C | A | U

Christian-Albrechts-Universität zu Kiel

Advanced Process Mining

Sommer term 2020

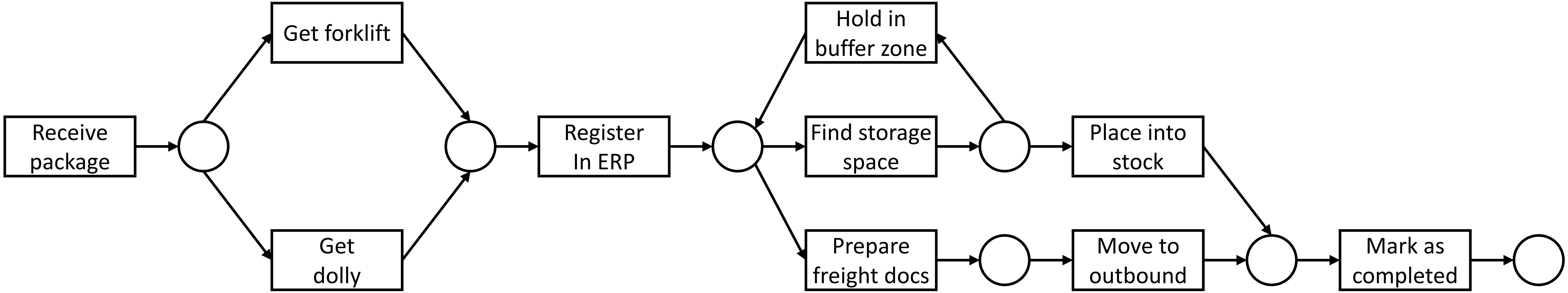
Exercise sheet 9

Event Log Clustering

Trace Clustering

Exercise 1a

Cluster the process model into 2 groups.

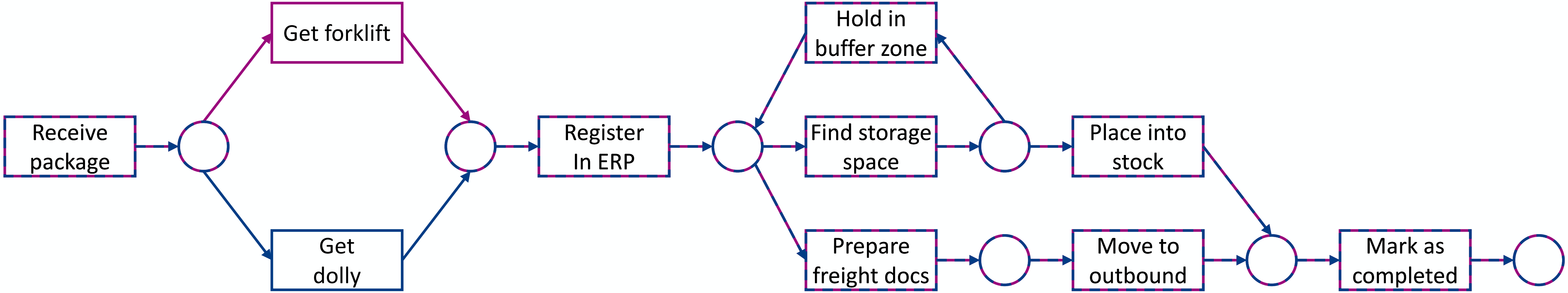


Trace Clustering

Exercise 1a

Christian-Albrechts-Universität zu Kiel

Cluster the process model into 2 groups.



Heavy goods

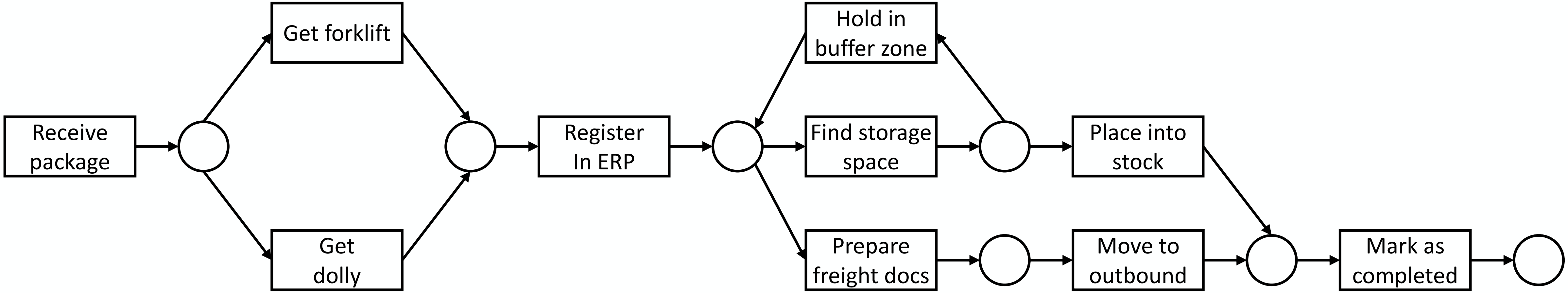
Regular goods

Trace Clustering

Exercise 1a

Christian-Albrechts-Universität zu Kiel

Cluster the process model into 3 groups.

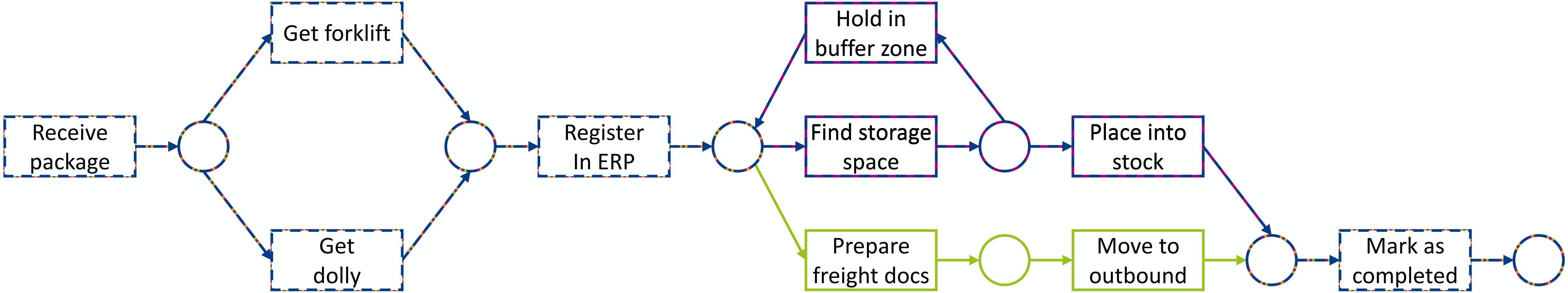


Trace Clustering

Exercise 1a

Christian-Albrechts-Universität zu Kiel

Cluster the process model into 3 groups.



Regular storing

Storage full

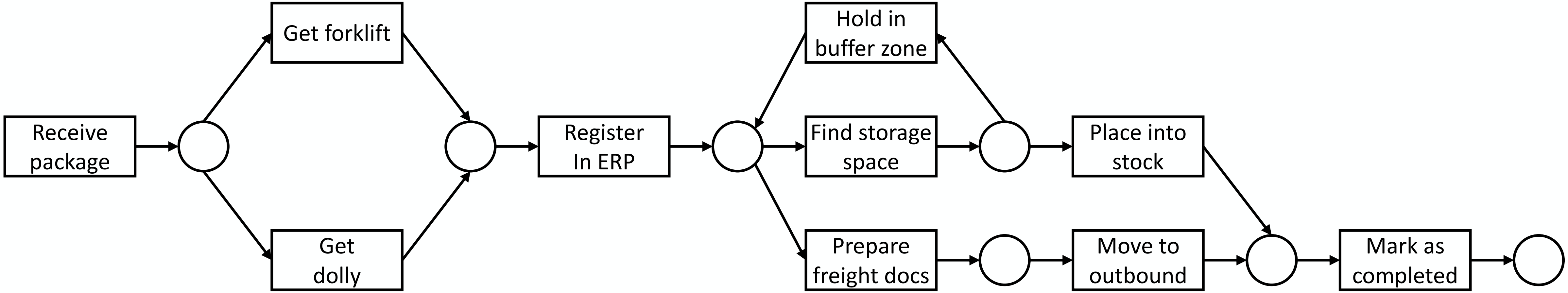
Ship directly

Trace Clustering

Exercise 1a

Christian-Albrechts-Universität zu Kiel

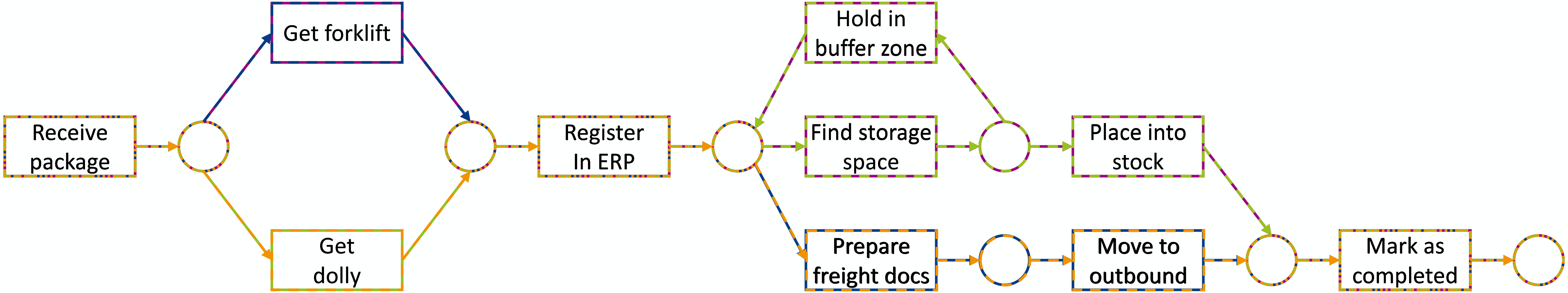
Cluster the process model into 4 groups.



Trace Clustering

Exercise 1a

Cluster the process model into 4 groups.

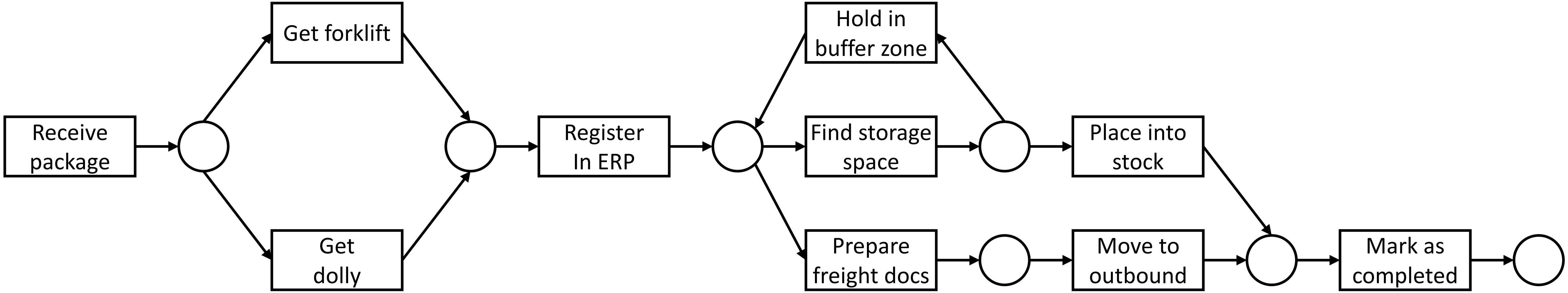


Store heavy goods Ship heavy goods directly Store regular goods Ship regular goods directly

Trace Clustering

Exercise 1b

What is the ideal number of clusters in this process model?



To determine an ideal number of clusters, the information given is not sufficient.
The number of clusters depends very strongly on the scenario and the intention of the clustering.

Distance Measures

Exercise 2a

Calculate the similarity between the following traces:

$$j = \langle A, B, C, D, E, F \rangle$$

$$k = \langle A, B, A, C, D, E, F, F \rangle$$

Using the Euclidean distance.

$$d_E(c_j, c_k) = \sqrt{\sum_{l=1}^{n} \left| i_{jl} - i_{kl} \right|^2}$$

What is the Euclidean distance between A and B?

Alternative notation, counting the occurrence of items:

$c_j = (1,1,1,1,1,1)$ $c_k = (2,1,1,1,1,2)$

Distance Measures

Exercise 2a

Calculate the similarity between the following traces:

$j = \langle A, B, C, D, E, F \rangle$ $\qquad c_j = (1,1,1,1,1,1)$

$k = \langle A, B, A, C, D, E, F, F \rangle$ $\qquad c_k = (2,1,1,1,1,2)$

Using the Euclidean distance.

$$d_E(c_j, c_k) = \sqrt{\sum_{l=1}^{n} \left| i_{jl} - i_{kl} \right|^2}$$

$$= \sqrt{|1-2|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-2|^2} = 1.4142$$

Distance Measures

Exercise 2a

Calculate the similarity between the following traces:

$$j = \langle A, B, C, D, E, F \rangle \qquad c_j = (1,1,1,1,1,1)$$

$$k = \langle A, B, A, C, D, E, F, F \rangle \qquad c_k = (2,1,1,1,1,2)$$

Using the Hamming distance [1].

$$d_H(c_j, c_k) = \frac{\sum_{l=1}^{n} \delta(i_{jl}, i_{kl})}{n} \qquad \delta(i_{jl}, i_{kl}) = \begin{cases} 0, & if\,(x > 0 \;\wedge y > 0) \vee (x = y = 0) \\ 1, & otherwise \end{cases}$$

$$= \frac{\delta(1,2) + \delta(1,1) + \delta(1,1) + \delta(1,1) + \delta(1,1) + \delta(1,2)}{6}$$

$$= \frac{0}{6} = 0$$

[1] According to *Trace Clustering in Process Mining* by Song *et al.*

Distance Measures

Exercise 2b

Calculate the similarity between the following traces:

$$j = \langle A, B, C, D, E, F \rangle \qquad c_j = (1,1,1,1,1,1)$$

$$m = \langle F, E, D, C, B, A \rangle \qquad c_m = (1,1,1,1,1,1)$$

Using the Euclidean distance.

$$d_E(c_j, c_k) = \sqrt{\sum_{l=1}^{n} \left| i_{jl} - i_{kl} \right|^2} = \sqrt{|1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2}$$
$$= 0$$

Does this result reflect the similarity of the two traces?

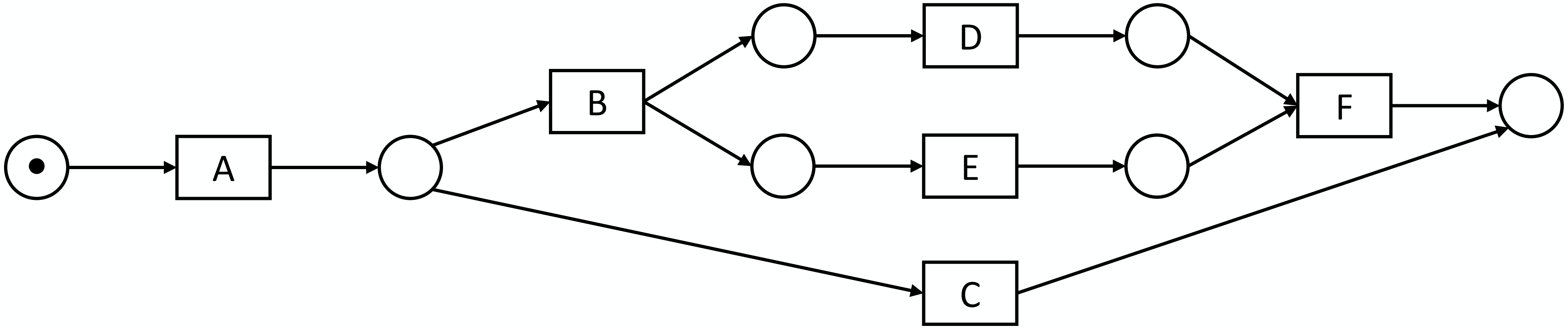
No.

According to the Euclidean distance in conjunction with only counting the appearances of certain items, the cases *j* and *m* are similar. Yet *m* is in reversed order and not similar at all to *j*.

Process Trees

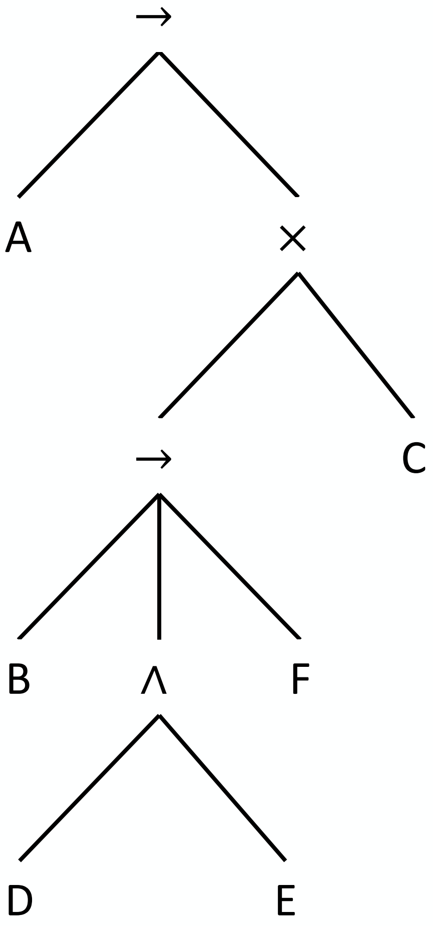
Exercise 3

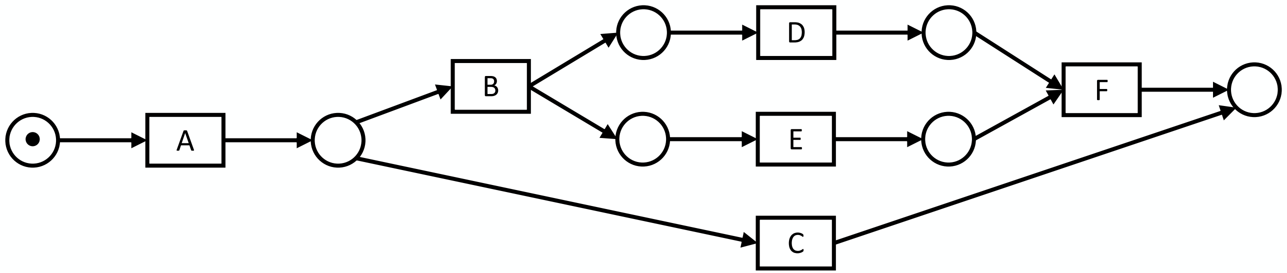
Derive a process tree from the Petri net:



Process Trees

Exercise 3





True or False

Exercise 4a

The k-Means clustering algorithm finds clusters of the size k.

False

The k stands for the number of clusters found by the k-means algorithm.
The size of the clusters themselves are irrelevant.

True or False

Exercise 4b

The advantage of the Levenshtein distance over the Hamming distance is that it can be applied to vectors of different lengths.

True

The Levenshtein distance is calculated by counting the single-character edits necessary to get from one vector to the other. The Hamming distance, on the other hand, only allows substitution and can therefore only be applied to compare vectors of the same size.

Calculating the Hamming distance can still be possible for traces of different lengths by converting the trace as it is done in exercise 2.