Business Process Intelligence

Exam II - 09.09.2022

Prof. Dr. Wil van der Aalst

Chair of Process and Data Science
RWTH Aachen University

First Name	Last Name	Matr. Nr.

Study Course:
☐ Master Informatik ☐ Master Data Science ☐ Master SSE
☐ Master Media Informatics ☐ Other: _____

- Duration of this exam: **120 Minutes**.
- Write your **matriculation number** on each sheet.
- Provide your solutions in a readable and traceable manner. **Solutions will be graded based on completeness and correctness of the description/application of the algorithm/method**.
- Provide your solutions on the exam sheets only. If you need extra paper, use only the paper provided by the exam supervisors.
- Additional paper can be found at the end of this document; make sure that it is clear to which question your solution belongs.
- Please cross out those things you do not wish to be graded. In case of multiple given answers, only the first one will be graded.
- In case of attempted deception, your exam will be graded as failed. This will also be reported and may result in severe sanctions.
- At the end of the exam, hand in your complete copy. Do not separate any sheets by removing the staples.
- This exam accounts for 60% of the final grade. The other 40% could have been obtained through the mandatory assignment.
- **You may only use a black or blue pen.**
- **During the exam, you may not communicate with other people! You have to work on this exam on your own!**
- **Only answers that are given in English will be graded**.
- **Please sign this first sheet indicating that you comply with the aforementioned.**

Signature: _____

Question:	1	2	3	4	5	6	7	8	9	Total
Points:	8	8	9	12	16	20	11	10	6	100
Score:										

Question 1: Petri Nets (8 points)

Consider the following event log.

$$L = [\langle a, b, c\rangle^{120}, \langle a, c, b, d\rangle^{50}, \langle a, c, b\rangle^{20}]$$

How do the models rank in terms of precision and fitness compared to L? Argue about precision and fitness based on the model and log language, i.e., the accepting traces. You do not need to provide calculations but a solid argumentation why you rank the models the way you do.

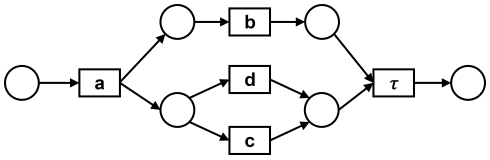


Figure 1: P_1

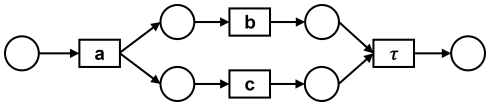


Figure 2: P_2

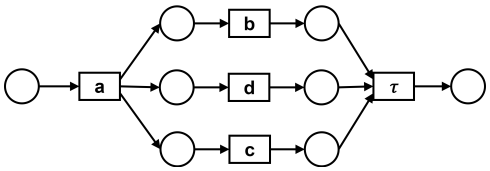


Figure 3: P_3

Fitness
1) P_3
2) $P_1 = P_2$

Precision
1) P_2
2) P_1
3) P_3

P_1 and P_2 can only replay traces $\langle a,b,c\rangle$ and $\langle a,c,b\rangle$ so their fitness is equal. But P_2 is more precise because it cannot replay any traces not in log. P_3 can replay all traces (best in fitness) but it is the least precise because b,c,d are concurrent it allows for many traces not in log

Total for Question 1: 8

Question 2: Alpha Miner (8 points)

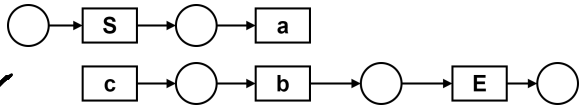
Connect the event log to the model that the Alpha Miner returns.

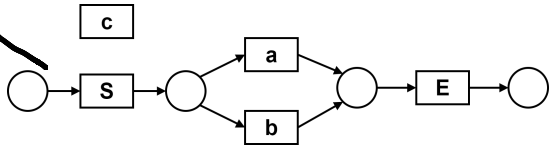
$L_1 = [\langle S, a, c, b, E\rangle, \langle S, b, c, a, E\rangle]$

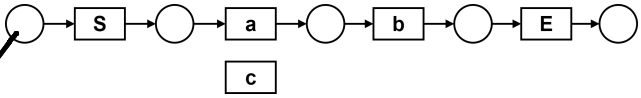
$L_2 = [\langle S, a, c, b, E\rangle, \langle S, c, a, b, E\rangle]$

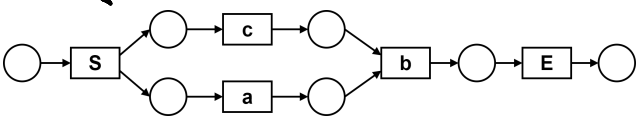
$L_3 = [\langle S, a, c, a, c, b, E\rangle, \langle S, a, c, b, E\rangle]$

$L_4 = [\langle S, a, c, c, b, E\rangle, \langle S, a, b, c, E\rangle]$







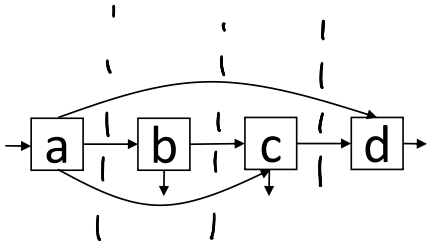


Total for Question 2: 8

Question 3: Inductive Miner (9 points)

(a) (4 points) Consider the directly-follows graphs given below. For each graph, indicate the first **maximal** cut the Inductive Miner would perform in the graph and give its type.

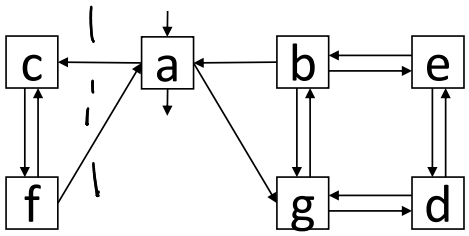
1.



sequence cut

2.

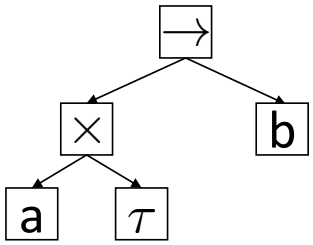
?



No ×, →, ∧

Loop cut

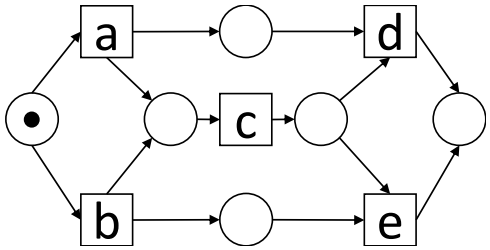
(b) (2 points) Consider the following process tree $(\rightarrow (\times(a, \tau), b))$:



Give an event log L, such that the basic Inductive Miner (no fall-throughs, no frequency filtering) with log projections would discover the given process tree from L.

$[\langle a, b\rangle, \langle b\rangle]$

(c) (3 points) Consider the following Petri net:



Is it possible to construct a process tree that models exactly the same behaviour? If yes, give such a process tree. If not, explain why.

It's not possible because of places after a and b which influence later stages of the net (dependency). This can't be represented by the process tree.

Total for Question 3: 9

Question 4: Heuristic Mining (12 points)

(a) (4 points) Consider the following event log L. The tables below show the corresponding direct successions frequencies and dependency measures as computed by the Heuristic Miner.

Construct the dependency graph with thresholds of at least 12 direct successions and a dependency value of at least 0.85 (no other heuristics, rule or modification should be applied).

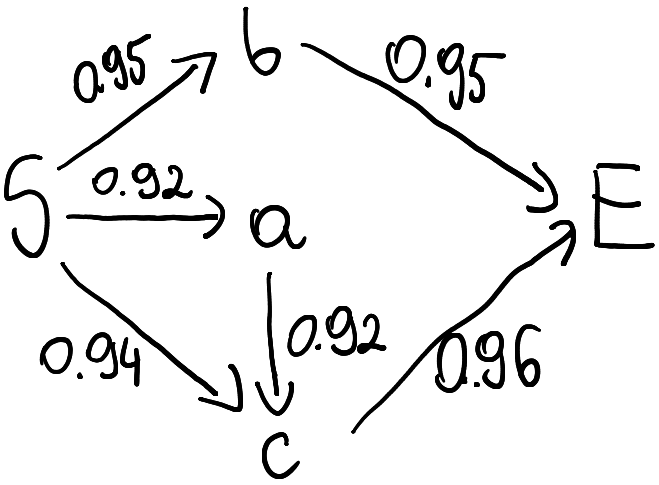
$$L = [\langle S, a, b, E\rangle^{10}, \langle S, b, a, c, E\rangle^{12}, \langle S, b, E\rangle^{8}, \langle S, c, E\rangle^{15}]$$

Direct Succession Frequencies:

	S	**a**	**b**	**c**	**E**
S	0	12	20	15	0
a	0	0	10	12	0
b	0	12	0	0	18
c	0	0	0	0	27
E	0	0	0	0	0

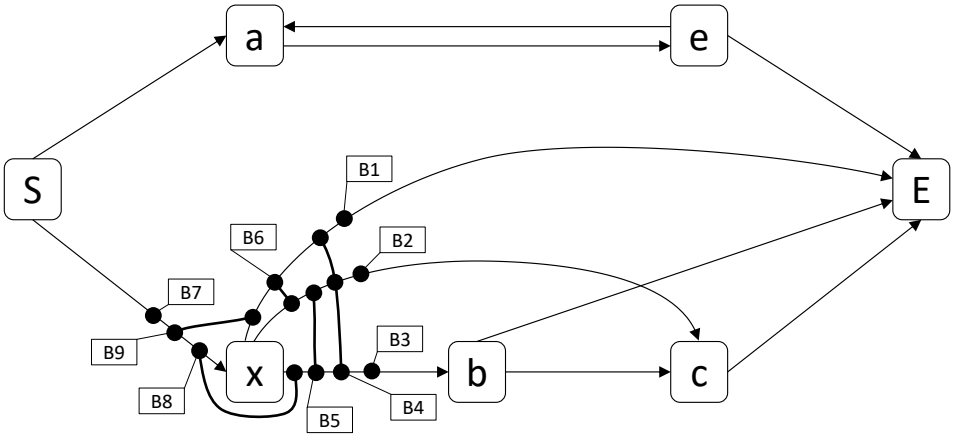
Dependency Measures:

	S	**a**	**b**	**c**	**E**
S	0	0.92	0.95	0.94	0
a	−0.91	0	−0.09	0.92	0
b	−0.95	0.09	0	0	0.95
c	−0.94	−0.92	0	0	0.96
E	0	0	−0.95	−0.96	0



(b) (4 points) Consider the event log and dependency graph given below. To convert such a dependency graph into a C-net, input and output bindings must be added. Take a look at the B1-B9 in the dependency graph. Answer the following questions.

1. (1 point) Which of B1-B9 are *not* valid input or output bindings? B8, B9,
2. (3 points) Based on the event log L and a window size of 3, there exist exactly 4 valid input or output bindings related to activity x that have a frequency of at least 10. Select those bindings out of B1-B9. **Give their exact frequencies**.



$$L = [\langle S, a, e, \mathbf{x}, b, c, E\rangle^{11},$$
$$\langle S, a, e, a, \mathbf{x}, e, c, E\rangle^{15},$$
$$\langle S, a, e, \mathbf{x}, a, b, e, e, c, E\rangle^{7},$$
$$\langle S, a, e, \mathbf{x}, a, e, b, c, E\rangle^{8}]$$

Input bindings B_7 (freq. 41) {a}

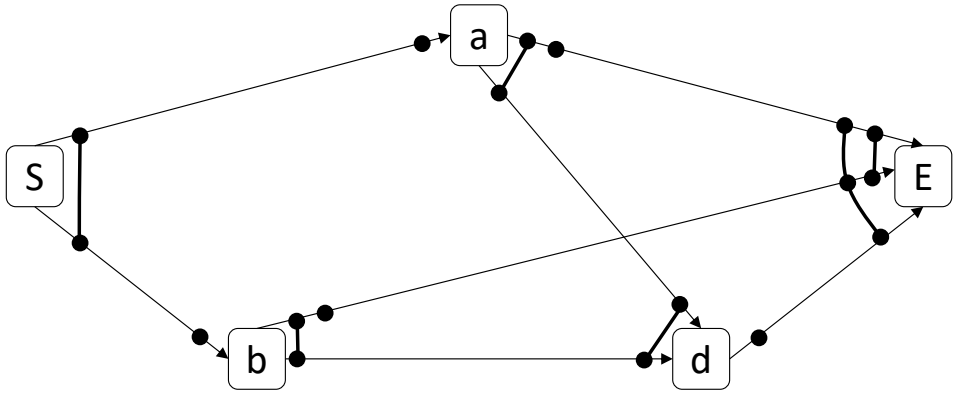
Output bindings

{b, c, E} B_4 (freq 11)

{c, E} B_6 (freq 15)

{b} B_3 (freq 15)

(c) (4 points) Consider the following event log and C-net.
$L = [\langle S, a, b, d, E\rangle^{22}, \langle S, b, a, d, E\rangle^{12}, \langle S, b, d, a, E\rangle^{8}]$



1. Give a trace σ_N that is part of the language of the C-net but is not part of L.
2. Give a trace σ_L that is part of the event log L but is not part of the language of the C-net.

1) $\sigma_N \langle S, a, b, E \rangle$

2) $\sigma_L \langle S, b, d, a, E \rangle$

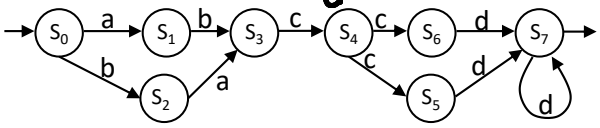
Total for Question 4: 12

Question 5: Region-based Mining (16 points)

(a) (6 points) Consider the event log $L = [\langle a, b, c, c, d\rangle, \langle b, a, c, c, d, d\rangle]$. Specify a combination of abstraction options (*abstraction, time, k-tail*) $\in$ {*sequence, set, multiset*} $\times$ {*past, future, past+future*} $\times [1, \infty)$ that could have been used to obtain the transition system below from L.

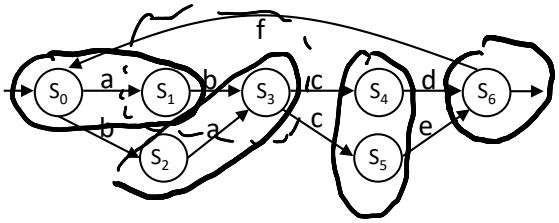
Hint: The solution is a triple of abstraction options from the given set of options, e.g., (sequence, past, 1-tail)

3-tail because it splits s_6, s_5

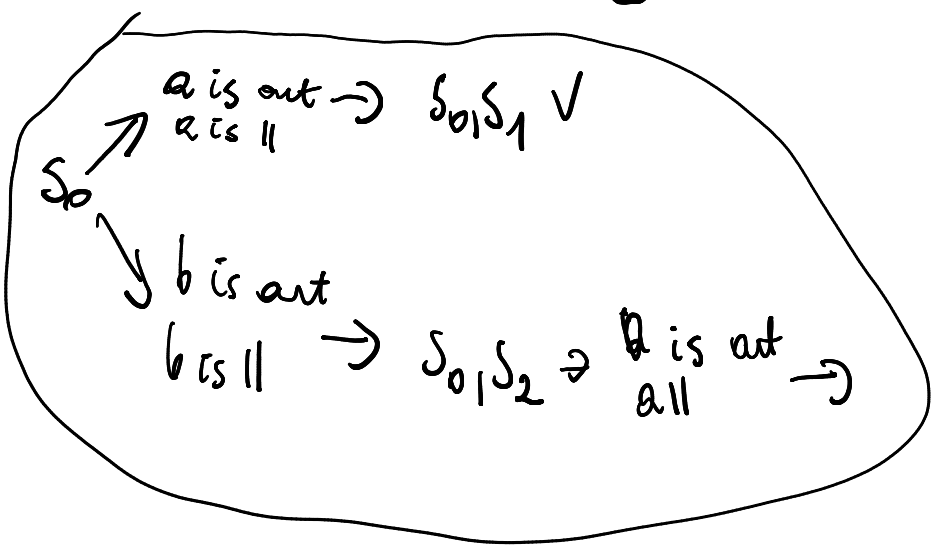


(Set, past, 3-tail)
Can't be multiset because of s_7
Can't be sequence because of multiple paths merging

(b) (2.5 points) Consider the following transition system. Give five different non-trivial, minimal regions.



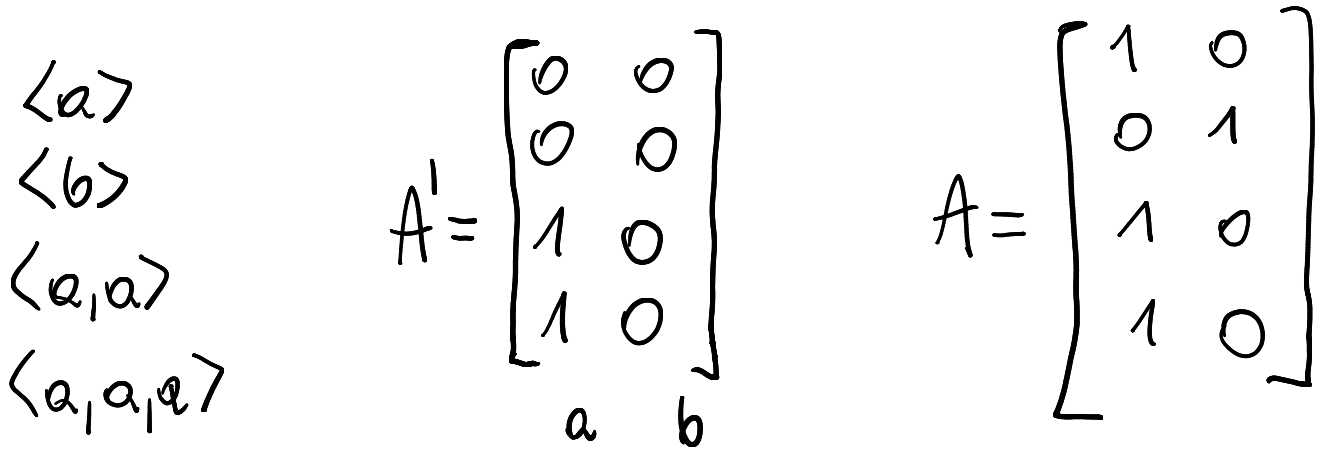
$\{\{s_0, s_1\}, \{s_2, s_3\}, \{s_6\}, \{s_4, s_5\}, \{s_1, s_3\}\}$



(c) (4 points) Consider the event log $L = [\langle a, a, a\rangle, \langle b\rangle]$ and the system of linear inequalities as constructed by the ILP Miner:

$$c \cdot \mathbf{1} + A' \cdot \mathbf{x} - A \cdot \mathbf{y} \geq \mathbf{0}.$$

1. Based on L, give the matrix A'.
2. Based on L, give the matrix A.

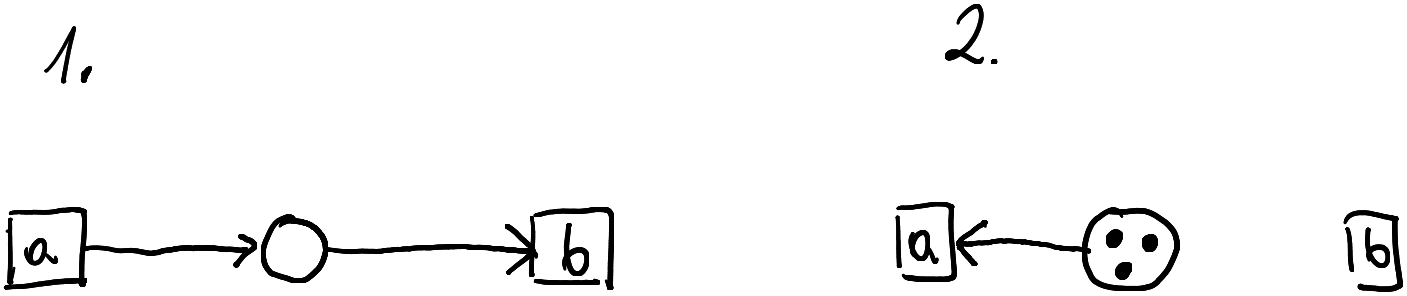


(d) (3.5 points) Consider a system of linear inequalities as constructed by the ILP Miner:

$$c \cdot \mathbf{1} + A' \cdot \mathbf{x} - A \cdot \mathbf{y} \geq \mathbf{0}.$$

Draw the Petri net places (and necessary related Petri net elements) corresponding to the following solutions of such a system of linear inequalities:

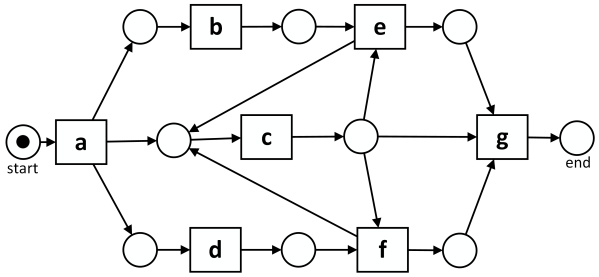
1. $c = 0, x_a = 1, x_b = 0, y_a = 0, y_b = 1$
2. $c = 3, x_a = 0, x_b = 0, y_a = 1, y_b = 0$



Total for Question 5: 16

Question 6: Conformance Checking (20 points)

(a) (4 points) Consider the following process model and event log L_1.



$L_1 = [\langle a, b, c, d, b, f, c, b, e, d, c, g\rangle, \langle a, d, e, c, f, b, d, e, c, g\rangle, \langle a, c, g\rangle]$
Complete the footprint matrix of the process model and event log.

Event Log:

	a	b	c	d	e	f	g
a	#	→	→	→	#	#	#
b	←	#	||	||	→	||	#
c	←	||	#	||	←	||	→
d	←	||	||	#	||	#	#
e	#	←	→	||	#	#	#
f	#	||	||	#	#	#	#
g	#	#	←	#	#	#	#

Process Model:

	a	b	c	d	e	f	g
a	#	→	→	→	#	#	#
b	←	#	||	||	→	||	#
c	←	||	#	||	||	||	→
d	←	||	||	#	||	→	#
e	#	←	||	||	#	#	#
f	#	||	||	←	#	#	#
g	#	#	←	#	#	#	#

fb, fd
ce, cf
ec, ef
bf, df

(b) (2 points) Consider the following footprint matrices that are extracted from an event log and a process model. Calculate the footprint-based conformance. Provide the formula that you use.

Event Log

	a	b	c	d
a	#	#	→	→
b	#	#	||	→
c	←	||	#	||
d	←	←	||	#

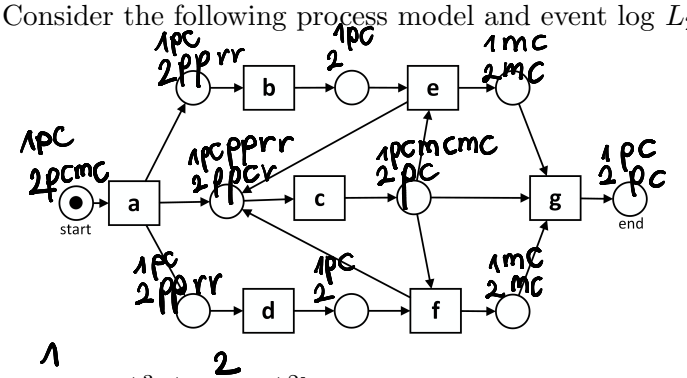
Process Model

	a	b	c	d
a	#	#	→	→
b	#	||	→	→
c	←	←	#	||
d	←	←	||	#

$$\text{Fitness} = 1 - \frac{\text{no differences}}{\text{no all possible diff}}$$

$$= 1 - \frac{3}{16} = \frac{13}{16}$$

(c) (9 points) Consider the following process model and event log L_2.



$L_2 = [\langle a, b, c, d, e, f, g\rangle^3, \langle a, a, c, g\rangle^2]$

Calculate the token-based replay fitness. Provide the formula that you use.

$$\text{Fitness} = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

$p + m = c + r$

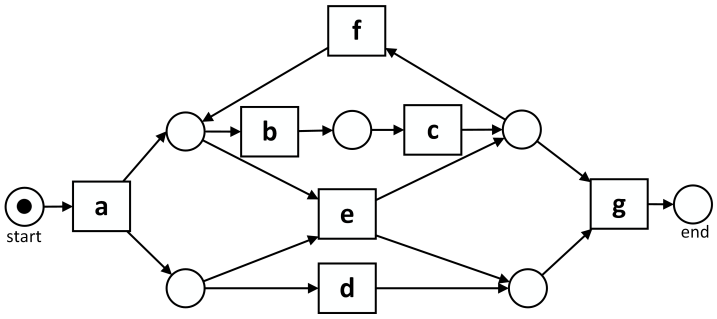
trace 1:		trace 2:	
p	10	p	9
c	12	c	7
m	4	m	3
r	2	r	5

$$\text{fitness}_{\sigma_1} = \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\cdot\frac{4}{5} = \frac{1}{3} + \frac{2}{5} \approx 0{,}73$$

$$\text{fitness}_{\sigma_2} = \frac{1}{2}\cdot\frac{4}{9} + \frac{1}{2}\cdot\frac{4}{7} = \frac{2}{9} + \frac{2}{7} \approx 0{,}51$$

$$\text{Fitness} = \frac{3\cdot \text{fitness}_{\sigma_1} + 2\cdot \text{fitness}_{\sigma_2}}{5} = 0.642$$

(d) (5 points) Consider the following process model and event log L_3.



$L_3 = [\langle a, c, b, f, c, g \rangle]$

Calculate the alignment-based fitness. Provide the formula that you use.

a c b f c ≫ g
a ≫ b ≫ c d g

Cost 3

Shortest path $\langle a, e, g \rangle$

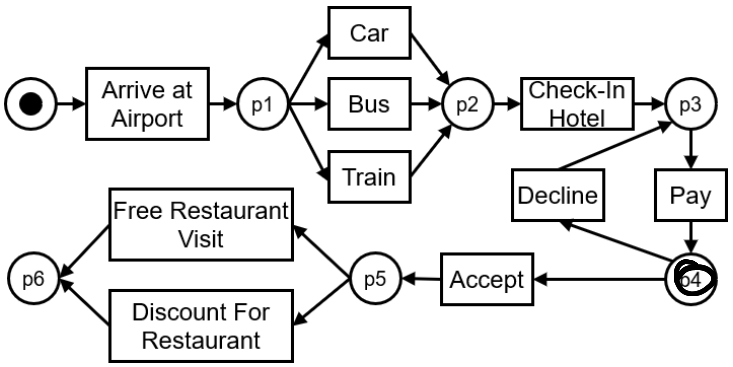
$$\text{fitness} = 1 - \frac{\text{cost}}{\text{len shortest path} + \text{len of trace}} =$$

$$= 1 - \frac{3}{3+6} = \frac{6}{9} = \frac{2}{3} \approx 0.66$$

Total for Question 6: 20

Question 7: Decision Mining (11 points)

A hotel records the actions of their guests and staff. The process from the local airport to the welcome dinner is visualized in the model depicted below.



(a) (2 points) The provided event log contains four traces from the process. Considering the event log, create an *event-based situation table* for decision point $p4$. Your situation table should consist of three columns: *Case ID*, *Card*, and the next executed activity (choice at $p4$).

(Decline vs. accept)

Case ID	Card	Next ex. activity
1	V	D
1	MC	D
1	V	A
1	GC	A
2	V	D
3	V	D
3	V	D
3	MC	A
3	V	D
4	MC	A
4		

Only consider events with decision!

Take event before lines

Case ID	Activity	Timestamp	Card
1	Arrive at Airport	6/1/2022	-
1	Car	6/2/2022	-
1	Check-In Hotel	6/3/2022	-
1	Pay	6/4/2022	Visa
1	Decline	6/5/2022	-
1	Pay	6/6/2022	MasterCard
1	Decline	6/7/2022	-
1	Pay	6/8/2022	Visa
1	Accept	6/9/2022	-
1	Free Restaurant Visit	6/10/2022	-
2	Arrive at Airport	6/11/2022	-
2	Bus	6/12/2022	-
2	Check-In Hotel	6/13/2022	-
2	Pay	6/14/2022	Girocard
2	Accept	6/15/2022	-
2	Discount For Restaurant	6/16/2022	-
3	Arrive at Airport	6/17/2022	-
3	Train	6/18/2022	-
3	Check-In Hotel	6/19/2022	-
3	Pay	6/29/2022	Visa
3	Decline	6/30/2022	-
3	Pay	7/1/2022	Visa
3	Decline	7/2/2022	-
3	Pay	7/3/2022	Visa
3	Decline	7/4/2022	-
3	Pay	7/5/2022	MasterCard
3	Accept	7/6/2022	-
3	Discount For Restaurant	7/7/2022	-
4	Arrive at Airport	6/1/2022	-
4	Car	6/2/2022	-
4	Check-In Hotel	6/3/2022	-
4	Pay	6/4/2022	Visa
4	Decline	6/5/2022	-
4	Pay	6/6/2022	MasterCard
4	Accept	6/9/2022	-
4	Free Restaurant Visit	6/10/2022	-

(b) (8 points) The given case-based situation table shows data concerning decision point *p1*. *Age*, *Booking Medium*, and *Nationality* are predictor variables; *Next Activity* is the response variable. Compute the initial entropy of the whole data set. Calculate for the predictor variables *Age* and *Nationality* the entropy and information gain. Round to the third decimal place. Which predictor variable maximizes information gain?

Case ID	Age	Booking Medium	Nationality	Next Activity
120	≥40	Online	German	Car
121	<40	Telephone	Dutch	Bus
122	≥40	Telephone	Dutch	Train
123	≥40	Telephone	German	Car
124	≥40	Online	English	Train
125	<40	Online	English	Bus
126	<40	App	German	Car
127	<40	App	Dutch	Bus

$\log_a b = \frac{\log b}{\log a}$

Initial entropy $= -\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) =$

$= -\left(-1.06 + \frac{1}{4}(-2)\right) = 1.56$

Age: ≥40 → 2C, 2T; <40 → 3B, 1C

Nationality: G → 3C; D → 2B, 1T; E → 1T, 1B

$\text{Entropy}_{Nat} = \frac{3}{8}\cdot\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{8}\left(-2\cdot\frac{1}{2}\cdot\log_2\frac{1}{2}\right) =$

$= \frac{3}{8}(0.38998 + 0.528) + \frac{1}{4} = 0.594$

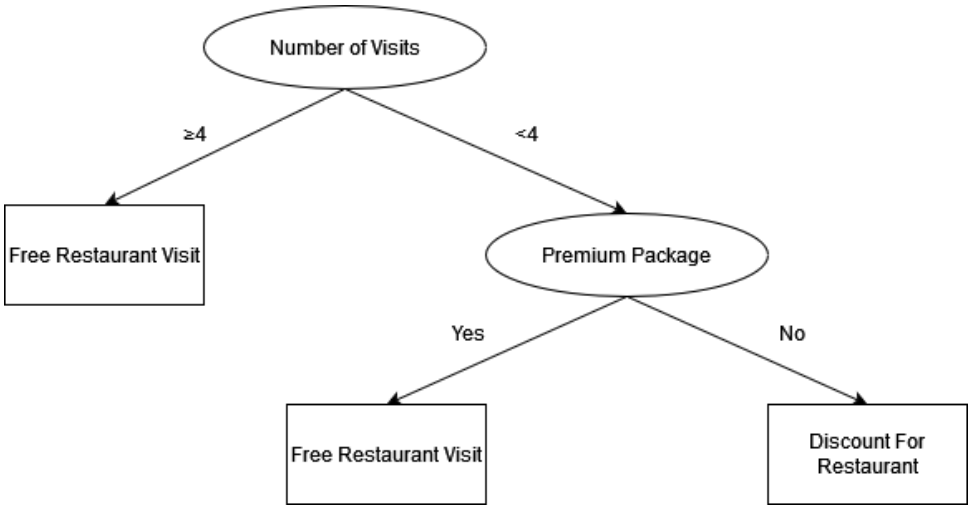
$IG_{Nat} = 1.56 - 0.594 = 0.966 \rightarrow$ Nationality maximizes IG

$\text{Entropy}_{Age} = \frac{1}{2}\cdot\left(-2\cdot\frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{2}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) =$

$= \frac{1}{2} + \frac{1}{2}(0.311 + 0.5) = 0.906$

$IG_{Age} = 0.654$

(c) (1 point) Consider the decision tree shown below for decision point *p5*.



Given the following two incomplete cases, predict which activity will be executed using the decision tree.

Case ID	Activity	Timestamp	Number of Visits	Premium Package
200	Arrive at Airport	6/1/2022	6	No
200	Car	6/2/2022	6	No
200	Check-In Hotel	6/3/2022	6	No
200	Accept	6/9/2022	6	No
911	Arrive at Airport	6/11/2022	2	Yes
911	Bus	6/12/2022	2	Yes
911	Check-In Hotel	6/13/2022	2	Yes
911	Pay	6/14/2022	2	Yes
911	Accept	6/15/2022	2	Yes

200 – Free restaurant visit

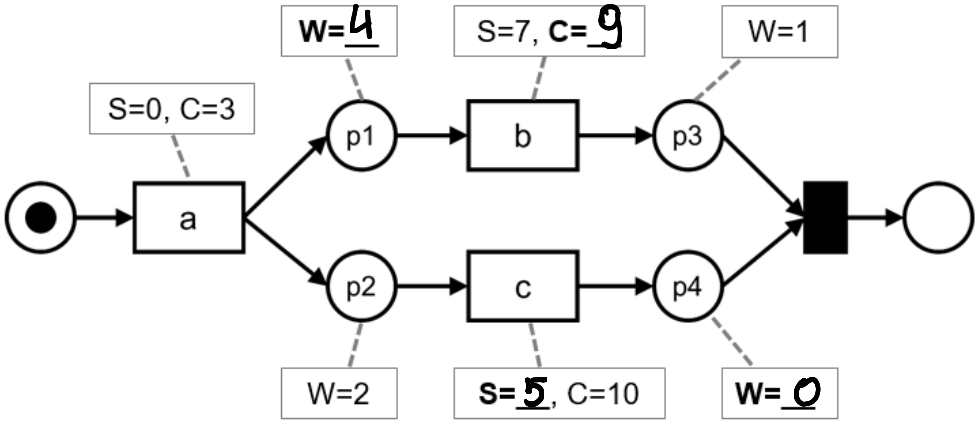
911 – Free rest. visit

Total for Question 7: 11

Question 8: Performance Analysis (10 points)

Below you are provided with the information recorded during the runs of two different cases through the process. The information beside each activity indicates when the activity started (S) and when it completed (C). The information beside each place indicates the waiting time (W) at that place. In the following, assume that silent transitions fire as soon as they are enabled.

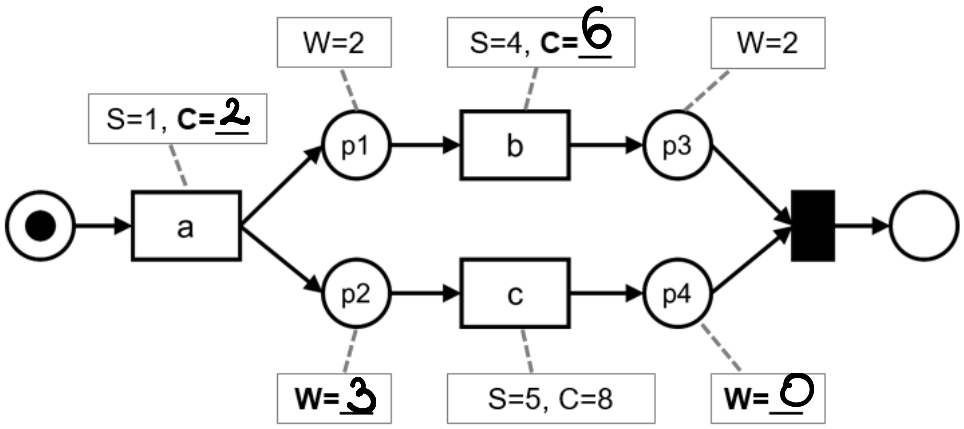
(a) (4 points) The given numbers refer to the process run of a case. Use the provided information to complete the missing entries.



(b) (1 point) What is the throughput time of the case shown in (a)?

10s

(c) (4 points) The given numbers refer to the process run of a case. Use the provided information to complete the missing entries.



(d) (1 point) What is the throughput time of the case shown in (c)?

8s

Total for Question 8: 10

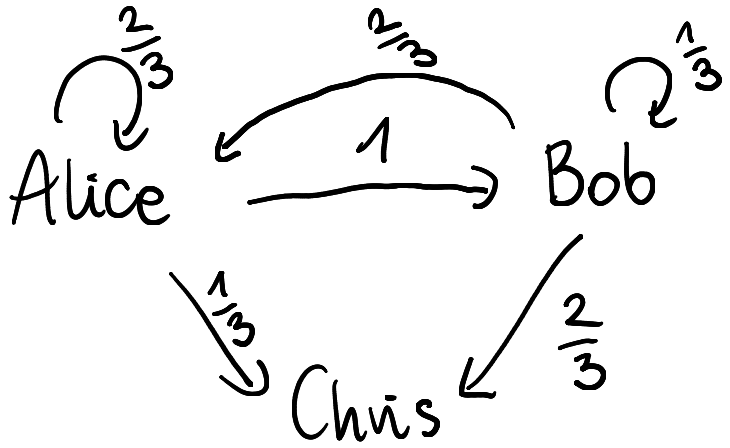
Question 9: Organizational Mining (6 points)

Given is the event log

$$L = [\langle a^{Alice}, b^{Bob}, a^{Alice}, b^{Bob}, c^{Chris}\rangle,$$
$$\langle a^{Alice}, a^{Alice}, b^{Bob}, b^{Bob}, c^{Chris}\rangle,$$
$$\langle b^{Bob}, a^{Alice}, a^{Alice}, c^{Chris}\rangle].$$

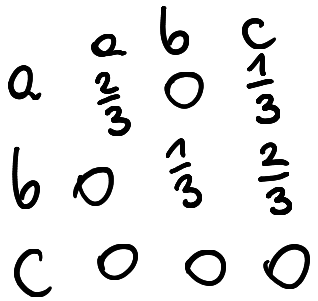
(a) (3 points) Create the handover of work matrix where you consider multiple transfers within the same case. Draw the corresponding social network containing all arcs that have a positive weight.

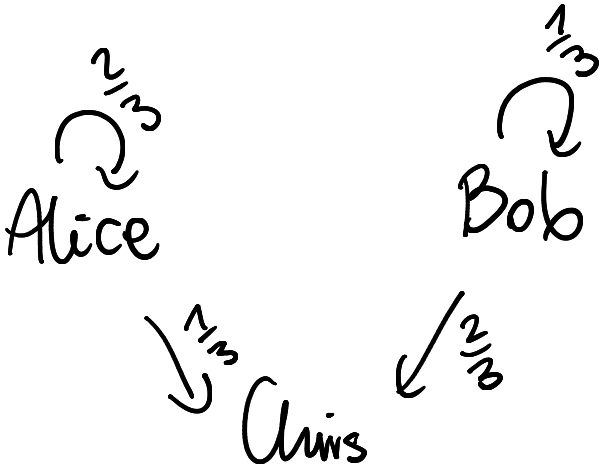
	Alice	Bob	Chris
Alice	2/3	1	1/3
Bob	2/3	1/3	2/3
Chris	0	0	0



(b) (3 points) The dependency matrix provided below is obtained from event log L. Create the real handover of work matrix using dependency threshold $\geq \frac{1}{2}$ and consider multiple transfers within the same case. Based on the real handover of work matrix, draw the corresponding social network containing all arcs that have a positive weight.

	a	b	c
a	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{2}$
b	$\frac{-1}{6}$	$\frac{1}{2}$	$\frac{2}{3}$
c	$\frac{-1}{2}$	$\frac{-2}{3}$	0





Total for Question 9: 6

Scratch paper: If you want something on this paper to be graded, clearly indicate to which tasks this belongs; otherwise, the following pages will **not** be graded.