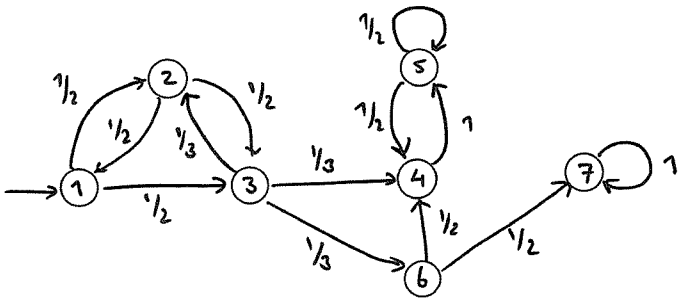
Exam April 18th 2024, 13:45-16:45

General information:

- Mark every sheet with your **student number**.
- Check that your copy of the exam consists of **four exercises**.
- Check that your copy of the exam consists of **seven pages**. This includes the front page (this page).
- You are allowed to bring one handwritten A4 page to the exam.
- You are not allowed to take any other material to the exam.
- No laptops, PDAs, mobile phones are allowed to use during the exam.
- Write with blue or black ink; do **not** use a pencil or red ink.
- You are neither allowed the help of anyone to complete your exam, nor is it allowed to help anyone else in completing this exam.
- Any attempt at deception leads to failure for this exam, even if detected later.

Question 1 (Markov chains) 31%

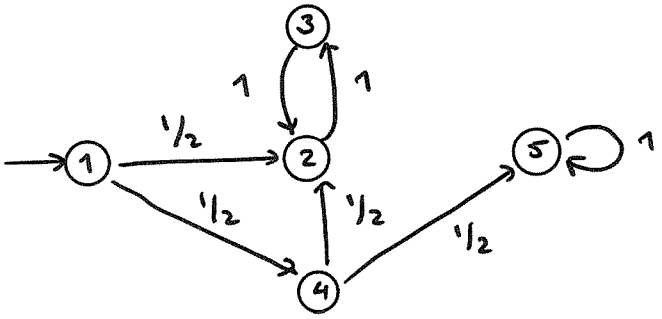
Consider the following Markov chain M_1:



(a) [7%] Provide the linear equation system $\mathbf{x} = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$ to reach the state 7 in M_1.

(b) [4%] Give the reachability probabilities to reach state 7, for all states in M_1.

(c) [4%] Give the reachability the probabilities to reach state 5, for all states in the following Markov chain M_2:



(d) [6%] Compare the reachability probability to reach state 7 from the initial state 1 in Markov chain M_1 and the reachability probability to reach state 5 from the initial state 1 in Markov chain M_2. Explain why these probabilities are equal or differ.

(e) [6%] Consider the following pGCL program R:

$n :\approx \text{uniform}[1,3];$
$s := 0;$
while$(n \geq 1)\{\, n := n-1; \{\, s := s+1 \,\}\ [1/2]\ \{\, \textbf{skip} \,\}\,\}$

Give the Markov chain that corresponds to this program. (Indicate the line number as well as the values of all program variables in each state of your Markov chain.)

(f) [4%] Determine $wp[\![R]\!]([s=0])$ using your Markov chain of part (e).

Question 2 (Probabilistic loop invariants) **34%**

Consider the following pGCL program P for non-negative integer x:

$$\textbf{while}(x > 0)\{\ x := x - 1\ \}\ [1/3]\ \{\ x := x + 1\ \}$$

(a) [3%] Give the characteristic function Φ_f of program P w.r.t. post-expectation $f = \mathbf{1}$.

(b) [6%] Determine $\Phi_f^k(\mathbf{0})$ for $k = 0, 1, 2, 3$.

(c) [7%] Let expectation $I = [x > 0] \cdot 2^{-x} + [x \leq 0]$. Prove that $wp[\![P]\!](\mathbf{1}) \sqsubseteq I$.

Hint: you may use Park's lemma: if $\Phi_f(I) \sqsubseteq I$ then $wp[\![\text{loop}]\!](f) \sqsubseteq I$ for characteristic function Φ_f of loop for post-expectation f.

(d) [3%] Give a non-trivial[1] upper bound on the probability that the program terminates when started with $x = 3$.

(e) [3%] Consider the program Q, the following variant of program P:

$$\textbf{while}(x > 0)\{\ x := x - 1\ \}\ [p]\ \{\ x := x + 1\ \}$$

for probability p with $0 < p \leq 1$.
Give the characteristic function Φ_f of program Q w.r.t. post-expectation $f = \mathbf{1}$.

(f) [8%] Let expectation $I_p = [x > 0] \cdot g(p)^{-x} + [x \leq 0]$ where $g(p)$ is a function that depends on p. Determine $\Phi_f(I_p)$ for $f = \mathbf{1}$.

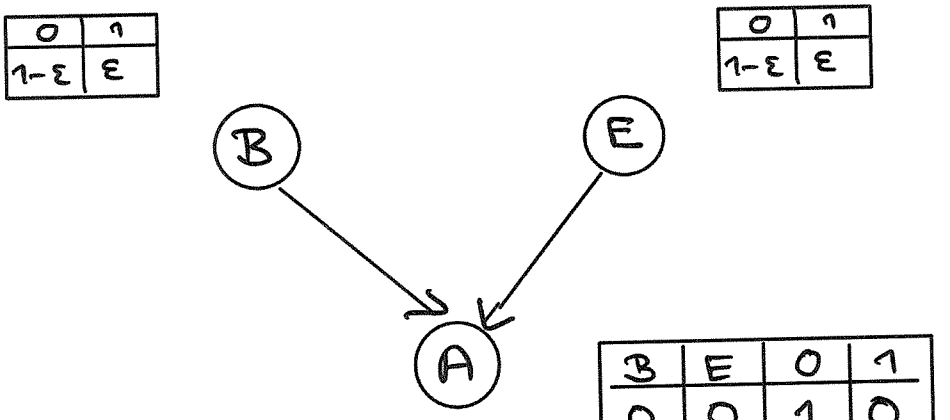
(g) [4%] Determine for which values of p, it holds that $\Phi_f(I_p) \sqsubseteq I_p$ for $f = \mathbf{1}$.

NB: this exercise should be solved by using the wp-calculus from the lecture.

[1] A *non-trivial* upper bound on a probability is a bound that is strictly smaller than 1.

Question 3 (Bayesian networks) 10%

Consider the following Bayesian network for $0 < \varepsilon < 1$:



(a) [4%]

1. Determine $\Pr\{ B = 1, E = 0, A = 1 \}$.
2. Determine $\Pr\{ B = 1 \mid A = 1 \}$.

(b) [4%] Give a pGCL program that corresponds to the above Bayesian network for the evidence $A = 1$.

(c) [2%] Explain how to obtain $\Pr\{ B = 1 \mid A = 1 \}$ by wp-reasoning on your program of item (b).
NB: you are not requested to calculate $\Pr\{ B = 1 \mid A = 1 \}$, but only need to indicate how to express this probability using wp on your program.

Question 4 (Probabilistic databases) 25%

The example data of Figure 1 is about the diagnosis results of blood sample analysis of cancer patients of a hypothethical cancer ward of some hospital.

There are 2 tables: patients with a patient identifier (pid), the patient's name (name), and their sex (sex: 'F'=female and 'M'=male). There is one bit of uncertainty among the patient data: one patient has been announced by phone to arrive later and it was unclear how to spell the patient's name: "Aleksandar" ($na = 1$) or "Alexander" ($na = 2$).

The other table is samples with a blood sample identifier (sid), the identifier of the patient the sample belongs to (pid), and the diagnosis of the kind of cancer detected in the sample (diagnosis). Obviously, blood analysis is an imperfect process, so for each sample i there is uncertainty about whether the particular cancer truly exists ($d_i = 1$) or not ($d_i = 0$). Furthermore, for sample 3, besides the uncertainty of whether cancer exists at all ($d_3 = 0$), it is unclear whether ovarian cancer ($d_3 = 1$) or leukemia ($d_3 = 2$) exists in the blood sample.

Disaster happened. The nurse fetching blood samples 1, 2, and 3 dropped the tray with the three blood samples after which the labels got separated from the samples. Therefore, it is not clear anymore which of the three samples belongs to which of the three patients. Figure 2 depicts all six possible ways how the diagnosed samples can belong to the patients.

(a) [3%] How many possible worlds does the probabilistic database of Figure 1 represent?

patients

⟨pid, name, sex⟩	φ
⟨1, Aleksandar, M⟩	$na = 1$
⟨1, Alexander, M⟩	$na = 2$
⟨2, Peter, M⟩	$\top$
⟨3, Mary, F⟩	$\top$
⟨4, John, M⟩	$\top$

samples

⟨sid, pid, diagnosis⟩	φ
⟨1, 1, Lymphoma⟩	$(sd = 1 \vee sd = 2) \wedge d_1 = 1$
⟨1, 2, Lymphoma⟩	$(sd = 3 \vee sd = 4) \wedge d_1 = 1$
⟨1, 3, Lymphoma⟩	$(sd = 5 \vee sd = 6) \wedge d_1 = 1$
⟨2, 1, Liver cancer⟩	$(sd = 3 \vee sd = 5) \wedge d_2 = 1$
⟨2, 2, Liver cancer⟩	$(sd = 1 \vee sd = 6) \wedge d_2 = 1$
⟨2, 3, Liver cancer⟩	$(sd = 2 \vee sd = 4) \wedge d_2 = 1$
⟨3, 1, Overian cancer⟩	$(sd = 4 \vee sd = 6) \wedge d_3 = 1$
⟨3, 2, Overian cancer⟩	$(sd = 2 \vee sd = 5) \wedge d_3 = 1$
⟨3, 3, Overian cancer⟩	$(sd = 1 \vee sd = 3) \wedge d_3 = 1$
⟨3, 1, Leukemia⟩	$(sd = 4 \vee sd = 6) \wedge d_3 = 2$
⟨3, 2, Leukemia⟩	$(sd = 2 \vee sd = 5) \wedge d_3 = 2$
⟨3, 3, Leukemia⟩	$(sd = 1 \vee sd = 3) \wedge d_3 = 2$
⟨4, 4, Pancreatic cancer⟩	$d_4 = 1$

dictionary

φ	prob
$na = 1$	$1/2$
$na = 2$	$1/2$
$sd = 1$	$1/6$
$sd = 2$	$1/6$
$sd = 3$	$1/6$
$sd = 4$	$1/6$
$sd = 5$	$1/6$
$sd = 6$	$1/6$
$d_1 = 0$	$5/8$
$d_1 = 1$	$3/8$
$d_2 = 0$	$5/6$
$d_2 = 1$	$1/6$
$d_3 = 0$	$3/9$
$d_3 = 1$	$5/9$
$d_3 = 2$	$1/9$
$d_4 = 0$	$3/4$
$d_4 = 1$	$1/4$

Figure 1: Example probabilistic data on diagnosis results of blood sample analysis of cancer patients.

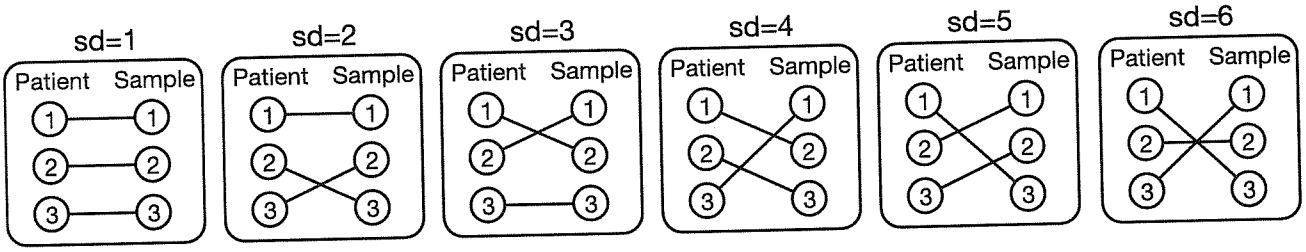


Figure 2: All possible ways how a diagnosed sample can belong to a patient after dropping the tray of samples.

(b) [3%] Calculate the probability of $(sd = 1 \vee sd = 2) \wedge d_1 = 1$.

Explain your answer by giving the complete calculation of the answer.

(c) [4%] Given the following DuBio query for the example database (the attribute containing the sentences is denoted with φ in the query)

```
SELECT p.name, s.diagnosis, p.φ & s.φ AS _sentence
FROM patients p, samples s, _dict d
WHERE p.pid=s.pid AND p.name='Alexander'
```

For each statement, indicate whether or not it is true *and provide an explanation why.*

- ○ The result of this query is probabilistic data.
- ○ The reason for the '&' is the SELECT-clause is because the query computes a join over two tables.
- ○ The query produces an error because we do not have a 'GROUP BY', which is always needed in DuBio queries.
- ○ The query constructs a sentence for each possible answer; the sum of the probabilities of all these sentences is the probability that patient 1 has cancer ("Overian cancer" or "Leukemia").

(d) [6%] Given the probabilistic algebra expression E below ('p.a' refers to attribute a of table "patients"; 's.a' refers to attribute a of table "samples").
Give the result of E.

$$E \equiv \pi_{\text{p.name,s.diagnosis}}(\bowtie_{\text{p.pid=s.pid}} (\sigma_{\text{p.name='Alexander'}}(\text{patients}), \text{samples}))$$

NB: I do not ask for a derivation, only the result. Note that an important part of your answer is to have it in *the right form*, so take care to provide all components that a result of a probabilistic algebra expression should have and omit those components that such a result should not have.

(e) [4%] As you can see in the samples table, one of the two possibly detected cancers for sample 3 is "Ovarian cancer". This is a type of cancer that can obviously only

exist in women. So, we as humans could provide the database with the evidence that the diagnosis of "Overian cancer" for sample 3 cannot possibly be correct if the sample belongs to a patient 1 or 2. This evidence can be used to update the probabilistic database to improve the quality of its data.

What is the technical term for this 'update the probabilistic database with evidence'? Moreover, describe what the evidence is in terms of random variables, alternatives, and sentences.

Explain your answer.

(f) [5%] In indeterministic duplicate detection, an M-graph is constructed from similarity match results of a duplicate detection tool which ran on tuples a="Aleksandar", b="Alexander", c="Alex", and d="Alexandra". The tool determines the following similarities: $s(a-b) = 0.8$, $s(a-c) = 0.6$, $s(a-d) = 0.2$, $s(b-c) = 0.6$, $s(b-d) = 0.2$, $s(c-d) = 0.6$. We set the upper threshold to 0.9 and the lower threshold to 0.3.

1. Draw the M-graph *after the thresholds have been applied.*
2. Which possible worlds does this produce? Use the following notation: $\{\ldots\}$ for the set of records comprising a possible world; ab for the merge of records a and b (other combinations analogously). Explain your answer.

NB: I do not ask for probabilities of possible worlds, so no need to compute them.