all unique)
  - 10478 events
  - 181 different events (activities)



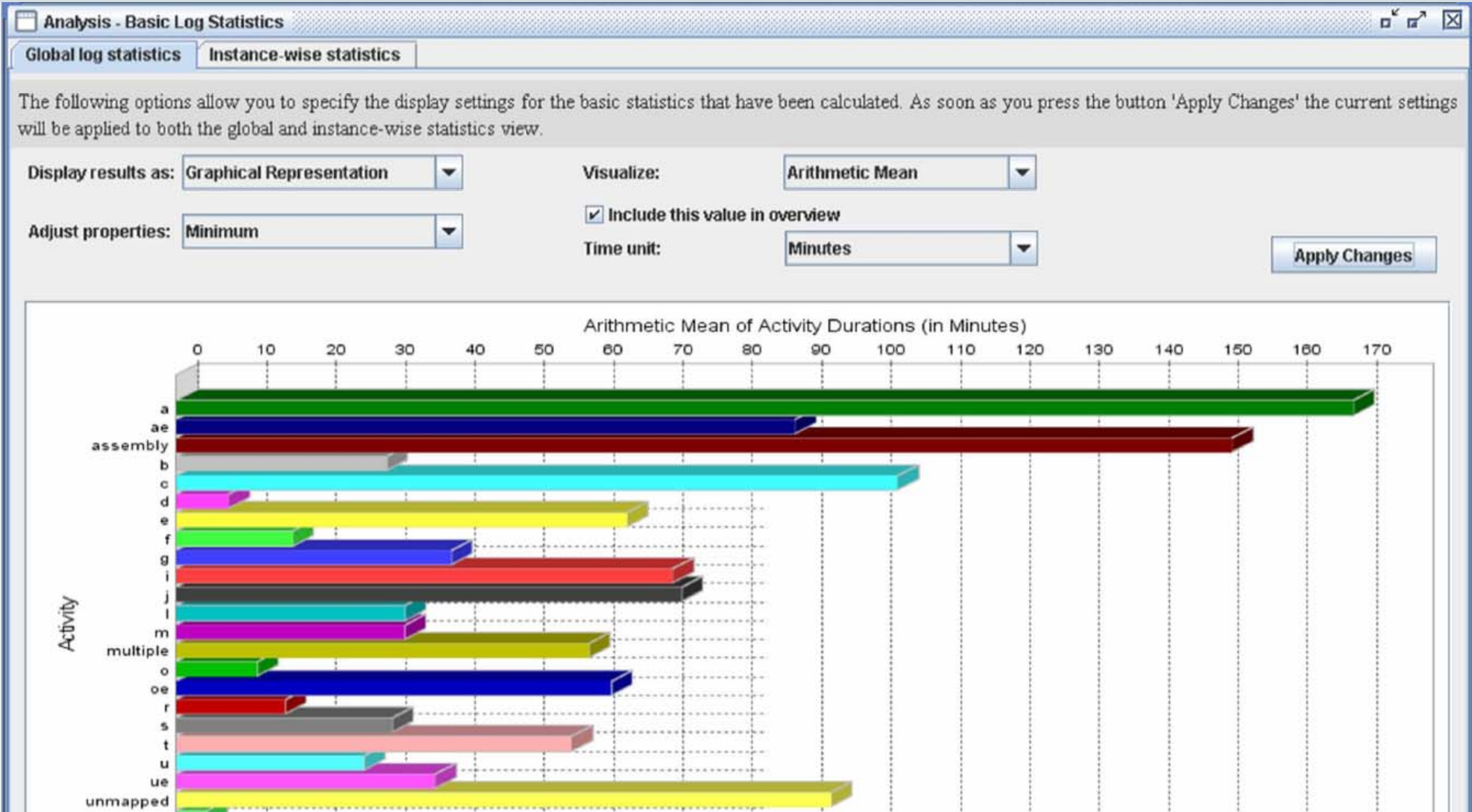
Second example: Test data from high-tech system manufacturer

- Information on testing process of high-tech systems.
- Data:
  - 24 comparable cases
  - 154966 events
  - +/- 6450 events per case
  - between 2820 and 16250 events per machine
  - 720 different events (start/complete activities)

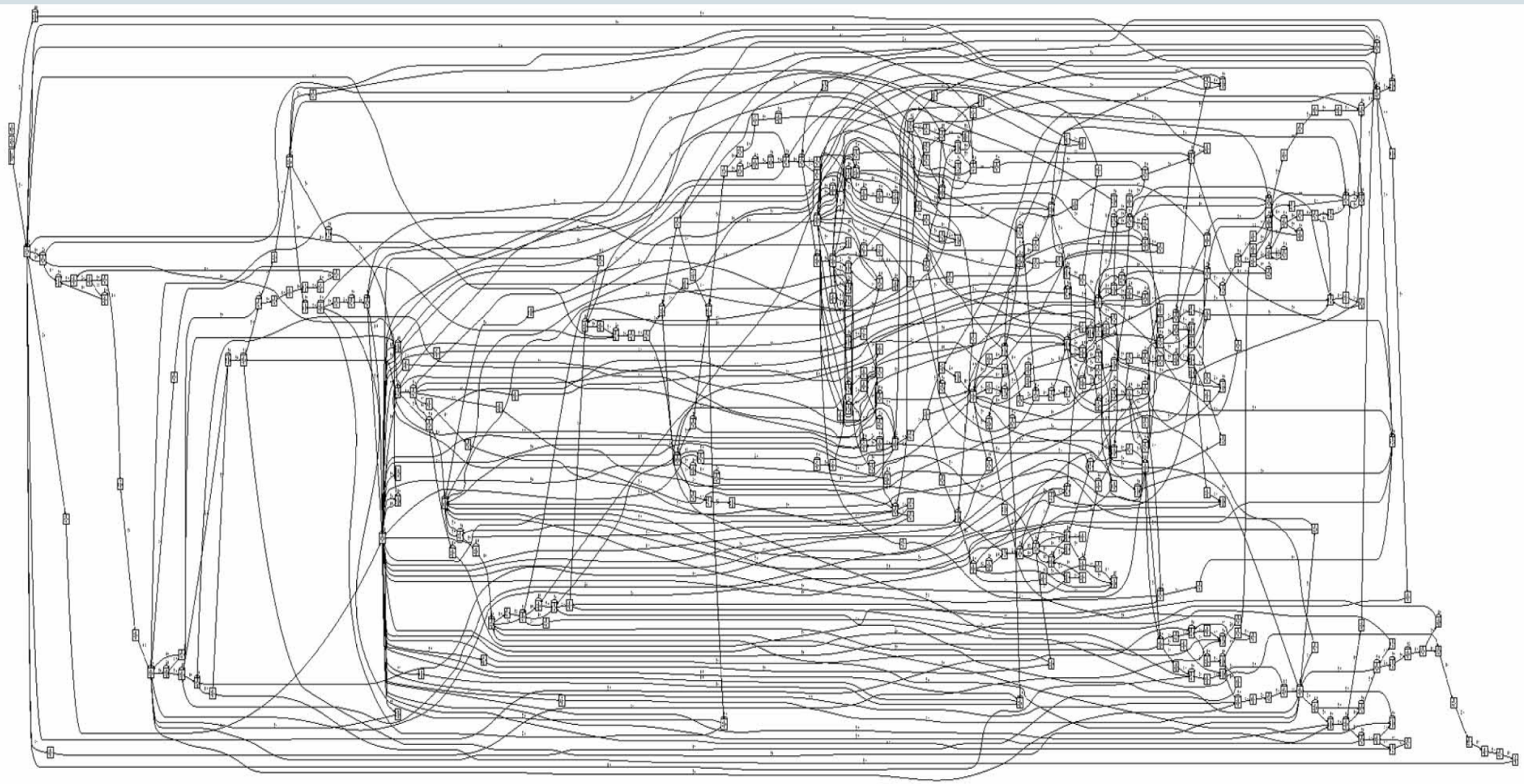
Helicopter view



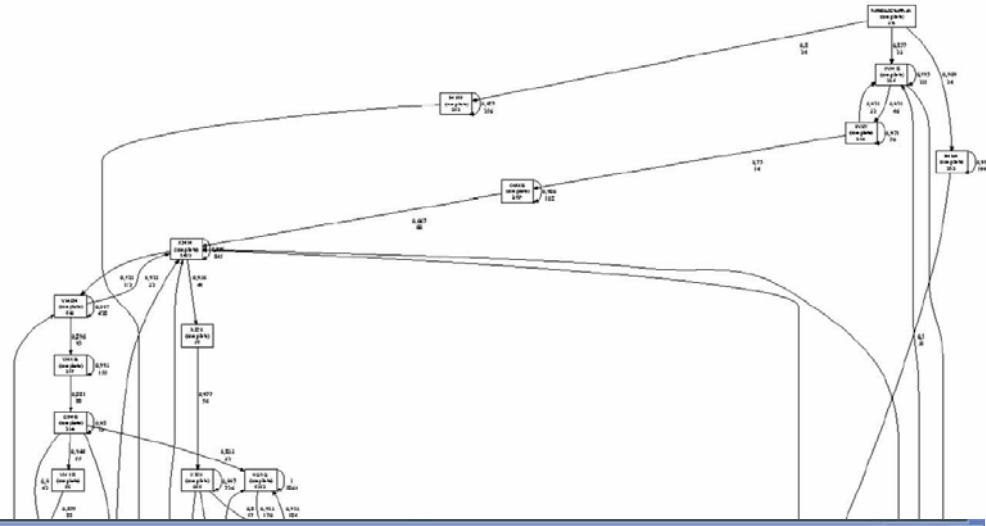
Average time spent in job-steps (aggregated events)

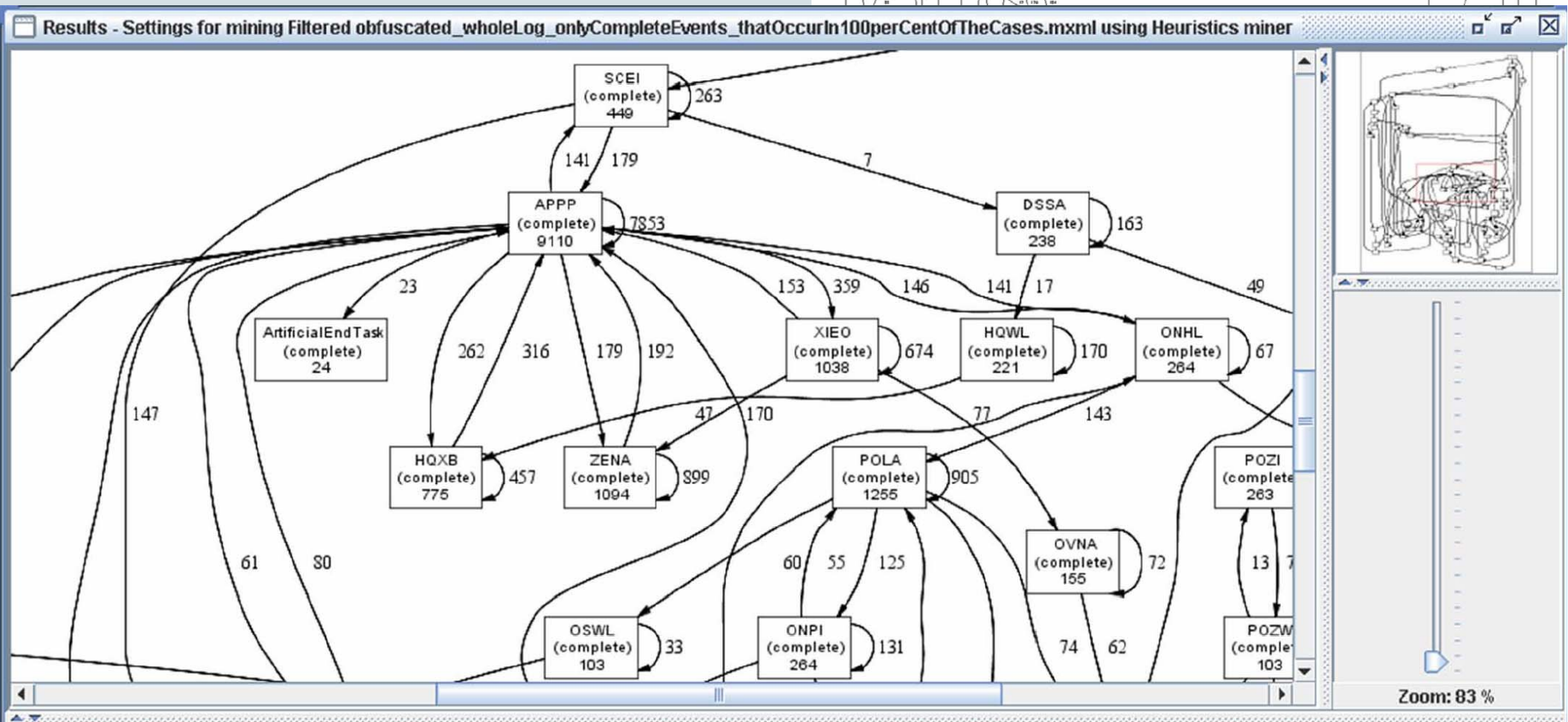


Mining just the complete events (# 360)…

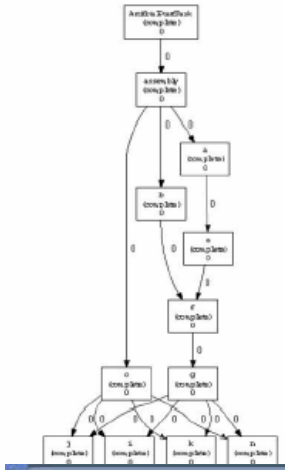


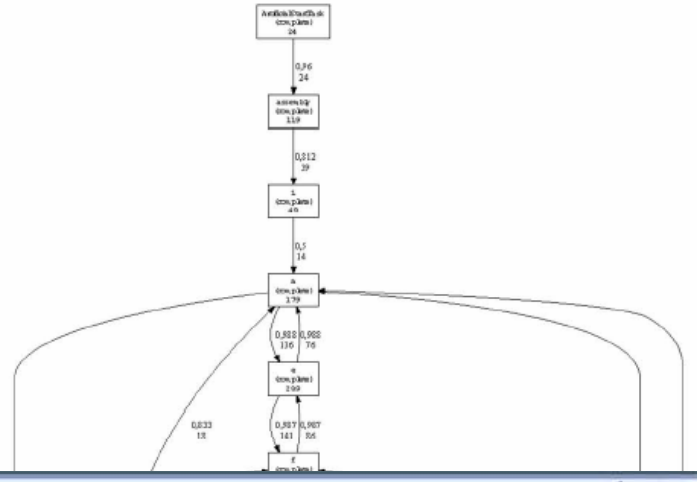
Common activities (#70)

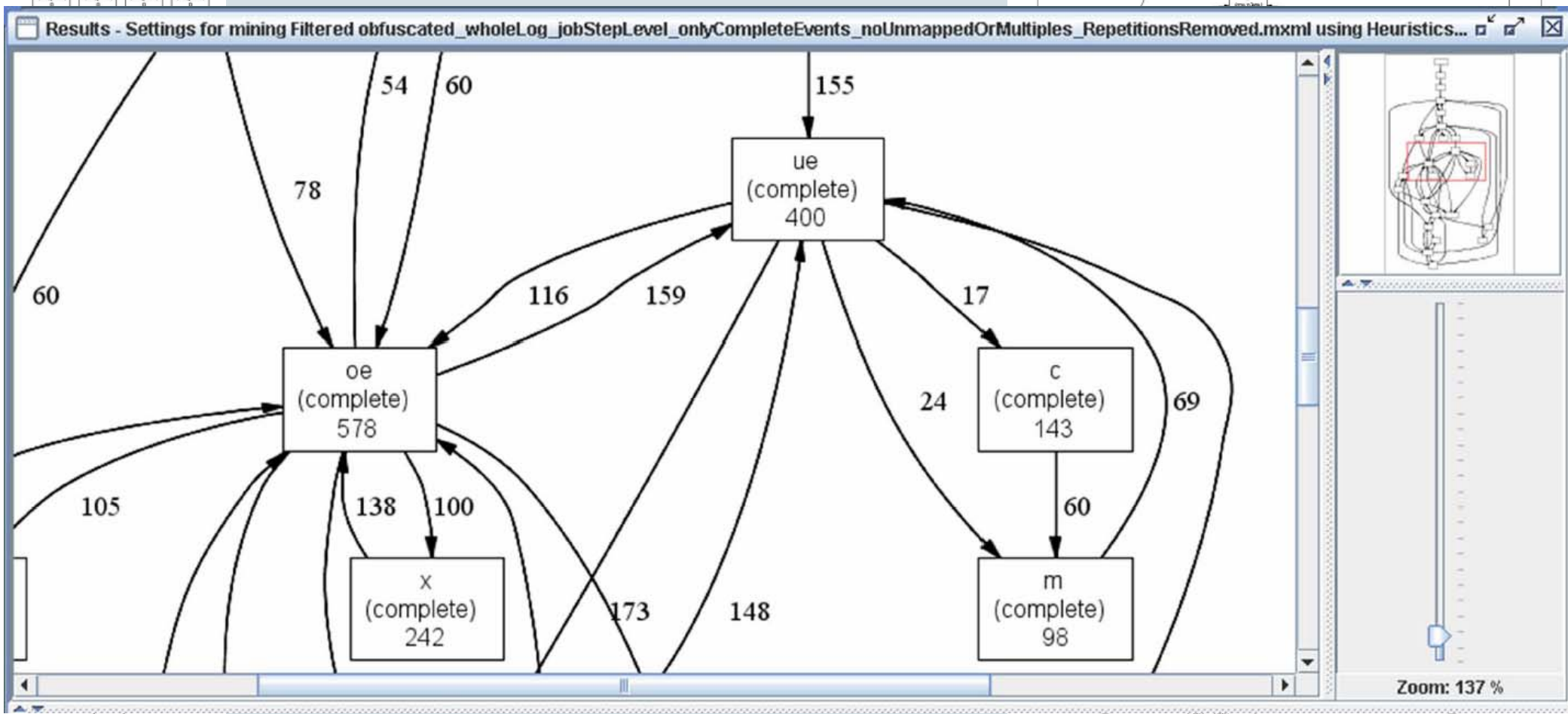




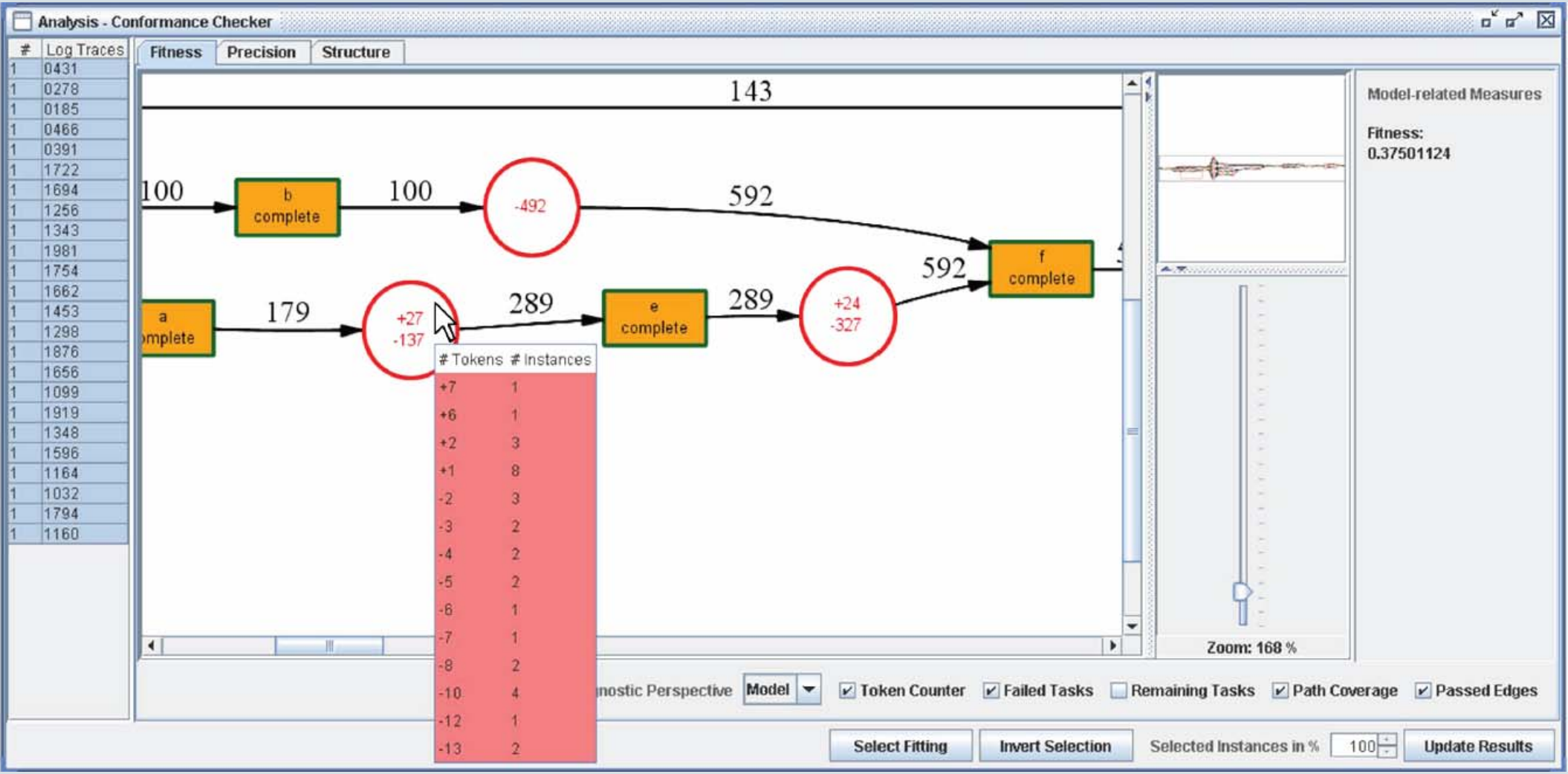
Job step level







Conformance checker (reference model – job steps)

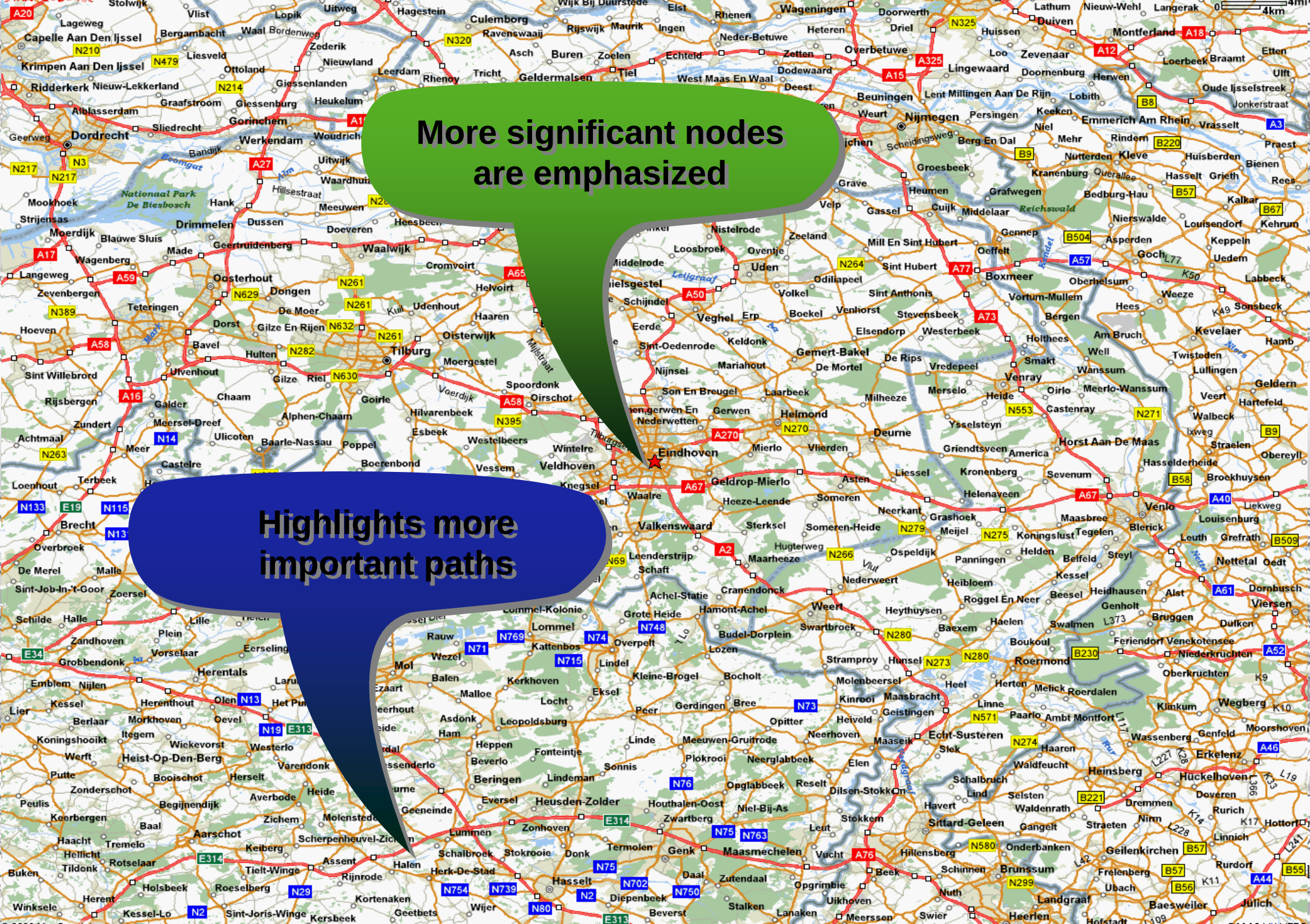


Discovered models fit better than reference model

Machine ID	*Fitness with respect to the reference process model*	*Fitness with respect to the discovered process model*
0431	$f = 0.30895045$	$f = 0.75113416$
0278	$f = 0.38491702$	$f = 0.82790506$
0185	$f = 0.37574032$	$f = 0.7171429$
0466	$f = 0.35643995$	$f = 0.74496365$
0391	$f = 0.38410193$	$f = 0.72710335$
1722	$f = 0.33359805$	$f = 0.7599035$
1694	$f = 0.39662793$	$f = 0.7821204$
1256	$f = 0.40988705$	$f = 0.7436544$
1343	$f = 0.3985147$	$f = 0.70074475$
1981	$f = 0.35723096$	$f = 0.6668054$
1754	$f = 0.401513$	$f = 0.77557445$
1662	$f = 0.4138763$	$f = 0.76897943$
1453	$f = 0.40458512$	$f = 0.5956065$
1298	$f = 0.37758428$	$f = 0.42357332$
1876	$f = 0.35556892$	$f = 0.7529762$
1656	$f = 0.36802232$	$f = 0.65629774$
1099	$f = 0.424476$	$f = 0.67167395$
1919	$f = 0.33690846$	$f = 0.72738254$
1348	$f = 0.41031277$	$f = 0.63753587$
1596	$f = 0.41015995$	$f = 0.58138484$
1164	$f = 0.37603533$	$f = 0.67173433$
1032	$f = 0.32361075$	$f = 0.7062931$
1794	$f = 0.39387232$	$f = 0.7337234$
1160	$f = 0.40484217$	$f = 0.7697767$

Research challenge

Mining less structured processes: the more unstructured, the more important it is to know what is going on!


More significant nodes are emphasized
Highlights more important paths

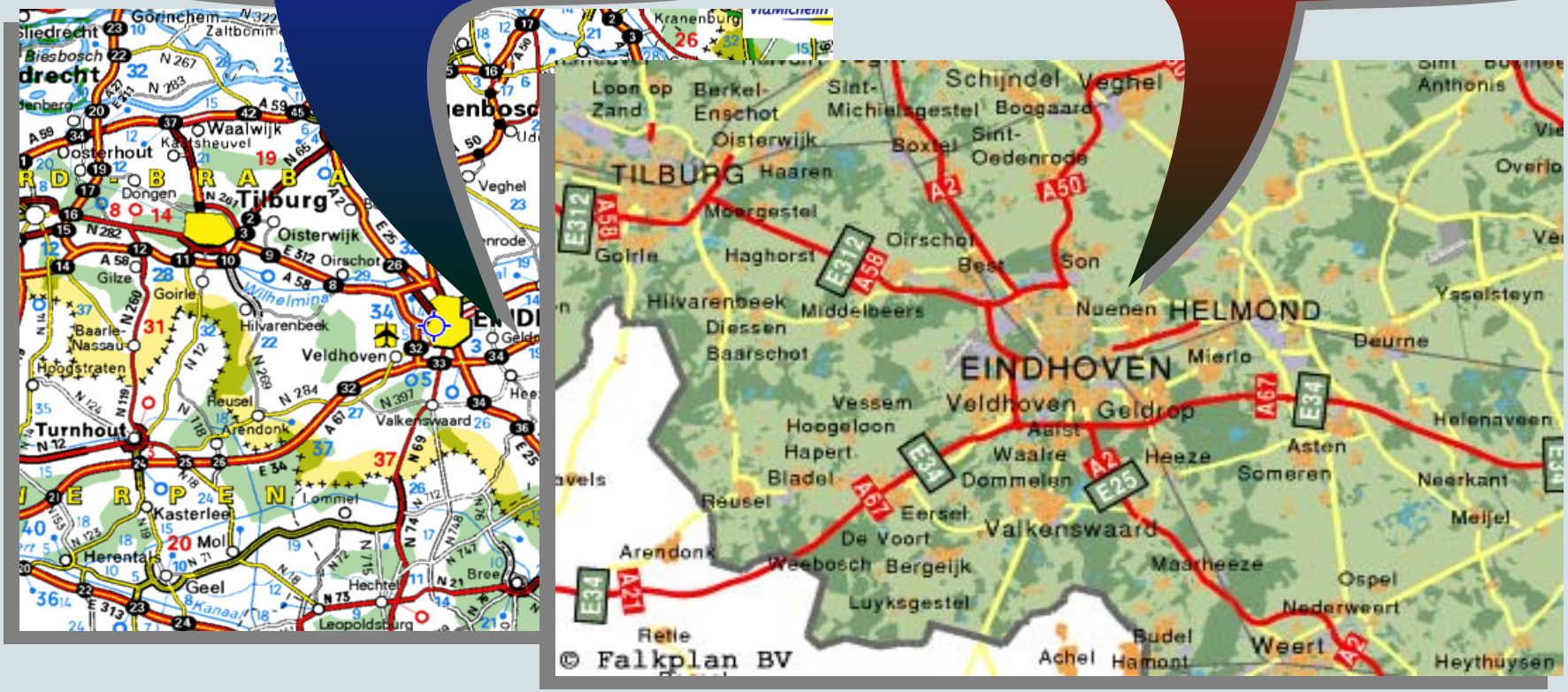
More to learn from maps...

Aggregation

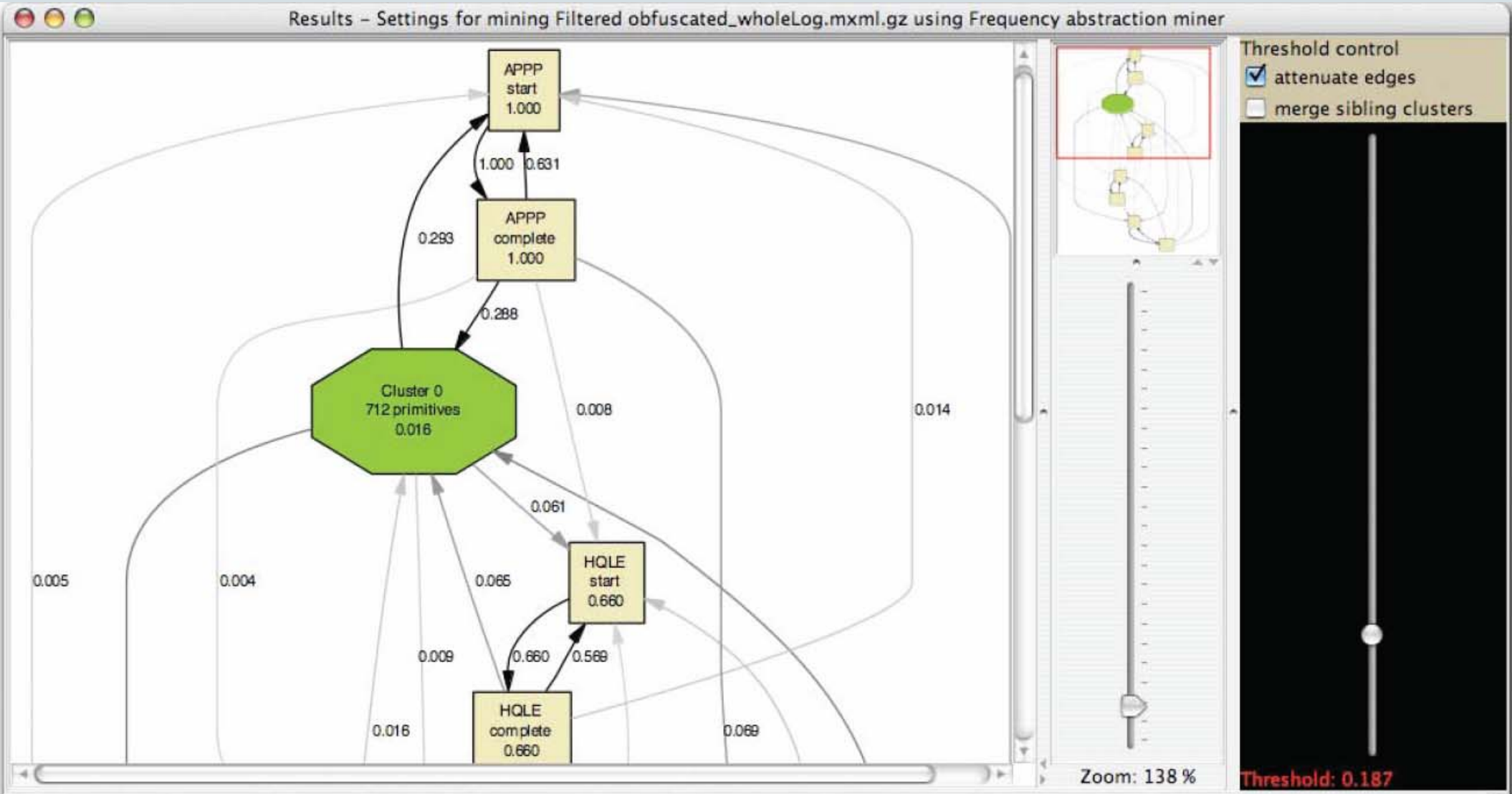
Clustering of coherent, less significant structures

Abstraction

Removing isolated, less significant structures

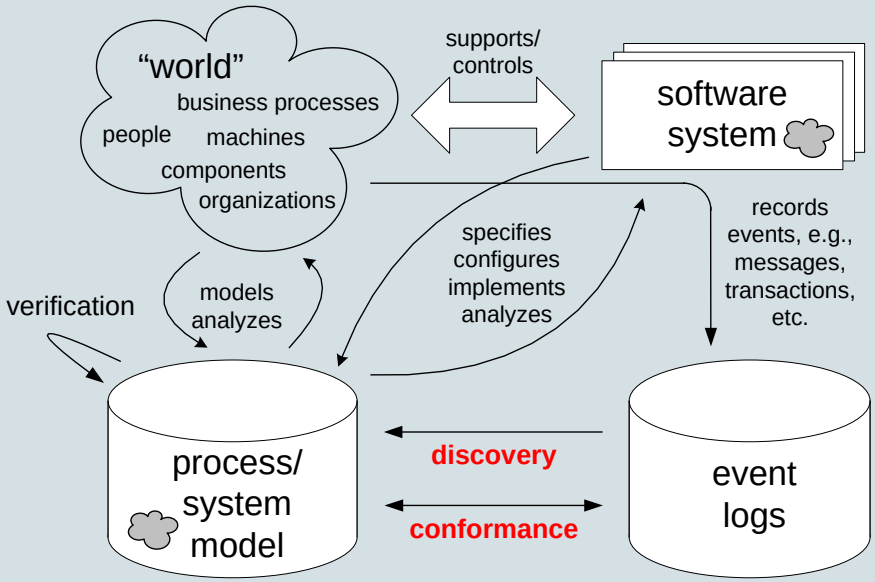


ProM's Frequency abstraction miner



Conclusion

- Reality is different from models!
- The existence of event data enables a wide variety of process mining techniques: discovery and conformance.
- ProM supports this (+150 plug-ins)
- Although quite successful for "structured processes", "spaghetti processes" remain a challenge (two examples were given).
- Research should aim to address this challenge.



Relevant WWW sites

- **http://www.processmining.org**
- http://promimport.sourceforge.net
- http://prom.sourceforge.net
- http://www.workflowpatterns.com
- http://www.workflowcourse.com
- http://www.win.tue.nl/is/
- http://is.tm.tue.nl/staff/wvdaalst