Student number:

Exercise 1

7 Points 2,5

a) What is the first stage of the process mining project methodology? (1 Point) ~~4,5~~ 1

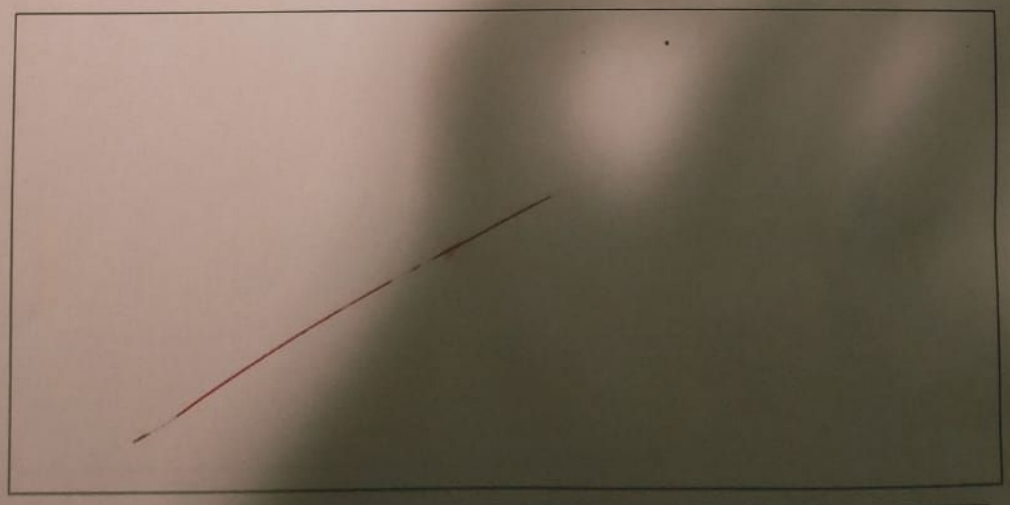
Planning ✓ elaborate..

b) To benchmark the quality of a process model in combination with an event log, four quality dimensions exist. Name two dimensions that directly oppose each other and explain why maximising those two at the same time is challenging. (2 Points) 2

Precision, Generalisation. ✓

Precision only allows behaviour that is seen in log whereas Generalisation requires the ✓ model to produce behaviour that may not ~~na~~ be seen in log.

c) In the lecture nine key learnings from industry settings have been discussed. Name two of them and explain both of them using an (artificial) use case. (4 Points) 0



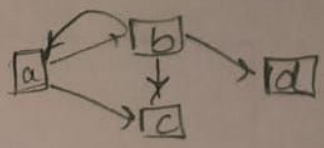
Student number:

Exercise 2 (DFG & EFG)

10 Points

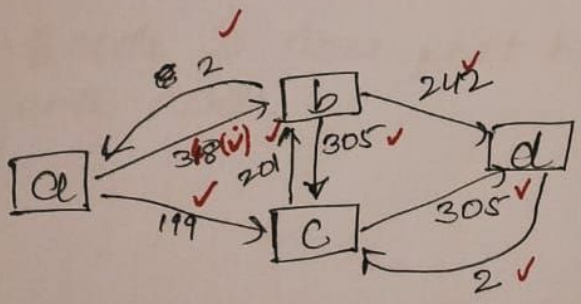
	a	b	c	d
a		348 199		
b	2		305	242
c		201		305
d			c	

#	Trace
305	abcd
199	acbd
43	abd
2	dcba



305+43 = 48 199 +43 242

a) Given the event-log above create a Directly-Follows Graph. (4 Points) 4
It is not necessary to use an auxiliary table.



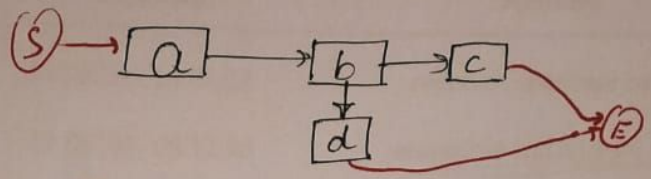
Student number:

b) True or false: Each unfiltered DFG is sound that has been discovered from an event log. Justify your answer using an example. (4 Points) 0

False!

#	Trace
m	ABC
n	ABD

a → b → c
↓
d



In the above DFG a is the start & c is the end, The node D does not have a path to C i.e. the end node.

c) Explain one of the parameters τ that can be applied during the DFG creation. (2 Points) 2

Tvar ✓ - It defines the threshold for minimum number of traces for a variant included. ✓ (based on $\#_a(e)$).

Student number:

Exercise 3 (Event Log Quality) **7 Points**

a) Inspect the following event log and identify possible event log imperfections. If possible correct them. Give a brief explanation why the identified parts are imperfections. (Highlighting the imperfections and correcting them in the event log above itself is allowed.) (4 Points) 4.

Case ID	Timestamp	Activity
1	01.08.20 - 14:01:22	receive customer order
1	02.08.20 - 08:12:10	assemble individual parts
2	02.08.20 - 11:02:22	receive customer order
2	02.08.20 - 17:05:03	collect raw materials from warehouse
1	02.08.20 - 17:05:03	collect raw materials from warehouse
2	03.08.20 - 09:11:48	assemble individual parts
1	03.08.20 - 14:34:02	quality check
1	03.08.20 - 17:06:02	ship order
2	04.08.20 - 04:45:43	quality check
2	04.08.20 - 07:35:55	shipping order ship order ✓

1. Imprecise data: For case 1 'ship order' is used whereas for case 2 'shipping order', both have the same meaning. ✓

2. Incorrect data: In case 1, the ordering of activities is incorrect. Without collection of raw materials, parts cannot be assembled. ✓

Student number:

b) Give two examples of *noise* and a short justification why your examples are not classified as *outliers* but as *noise*. (3 Points) 2.5

~~When in an activity two timestat~~

1. When two activity have same timestamp it can be counted as noise and not outlier as it is an error but not a deviating value. 1/1.

2. ~~Synonmw~~ Synonyms used for names can be counted as an example of noise and not content outliers. (1)

Better explanation, what those "Synonyms" are, necessary

Student number:

Exercise 4 (Heuristic Miner)

9 Points

a) What are the challenges of applying a filter in the Heuristic Miner algorithm? Describe a scenario where it is suitable to apply a very strict threshold and a scenario where an (almost) unfiltered process model can be helpful. (3 Points) 0

When applying a filter, the DFG can change in terms of ~~fit~~ precision. It can become very low due to missing arcs.

b) Construct a dependency matrix from the following Directly-Follows Matrix: (6 Points) 5

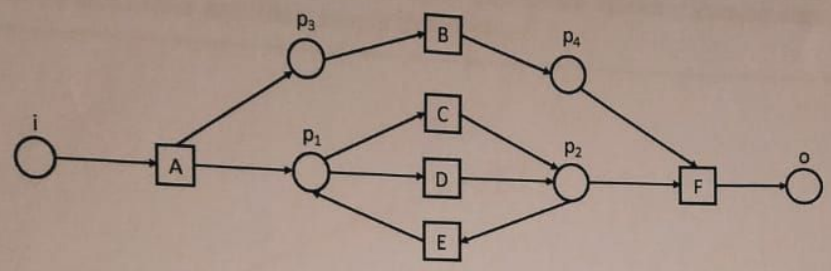
$>_L$	a	b	c	d
a	5	10	11	
b				10
c	3			11
d				

	a	b	c	d
a	0.83	0.909	0.53	
b	-0,91			0.909
c	-0.53			0.916
d		-0,91	-0,92	

5×0,9
1×0,4

Student number:

Exercise 5 (Conformance Checking) **17 Points**



a) Apply alignment-based conformance checking and list all optimal alignments for the following traces: (7 Points) 1

#	Trace
189	ADBF
100	ACF
32	ABEF

A D B F
A D B F ✓

A C F
A C F

~~A B~~

A B D E F
A B D E F

1. Trace ADBF

σ	A	D	B	F
π	A	D	B	F

✓ +1

every move is synchronous, this is the optimal alignment. (Trace)

2. Trace ACF

σ	A	C	F
π	A	C	F

B is missing

The trace is the optimal alignment as it is synchronous. no.

3. Trace ABEF

→ after E there has to be either C or D

σ	A	B	≫	E	F
π	A	B	D	E	F

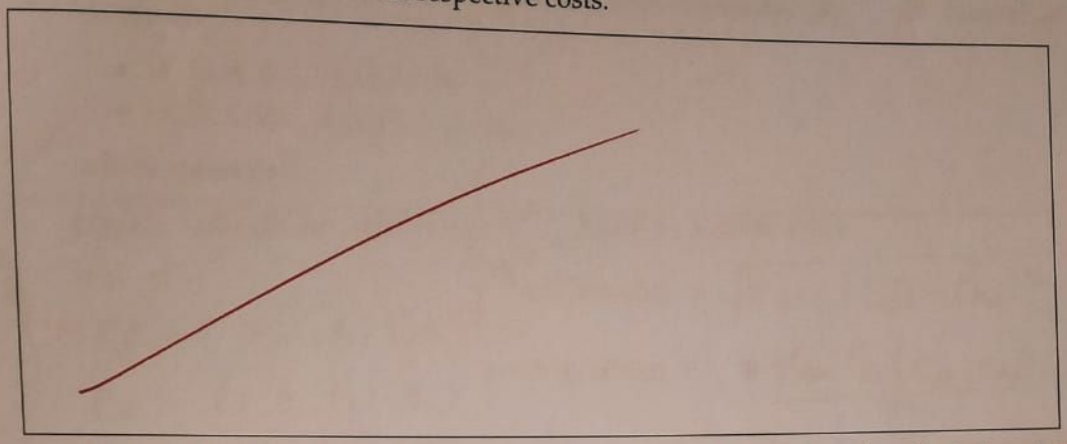
,

σ	A	B	≫	E	F
π	A	B	C	E	F

both alignment shown are optimal with cost = 1

Student number:

b) In order to quantify the quality of the alignments a standard cost function was introduced in the lecture. List all the relevant combinations of log- and model move and their respective costs. (4 Points) 0



c) Explain briefly what an alignment search space is and if finding the optimal alignment is a trivial task. Justify your answer. (3 Points) 1.5

Alignment search space is the product of state space & the log and trace. ✓
Finding the optimal alignment is ~~a~~ not a trivial task.

d) If we want to compute the fitness of a process model and the corresponding event log with alignment based conformance checking. Is it sufficient to know the cost of the optimal alignments $\delta(\lambda^N_{opt}(\sigma))$? If not, what kind of additional information is required? Justify your answer and explain why an additional value is or is not needed. (3 Points) 2

It is not sufficient as we need the ~~optimal~~ worst cost ✓ & all the executed sequences and the number & occurence & trace.

$$Fitness(L, N) = 1 - \frac{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda^N_{opt}(\sigma))}{\sum_{\sigma \in L} L(\sigma) \times \delta(\lambda^N_{worst}(\pi))}$$ ✓

But why?

Student number:

Exercise 6 (Clustering)

a) Explain how similarity between two traces with non-numerical values like: **6 Points** (2 Points) 1,5

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

can be quantified.

How to get from here to there

Using euclidian distance. ✓ Hamming distance.

$n = 5$

$C_j = (3, 1, 2, 1, 1)$

$C_k = (1, 3, 1, 1, 2)$

$\text{distance} = \sqrt{\sum_{l=1}^{n} |C_{jl} - C_{kl}|^2}$

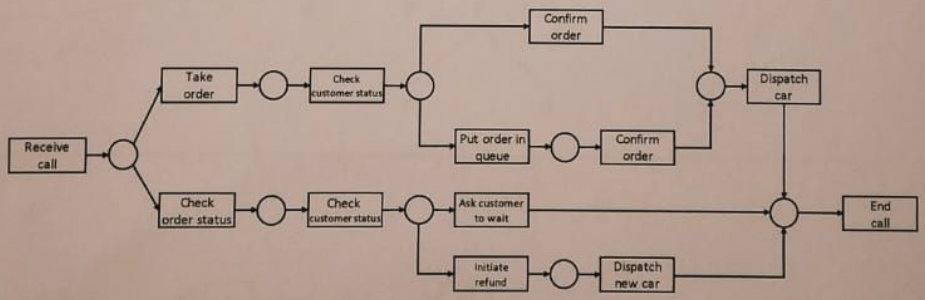
$\text{distance} = \frac{\sum_{l=1}^{n} \delta(C_{jl}, C_{kl})}{n}$.

~~Distance~~ $= \sqrt{\sum_{l=1}^{n} |C_{jl} - C_{kl}|^2}$ where $\delta(x,y) = \begin{cases} 0 & (x \wedge y > 0) \vee (x = y = 0) \\ 1 & \text{otherwise} \end{cases}$

$= \sqrt{4 + 4 + 1 + 1 + 1}$ $\sqrt{5 + 2 + 1 + 1 + 1}$

$= \sqrt{11}$ ✓

b) Find two different ways to cluster the following process model in two groups: (2 Points) 1



Group 1: New customer (new order) ✓

Group 2: Old customer (order placed) ✓

Student number:

c) Name one of the clustering algorithms presented in the lecture and explain briefly how it roughly works. (2 Points) 2

K-means clustering. ✓
- It is an unsupervised clustering method used for unlabelled data. ✓
- It ~~from~~ forms groups based on some similarity and number of groups (K) is defined by user. ✓

Exercise 7 (Process Tree) **4 Points** 2

Derive a Petri net from the following process tree: