Stochastic Process Model-Log Quality Dimensions

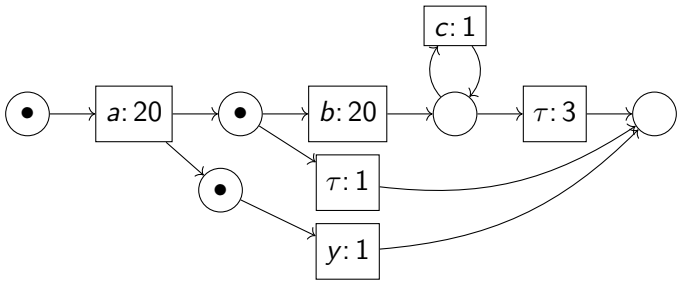
an experimental study

Adam Burke, *Sander Leemans*, Moe Wynn,
Wil van der Aalst and Arthur ter Hofstede

The Stochastic Perspective

- Event logs have stochastic information
  $[\langle a, b\rangle^{20}, \langle a, b, c\rangle^{2}, \langle a, b, c, c\rangle^{1}, \langle e, f\rangle^{1}]$
  ... already has frequency information
- Control-flow models discard stochastic information
- Stochastic process models retain stochastic information
- Simulation, analysis and recommendation need stochastic information

A Stochastic Model



Stochastic Conformance Checking Measures

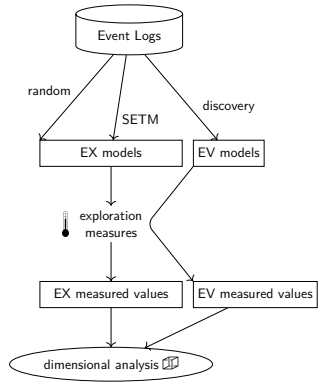
Evaluation measures

- Earth-Movers' Stochastic Conformance
- Entropy recall
- Entropy precision

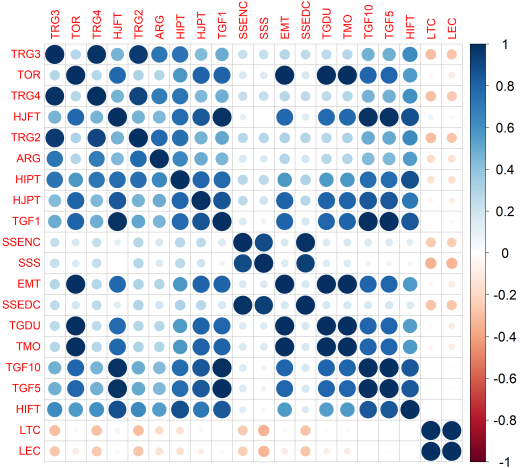
What dimensions describe the quality of stochastic process models?

Discovering the Dimensions

1. Use 6 public logs
2. 9301 stochastic process models
   random, new genetic algorithm & discovered
3. 18 exploration measures
   - Earth Movers' trace-wise (1)
   - Probability mass (2)
   - Fitness (6)
   - Precision (2)
   - Simplicity (3)
   - Generalisation (4)
4. Dimensional analysis



Dimensional Analysis 1: Correlations



- Baselines: 2 log-only measures
- Remove 3 too-correlated measures:
  Trace Overlap Ratio,
  Trace Generalization Floor-1,
  Trace Generalization Floor-10

Dimensional Analysis 2: Principal Component Analysis

- Find linear relation that best describes the data
- Find linear relation that best describes the data, orthogonal to first relation
- ...(15 times)

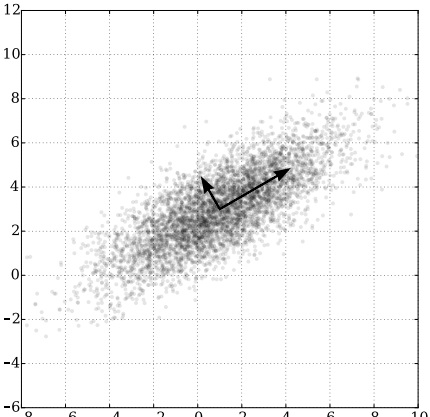


image source: wikipedia

Dimensional Analysis 2: Principal Component Analysis

- 15 linear combinations of measures
- Scree plot: we choose 3
  Covers 89% of variance

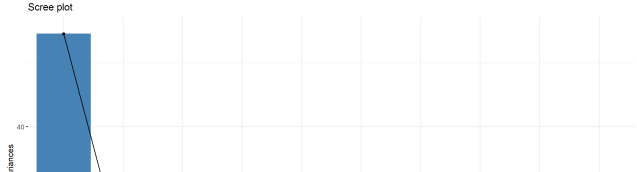
Standard deviations (1, .., p=15):

[1] 2.8598595882 1.6990123156 1.5246609206 0.7865665324 0.6297257503 0.4600371044 0.3943236321 0.2941391

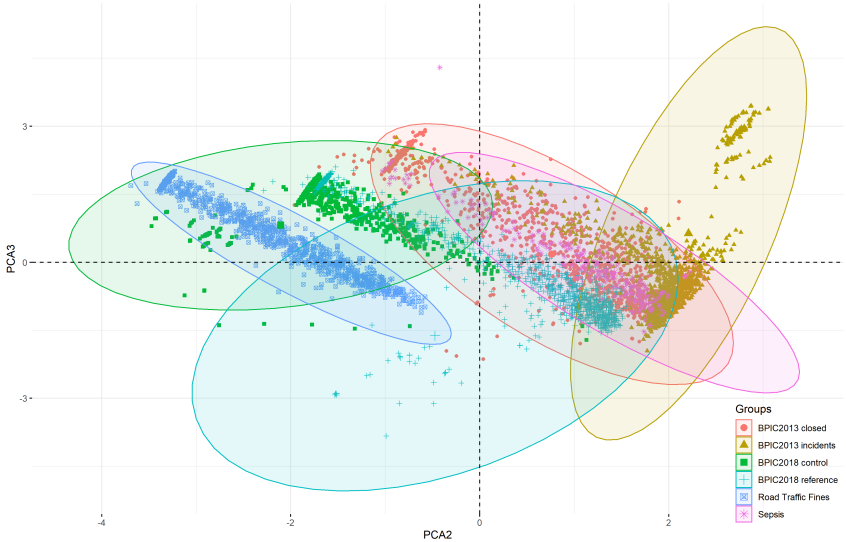
[14] 0.0149595210 0.0001853429

Rotation (n x k) = (15 x 15):

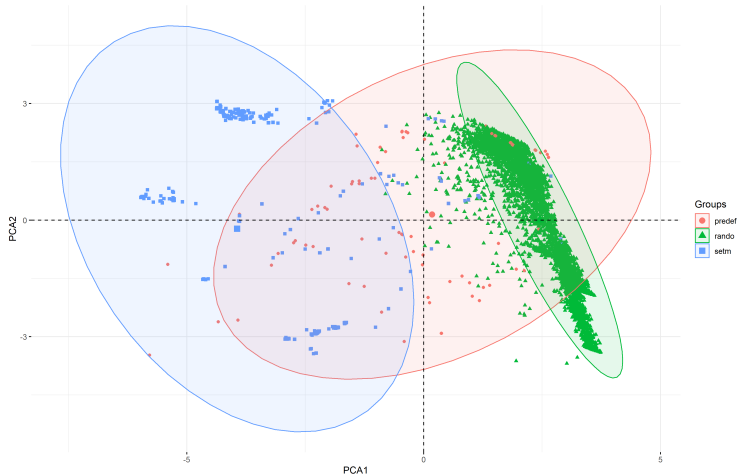
	PC1	PC2	PC3	PC4	PC5
ACTIVITY_RATIO_GOWER	-0.2313639	0.12995624	-0.28884572	-0.20701081	-0.83508956
TRACE_RATIO_GOWER_2	-0.2613284	0.21783178	-0.31657056	-0.17945506	0.09996934
TRACE_RATIO_GOWER_3	-0.25505	0.2061479	-0.34733325	-0.19427176	0.29392609
TRACE_RATIO_GOWER_4	-0.2508668	0.20311204	-0.33629168	-0.19388809	0.3374191
STRUCTURAL_SIMPLICITY_STOCHASTIC	-0.108313	0.45407014	0.32417104	-0.07154331	0.03656602
STRUCTURAL_SIMPLICITY_ENTITY_COUNT	-0.1220488	0.44260934	0.34036722	0.11738966	-0.03241646
STRUCTURAL_SIMPLICITY_EDGE_COUNT	-0.1306636	0.44564103	0.33960221	0.11777642	-0.0214105
TRACE_GENERALIZATION_DIFF_UNIQ	-0.2740367	-0.23386526	0.2618071	-0.31910582	0.05324514
EARTH_MOVERS_TRACEWISE	-0.2773889	-0.22840527	0.25235332	-0.3289332	0.05816939
TRACE_PROBMASS_OVERLAP	-0.2773891	-0.22840708	0.25235245	-0.32892646	0.05817131
ENTROPY_PRECISION_TRACEWISE	-0.3168787	0.04583473	-0.11641767	0.24698472	-0.02870561
ENTROPY_FITNESS_TRACEWISE	-0.2990944	-0.07831206	-0.13219526	0.49549231	0.05756591
ENTROPY_PRECISION_TRACEPROJECT	-0.3075119	-0.11760867	0.12287367	0.08807043	-0.26170812
TRACE_GENERALIZATION_FLOOR_5	-0.3131718	-0.17548532	0.05453552	0.26915487	0.07638127
ENTROPY_FITNESS_TRACEPROJECT	-0.308217	-0.19833903	0.03667083	0.3345224	0.03841023



Principal Components - Variation By Log

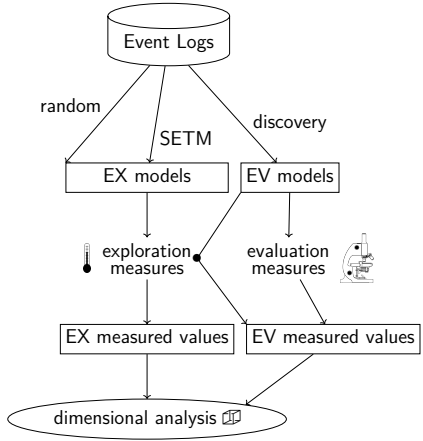


Principal Components - By Model Generator



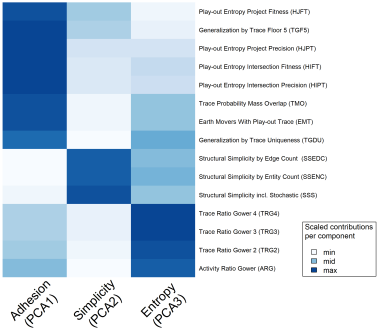
Identifying the Dimensions

- Remove random & genetic models
- Add the 3 evaluation measures *on EV models only*
- Redo principal component analysis

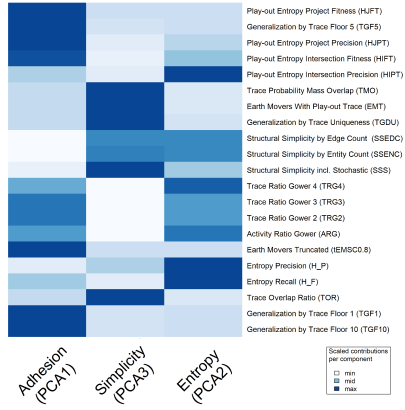


Identifying the Dimensions: Comparison

Discovered dimensions



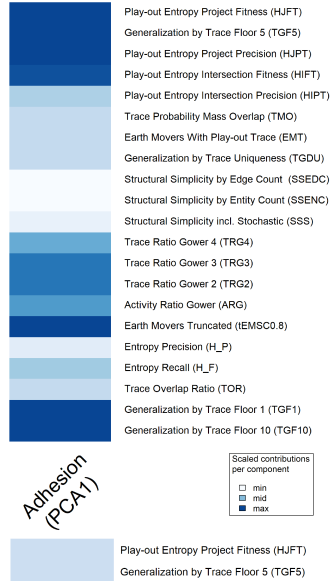
Identified dimensions



Three Empirical Dimensions

- Adhesion
  How little effort is required to transform one stochastic language into another
- Entropy
  The amount of information in a system

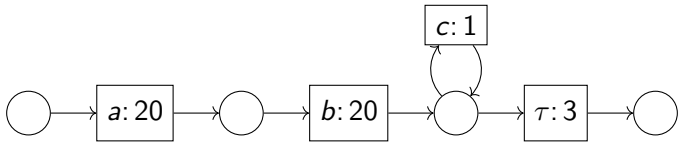
  In this case, the combination of log and model
- Simplicity
  Structural simplicity of the model



Example

Adhesion + entropy + simplicity +

$$[\langle a, b\rangle^{20}, \langle a, b, c\rangle^{2}, \langle a, b, c, c\rangle^{1}, \langle e, f\rangle^{1}]$$



Limitations

- Models block-structured
- SETM evolutionary fitness function may tend to correlate measures
  - Robustness tests excluding SETM still show the effect, though
- Largest log 200 000 traces

→ [the set-up]

You have been watching...

- Three empirically derived dimensions
- Empirical and orthogonal
- Measures may be *non-orthogonal* but still *useful*
- Future work
  - Theoretical grounded measures for these dimensions
  - Further tests

Example 2

Adhesion + entropy - simplicity +

$$[\langle a, b\rangle^{20}, \langle a, b, c\rangle^{2}, \langle a, b, c, c\rangle^{1}, \langle e, f\rangle^{1}]$$