Process Mining

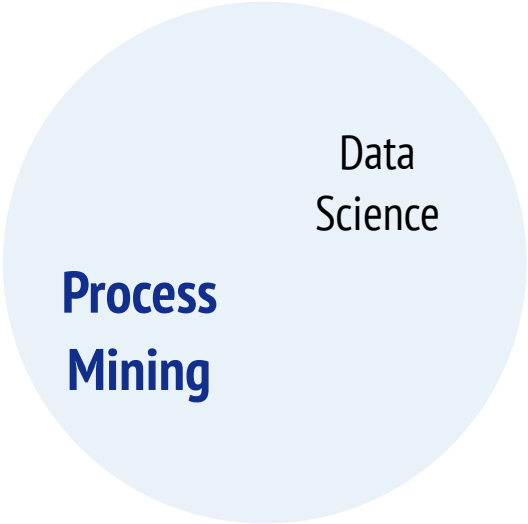
Generalized Alignment-Based Trace Clustering for Process Behavior

Mathilde Boltenhagen[1], Thomas Chatain[1], Josep Carmona[2]

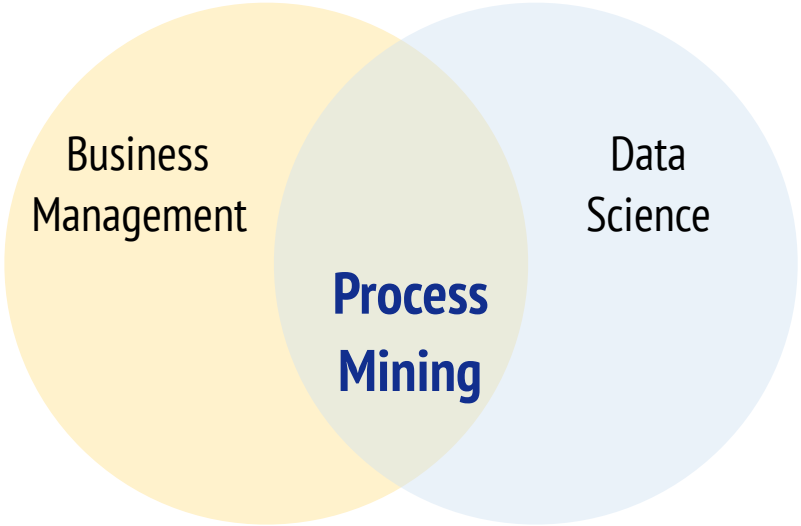
[1]LSV, CNRS, ENS Paris-Saclay, Inria, Université Paris-Saclay
[2]Universitat Politècnica de Catalunya

Process Mining

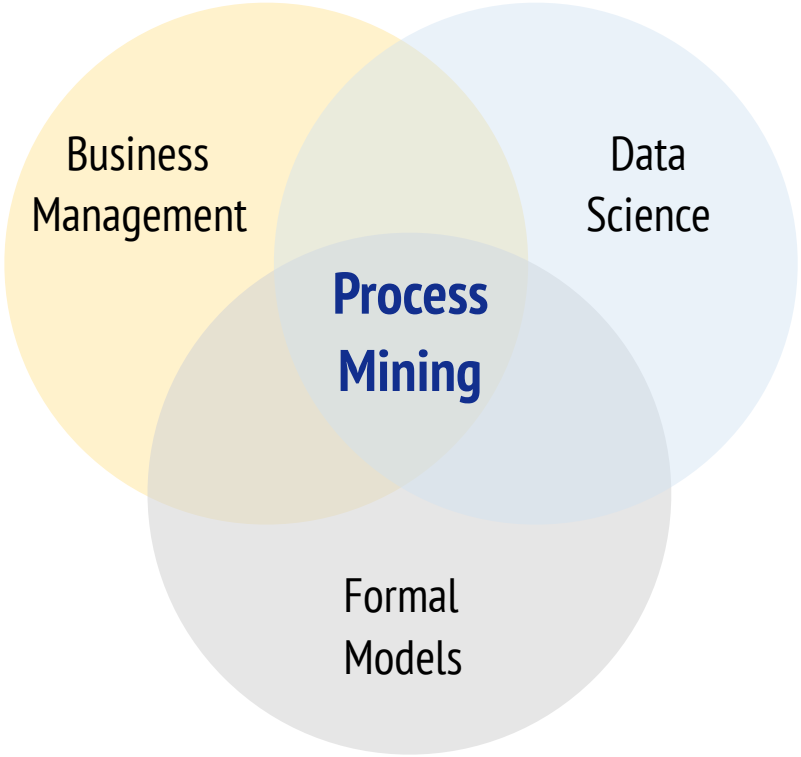
Process Mining



Process Mining



Process Mining



Motivation example

Motivation example

Source: https://www.pexels.com/zh-tw/

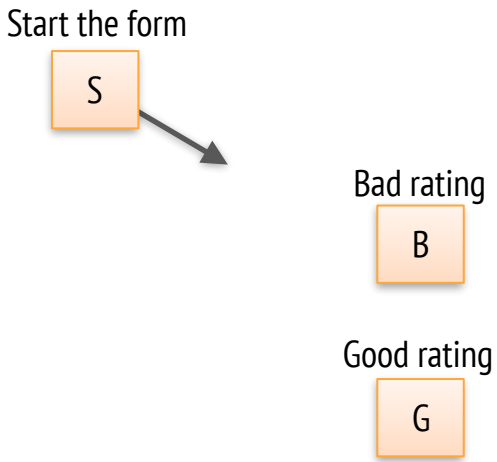
Motivation example

Start the form

S

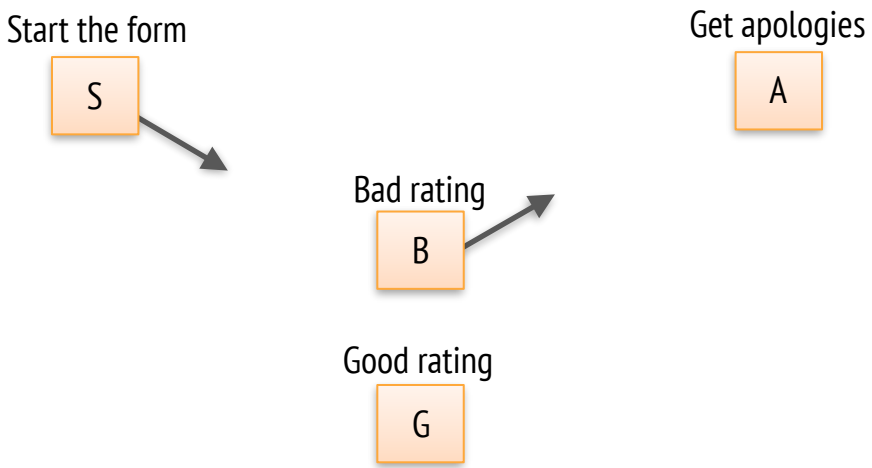
Source: https://www.pexels.com/zh-tw/

Motivation example



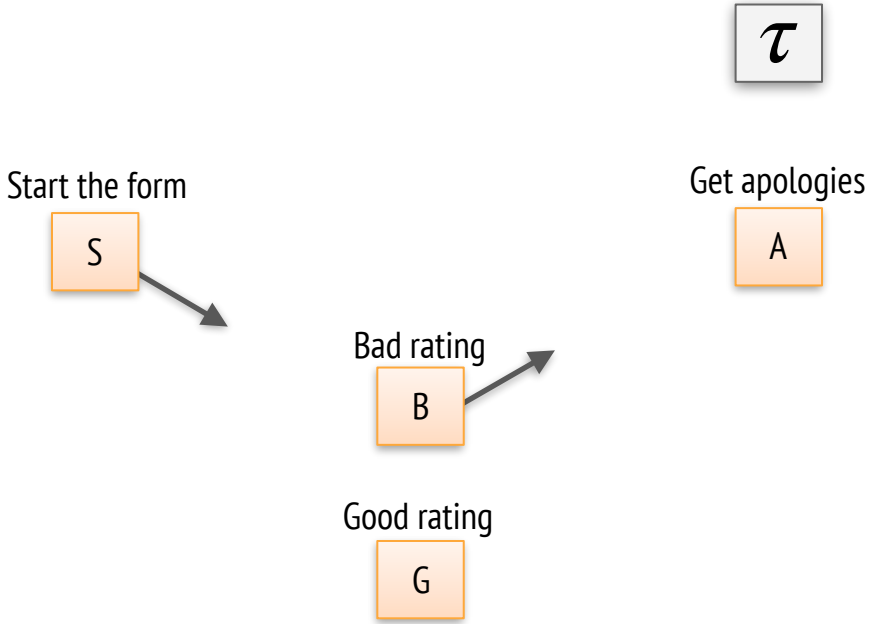
Source: https://www.pexels.com/zh-tw/

Motivation example



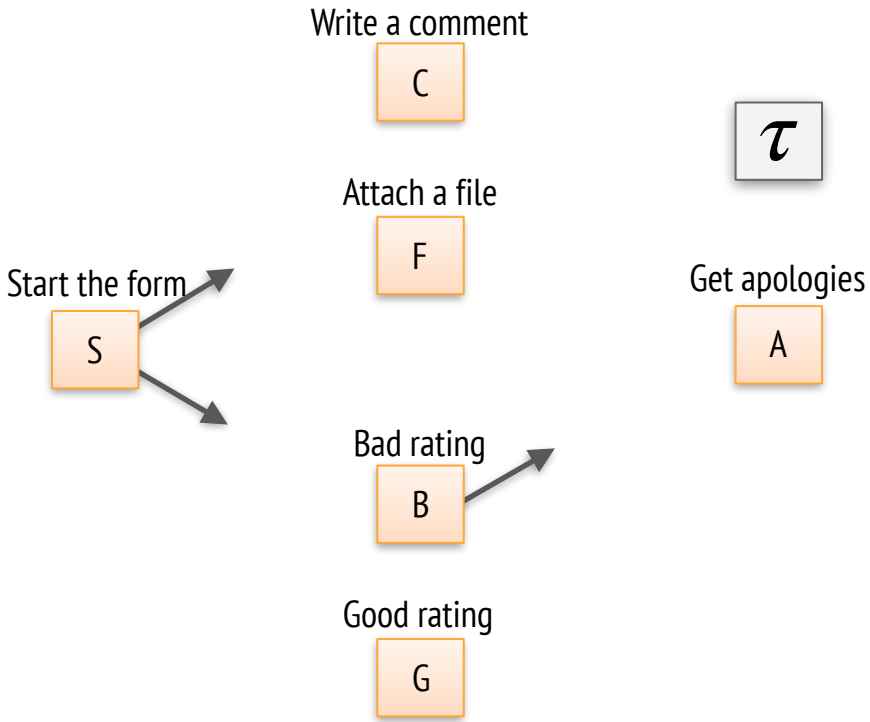
Source: https://www.pexels.com/zh-tw/

Motivation example



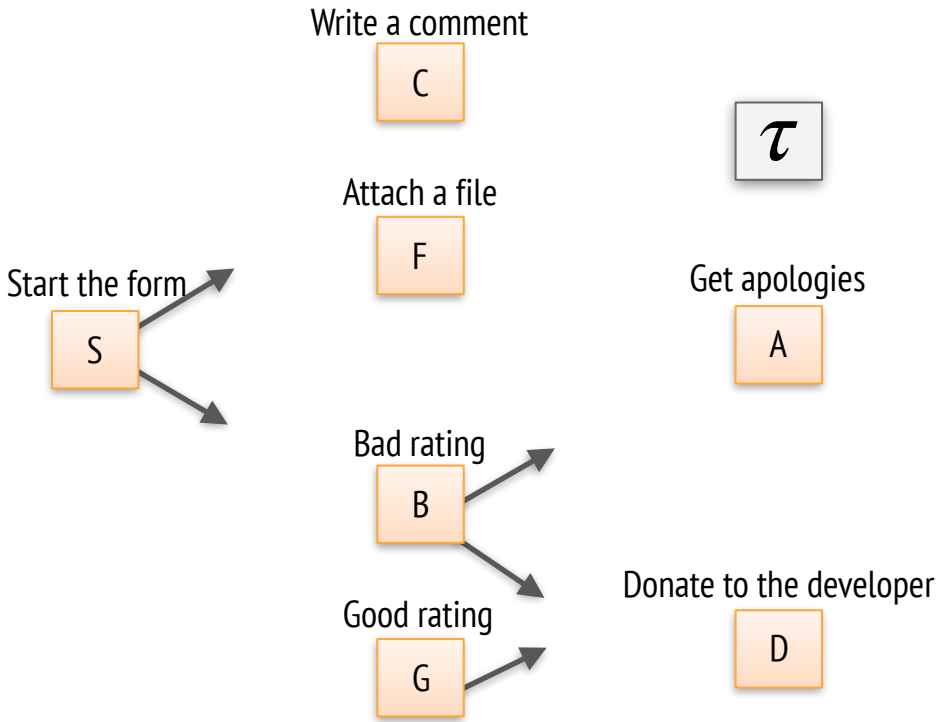
Source: https://www.pexels.com/zh-tw/

Motivation example



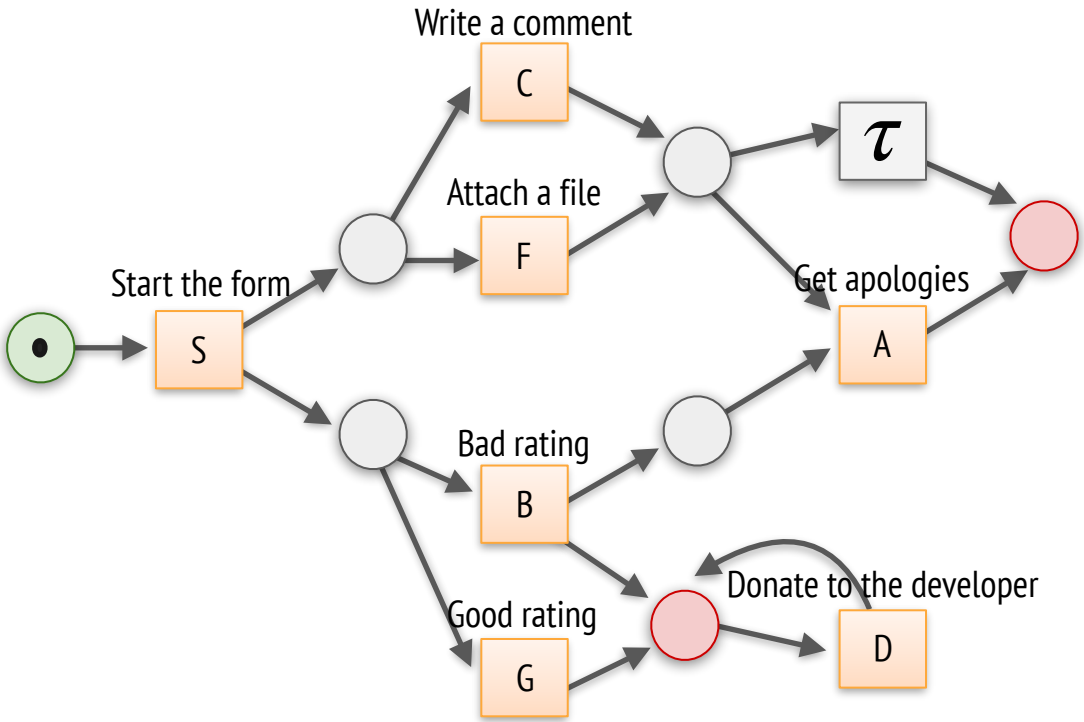
Source: https://www.pexels.com/zh-tw/

Motivation example



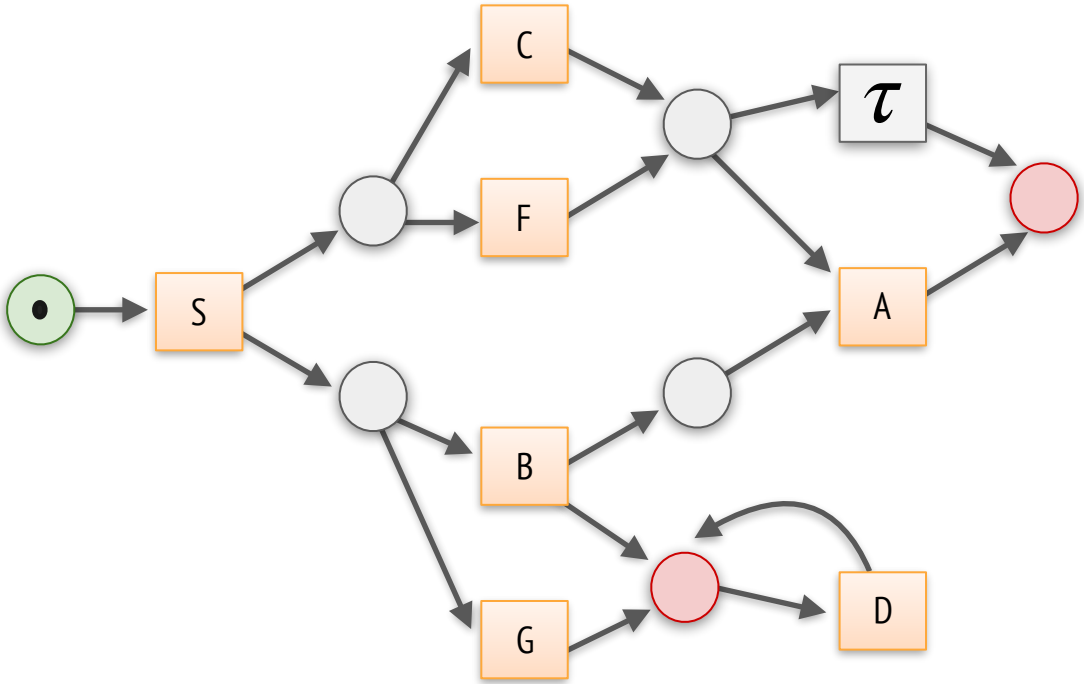
Source: https://www.pexels.com/zh-tw/

Motivation example



Source: https://www.pexels.com/zh-tw/

Motivation example



Source: https://www.pexels.com/zh-tw/

Trace clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

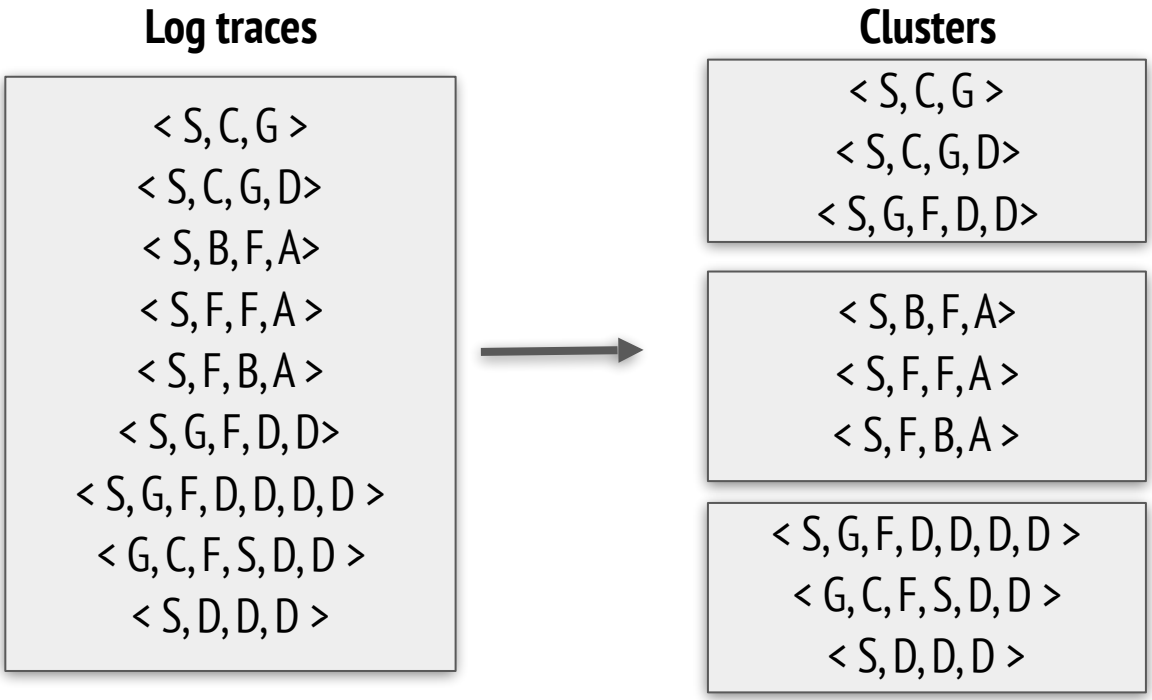
Trace clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

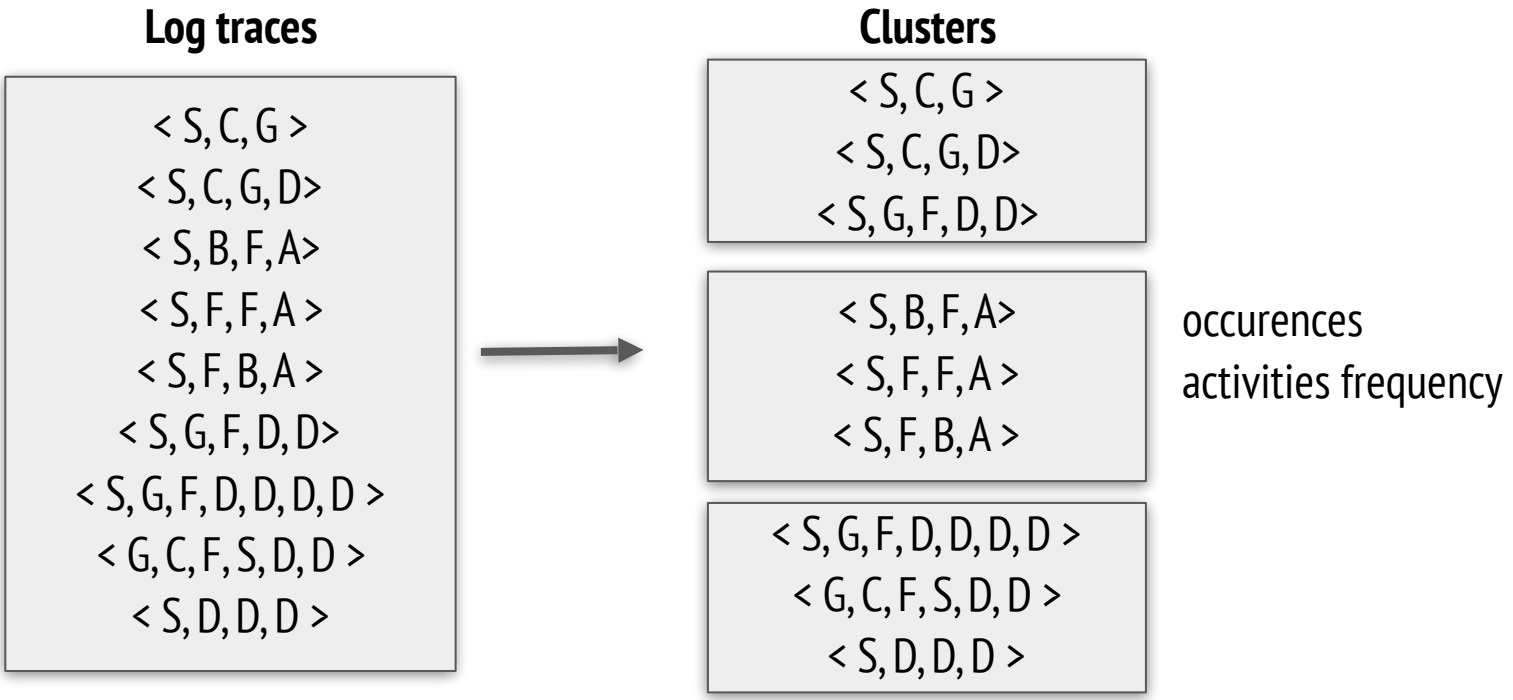
Data clustering is the task of grouping objects by similarity.

Trace clustering



[Greco et al. 2006] ; [Ferreira et al. 2007] ; [Bose et al. 2009]

Trace clustering



[Greco et al. 2006] ; [Ferreira et al. 2007] ; [Bose et al. 2009]

Model-based Trace Clustering

New idea : to cluster data based on an existing process model

> highlight parts of models that are executed

> show deviating traces

> model repair

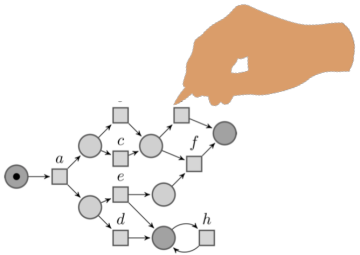
[Chatain et al. 2017]

Model-based Trace Clustering

Existing models

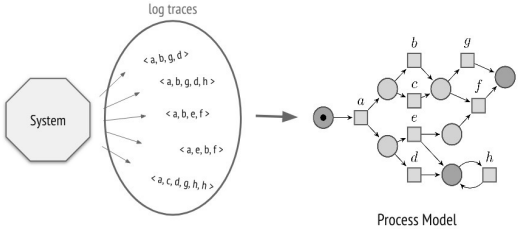
> Available process model

- system modelled by hand
- model used at design phase



> Discovered process model

- data are used to create the process model : process discovery (Alpha Miner, ILP miner, Inductive miner..)



Alignment-based Trace Clustering

Log traces

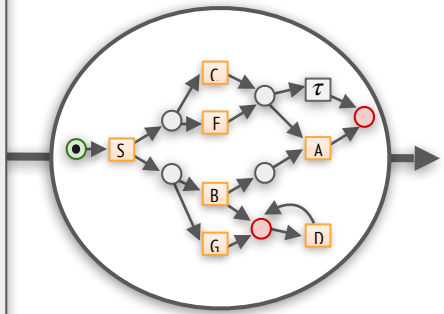
< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

[Chatain et al. 2017]

Alignment-based Trace Clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

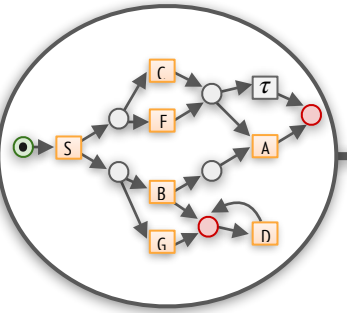


[Chatain et al. 2017]

Alignment-based Trace Clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



Clusters

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >

< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

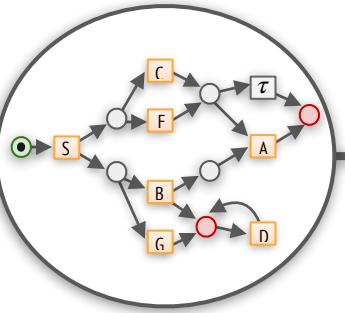
< G, C, F, S, D, D >
< S, D, D, D >

[Chatain et al. 2017]

Alignment-based Trace Clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



Clusters

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >

< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

< G, C, F, S, D, D >
< S, D, D, D >

Centroids : runs

< S, C, τ, G >

< S, B, F, A >

< S, F, B, A >

< S, G, F, D, D>

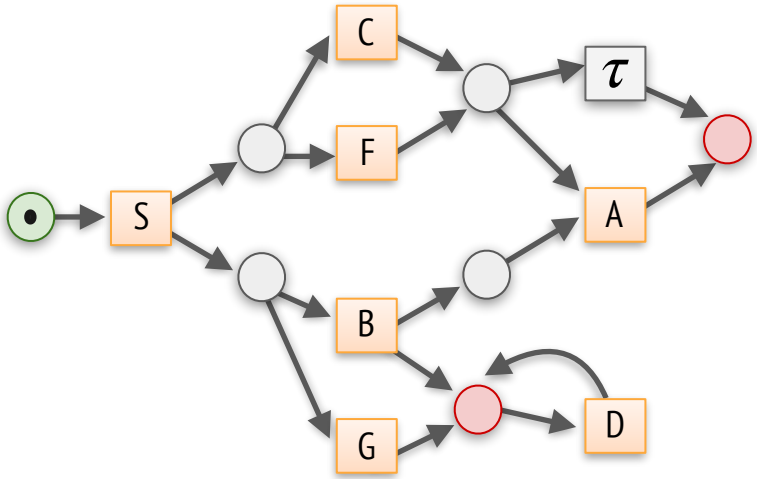
< S, G, F, D, D, D, D >

Non-clustered

[Chatain et al. 2017]

Alignment-based Trace Clustering

Full runs as centroids



Example of full run : < S, C, τ ,G>

Alignment-based Trace Clustering

Full runs as centroids

$\mathcal{N}$ a process model, $\mathcal{L}$ a log :

Find $\mathcal{C} = (\{u_1 \dots u_n\}, \mathcal{X})$:

$\{u_1 \dots u_n\} \in \text{Runs}(\mathcal{N})$

$\mathcal{X} : \mathcal{L} \rightarrow \{nc, u_1 \dots u_n\}$

Such that for every $\sigma \in \mathcal{L}$, $dist(\sigma, \mathcal{X}(\sigma))$ is small

Alignment-based Trace Clustering

Full runs as centroids

$\mathcal{N}$ a process model, $\mathcal{L}$ a log :

Find $\mathcal{C} = (\{u_1 \dots u_n\}, \mathcal{X})$:

$\{u_1 \dots u_n\} \in \text{Runs}(\mathcal{N})$

$\mathcal{X} : \mathcal{L} \longrightarrow \{nc, u_1 \dots u_n\}$

Such that for every $\sigma \in \mathcal{L}$, $dist(\sigma, \mathcal{X}(\sigma))$ is small

Example :

$\sigma 1 = < S, C, G >$

$\sigma 2 = < S, C, G, D >$

Alignment-based Trace Clustering

Full runs as centroids

$\mathcal{N}$ a process model, $\mathcal{L}$ a log :

Find $\mathcal{C} = (\{u_1 \dots u_n\}, \mathcal{X})$:

$\{u_1 \dots u_n\} \in \text{Runs}(\mathcal{N})$

$\mathcal{X} : \mathcal{L} \longrightarrow \{nc, u_1 \dots u_n\}$

Such that for every $\sigma \in \mathcal{L}$, $dist(\sigma, \mathcal{X}(\sigma))$ is small

Example :

$\sigma 1 = < S, C, G >$

$\sigma 2 = < S, C, G, D >$

$\mathbf{u = < S, C, \tau, G >}$

$dist(\sigma 1, u) = 0$

$dist(\sigma 2, u) = 1$

Alignment-based Trace Clustering

Full runs as centroids

$\mathcal{N}$ a process model, $\mathcal{L}$ a log :

Find $\mathcal{C} = (\{u_1 \dots u_n\}, \mathcal{X})$:

$\{u_1 \dots u_n\} \in \text{Runs}(\mathcal{N})$

$\mathcal{X} : \mathcal{L} \longrightarrow \{nc, u_1 \dots u_n\}$

Such that for every $\sigma \in \mathcal{L}$, $dist(\sigma, \mathcal{X}(\sigma))$ is small

Example :

$\sigma 1 = < S, C, G >$

$\sigma 2 = < S, C, G, D >$

$\mathbf{u = < S, C, \tau, G >}$

$dist(\sigma 1, u) = 0$

$dist(\sigma 2, u) = 1$

dist is a distance between words (Hamming distance, Edit distance..)

Alignment-based Trace Clustering

> Pseudo SAT formulas

> solver : minisat+

Alignment-based Trace Clustering

> Pseudo SAT formulas

> solver : minisat+

Example of parts of the formula :

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

transition a fires at instant i

Alignment-based Trace Clustering

> Pseudo SAT formulas

> solver : minisat+

Example of parts of the formula :

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

transition a fires at instant i

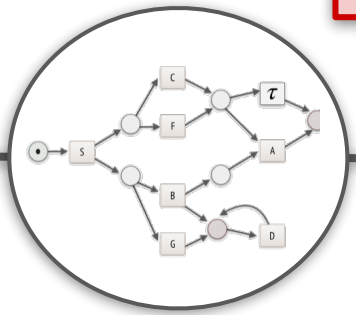
$$\bigwedge_{n=1}^{k} \bigwedge_{\sigma \in L} (\chi_{n,\sigma} \wedge \bigwedge_{n \neq m} \neg \chi_{m,\sigma})$$

trace σ is in cluster u

Alignment-based Trace Clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



Clusters

The use of full runs as centroids does not allow concurrency

< S, B, F, A>
< S, F, F, A >

< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

< G, C, F, S, D, D >
< S, D, D, D >

ntroids : runs

τ, G >

< S, B, F, A >

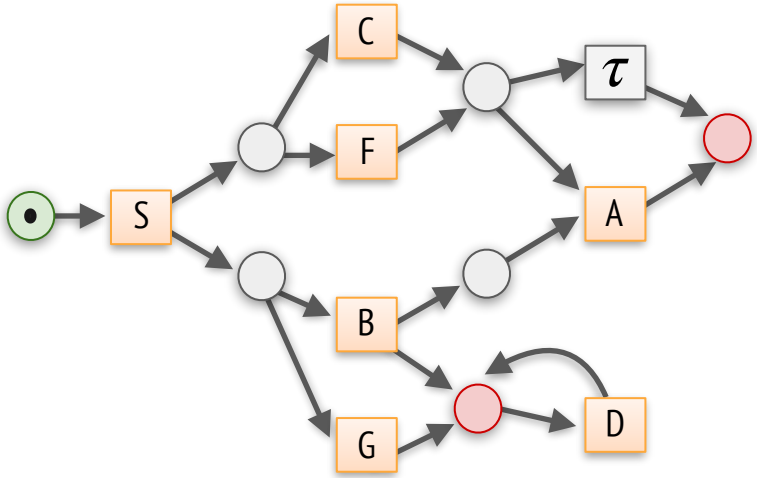
< S, F, B, A >

< S, G, F, D, D>

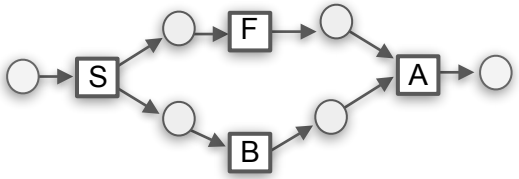
< S, G, F, D, D, D, D >

Non-clustered

Processes as centroids



Example of process :



Linearizations of the process :

< S, B, F, A >

< S, F, B, A >

[Engelfriet 1991]

Processes as centroids

$\mathcal{N}$ a process model, $\mathcal{L}$ a log :

Find $\mathcal{C} = (\{P_1 \dots P_n\}, \mathcal{X})$:

$\{P_1 \dots P_n\} \in \text{Processes}(\mathcal{N})$

$\mathcal{X} : \mathcal{L} \rightarrow \{nc, P_1 \dots P_n\}$

Such that for every $\sigma \in \mathcal{L}$, $dist(\sigma, \mathcal{X}(\sigma))$ is small

Processes as centroids

$\mathcal{N}$ a process model, $\mathcal{L}$ a log :

Find $\mathcal{C} = (\{P_1 \dots P_n\}, \mathcal{X})$:

$\{P_1 \dots P_n\} \in \text{Processes}(\mathcal{N})$

$\mathcal{X} : \mathcal{L} \rightarrow \{nc, P_1 \dots P_n\}$

Such that for every $\sigma \in \mathcal{L}$, $dist(\sigma, \mathcal{X}(\sigma))$ is small

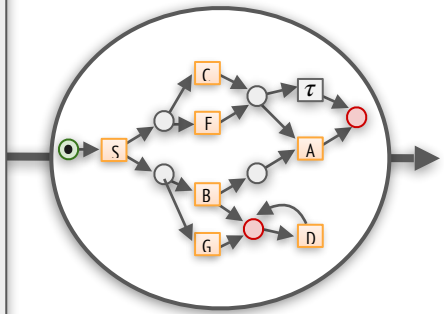
dist is the minimal distance between a **linearization of P** and the trace (computed as distance between words : Hamming distance, Edit distance..)

Alignment and Model Partial-Order-based Trace Clustering

Processes as centroids

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



Clusters

< S, C, G >
< S, C, G, D>

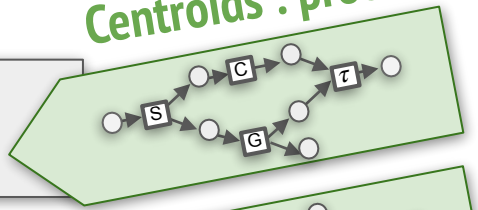
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >

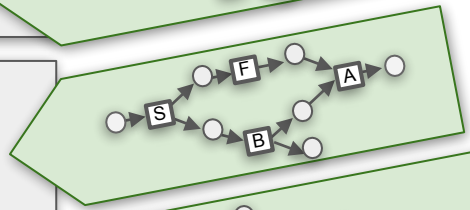
< S, G, F, D, D>

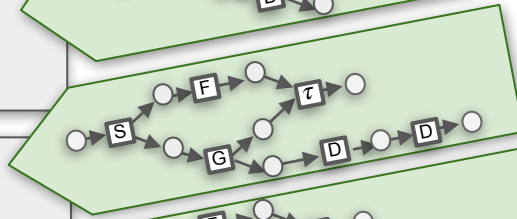
< S, G, F, D, D, D, D >

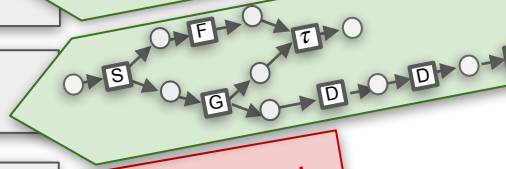
< G, C, F, S, D, D >
< S, D, D, D >

Centroids : processes









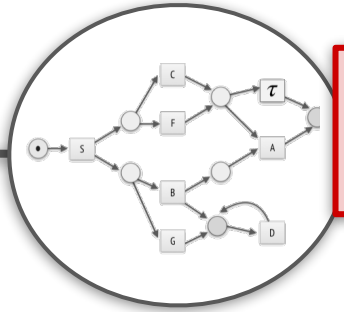
Non-clustered

[Boltenhagen 2019]

Alignment and Model Partial-Order-based Trace Clustering

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



Clusters

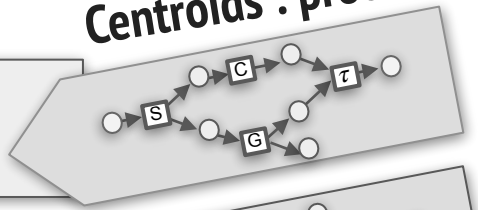
< S, C, G >
< S, C, G, D>

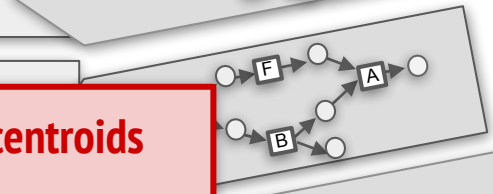
< S, G, F, D, D>

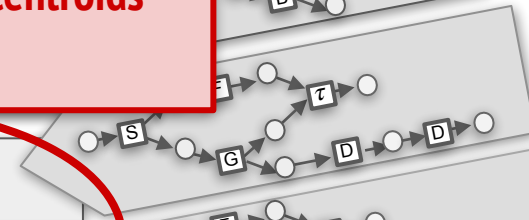
< S, G, F, D, D, D, D >

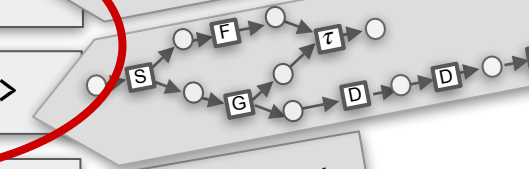
< G, C, F, S, D, D >
< S, D, D, D >

Centroids : processes







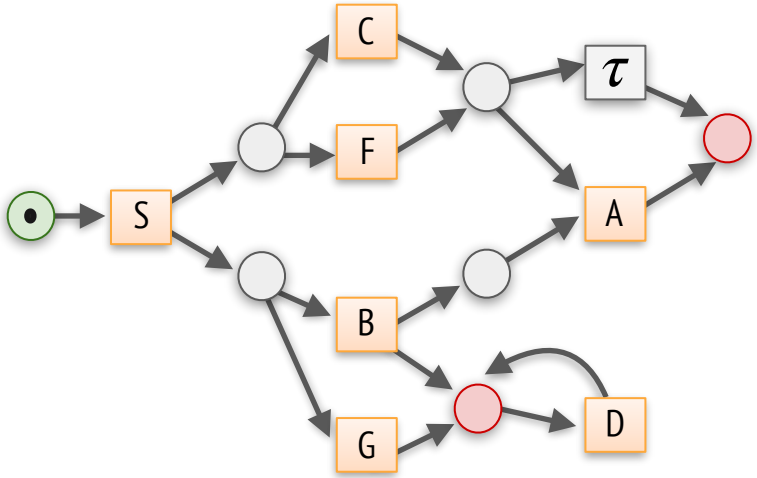


Non-clustered

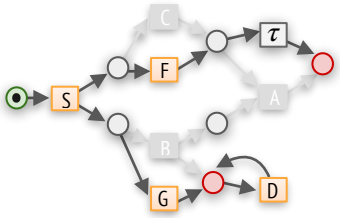
The use of processes as centroids does not allow loops

Alignment and Model Subnet-based Trace Clustering

Subnets as centroids



Example of subnet :



Runs of the subnet :

< S, G, F, τ, D>

< S, F, τ, G, D, D>

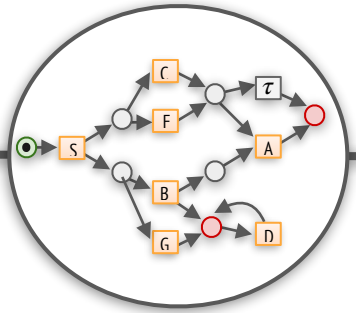
< S, G, D, D, D, D, F, τ>

Alignment and Model Subnet-based Trace Clustering

Subnets as centroids

Log traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



Clusters

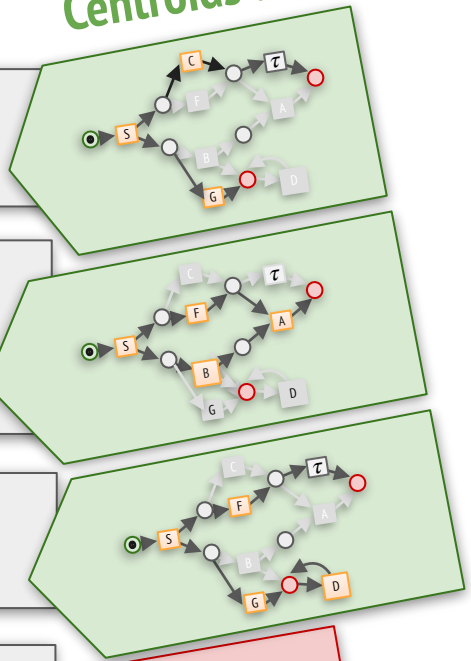
< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >
< S, F, B, A >

< S, G, F, D, D>
< S, G, F, D, D, D, D >

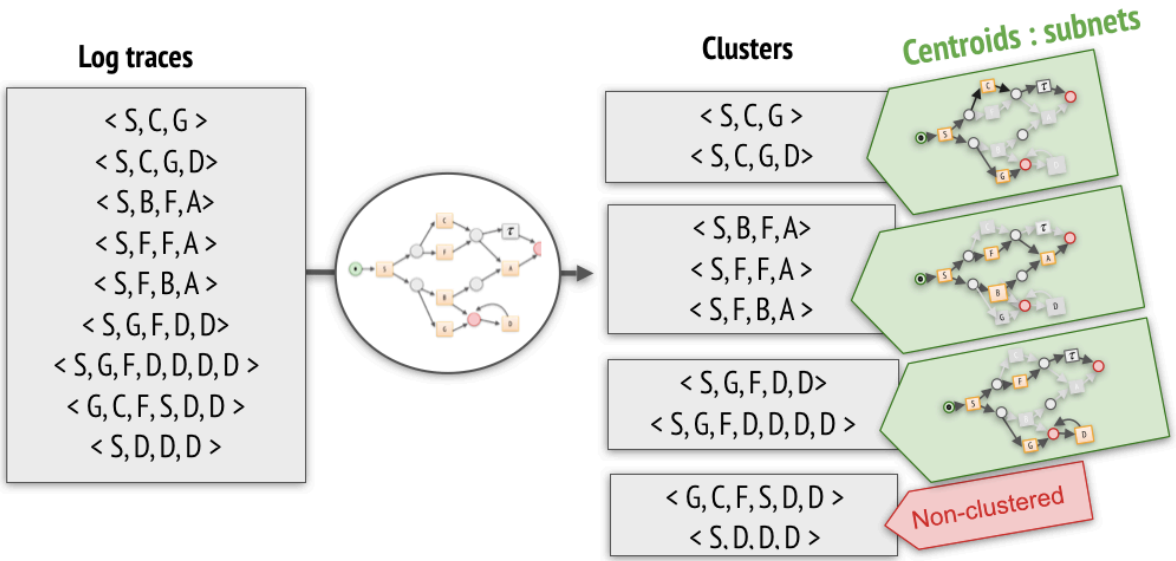
< G, C, F, S, D, D >
< S, D, D, D >

Centroids : subnets



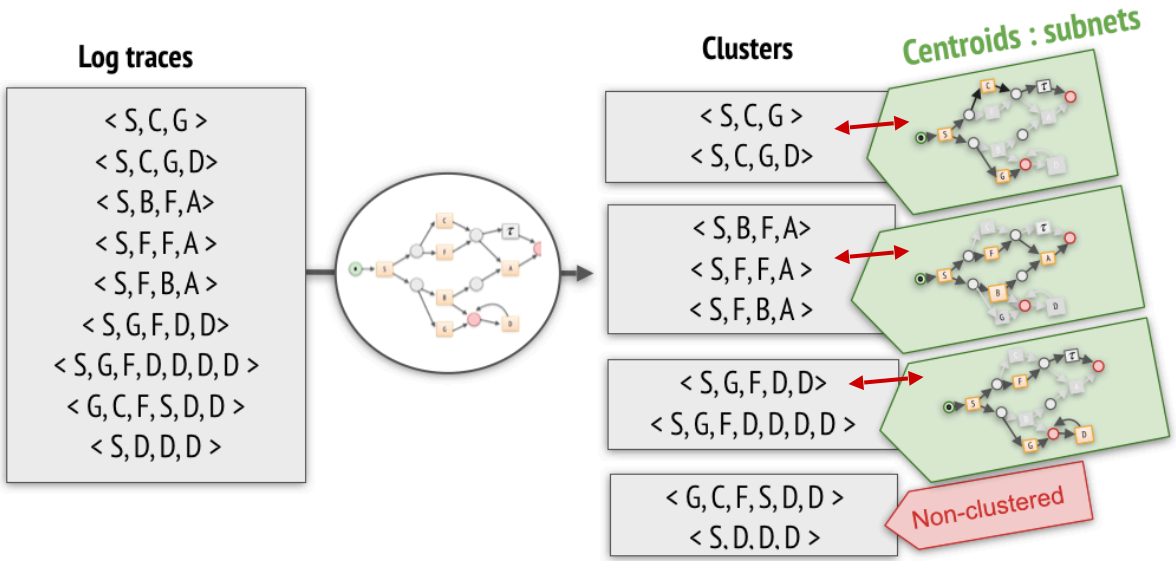
Non-clustered

Quality criteria



Quality criteria

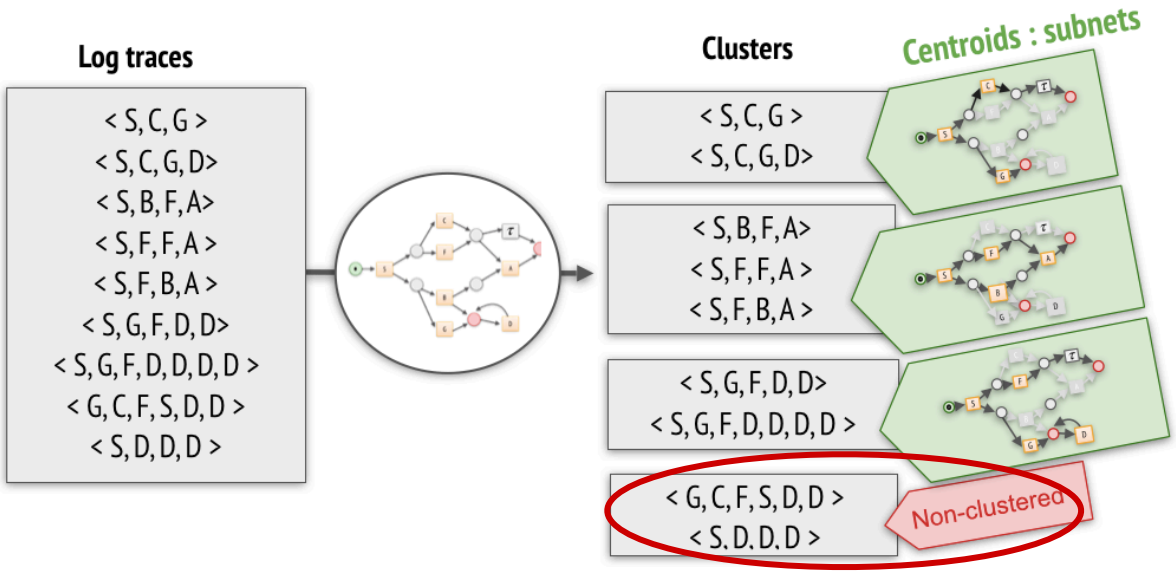
> Distance between traces and their centroid



Quality criteria

> Distance between traces and their centroid

> Number of non-clustered traces

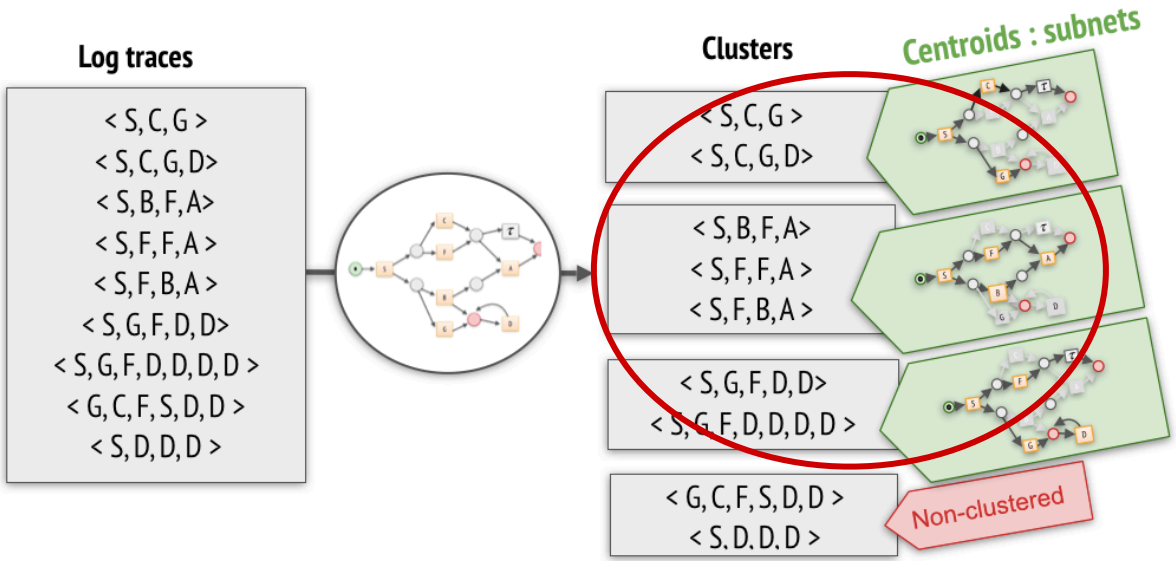


Quality criteria

> Distance between traces and their centroid

> Number of non-clustered traces

> Number of clusters



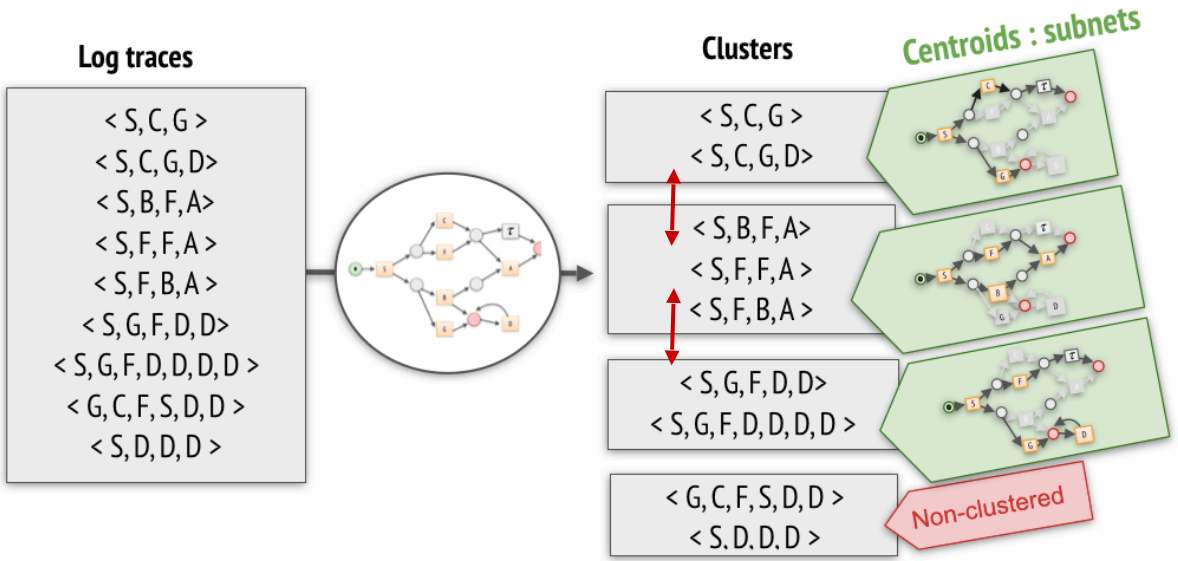
Quality criteria

> Distance between traces and their centroid

> Number of non-clustered traces

> Number of clusters

> Inter-cluster distance



Quality criteria

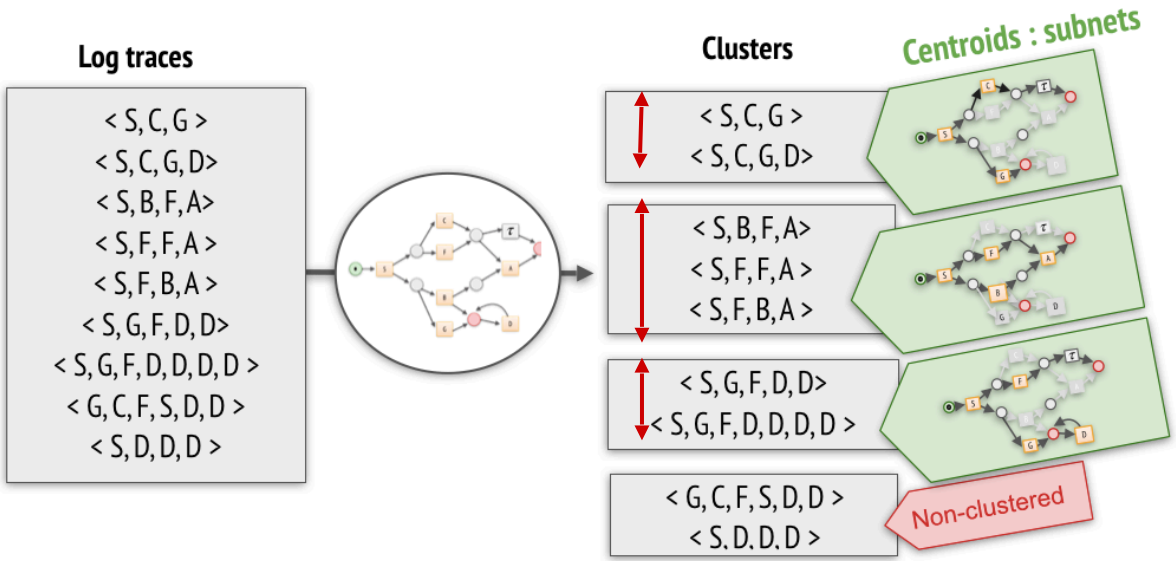
> Distance between traces and their centroid

> Number of non-clustered traces

> Number of clusters

> Inter-cluster distance

> Intra-cluster distance



Quality criteria

Inter-cluster distance

> maximise differences between clusters

Full runs centroids

$$\Phi(\mathcal{C}) \stackrel{\text{def}}{=} \min_{i \neq j} dist(u_i, u_j)$$

distance between centroids computed as distance between words (Hamming distance, Edit distance..)

Quality criteria

Inter-cluster distance

> maximise differences between clusters

Full runs centroids

$$\Phi(\mathcal{C}) \stackrel{\text{def}}{=} \min_{i \neq j} dist(u_i, u_j)$$

distance between centroids computed as distance between words (Hamming distance, Edit distance..)

Processes centroids

$$\Phi(\mathcal{C}) = \min_{i \neq j} dist(\mathcal{P}_i, \mathcal{P}_j)$$

$$dist(\mathcal{P}, \mathcal{P}') \stackrel{\text{def}}{=} \min_{\substack{u \in Runs(\mathcal{P}) \\ u' \in Runs(\mathcal{P}')}} dist(u, u')$$

distance between centroids is the minimal distance for any linearizations of the processes

Subnets centroids

$$\Phi(\mathcal{C}) = \min_{i \neq j} dist(\mathcal{N}_i, \mathcal{N}_j)$$

$$dist(\mathcal{N}, \mathcal{N}') \stackrel{\text{def}}{=} \min_{\substack{u \in Runs(\mathcal{N}) \\ u' \in Runs(\mathcal{N}')}} dist(u, u').$$

distance between centroids is the minimal distance for any full runs of the subnets

Quality criteria

Intra-cluster distance

> only for subnet centroids

> minimize differences in a cluster

$$\Theta(\mathcal{C}) \stackrel{\text{def}}{=} \max_{k} \left(\sup_{\mathcal{P},\mathcal{P}' \in Proc(\mathcal{N}_k)} dist(\mathcal{P}, \mathcal{P}') \right)$$

Quality criteria

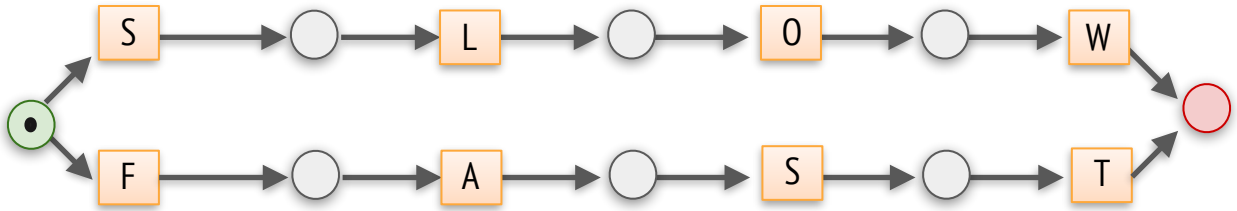
Intra-cluster distance

> only for subnet centroids

> minimize differences in a cluster

$$\Theta(\mathcal{C}) \stackrel{\text{def}}{=} \max_{k} \left(\sup_{\mathcal{P},\mathcal{P}' \in Proc(\mathcal{N}_k)} dist(\mathcal{P}, \mathcal{P}') \right)$$

A subnet with very different behaviors :



Quality criteria

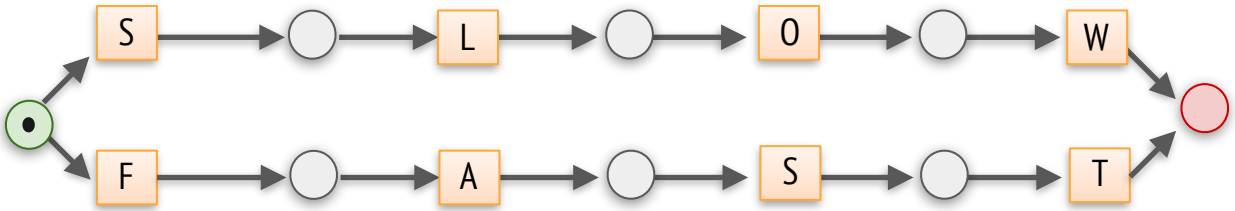
Intra-cluster distance

> only for subnet centroids

> minimize differences in a cluster

$$\Theta(\mathcal{C}) \stackrel{\text{def}}{=} \max_{k} \left(\sup_{\mathcal{P},\mathcal{P}' \in Proc(\mathcal{N}_k)} \frac{dist(\mathcal{P},\mathcal{P}')}{(1+\epsilon)^{max(|\mathcal{P}|,|\mathcal{P}'|)}} \right)$$

A subnet with very different behaviors :



Relating clusterings

> Relating clustering by generalizing centroids

Full runs centroids → Processes centroids → Subnets centroids

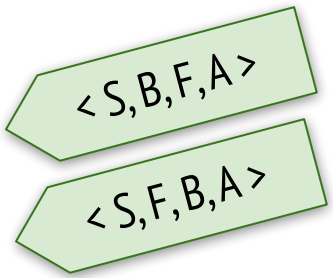
> Decrease the distance between the traces and the centroids

> Decrease the inter-cluster distance

Relating clusterings

> Relating clustering by generalizing centroids

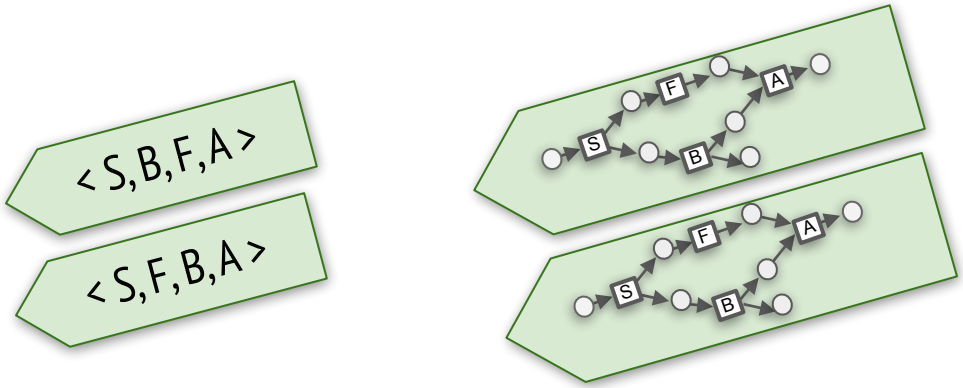
Full runs centroids → Processes centroids → Subnets centroids



Relating clusterings

> Relating clustering by generalizing centroids

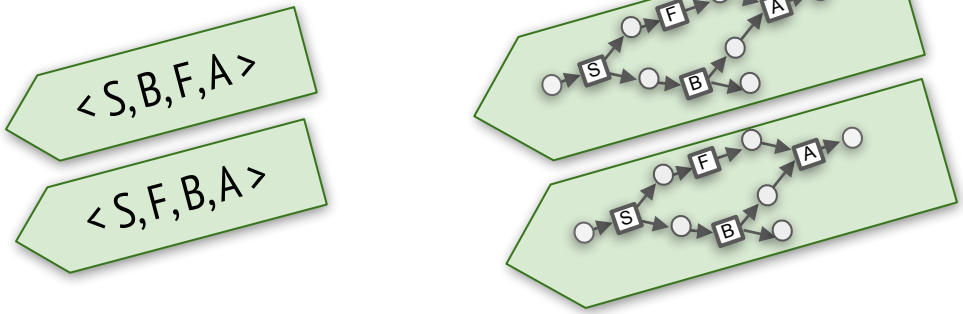
Full runs centroids → Processes centroids → Subnets centroids



Relating clusterings

> Relating clustering by generalizing centroids

Full runs centroids → Processes centroids → Subnets centroids



$$\Phi(C) = 0$$

Implementation

> Tool DARK SIDER*

> SAT formulas

> Optimal clusterings

```
|T|: 18
|P|: 14
|L|: 9
clusters : full runs
nb variables : 75438
nb constrains : 165569
total time :5.099369
d(c) : 2
delta(c) : 16
n(c) : 3
ĉ(c) : 0.666667

(A C E X3 I J X1 F D B , [A C E X3 I J X1 F D B   A C E X3 I J X1 F D B   ],0.222222)
(A C E L N M F D B , [A C E L N M F D B   A C E L O M F D B   ],0.222222)
(A C E G H F D B , [A C E G H F D B   A C E G H F F B   ],0.222222)
(nc, [ A C E X3 I J K I J X1 F D B   A C E X3 I J K I J K I J X1 F D B   A A C E G H F D B   ],0.333333)
```

*https://github.com/BoltMaud/darksider

Conclusion

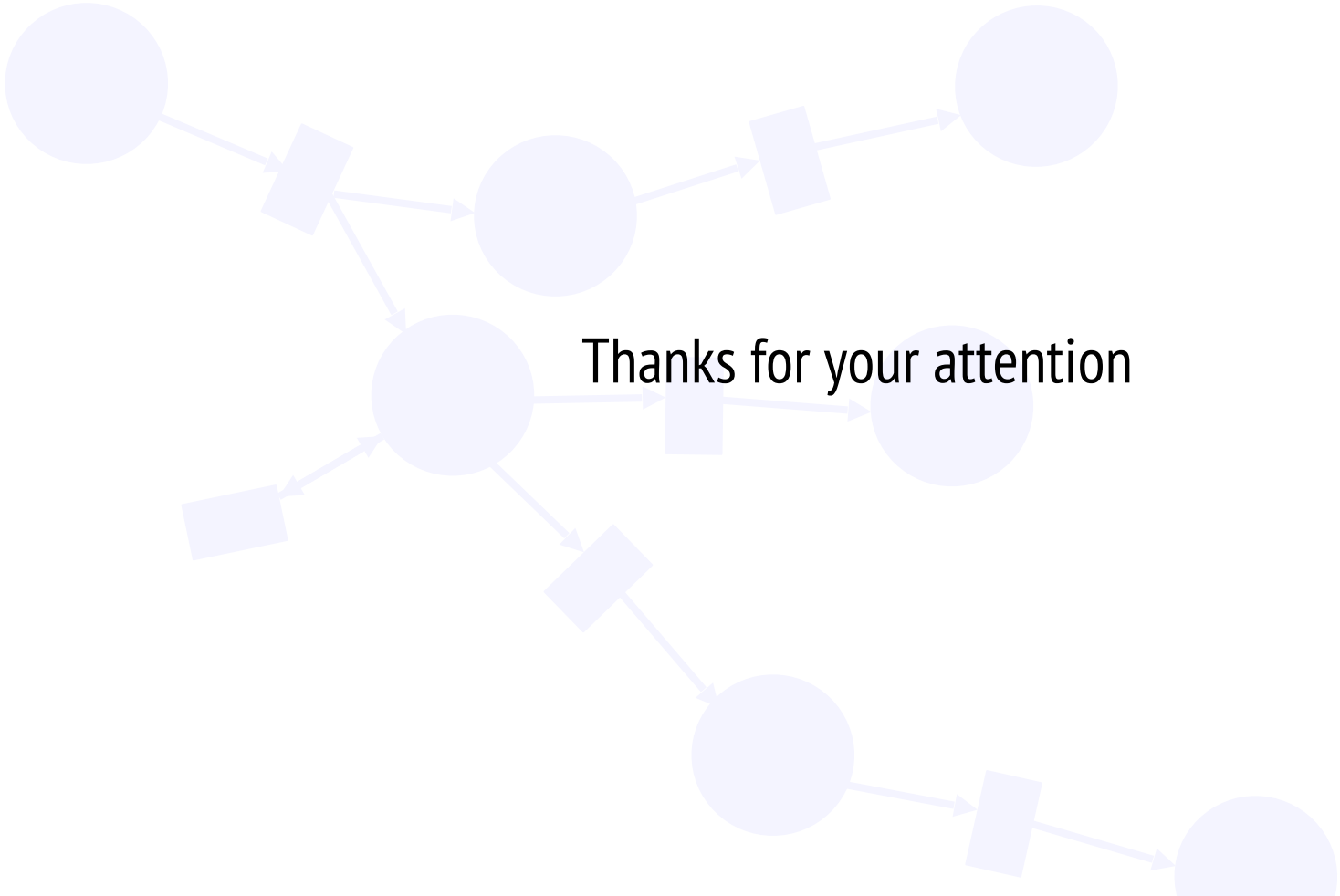
> 2 novel clustering methods for Process Mining problems

> Quality criteria

> Tool and experimentation

Limits : execution time

Future works : non-clustered traces



Thanks for your attention

Generalized Alignment-Based Trace Clustering for Process Behavior

Mathilde Boltenhagen[1], Thomas Chatain[1], Josep Carmona[2]

[1]LSV, CNRS, ENS Paris-Saclay, Inria, Université Paris-Saclay
[2]Universitat Politècnica de Catalunya