A Principled Approach to the Analysis of Process Mining Algorithms

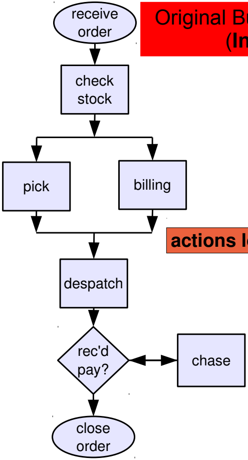
Phil Weber, Behzad Bordbar, Peter Tiňo
University of Birmingham

IDEAL 2011, Norwich, 8 Sept 2011

A Problem of Choice

Process Mining

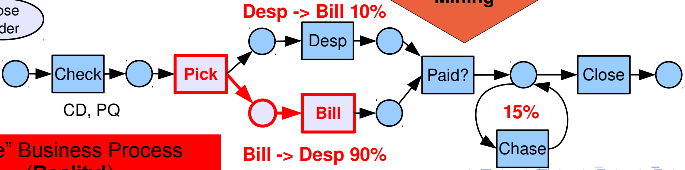
Original Business Process (**Intention**)



actions logged

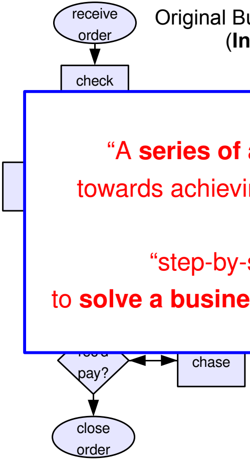
Date	Case ID	User	Task	Other Data...
20100714	0001	AB	Rec	orderno
20100714	0001	CD	Check	---
...	0001	XY	Pick	---
...	0002	AB	Rec	orderno
...	0001	MN	Billing	BACSxxxx
...	0002	PQ	Check	fail
...	0003	AB	Rec	orderno

Process Mining



"True" Business Process (**Reality!**)

Process Mining

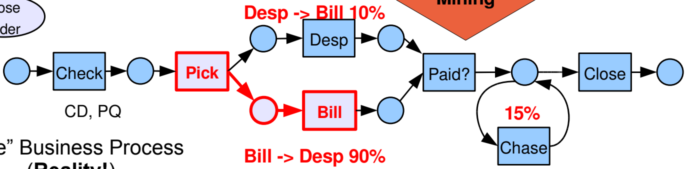


Original Business Process
(**Intention**)

“A **series of actions or steps** towards achieving a particular end.”

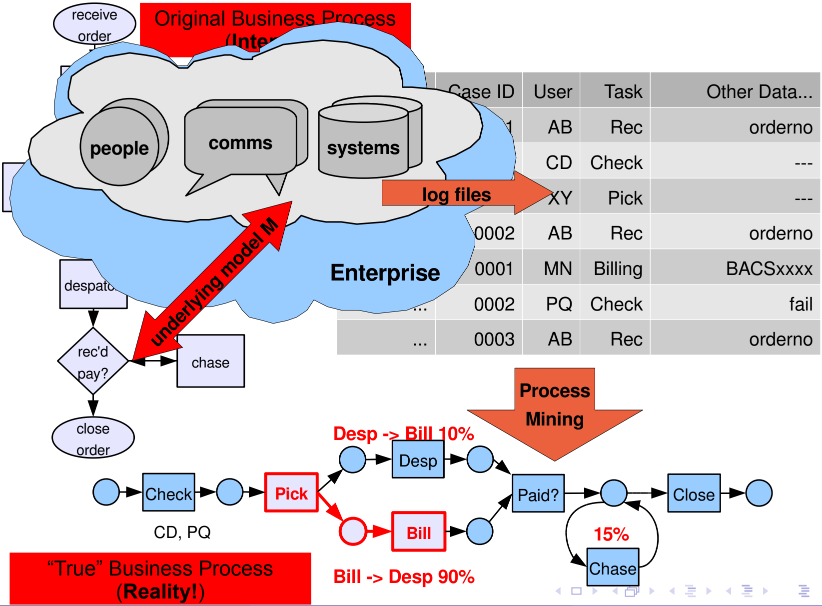
“step-by-step activities to **solve a business problem or need**.”

Date	Case ID	User	Task	Other Data...
		AB	Rec	orderno
		CD	Check	---
		XY	Pick	---
		AB	Rec	orderno
		MN	Billing	BACSxxxx
		PQ	Check	fail
		AB	Rec	orderno



“True” Business Process
(**Reality!**)

Process Mining



	Case ID	User	Task	Other Data...
	1	AB	Rec	orderno
		CD	Check	---
		XY	Pick	---
	0002	AB	Rec	orderno
	0001	MN	Billing	BACSxxxx
...	0002	PQ	Check	fail
...	0003	AB	Rec	orderno

Representations

- Petri nets
- Heuristic nets
- Activity Graphs
- BPMN
- ...

Algorithms

- formal / heuristic
- natural / neural / genetic
- slow / fast
- restricted / general
- cycles / acyclic
- ...

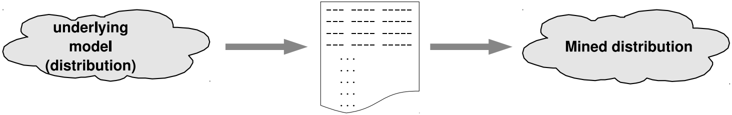
Issues

Choice! — Non-probabilistic — How much data?

Probabilistic View of Processes

Representation-free:

1. process = distribution ⇒ to learn.
2. *Secondary*: represent, analyse, cluster, abstract ...



underlying model,	workflow log,	distribution
activities $a, b, \ldots$.	traces $abdefggh, \ldots$,	$P_{\mathcal{M}'} \approx P_{\mathcal{M}}$
fixed distribution $P_{\mathcal{M}}$.	finite sample, *i.i.d.*.	

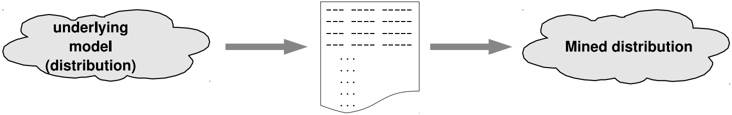
1. common basis for analysis and comparison,
2. consider convergence behaviour of algorithms,

Representation-free, but we do use a representation!

Probabilistic View of Processes

Representation-free:

1. process = distribution ⇒ to learn.
2. *Secondary*: represent, analyse, cluster, abstract ...

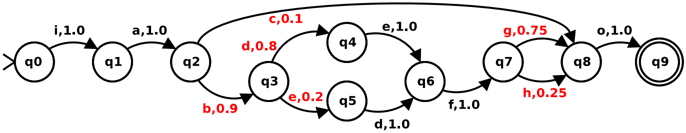


underlying model,	workflow log,	distribution
activities $a, b, \ldots$.	traces $abdefggh, \ldots$,	$P_{\mathcal{M}'} \approx P_{\mathcal{M}}$
fixed distribution $P_{\mathcal{M}}$.	finite sample, *i.i.d.*.	

1. common basis for analysis and comparison,
2. consider convergence behaviour of algorithms,

Representation-free, but we do use a representation!

Probabilistic Deterministic Finite Automata



PDFA $A = (Q_A, \Sigma, \delta_A, q_0, q_F)$,

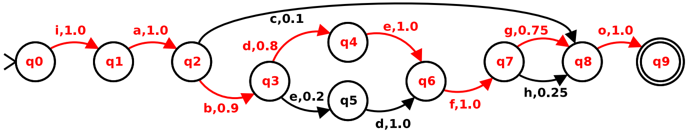
- Probabilistic: transition probability function

$$\delta_A : Q_A \times \Sigma \times Q_A \rightarrow [0, 1],$$

- Deterministic: single paths,
- Finite: finite Q_A, single q_0, q_F,

Represent single probability distribution.
Common denominator.

Probabilistic Deterministic Finite Automata



PDFA $A = (Q_A, \Sigma, \delta_A, q_0, q_F)$,

- Probabilistic: transition probability function

$$\delta_A : Q_A \times \Sigma \times Q_A \to [0, 1],$$

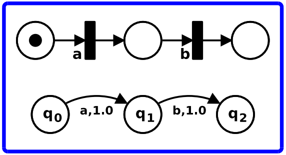
- Deterministic: single paths,
- Finite: finite Q_A, single q_0, q_F,

Represent single probability distribution. $p(iabdefgo) = 0.54$.
Common denominator.

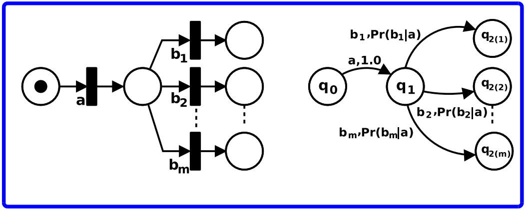
Process Structures

(Business) processes can be split into basic building blocks.

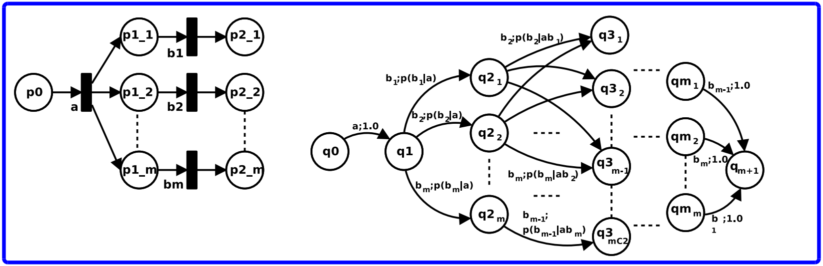
Sequence



Exclusive Or (XOR) split

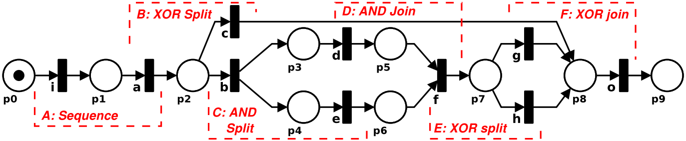


Parallel (AND) split



A Framework for Analysis

Example:



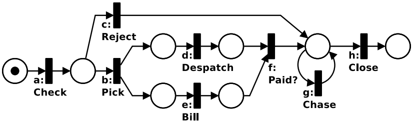
1. Probability formulae: $p(structure)$,
2. aggregate,
3. investigate,
4. experimental confirmation:
   - design and simulate,
   - mine,
   - convert to PDFA,
   - compare distributions.

Next we apply to the well-known Alpha Algorithm.

Application to the Alpha Algorithm

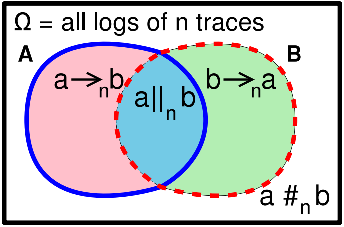
Alpha mines a Petri Net:

- activities → transitions,
- local relationships → places.



Scan log → basic 'relations':

- $a > b$: 'saw' ab in log,
- $a \rightarrow b$: causal relation,
- $a \# b$: no relation,
- $a \parallel b$: parallel.



Partition the set of activities.
Partition the set of logs of n traces.

Probability Formulae for Alpha

Notation:
$\pi(ab)$ = 'probability of ab occurring in a trace'.
$P_\alpha(a \rightarrow_n b)$ = 'probability that Alpha infers $a \rightarrow b$ from log of n traces.
Example formulae for basic relations:

$$P_\alpha(a >_n b) = 1 - (1 - \pi(ab))^n$$

$$P_\alpha(a \rightarrow_n b) = (1 - \pi(ba))^n - (1 - \pi(ab) - \pi(ba))^n$$

But general splits and joins lead to complex formulae...

Probability Formulae for Alpha

Notation:
$\pi(ab)$ = 'probability of *ab* occurring in a trace'.
$P_\alpha(a \rightarrow_n b)$ = 'probability that Alpha infers $a \rightarrow b$ from log of n traces.
Example formulae for basic relations:

$$P_\alpha(a >_n b) = 1 - (1 - \pi(ab))^n$$

$$P_\alpha(a \rightarrow_n b) = (1 - \pi(ba))^n - (1 - \pi(ab) - \pi(ba))^n$$

But general splits and joins lead to complex formulae...

Probability Formulae for Alpha

Notation:
$\pi(ab)$ = 'probability of ab occurring in a trace'.
$P_\alpha(a \to_n b)$ = 'probability that Alpha infers $a \to b$ from log of n traces.
Example formulae for basic relations:

$$P_\alpha(a >_n b) = 1 - (1 - \pi(ab))^n$$

$$P_\alpha(a \to_n b) = (1 - \pi(ba))^n - (1 - \pi(ab) - \pi(ba))^n$$

But general splits and joins lead to complex formulae...

Probability Formulae for Alpha

Notation:
$\pi(ab)$ = 'probability of ab occurring in a trace'.
$P_\alpha(a \rightarrow_n b)$ = 'probability that Alpha infers $a \rightarrow b$ from log of n traces.
Example formulae for basic relations:

$$P_\alpha(a >_n b) = 1 - (1 - \pi(ab))^n$$

$$P_\alpha(a \rightarrow_n b) = (1 - \pi(ba))^n - (1 - \pi(ab) - \pi(ba))^n$$

But general splits and joins lead to complex formulae...

Probability Formulae for Alpha

Notation:
$\pi(ab)$ = 'probability of ab occurring in a trace'.
$P_\alpha(a \rightarrow_n b)$ = 'probability that Alpha infers $a \rightarrow b$ from log of n traces.
Example formulae for basic relations:

$$P_\alpha(a >_n b) = 1 - (1 - \pi(ab))^n$$

$$P_\alpha(a \rightarrow_n b) = (1 - \pi(ba))^n - (1 - \pi(ab) - \pi(ba))^n$$

But general splits and joins lead to complex formulae...

Probability Formulae for Alpha

Notation:
$\pi(ab)$ = ‘probability of ab occurring in a trace’.
$P_\alpha(a \rightarrow_n b)$ = ‘probability that Alpha infers $a \rightarrow b$ from log of n traces.
Example formulae for basic relations:

$$P_\alpha(a >_n b) = 1 - (1 - \pi(ab))^n$$

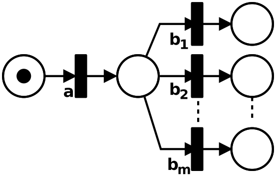
$$P_\alpha(a \rightarrow_n b) = (1 - \pi(ba))^n - (1 - \pi(ab) - \pi(ba))^n$$

But general splits and joins lead to complex formulae...

Probability Formulae for Alpha

But general splits and joins lead to complex formulae...
Require:

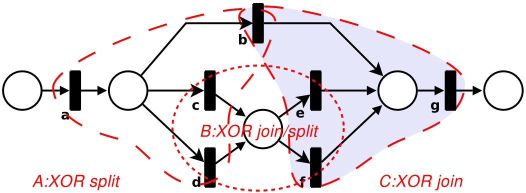
$$a \rightarrow_n b_1$$
$$a \rightarrow_n b_2$$
$$\ldots$$
$$b_1 \#_n b_2$$
$$b_1 \#_n b_3$$
$$\ldots$$

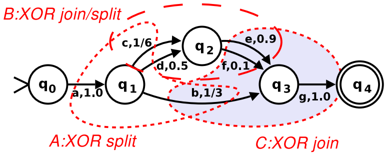


- NOT independent,
- so need probability of 'seeing' all pairs $ab_1, ab_2, \ldots$,
    - but not $b_1 b_2, b_1 b_3, \ldots$, 'Inclusion-exclusion principle'.

Simplify: assume independent:
intuitive and exponentially-decreasing error.

Very simple model, identified structures:





170 traces for 95% probability of correct discovery.

Use Reachability Graph and Maximum Likelihood probability estimation.

Graphs next.

Convergence of Mined Model with Ground Truth.

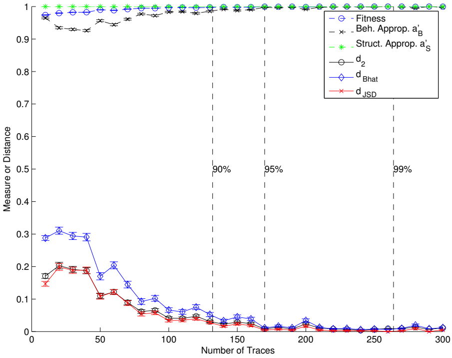


Figure 1: Probability of Approximately Correct Model

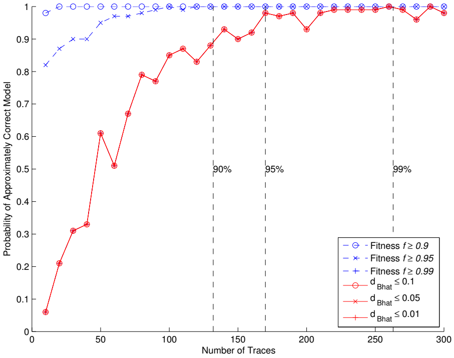


Figure 2: Probability of Approximately Correct Model

Initial results show that the amount of data needed for mining can indeed be successfully predicted.

To Conclude

Framework for analysis of process mining algorithms
— behaviour — data requirements.

$\Rightarrow$ Probability distributions over strings of symbols.
$\Rightarrow$ Probabilistic discovery of process structures.
$\Rightarrow$ Representation-free.

Initial results:

1. viable at least for Alpha,
2. distance measures more discerning,
3. separates learning behaviour from representation.

Thank You!

Phil Weber

http://www.cs.bham.ac.uk/~pxw869/
p.weber@cs.bham.ac.uk

References

[1] van der Aalst, W. M. P. and Weijters, A. J. M. M.
Process mining: a research agenda.
Computers and Industry, vol. 53, no. 3, pp. 231–244, 2004.

[2] Tiwari, A., Turner, C. J., and Majeed, B.
A Review of Business Process Mining: State-of-the-Art and Future Trends.
Bus. Process Manage. J., 14(1):5 – 22, 2008.

[3] van der Aalst, W. M. P., Weijters, T., and Maruster, L.
Workflow Mining: Discovering Process Models from Event Logs.
IEEE Trans. Knowl. Data Eng., 16(9):1128–42, 2004.

References

[1] Vidal, E., Thollard, F., de la Higuera, F., Casacuberta, F., and Carrasco, R. C.
Probabilistic Finite-State Machines - Part I.
IEEE Trans. Pattern Anal., 27(7):1013 – 25, 2005.

[2] Ferreira, D. R. and Gillblad, D.
Discovering Process Models from Unlabelled Event Logs.
In Dayal, U., Eder, J., Koehler, J., and Reijers, H. A. (eds.), *BPM 2009*. LNCS, vol. 5701, pp. 143–158. Springer, 2009.

[3] Rozinat, A., de Medeiros, A. A., Günther, C., Weijters, T., and van der Aalst, W. M. P.
Towards an evaluation framework for process mining algorithms.
BPM Center Report BPM-07-06, 2007.