Stochastic Process Discovery: can it be done optimally?

Sander J.J. Leemans, Tian Li, Marco Montali, Artem Polyvyanyy

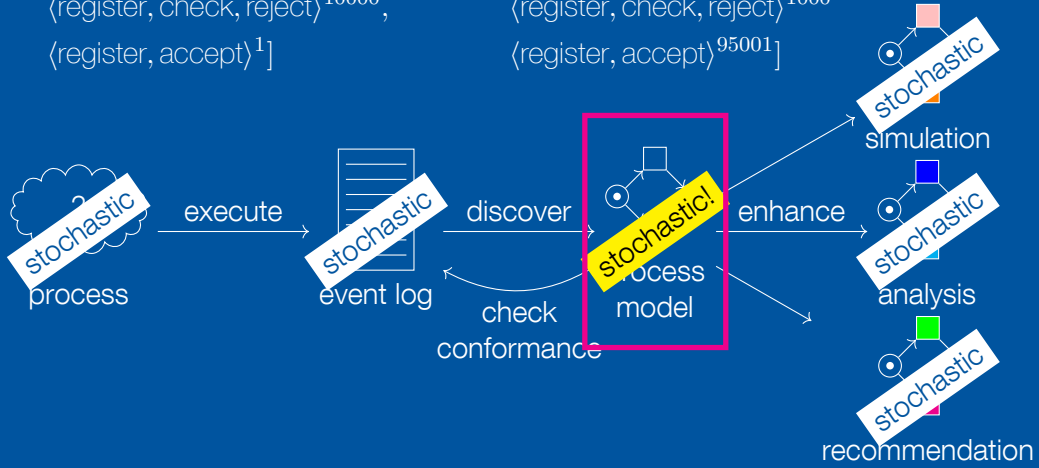
Teaching and Research Area of Business Process Management Foundations and Engineering

RWTH AACHEN UNIVERSITY

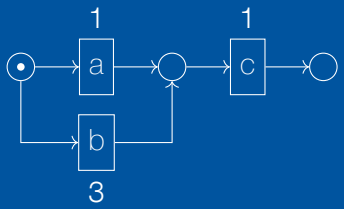
Process mining & frequencies

$L_1 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{10000}, \langle \text{register}, \text{check}, \text{reject} \rangle^{10000}, \langle \text{register}, \text{accept} \rangle^{1}]$

$L_2 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{9500}, \langle \text{register}, \text{check}, \text{reject} \rangle^{1000} \langle \text{register}, \text{accept} \rangle^{95001}]$

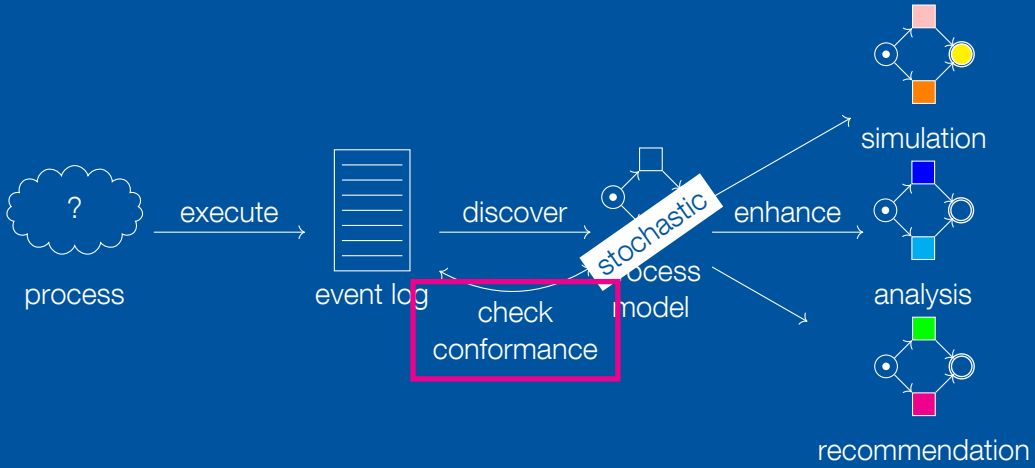


Stochastic labelled Petri nets



Stochastic language:

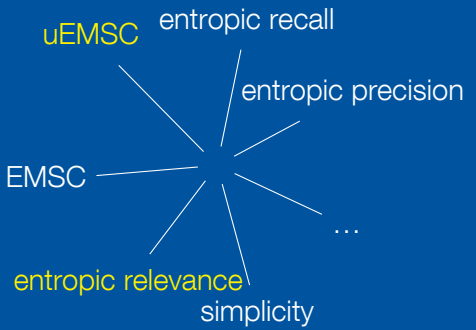
$$[\langle a, c\rangle^{0.25}$$
$$\langle b, c\rangle^{0.75}]$$

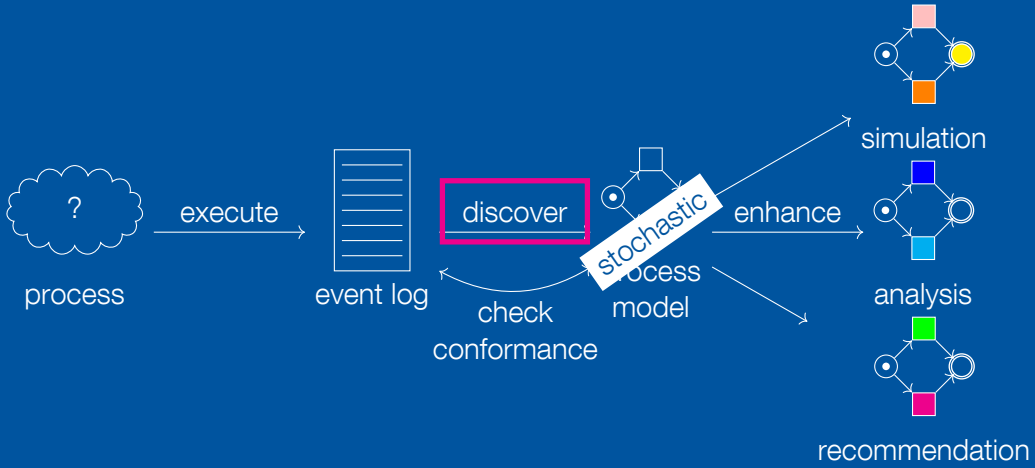

?
execute
process
event log
discover
stochastic
process
model
enhance
check
conformance
simulation
analysis
recommendation

Stochastic model quality

unit earth movers' stochastic conformance

$$1 - \sum_{\text{trace } t} \frac{|M(t) - L(t)|}{2}$$



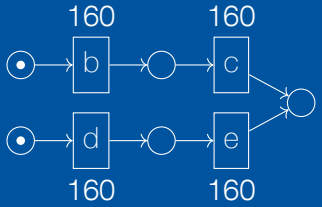

?
execute
process
event log
discover
check
conformance
stochastic
process
model
enhance
simulation
analysis
recommendation

Why we cannot just use frequencies/alignments

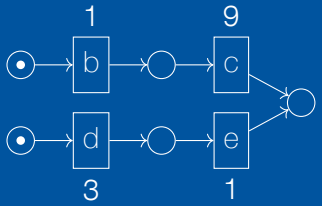
Event log:

$$[\langle b, c, d, e\rangle^{30}, \langle b, d, c, e\rangle^{9}, \langle b, d, e, c\rangle^{1}, \langle d, b, c, e\rangle^{54}, \langle d, b, e, c\rangle^{6}, \langle d, e, b, c\rangle^{60}]$$

Using frequencies/alignments:



This paper:

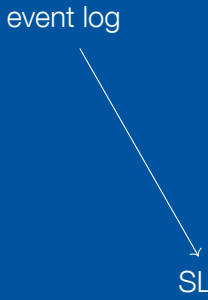


Optimal stochastic process discovery

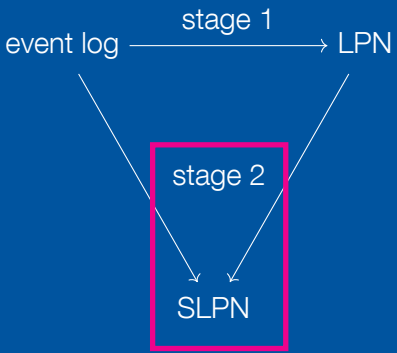
such that either uEMSC or entropic relevance is as high as possible.

Four optimal stochastic discovery techniques

One-stage approach



Two-stage approach



"stochastic ILP miners"
please refer to the paper

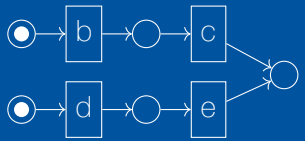
Two-stage discovery

- Given an event log, construct a control-flow model
- Annotate model with stochastic information:
  - Describe trace probability symbolically
  - Solve an optimisation problem that maximizes uEMSC

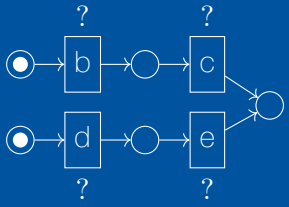
Event log:

$$[\langle b, c, d, e\rangle^{30}, \langle b, d, c, e\rangle^{9}, \langle b, d, e, c\rangle^{1}, \langle d, b, c, e\rangle^{54}, \langle d, b, e, c\rangle^{6}, \langle d, e, b, c\rangle^{60}]$$

Control-flow model:



SLPN:



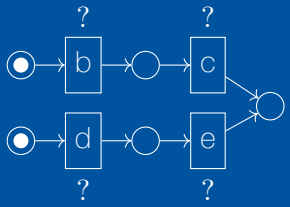
Working example

Describe trace probability symbolically

Event log:

$$[\langle b, c, d, e\rangle^{30}, \langle b, d, c, e\rangle^{9}, \langle b, d, e, c\rangle^{1}, \langle d, b, c, e\rangle^{54}, \langle d, b, e, c\rangle^{6}, \langle d, e, b, c\rangle^{60}]$$

SLPN:



The probability of trace $\sigma_1 = \langle b, c, d, e\rangle$ in the stochastic model is:

$$M(\sigma_1) = \frac{b}{b+d} \cdot \frac{c}{c+d} \cdot \frac{d}{d} \cdot \frac{e}{e}$$

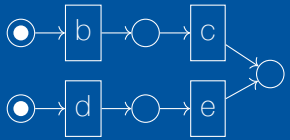
Working example

Describe trace probability symbolically

Event log:

$$[\langle b, c, d, e\rangle^{30}, \langle b, d, c, e\rangle^{9}, \langle b, d, e, c\rangle^{1}, \langle d, b, c, e\rangle^{54}, \langle d, b, e, c\rangle^{6}, \langle d, e, b, c\rangle^{60}]$$

Stochastic model:



The probability of trace $\sigma_2 = \langle b, d, c, e\rangle$ in the stochastic model is:

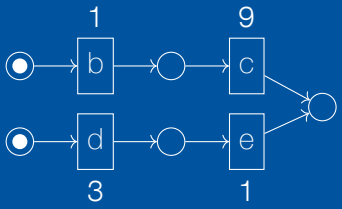
$$M(\sigma_2) = \frac{b}{b+d} \cdot \frac{d}{c+d} \cdot \frac{c}{c+e} \cdot \frac{e}{e}$$

Working example

Solve an optimisation problem:
Maximize $1 - \sum_{\sigma \in \mathrm{L}} |\mathrm{L}(\sigma) - \mathrm{M}(\sigma)|/2$ where:

- $\mathrm{L}(\sigma)$ is the probability of trace in log
- $\mathrm{M}(\sigma)$ is the probability of trace in model

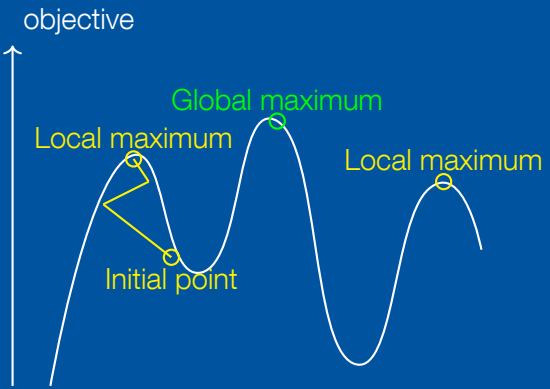
The solution to $\mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}$ is the weight of transitions in SLPN.



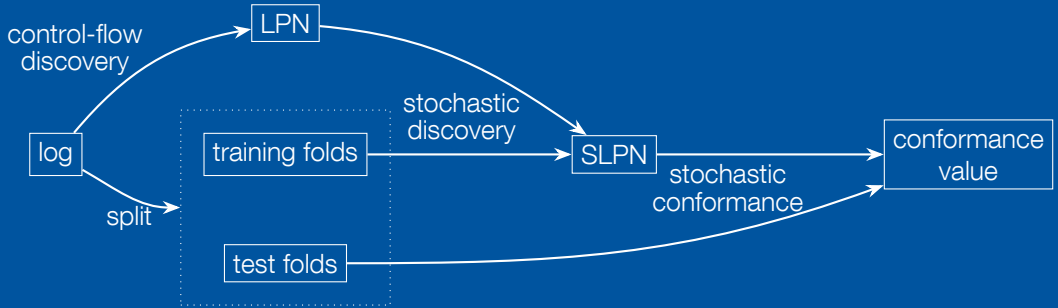
However, our optimisation problem is non-convex

Challenges in non-convex optimisation

- Multiple local optima
- Sensitivity to initialization
- Lack of convergence guarantees



Evaluation setup



Evaluation results

Log	Measure	d-uEMSC	d-ER	Frequency	Alignment	Scaled
17_application	uEMSC	0.5832	0.5072	0.2785	0.4117	0.3214
	EMSC	0.6804	0.8673	0.8605	0.8672	0.8672
	ER	0.1134	0.1106	0.1110	0.1137	0.1110
17_offer	uEMSC	0.6563	0.5828	0.5388	0.5811	0.5373
	EMSC	0.9155	0.9101	0.9026	0.9102	0.9010
	ER	0.2312	0.2330	0.1091	0.3115	0.1092
20_domestic	uEMSC	0.8079	0.8064	0.0000	0.3575	0.0000
	EMSC	0.9158	0.8596	timeout	timeout	timeout
	ER	0.1543	0.1569	0.0428	0.1144	0.0384
20_request	uEMSC	0.7537	0.7151	0.0000	0.6256	0.0000
	EMSC	0.2830	0.1978	0.4116	0.4026	0.4006
	ER	0.1610	0.1538	0.0386	0.1220	0.0389
traffic fines	uEMSC	0.8196	0.3048	0.0139	0.2940	0.0000
	EMSC	0.9061	0.7054	0.5159	0.5311	0.5403
	ER	0.1938	0.1955	0.0627	0.1675	0.0590

Evaluation results

Log	Measure	d-uEMSC	d-ER	Frequency	Alignment	Scaled
17_application	uEMSC	0.5832	0.5072	0.2785	0.4117	0.3214
	EMSC	0.6804	0.8673	0.8605	0.8672	0.8672
	ER	0.1134	0.1106	0.1110	0.1137	0.1110
17_offer	uEMSC	0.6563	0.5828	0.5388	0.5811	0.5373
	EMSC	0.9155	0.9101	0.9026	0.9102	0.9010
	ER	0.2312	0.2330	0.1091	0.3115	0.1092
20_domestic	uEMSC	0.8079	0.8064	0.0000	0.3575	0.0000
	EMSC	0.9158	0.8596	timeout	timeout	timeout
	ER	0.1543	0.1569	0.0428	0.1144	0.0384
20_request	uEMSC	0.7537	0.7151	0.0000	0.6256	0.0000
	EMSC	0.2830	0.1978	0.4116	0.4026	0.4006
	ER	0.1610	0.1538	0.0386	0.1220	0.0389
traffic fines	uEMSC	0.8196	0.3048	0.0139	0.2940	0.0000
	EMSC	0.9061	0.7054	0.5159	0.5311	0.5403
	ER	0.1938	0.1955	0.0627	0.1675	0.0590

Evaluation results

Log	Measure	d-uEMSC	d-ER	Frequency	Alignment	Scaled
17_application	uEMSC	0.5832	0.5072	0.2785	0.4117	0.3214
	EMSC	0.6804	0.8673	0.8605	0.8672	0.8672
	ER	0.1134	0.1106	0.1110	0.1137	0.1110
17_offer	uEMSC	0.6563	0.5828	0.5388	0.5811	0.5373
	EMSC	0.9155	0.9101	0.9026	0.9102	0.9010
	ER	0.2312	0.2330	0.1091	0.3115	0.1092
20_domestic	uEMSC	0.8079	0.8064	0.0000	0.3575	0.0000
	EMSC	0.9158	0.8596	timeout	timeout	timeout
	ER	0.1543	0.1569	0.0428	0.1144	0.0384
20_request	uEMSC	0.7537	0.7151	0.0000	0.6256	0.0000
	EMSC	0.2830	0.1978	0.4116	0.4026	0.4006
	ER	0.1610	0.1538	0.0386	0.1220	0.0389
traffic fines	uEMSC	0.8196	0.3048	0.0139	0.2940	0.0000
	EMSC	0.9061	0.7054	0.5159	0.5311	0.5403
	ER	0.1938	0.1955	0.0627	0.1675	0.0590

Evaluation results

Log	Measure	d-uEMSC	d-ER	Frequency	Alignment	Scaled
17_application	uEMSC	0.5832	0.5072	0.2785	0.4117	0.3214
	EMSC	0.6804	0.8673	0.8605	0.8672	0.8672
	ER	0.1134	0.1106	0.1110	0.1137	0.1110
17_offer	uEMSC	0.6563	0.5828	0.5388	0.5811	0.5373
	EMSC	0.9155	0.9101	0.9026	0.9102	0.9010
	ER	0.2312	0.2330	0.1091	0.3115	0.1092
20_domestic	uEMSC	0.8079	0.8064	0.0000	0.3575	0.0000
	EMSC	0.9158	0.8596	timeout	timeout	timeout
	ER	0.1543	0.1569	0.0428	0.1144	0.0384
20_request	uEMSC	0.7537	0.7151	0.0000	0.6256	0.0000
	EMSC	0.2830	0.1978	0.4116	0.4026	0.4006
	ER	0.1610	0.1538	0.0386	0.1220	0.0389
traffic fines	uEMSC	0.8196	0.3048	0.0139	0.2940	0.0000
	EMSC	0.9061	0.7054	0.5159	0.5311	0.5403
	ER	0.1938	0.1955	0.0627	0.1675	0.0590

You have been watching

- Optimal stochastic process discovery
- Four optimal approaches
  - uEMSC / ER
  - one-stage / two-stage

Future work

- Implement one-stage approach
- Improve implementation of two-stage approach
- Optimality on combinations of measures

Sander Leemans & Tian Li
s.leemans@bpm.rwth-aachen.de
https://bpm.rwth-aachen.de

Teaching and Research Area of Business Process Management Foundations and Engineering

RWTH AACHEN UNIVERSITY

One more thing…

A one-stage discovered model