An Introduction to Process Mining and Conformance Checking

Thomas Chatain

LSV, ENS Paris-Saclay, Cachan, France
chatain@lsv.fr

Collaborations with:

Mathilde Boltenhagen, Josep Carmona, Boudewijn van Dongen

June 6, 2019

Process Mining

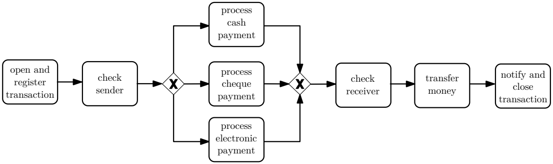
Process Mining

Discovery of process models from real process executions

Input: Event Logs Data recorded from process executions, e.g.:

- analyze usage of an e-commerce web site
- analyze medical processes in hospitals
- improve user interface
- detect deviant behavior

Output: Process Models



Process Mining

- At the interface between
  - Data science
  - Business Process Management
  - Machine learning
  - Formal models:
    models used as representation for data
- Young and very active research domain
- New conference ICPM
  - 50 submissions. . .

Many (Industrial) Process Mining Tools

- Celonis
- Disco
- Minit
- ProM
- ...

Event Logs and Data Extraction[1]

patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014:10.30	Dr. Jones	45	70.00
5541	blood test	23-1-2014:10.18	Dr. Scott	61	40.00
5833	blood test	23-1-2014:10.27	Dr. Scott	24	40.00
5781	blood test	23-1-2014:10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014:11.10	Dr. Fox	45	1200.00
5833	surgery	23-1-2014:12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014:12.41	Carol Hope	45	0.00
5541	radiation therapy	23-1-2014:13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014:13.08	Dr. Jones	61	140.00

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
5781	make X-ray	23-1-2014:10.30
5541	blood test	23-1-2014:10.18
5833	blood test	23-1-2014:10.27
5781	blood test	23-1-2014:10.49
5781	CT scan	23-1-2014:11.10
5833	surgery	23-1-2014:12.34
5781	handle payment	23-1-2014:12.41
5541	radiation therapy	23-1-2014:13.57
5541	radiation therapy	23-1-2014:13.08

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
5781	make X-ray	23-1-2014:10.30
5541	blood test	23-1-2014:10.18
5833	blood test	23-1-2014:10.27
5781	blood test	23-1-2014:10.49
5781	CT scan	23-1-2014:11.10
5833	surgery	23-1-2014:12.34
5781	handle payment	23-1-2014:12.41
5541	radiation therapy	23-1-2014:13.57
5541	radiation therapy	23-1-2014:13.08

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
5781	make X-ray	23-1-2014:10.30
5781	blood test	23-1-2014:10.49
5781	CT scan	23-1-2014:11.10
5781	handle payment	23-1-2014:12.41
5541	blood test	23-1-2014:10.18
5541	radiation therapy	23-1-2014:13.57
5541	radiation therapy	23-1-2014:13.08
5833	blood test	23-1-2014:10.27
5833	surgery	23-1-2014:12.34

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
5781	make X-ray	23-1-2014:10.30
5781	blood test	23-1-2014:10.49
5781	CT scan	23-1-2014:11.10
5781	handle payment	23-1-2014:12.41
5541	blood test	23-1-2014:10.18
5541	radiation therapy	23-1-2014:13.08
5541	radiation therapy	23-1-2014:13.57
5833	blood test	23-1-2014:10.27
5833	surgery	23-1-2014:12.34

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
	make X-ray	
	blood test	
	CT scan	
	handle payment	
	blood test	
	radiation therapy	
	radiation therapy	
	blood test	
	surgery	

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
	X	
	B	
	C	
	P	
	B	
	R	
	R	
	B	
	S	

[1] Acknowledgements to Wil van der Aalst

Event Logs and Data Extraction[1]

patient	activity	timestamp
	X	
	B	
	C	
	P	
	B	
	R	
	R	
	B	
	S	

$\langle X, B, C, P \rangle$
$\langle B, R, R \rangle$
$\langle B, S \rangle$

[1]Acknowledgements to Wil van der Aalst

Process Discovery

Automatic construction of a model N from an event log L that represents a partial observation of a system $\mathcal{S}$.

Process Discovery

Automatic construction of a model N from an event log L that represents a partial observation of a system $\mathcal{S}$.

$\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$

L

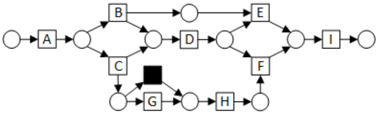
Process Discovery

Automatic construction of a model N from an event log L that represents a partial observation of a system $\mathcal{S}$.

$\langle A, B, D, E, I\rangle$
$\langle A, C, D, G, H, F, I\rangle$
$\langle A, C, G, D, H, F, I\rangle$
$\langle A, C, H, D, F, I\rangle$
$\langle A, C, D, H, F, I\rangle$

L

$\longrightarrow$



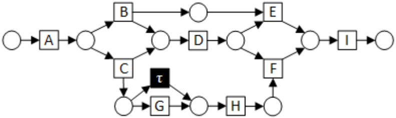
N

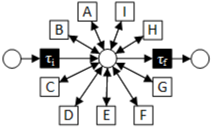
One Process Discovery Technique: Inductive Mining

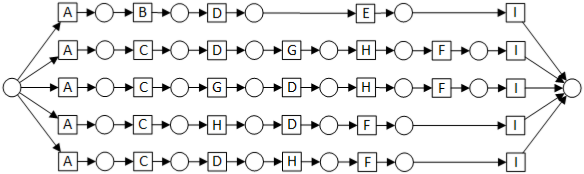
Credits: Wil van der Aalst

Process Discovery: Several Solutions

$$\text{Log: } \begin{array}{l} \langle A, B, D, E, I \rangle \\ \langle A, C, D, G, H, F, I \rangle \\ \langle A, C, G, D, H, F, I \rangle \\ \langle A, C, H, D, F, I \rangle \\ \langle A, C, D, H, F, I \rangle \end{array}$$







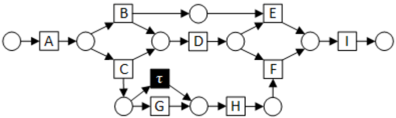
Conformance Checking

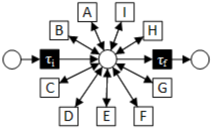
Define quality criteria to evaluate models:

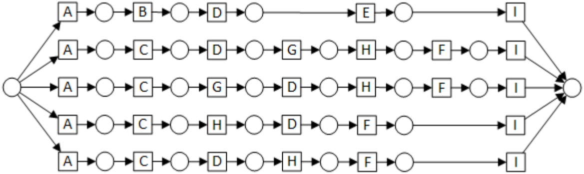
- N fits L if $L \subseteq \mathcal{L}(N)$
- N is precise if $\mathcal{L}(N) \backslash L$ is small
- N generalizes L with respect to $\mathcal{S}$ if $\mathcal{L}(N)$ contains some unobserved behavior in $\mathcal{L}(\mathcal{S}) \backslash L$
- simplicity...

Conformance Checking: Example

Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$

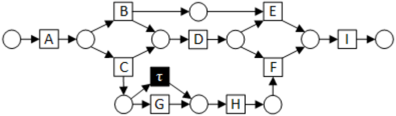


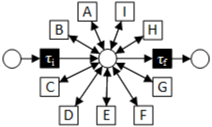


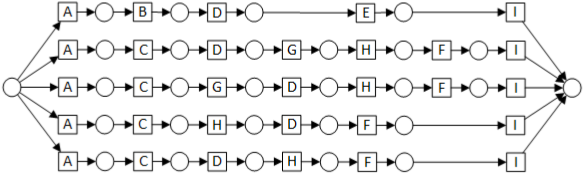


Conformance Checking: Example

Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$

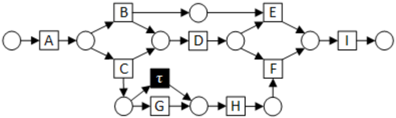




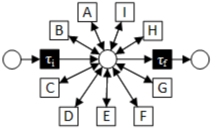


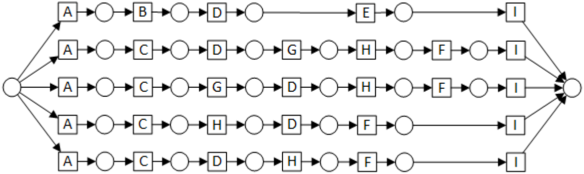
Conformance Checking: Example

Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



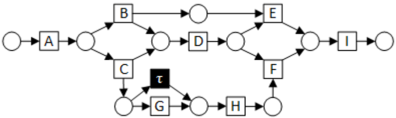
fitting
fairly precise
simple
generalizing



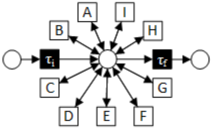


Conformance Checking: Example

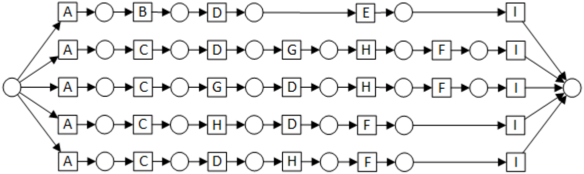
Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



fitting
fairly precise
simple
generalizing

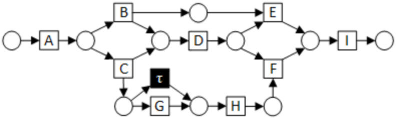


fitting
very imprecise
simple
generalizing

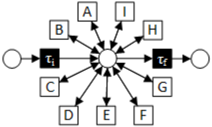


Conformance Checking: Example

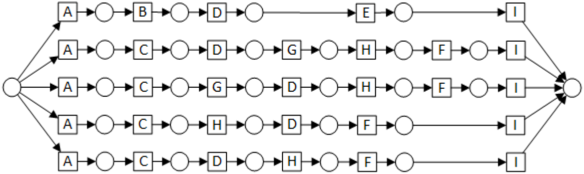
Log:
$\langle A, B, D, E, I\rangle$
$\langle A, C, D, G, H, F, I\rangle$
$\langle A, C, G, D, H, F, I\rangle$
$\langle A, C, H, D, F, I\rangle$
$\langle A, C, D, H, F, I\rangle$



fitting
fairly precise
simple
generalizing



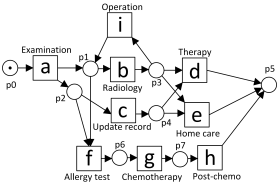
fitting
very imprecise
simple
generalizing

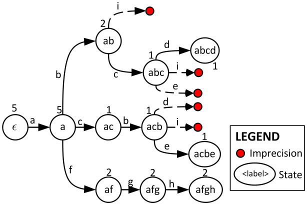


fitting
very precise
not simple
not generalizing

Measuring Precision – State of the Art

Log: $\langle a, b, c, d \rangle$
$\langle a, c, b, e \rangle$
$\langle a, f, g, h \rangle$



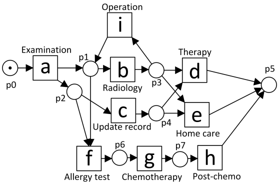


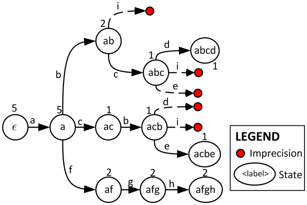
Alignment-based precision metrics [Adriansyah *et al.*]

- Build a representation $\mathcal{A}_{\Gamma(N,L)}$ of the part of the behaviour of the model which is covered by the log
- Count escaping points in $\mathcal{A}_{\Gamma(N,L)}$

Measuring Precision – State of the Art

Log: $\langle a, b, c, d \rangle$
$\langle a, c, b, e \rangle$
$\langle a, f, g, h \rangle$





Alignment-based precision metrics [Adriansyah *et al.*]

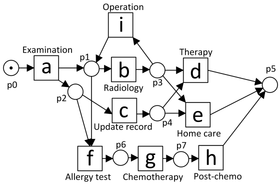
- Build a representation $\mathcal{A}_{\Gamma(N,L)}$ of the part of the behaviour of the model which is covered by the log
- Count escaping points in $\mathcal{A}_{\Gamma(N,L)}$

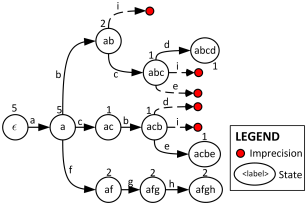
Drawbacks of alignment-based precision:

- Short sighted: only a step ahead of log behavior is considered

Measuring Precision – State of the Art

Log: $\langle a, b, c, d\rangle$
$\langle a, c, b, e\rangle$
$\langle a, f, g, h\rangle$





Alignment-based precision metrics [Adriansyah *et al.*]

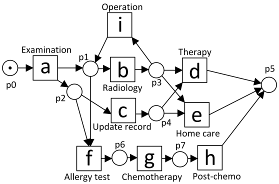
- Build a representation $\mathcal{A}_{\Gamma(N,L)}$ of the part of the behaviour of the model which is covered by the log
- Count escaping points in $\mathcal{A}_{\Gamma(N,L)}$

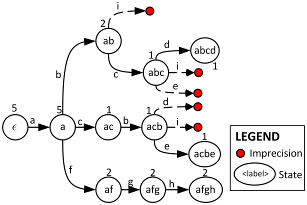
Drawbacks of alignment-based precision:

- Short sighted: only a step ahead of log behavior is considered
- Non-monotonic: observing a new trace may unveil new imprecisions

Measuring Precision – State of the Art

Log: $\langle a, b, c, d \rangle$
$\langle a, c, b, e \rangle$
$\langle a, f, g, h \rangle$
$\langle a, b, i, b, c, d \rangle$





Alignment-based precision metrics [Adriansyah *et al.*]

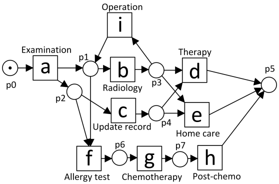
- Build a representation $\mathcal{A}_{\Gamma(N,L)}$ of the part of the behaviour of the model which is covered by the log
- Count escaping points in $\mathcal{A}_{\Gamma(N,L)}$

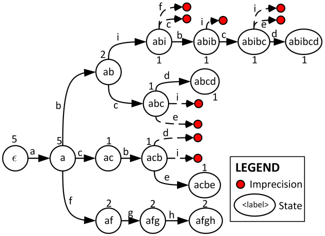
Drawbacks of alignment-based precision:

- Short sighted: only a step ahead of log behavior is considered
- Non-monotonic: observing a new trace may unveil new imprecisions

Measuring Precision – State of the Art

Log: $\langle a, b, c, d\rangle$
$\langle a, c, b, e\rangle$
$\langle a, f, g, h\rangle$
$\langle a, b, i, b, c, d\rangle$





Alignment-based precision metrics [Adriansyah *et al.*]

- Build a representation $\mathcal{A}_{\Gamma(N,L)}$ of the part of the behaviour of the model which is covered by the log
- Count escaping points in $\mathcal{A}_{\Gamma(N,L)}$

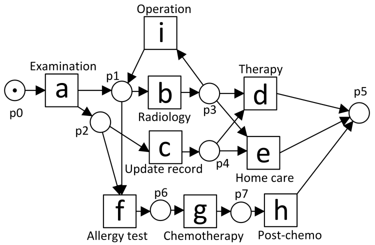
Drawbacks of alignment-based precision:

- Short sighted: only a step ahead of log behavior is considered
- Non-monotonic: observing a new trace may unveil new imprecisions

Alignments

Alignment

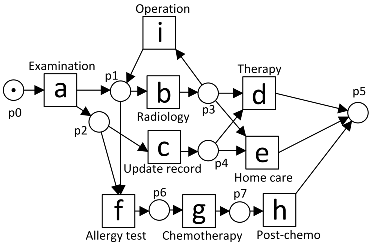
Given a trace σ and a model N,
an alignment is a full run u of N which minimizes its distance to σ.



Alignments

Alignment

Given a trace σ and a model N,
an alignment is a full run u of N which minimizes its distance to σ.

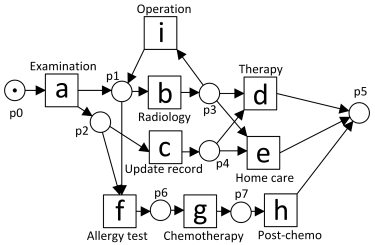


Example:
For trace $\langle a, f, c, h \rangle$,
best alignment: $\langle a, f, g, h \rangle$

Alignments

Alignment

Given a trace σ and a model N,
an alignment is a full run u of N which minimizes its distance to σ.



Example:
For trace $\langle a, f, c, h \rangle$,
best alignment: $\langle a, f, g, h \rangle$

Important notion in process mining:

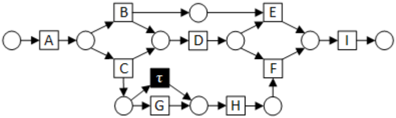
- for computing fitness and precision,
- for detecting deviations,
- for model enhancement techniques.

Anti-alignments and Precision

Anti-alignments – Motivation

Log L:

$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$



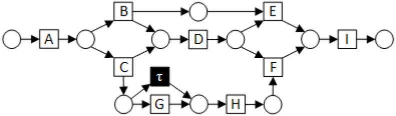
Motivation

In order to measure precision, find the run of N which is most misaligned with the log L.

Anti-alignments – Motivation

Log L:

$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$



Motivation

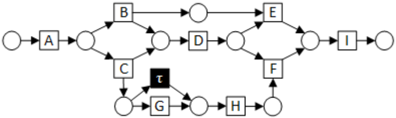
In order to measure precision, find the run of N which is most misaligned with the log L.

Here: $\langle A, B, D, E, I \rangle$

Anti-alignments

Log L:

$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$



- $L \subset \Sigma^*$: a log (set of traces) of an observed system
- N: a (labeled) Petri net model (constructed by process discovery)

Definition (Anti-alignment)

An (n, m)-anti-alignment of a model N w.r.t. a log L is a run $\gamma \in \mathcal{L}(N)$ such that

- $|\gamma| \leq n$ and
- for every $\sigma \in L$, $dist(\gamma, \sigma) \geq m$.

Which distance *dist*?

Definition (Levenshtein's edit distance $dist(\gamma, \sigma)$)

Number of letter replacements/deletions/insertions needed to edit γ to σ.

- Example: $dist_{\mathrm{Levenshtein}}(\langle ababababab \rangle, \langle bababababa \rangle) = 2$

Definition (Hamming distance)

For two traces $\gamma = \gamma_1 \ldots \gamma_n$ and $\sigma = \sigma_1 \ldots \sigma_n$, of same length n, define $dist(\gamma, \sigma) \stackrel{\mathrm{def}}{=} \left| \left\{ i \in \{1 \ldots n\} \mid \gamma_i \neq \sigma_i \right\} \right|$.

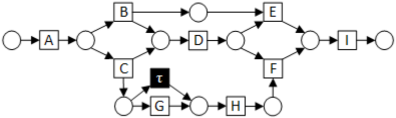
Pad when different lengths

- Example: $dist_{\mathrm{Hamming}}(\langle ababababab \rangle, \langle bababababa \rangle) = 10$

Anti-alignments: Example

Log L:

$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$



$(5, 3)$-anti-alignment $\langle A, B, D, E, I \rangle$

NP-completeness

Lemma

The problem of existence of (n, m)-anti-alignment is NP-complete.
(with n and m represented in unary.)

Proof.

The problem is clearly in NP: checking that a run γ is a (n, m)-anti-alignment for a net N and a log L takes polynomial time.

For NP-hardness, reduction from the problem of reachability of a marking M in a safe acyclic Petri net N, known to be NP-complete [a]. □

[a] Cheng, A., Esparza, J., Palsberg, J.: Complexity results for safe nets. Theor. Comput. Sci. 147(1&2) (1995) 117–136

Anti-alignments to Measure Precision

- $L \subset \Sigma^*$: a log (set of traces) of an observed system
- N: a (labeled) Petri net model (constructed by process discovery)

Anti-alignment-based precision metrics

$$P^n(N, L) = 1 - \frac{max^n(N, L)}{n}$$

with

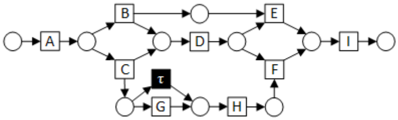
- n: (in the order of) the maximal length for a trace in the log
- $max^n(N, L)$: the largest m for which there exists a (n, m)-anti-alignment

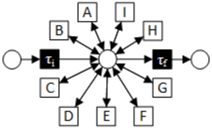
Clearly, $max^n(N, L) \in [0 \ldots n]$ which implies $P^n(N, L) \in [0 \ldots 1]$.

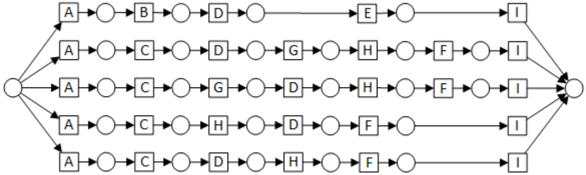
Anti-alignments to Measure Precision – Exercise

Sort the models by decreasing precision.

Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



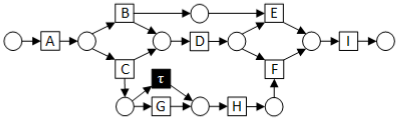


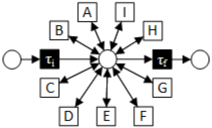


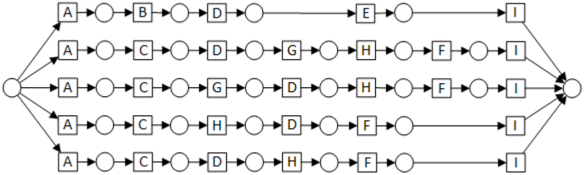
Anti-alignments to Measure Precision – Exercise

Sort the models by decreasing precision.
For each model, find the best anti-alignment of length ≤ 7.

Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



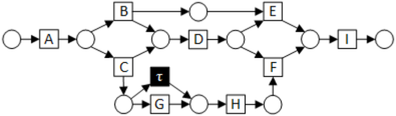




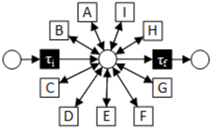
Anti-alignments to Measure Precision – Exercise

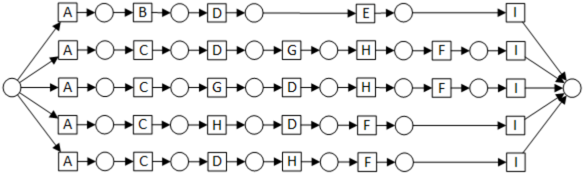
Sort the models by decreasing precision.
For each model, find the best anti-alignment of length ≤ 7.

Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



Anti-alignment $\langle A, C, G, H, D, F, I \rangle$
$P^7(N_1, L) = 0.857$

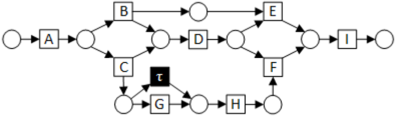




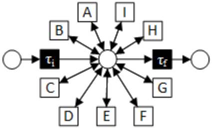
Anti-alignments to Measure Precision – Exercise

Sort the models by decreasing precision.
For each model, find the best anti-alignment of length ≤ 7.

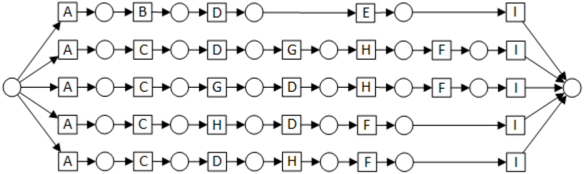
Log:
$\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



Anti-alignment $\langle A, C, G, H, D, F, I \rangle$
$P^7(N_1, L) = 0.857$



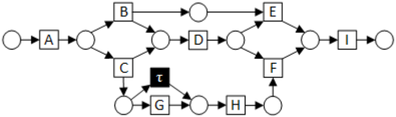
Anti-alignment
$\langle I, I, I, A, A, A, A \rangle$
$P^7(N_2, L) = 0$



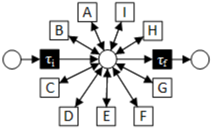
Anti-alignments to Measure Precision – Exercise

Sort the models by decreasing precision.
For each model, find the best anti-alignment of length ≤ 7.

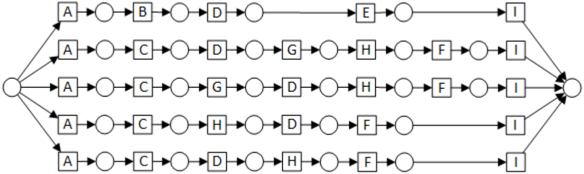
Log: $\langle A, B, D, E, I \rangle$
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$
$\langle A, C, D, H, F, I \rangle$



Anti-alignment $\langle A, C, G, H, D, F, I \rangle$
$P^7(N_1, L) = 0.857$



Anti-alignment
$\langle I, I, I, A, A, A, A \rangle$
$P^7(N_2, L) = 0$



No $(7, 1)$-anti-alignment
$P^7(N_3, L) = 1$

Handling Models with Loops

A model with an executable loop has

- arbitrary long runs
- runs arbitrary far from any finite log

Drop the bound n, but penalize long runs when looking for the optimal.

$$P^{\epsilon}(N, L) \stackrel{\text{def}}{=} 1 - \sup_{\gamma \in \mathcal{L}(N)} \frac{dist(\gamma, L)}{(1+\epsilon)^{|\gamma|}}$$

with some $\epsilon \geq 0$ which is a parameter of this definition.

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

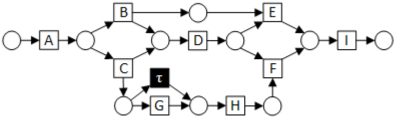
For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$



Best anti-alignment	$max^7(N, L)$	$P^7(N, L)$
$\langle A, B, D, E, I \rangle$	4	$\frac{3}{7}$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

For every N, L and for every $\sigma \in \mathcal{L}(N)$,

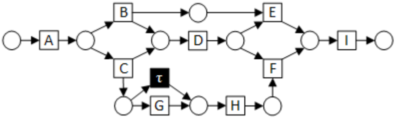
$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:

$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, B, D, E, I \rangle$



Best anti-alignment $max^7(N, L)$ $P^7(N, L)$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

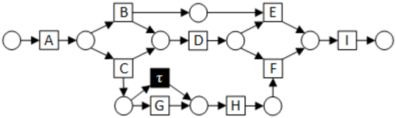
For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, B, D, E, I \rangle$



Best anti-alignment	$max^7(N, L)$	$P^7(N, L)$
$\langle A, C, H, D, F, I \rangle$	2	$\frac{5}{7}$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

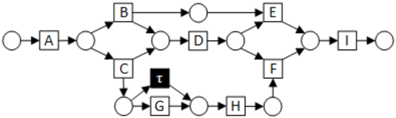
For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:
$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, B, D, E, I \rangle$
$\langle A, C, D, H, F, I \rangle$



Best anti-alignment $max^7(N, L)$ $P^7(N, L)$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

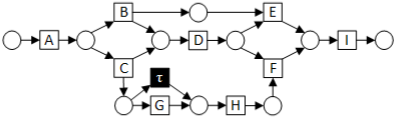
For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:
$\langle A, C, D, G, H, F, I\rangle$
$\langle A, C, G, D, H, F, I\rangle$
$\langle A, B, D, E, I\rangle$
$\langle A, C, D, H, F, I\rangle$



Best anti-alignment	$max^7(N, L)$	$P^7(N, L)$
$\langle A, C, H, D, F, I\rangle$	2	$\frac{5}{7}$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

For every N, L and for every $\sigma \in \mathcal{L}(N)$,

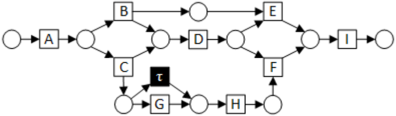
$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:

$\langle A, C, D, G, H, F, I \rangle$
$\langle A, C, G, D, H, F, I \rangle$
$\langle A, B, D, E, I \rangle$
$\langle A, C, D, H, F, I \rangle$
$\langle A, C, H, D, F, I \rangle$



Best anti-alignment $max^7(N, L)$ $P^7(N, L)$

Monotonicity w.r.t. New Observations

Observing a new trace which happens to be already a run of the model, can only increase the precision measure.

Theorem

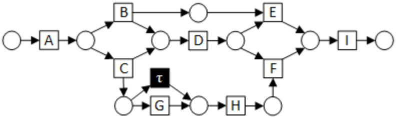
For every N, L and for every $\sigma \in \mathcal{L}(N)$,

$$P^n(N, L \cup \{\sigma\}) \geq P^n(N, L)$$

Hint: every (n, m)-anti-alignment for $(N, L \cup \{\sigma\})$ is also a (n, m)-anti-alignment for (N, L).

Example

Log L:
$\langle A, C, D, G, H, F, I\rangle$
$\langle A, C, G, D, H, F, I\rangle$
$\langle A, B, D, E, I\rangle$
$\langle A, C, D, H, F, I\rangle$
$\langle A, C, H, D, F, I\rangle$



Best anti-alignment	$max^7(N, L)$	$P^7(N, L)$
$\langle A, C, G, H, D, F, I\rangle$	1	$\frac{6}{7}$

Monotonicity w.r.t. Model Language

Theorem

Given two models N_1 and N_2, if $\mathcal{L}(N_1) \subseteq \mathcal{L}(N_2)$, then N_1 is more precise than N_2.

$$\mathcal{L}(N_1) \subseteq \mathcal{L}(N_2) \implies P^n(N_1, L) \geq P^n(N_2, L)$$

Implementation

Formula $\Phi_m^n(N, L)$ states that γ is a (n, m)-anti-alignment:

- $\gamma = \lambda(t_1) \dots \lambda(t_n) \in \mathcal{L}(N)$, and
- for every $\sigma \in L$, $dist(\gamma, \sigma) \geq m$.

Encoding in SAT

$\Phi_m^n(N, L)$ is coded using the following Boolean variables:

- $\tau_{i,t}$ for $i = 1 \dots n$, $t \in T$ means that transition $t_i = t$.
- $m_{i,p}$ for $i = 0 \dots n$, $p \in P$ means that place p is marked in marking M_i (safe Petri nets: Boolean variables)
- $\delta_{i,j,\sigma}$ to encode the distances $dist(\gamma, \sigma)$.

Total size for the SAT encoding of the formula $\Phi_m^n(N, L)$:

$$O\left(n \times |T| \times \left(|N| + m^2 \times |L|\right)\right)$$

Encoding in SAT (1) $\gamma = \lambda(t_1)\ldots\lambda(t_n) \in \mathcal{L}(N)$

- Initial marking:

$$\left(\bigwedge_{p\in M_0} m_{0,p}\right) \wedge \left(\bigwedge_{p\in P\setminus M_0} \neg m_{0,p}\right)$$

- One and only one t_i for each i:

$$\bigwedge_{i=1}^{n} \bigvee_{t\in T} (\tau_{i,t} \wedge \bigwedge_{t'\in T} \neg\tau_{i,t'})$$

- The transitions are enabled when they fire:

$$\bigwedge_{i=1}^{n} \bigwedge_{t\in T} (\tau_{i,t} \implies \bigwedge_{p\in{}^\bullet t} m_{i-1,p})$$

- Token game (for safe Petri nets):

$$\bigwedge_{i=1}^{n} \bigwedge_{t\in T} \bigwedge_{p\in t^\bullet} (\tau_{i,t} \implies m_{i,p})$$

$$\bigwedge_{i=1}^{n} \bigwedge_{t\in T} \bigwedge_{p\in {}^\bullet t\setminus t^\bullet} (\tau_{i,t} \implies \neg m_{i,p})$$

$$\bigwedge_{i=1}^{n} \bigwedge_{t\in T} \bigwedge_{p\in P, p\notin {}^\bullet t, p\notin t^\bullet} (\tau_{i,t} \implies (m_{i,p} \iff m_{i-1,p}))$$

Encoding in SAT (2) $\mathit{dist}(\gamma, \sigma) \geq m$

Encoding in SAT (2) $\mathit{dist}(\gamma, \sigma) \geq m$

- For Hamming distance: easy

Encoding in SAT (2) $dist(\gamma, \sigma) \geq m$

- For Hamming distance: easy
- For Levenshtein's distance:
  Use same relations as the classical algorithm:

$$
\begin{aligned}
&dist(\langle u_1, \ldots, u_i \rangle, \epsilon) = i \\
&dist(\epsilon, \langle v_1, \ldots, v_j \rangle) = j \\
&dist(\langle u_1, \ldots, u_{i+1} \rangle, \langle v_1, \ldots, v_{j+1} \rangle) = \\
&\quad \begin{cases}
dist(\langle u_1, \ldots, u_i \rangle, \langle v_1, \ldots, v_j \rangle) & \text{if } u_{i+1} = v_{j+1} \\
1 + \min(dist(\langle u_1, \ldots, u_{i+1} \rangle, \langle v_1, \ldots, v_j \rangle), \; dist(\langle u_1, \ldots, u_i \rangle, \langle v_1, \ldots, v_{j+1} \rangle)) & \text{if } u_{i+1} \neq v_{j+1}
\end{cases}
\end{aligned}
$$

Encoding as SAT formula using variables $\delta_{i,j,d}$

$\delta_{i,j,d} = \texttt{true}$ means $dist(\langle u_1 \ldots u_i \rangle, \langle v_1 \ldots v_j \rangle) \geq d$.

$$\delta_{0,0,0} \quad \wedge \quad \bigwedge_{d>0} \neg\delta_{0,0,d} \tag{1}$$

$$\bigwedge_d \bigwedge_{i=0}^{n} \quad (\delta_{i+1,0,d+1} \Leftrightarrow \delta_{i,0,d}) \tag{2}$$

$$\bigwedge_d \bigwedge_{j=0}^{n} \quad (\delta_{0,j+1,d+1} \Leftrightarrow \delta_{0,j,d}) \tag{3}$$

$$\bigwedge_d \bigwedge_{i,j \text{ s.t. } u_{i+1}=v_{j+1}} \delta_{i+1,j+1,d} \Leftrightarrow \delta_{i,j,d} \tag{4}$$

$$\bigwedge_d \bigwedge_{i,j \text{ s.t. } u_{i+1}\neq v_{j+1}} \delta_{i+1,j+1,d+1} \Leftrightarrow (\delta_{i+1,j,d} \wedge \delta_{i,j+1,d}) \tag{5}$$

Experiments: Alignments (showing averages)

Model			$\|L\|$	Size of run	Maximal number of editions	Formula construction time (sec)	Total execution time (sec)
Reference	$\|T\|$	$\|P\|$					
Fig. 2	8	7	100	7	5	0.239	0.349
M8 of [25]	15	17	100	PRE: 20	LIM:10	10.139	15.530
M1 of [25]	40	39	100	PRE: 7	LIM:10	4.924	7.16
Loan [10]	15	16	100	PRE: 19	LIM: 10	14.047	20.915

Experiments: Anti-alignments

Model			$\|L\|$	Size of run	Maximal number of editions	Formula construction time (sec)	Total execution time (sec)
Reference	$\|T\|$	$\|P\|$					
Fig. 2	8	7	10	8	LIM: 10	13.802	21.502
			100	8	LIM: 10	137.213	243.842
M8 of [25]	15	17	10	18	LIM:10	103.812	148.271
			100	PRE: 10	LIM: 10	343.529	496.733
M1 of [25]	40	39	10	39	LIM:10	1337.806	2069.505
			100	PRE:13	LIM:5	680.556	995.361
Loan [10]	15	16	10	PRE: 19	LIM: 10	140.840	203.257
			100	PRE:19	LIM: 10	1526.048	2185.785

Experiments: Anti-alignments (Hamming distance)

benchmark	$\|P\|$	$\|T\|$	$\|L\|$	$\|A_L\|$	n	m	$\Phi^n_m(N,L)$	$\min_m(N,L)$	$\max^n(N,L)$
prAm6	347	363	761	272	41	1	✓	3	39
						5	✓	7	
					21	1	✓	3	19
						5	✓	7	
			1200	363	41	1	✓	4	19
						5	✓	8	
					21	1	✓	4	15
						5	✓	8	
BankTransfer	121	114	989	101	51	1	✓	8	32
						10	✓	17	
					21	1	✓	8	14
						10	✓	17	
			2000	113	51	1	✓	15	16
						10	✓	37	
					21	1	✓	15	5
						10	✗	37	

Experiments: Multi-alignments

Model			$\lvert L\rvert$	Size of run	Maximal number of editions	Formula construction time (sec)	Total execution time (sec)
Reference	$\lvert T\rvert$	$\lvert P\rvert$					
Fig. 2	8	7	10	8	7	10.101	15.362
			100	8	7	99.602	200.569
M8 of [25]	15	17	10	18	LIM:6	252.471	414.174
			100	PRE:15	LIM:6	516.391	741.162
M1 of [25]	40	39	10	PRE: 13	LIM:10	115.706	172.500
			100	PRE: 13	LIM: 5	681.95	1066.94
Loan [10]	15	16	10	PRE: 19	15	252.572	373.683
			100	PRE: 9	LIM:10	359.982	508.542

Conclusion

Anti-alignment

- Run of the model which maximizes its distance to the observed traces
- New metric for precision in process mining
  - monotonic w.r.t. new observations

Implementations

- DarkSider (using SAT encoding)
  `www.lsv.ens-cachan.fr/~chatain/darksider`
- Also available in ProM
  `www.promtools.org`

SAT-based approach for conformance checking

- Very flexible
- Good for prototyping
- Efficiency depends a lot on precise problem and encoding

Thank you!