Exam Probabilistic Programming 2021

April 15, 2021, 09:00–12:00

General Information:

- Mark every sheet with your **student number**.
- Check that your copy of the exam consists of **five exercises**.
- This is an **open exam**, i.e. all lecture material (slides, exercises, solutions to the exercises) is permitted.
- 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.
- Your exam is only valid if the integrity statement on the next page is signed by ticking the box.
- You are supposed to send your solutions to the exam via e-mail to `c.kolb@utwente.nl`.
- Your solutions need to be send in a **single** pdf-file and must be received ultimately on **Thursday April 15, 12:15 (CEST).**
- During the exam, Christina Kolb is on-line available for questions via Canvas (conference).

Please read the following paragraph carefully, and tick the box to acknowledge that you have done so. To find more information, please consult the Canvas page of the course Probabilistic Programming 2020–2021.

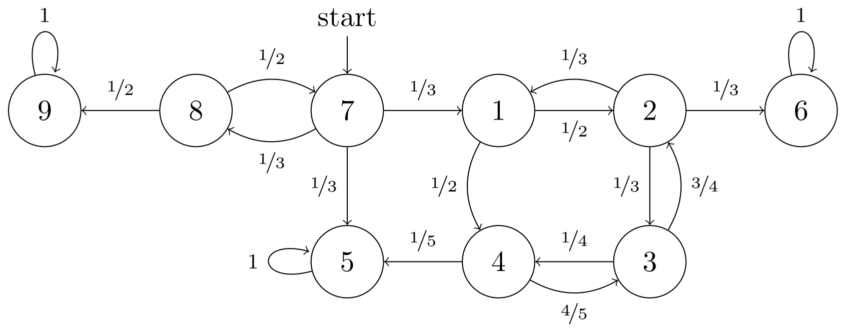
By testing you remotely in this fashion, we express our trust that you will adhere to the ethical standard of behaviour expected of you. This means that we trust you to answer the questions and perform the assignments in this test to the best of your own ability, without seeking or accepting the help of any source that is not explicitly allowed by the conditions of this test.

Please tick: ☐

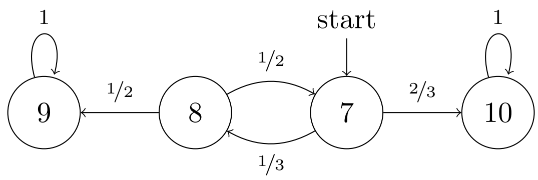
(If you are not able to tick this box, copy the above statement on your solution sheet, and tick the box drawn by you.)

Exercise 1 (Reachability probabilities) 10%

Determine $\Pr(\Diamond G)$ in the following Markov chain where $G = \{5, 6\}$. Compute and justify your answer.



Solution: We observe that the goal $G = \{5, 6\}$ is reached with probability 1 from all states $\{1, 2, 3, 4, 5, 6\}$ which can therefore be collapsed into a single goal state $G' = \{10\}$:



To compute the reachability probability in the simplified Markov chain, we use the equation system. It is $\Sigma_? = \{7, 8\}$, and

$$\mathbf{A} = \begin{bmatrix} 0 & 1/3 \\ 1/2 & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2/3 \\ 0 \end{bmatrix}.$$

Solving the system $(\mathbf{I} - \mathbf{A}) \cdot \mathbf{x} = \mathbf{b}$ yields $\mathbf{x} = \begin{bmatrix} 4/5 \\ 2/5 \end{bmatrix}$, thus $\Pr(\Diamond G) = \Pr(\Diamond G') = 4/5$.

Exercise 2 (Weakest pre-expectation calculus) 16%

Prove or disprove in each of the following two questions whether the given `pGCL` programs P_1 and P_2 are equivalent w.r.t. the post-expectation $f = x$:

(a) [6%]

P_1: $y := 5;$ `if` $(y < 0)$ { `skip` } `else` { { $x := 1$ } [1/2] { `skip` } }

P_2: { { $x := x + 3$ } [1/3] { $x := x$ } } [1/2] { $x := 0$ }

Solution:

$$
\begin{aligned}
& k(x) = wp(x := 1\,[1/2]\,\texttt{skip},\, x) \\
= \;& 1/2 \cdot wp(x := 1,\, x) + 1/2 \cdot wp(\texttt{skip},\, x) \\
= \;& 1/2 \cdot x[x := 1] + 1/2 \cdot x \\
= \;& 1/2 \cdot (1 + x)
\end{aligned}
$$

$$
\begin{aligned}
& g(x) = wp(\{\texttt{if}(\,y < 0\,)\{\,\texttt{skip}\,\}\texttt{else}\,\{\,x := 1\,[1/2]\,\texttt{skip}\}\,\},\, x) \\
= \;& [y < 0] \cdot wp(\texttt{skip},\, x) + [y \geq 0] \cdot wp(x := 1\,[1/2]\,\texttt{skip},\, x) \\
= \;& [y < 0] \cdot x \;+\; [y \geq 0] \;\cdot k(x) \\
= \;& [y < 0] \cdot x + [y \geq 0] 1/2 \cdot (1 + x)
\end{aligned}
$$

$$
\begin{aligned}
& wp(P_1,\, x) \\
= \;& wp(y := 5, wp(\texttt{if}(y < 0)\{\,\texttt{skip}\,\}\texttt{else}\,\{\,x := 1\,[1/2]\,\texttt{skip}\,\},\, x)) \\
= \;& wp(y := 5, g(x)) \\
= \;& \big([y < 0] \cdot x + [y \geq 0] \cdot 1/2 \cdot (1 + x)\big)\,[y := 5] \\
= \;& [5 < 0] \cdot x + [5 > 0] \cdot 1/2(1 + x) \\
= \;& 1/2 \cdot (1 + x)
\end{aligned}
$$

$$
\begin{aligned}
& wp(P_2,\, x) \\
= \;& wp(\{\,x := x + 3\,[1/3]\,x := x\,\}\,[1/2]\,\{x := 0\},\, x) \\
= \;& 1/2 \cdot (wp(x := x + 3\,[1/3]\,x := x\},\, x)) + 1/2 \cdot wp(x := 0,\, x) \\
= \;& 1/2 \cdot (1/3 \cdot wp(x := x + 3,\, x) + 2/3 \cdot wp(x := x,\, x)) + 1/2 \cdot 0 \\
= \;& 1/2 \cdot (1/3 \cdot (x + 3) + 2/3 \cdot x) + 0 \\
= \;& 1/2 \cdot (x + 1)
\end{aligned}
$$

We observe that $wp(P_1,\, x) = wp(P_2,\, x) = 1/2 \cdot (\, x{+}1\,)$. Therefore P_1 and P_2 are equivalent w.r.t. $f = x$.

(b) [10%]

P_1: $\texttt{while}\ (x \neq x)\ \{\ \{\ x := y + 1\ \}\ [1/2]\ \{\ x := y - 1\ \}\ \}$

P_2: $\texttt{while}\ (\mathit{true})\ \{\ \texttt{skip}\ \}$

Solution:

$$
\begin{aligned}
& wp(P_1, x) \\
=\ & wp(\texttt{while}(x \neq x)\{\, x := y + 1\, [1/2]\, x := y - 1\, \},\, x) \\
=\ & \text{lfp}X.\, (([x \neq x] \cdot wp(x := y + 1\, [1/2]\, x := y - 1, X)) + ([x = x] \cdot x)) \\
& \Psi(X) = (([x \neq x] \cdot wp(x := y + 1\, [1/2]\, x := y - 1, X)) + ([x = x] \cdot x)) \\
=\ & 0 \cdot wp(x := y + 1\, [1/2]\, x := y - 1,\, X) + 1 \cdot x \\
=\ & x \\
& \Psi^0(0) = x, \quad \Psi^1(0) = x, \cdots, \Psi^n(0) = x \\
& wp(P_1,\, x) = \text{lfp}X.\Psi(X) = x
\end{aligned}
$$

$$
\begin{aligned}
& wp(P_2, x) \\
=\ & wp(\texttt{while}(\texttt{true})\{\, \texttt{skip}\, \},\, x) \\
=\ & \text{lfp}X.\, ([\texttt{true}] \cdot wp(\, \texttt{skip},\, X) + [\, \texttt{false}\,] \cdot x) \\
& \Psi(X) = ([\texttt{true}] \cdot wp(\, \texttt{skip}\, ,\, X) + [\texttt{false}] \cdot x) \\
=\ & 1 \cdot X + 0 \cdot x = X \\
& \Psi^0(0) = 0, \Psi^1(0) = 0, \cdots, \Psi^n(0) = 0 \\
& wp(P_2,\, x) = \text{lfp}X.\Psi(X) = 0
\end{aligned}
$$

We observe that $wp(P_1,\, x) \neq wp(P_2,\, x)$. Therefore P_1 and P_2 are not equivalent w.r.t the post-expectation x.

Exercise 3 (Markov chains with rewards) 16%

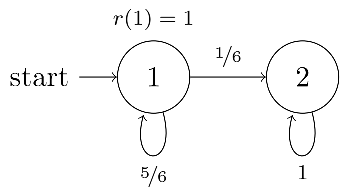
In this question, you are requested to provide a Markov chain with rewards for a certain problem statement. It is useful to find a Markov chain with a minimal number of states.

(a) [10%] Consider a six-sided fair die. Compute the *expected* number of trials needed to see a six for the first time by modelling the problem by a Markov chain with rewards.

Hint: The following variant of the geometric series might be helpful to compute the expected reward: For $0 < q < 1$,

$$\sum_{k \geq 0} k q^k = \frac{q}{(1-q)^2} \ .$$

Solution: We use the following Markov chain with rewards and goal set $G = \{2\}$:



The idea is to count every trial (= visit to state 1) by assigning it a reward $r(1) = 1$. The expected number of trials is then given as $ER(1, \Diamond G)$. To compute this quantity, consider all finite paths from 1 to $G = \{2\}$ in the Markov chain:

1 2

1 1 2

1 1 1 2

...

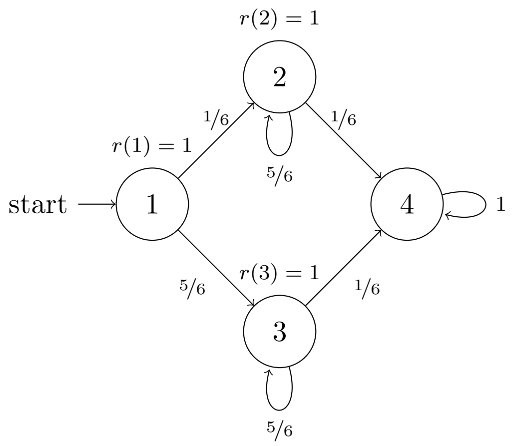
All these paths are of the form $1^k\ 2$, where 1^k denotes the string $\underbrace{1 \ldots 1}_{k \text{ times}}$ and $k \geq 1$. The reward of these paths is $r_G(1^k\ 2) = k$ and they occur with probability $Pr(1^k\ 2) = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}$.

Thus,

$$\begin{aligned}
ER(1, \Diamond G) &= \sum_{k \geq 1} k \left(\frac{5}{6}\right)^{k-1} \frac{1}{6} \\
&= \frac{1}{6} \left(\sum_{k \geq 0} (k+1) \left(\frac{5}{6}\right)^{k} \right) \qquad \text{(index shift)} \\
&= \frac{1}{6} \left(\sum_{k \geq 0} k \left(\frac{5}{6}\right)^{k} + \sum_{k \geq 0} \left(\frac{5}{6}\right)^{k} \right) \\
&= \frac{1}{6} \left(\frac{5/6}{(1 - 5/6)^2} + \frac{1}{1 - 5/6} \right) \qquad \text{(formula from hint + standard geom. series)} \\
&= \frac{1}{6} \left(\frac{5/6}{1/6 \cdot 1/6} + 6 \right) \\
&= 5 + 1 = 6 \ .
\end{aligned}$$

(b) [6%] Modify your reward Markov chain from part (a) such that it models the *conditional* expected number of trials needed to see a six for the first time *given that the first trial was no six*. Indicate the set F of "forbidden" states. (You do **not** have to compute the conditional expected reward.)

Solution: The following Markov chain "remembers" if the first trial was a six (state 2) or not (state 3):



To express that the first trial was no six, we let the forbidden states be $F = \{2\}$. The desired conditional expected reward is then $ER(1, \Diamond G \mid \neg \Diamond F)$, where $G = \{4\}$.

Exercise 4 (From pGCL to conditional reward Markov chains) 15%

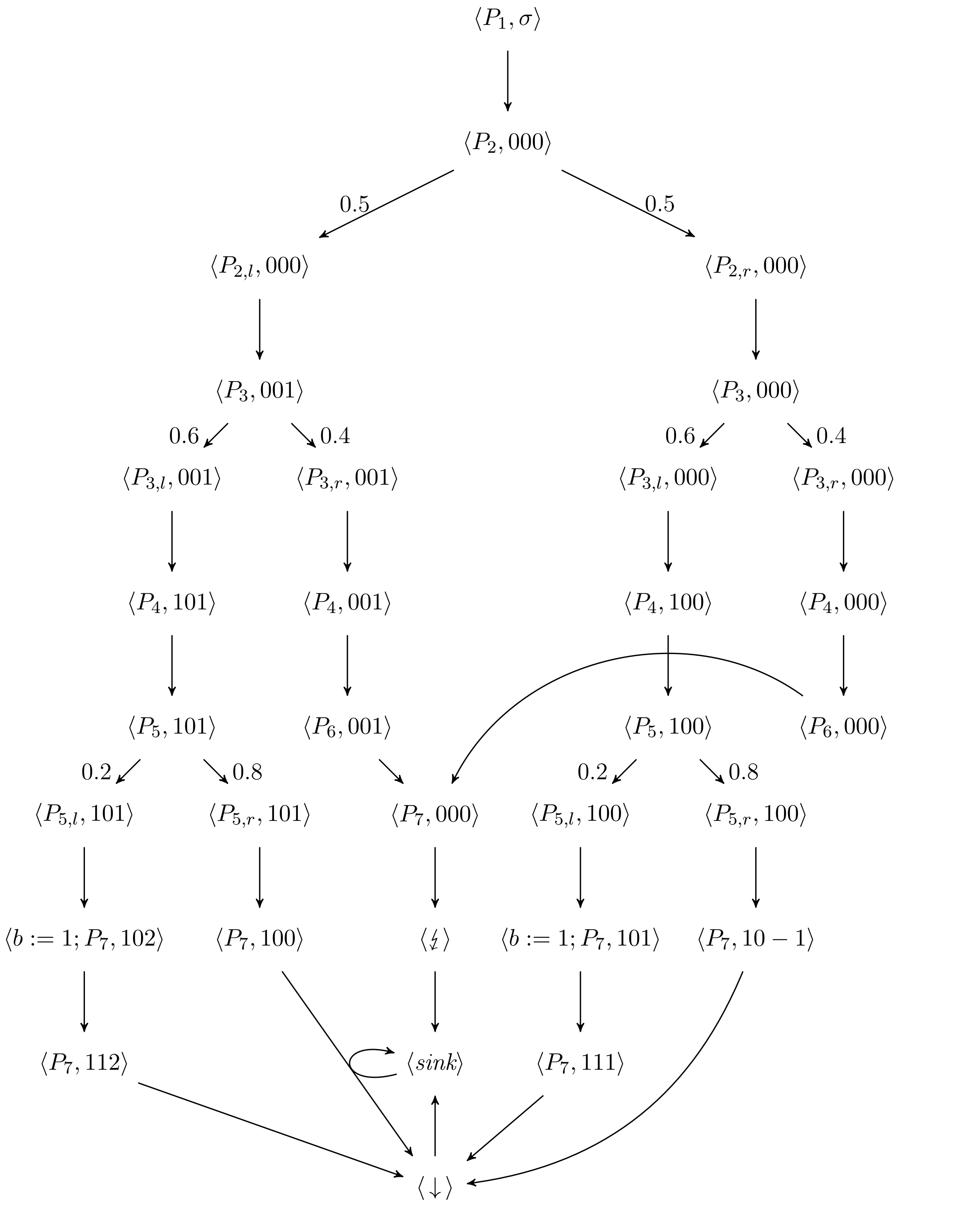
Consider the following `pGCL` program P:

$$
\begin{array}{l}
a, b, c := 0, 0, 0; \\
\{c := c + 1; \} \; [0.5] \; \{\mathsf{skip}\}; \\
\{a := 1\} \; [0.6] \; \{\mathsf{skip}\}; \\
\mathsf{if}(a = 1)\{ \\
\qquad \{c := c + 1; b := 1\} \; [0.2] \; \{c := c - 1\}; \\
\} \; \mathsf{else} \; \{ \\
\qquad c := 0 \\
\}; \\
\mathsf{observe} \; (a \neq 0 \;\lor\; b \neq 0)
\end{array}
$$

(a) [10%] Construct the reward Markov chain corresponding to P.

(b) [5%] Compute the expected value of c after termination of P.

Solution: We write P_i to denote the sub-program of P that starts with line i. Further, $P_{i,l}$ and $P_{i,r}$ denote the left and right branch of such a program in case of probabilistic choice or conditional statements. To simplify notation, we write $1, 2, 3$ instead of $\sigma[a \mapsto 1, b \mapsto 2, c \mapsto 3]$.



We assign a reward corresponding to the value of c to every state with first component P_7. All other states have a reward of 0.

Then the expected value of c after eventually reaching $\langle sink \rangle$ under the condition $\neg\langle \lightning \rangle$ is

$$\begin{aligned}
\mathit{ExpRew}(\Diamond\langle sink \rangle \mid \neg\langle \lightning \rangle) &= \frac{\mathit{ExpRew}(\Diamond\langle sink \rangle \cap \neg\Diamond\langle \lightning \rangle)}{\mathit{Pr}(\neg\Diamond\langle \lightning \rangle)} \\
&= \frac{0.5 \cdot 0.6 \cdot 0.2 \cdot 2 + 0.5 \cdot 0.6 \cdot 0.2 \cdot 1 + 0.5 \cdot 0.6 \cdot 0.8 \cdot (-1)}{1 - (0.5 \cdot 0.4 + 0.5 \cdot 0.4)} \\
&= \frac{0.12 + 0.06 - 0.24}{0.6} \;=\; -0.1
\end{aligned}$$

Exercise 5 (I.i.d-loops) 18%

For each of the following expectation-program pairs (f, P) prove or disprove that the loop P is i.i.d. with respect to expectation f. If applicable, give the exact runtime $ert(P, 0)$ using the closed-form formulation of the last lecture.

(a) [6%] $f = x + y + z$ and P is the following loop :

```
while (x + 2y + 4z >= 6) {
    { x:=0 } [1/2] { x:=1 };
    { y:=0 } [1/2] { y:=1 };
    { z:=0 } [1/2] { z:=1 }
}
```

Solution: The loop is i.i.d. with respect to f. Let P' be the loop body of P:

- $wp(P', [G]) = \frac{1}{8}\sum_{(x',y',z')\in\{0,1\}^3}[x'+2y'+4z' \geq 6] = \frac{1}{4}$. This expectation is constant and thus unaffected by the body of P.
- $wp(P', [\neg G]\cdot f) = \frac{1}{8}\sum_{(x',y',z')\in\{0,1\}^3}[x'+2y'+4z' < 6]\cdot(x'+y'+z')$. This expectation is also constant and thus unaffected by the body P'.

Since the loop is clearly a.s.-terminating (in each iteration, the probability to violate the guard is $\frac{3}{4}$) and each iteration runs in the same expected time 6, we can apply the rule from the lecture:

$$ert(P, 0) = 1 + [G]\frac{1 + ert(P', 0)}{1 - wp(P', [G])} = 1 + [x+2y+4z >= 6]\frac{7}{1 - \frac{1}{4}} = 1 + [x+2y+4z >= 6]\frac{28}{3}$$

Thus for an initial state that satisfies the loop guard, the expected runtime is $\frac{31}{3} \approx 10.33$.

(b) [6%] $f = x$ and P is the following loop:

```
while (x > 0) {
    { x:= x - 1 } [1/2] { skip };
}
```

Solution: This loop is not i.i.d. with respect to expectation $f = x$. Let P' be the body. It is

$$wp(P', [G]) = \frac{1}{2}[x - 1 > 0] + \frac{1}{2}[x > 0] =: g\ .$$

Clearly, x is a variable (in the sense of Lecture 19, slide 7) of the above pre-expectation g which *is* affected by the loop body P'. Therefore the loop P cannot be i.i.d.

(c) [6%] $f = 1$ and P is the following loop:

```
while (y = 42) {
    { x:= x + 1 } [1/2] { x := x + 2 };
}
```

Solution: Formally, the loop is i.i.d. with respect to $f = 1$. Let P' be the loop body of P:

- $wp(P', [G]) = \frac{1}{2}[y = 42](x/x + 1) + \frac{1}{2}[y = 42](x/x + 2) = [y = 42]$ because x does not appear in $[G] = [y = 42]$. So $wp(P', [G])$ is unaffected by P'.
- $wp(P', [\neg G] \cdot 1) = \frac{1}{2}[y \neq 42](x/x + 1) + \frac{1}{2}[y \neq 42](x/x + 2) = [y \neq 42]$. So $wp(P', [\neg G] \cdot f)$ is also unaffected by P'.

However, since P diverges with probability 1, we cannot formally apply the rule for inferring its runtime (which is of course ∞).

Exercise 6 (Probabilistic databases) **25%**

The example data of Figure 1 is about a few kids and their furry friends (stuffed animals). Every kid has exactly one furry friend. The "ff" attribute in the "kids" table refers to the 'fid' of the furry friend in the "furryfriends" table. The "dictionary" table contains all alternatives and their probabilities. The data was constructed from the observations of a certain person, but this person is not sure what he has seen or heard. For example, he heard two kids names "Tom" and "Tommy" playing with a grey furry friend and saying "peep peep", so it has to be a mouse or a rat. He is unsure whether this is one kid or two different kids. Furthermore, "Lindsay" and "Betty" have a brown furry friend. We know that a brown furry friend can only be a bear, a dog, or a monkey and that they have a different one.

(a) [2%] How many possible worlds does the probabilistic database of Figure 1 contain? Explain your answer by giving the complete calculation of the answer.

> **Solution:**
>
> There are 6 random variables with varying numbers of alternatives, so by independence of random variables $2 \times 2 \times 2 \times 2 \times 3 \times 2 = 96$ possible worlds.

(b) [2%] The sentence of the first record $\langle 1, \text{bear}, \text{brown} \rangle$ in the "furryfriends" table is not given. Provide this sentence, so that the database contents correspond with the above story of what was seen and heard and what we know to be true.

kids

⟨id, name, ff⟩	φ
⟨1, lindsay, 1⟩	$l = 1$
⟨1, lindsay, 2⟩	$l = 2$
⟨1, lindsay, 3⟩	$l = 3$
⟨2, betty, 1⟩	$\neg(l = 1) \wedge b = 1$
⟨2, betty, 2⟩	$l = 1 \wedge b = 1$
⟨2, betty, 2⟩	$l = 3 \wedge b = 2$
⟨2, betty, 3⟩	$\neg(l = 3) \wedge b = 2$
⟨3, tommy, 4⟩	$s = 1 \wedge i = 1$
⟨3, tom, 4⟩	$s = 1 \wedge i = 2$
⟨3, tom, 4⟩	$s = 0$
⟨3, tommy, 5⟩	$s = 0$

furryfriends

⟨fid, animal, color⟩	φ
⟨1, bear, brown⟩	**Question (b)**
⟨2, monkey, brown⟩	$l = 2$
⟨2, monkey, brown⟩	$l = 1 \wedge b = 1$
⟨2, monkey, brown⟩	$l = 3 \wedge b = 2$
⟨3, dog, brown⟩	$l = 3 \vee b = 2$
⟨4, mouse, grey⟩	$f_1 = 1$
⟨4, rat, grey⟩	$f_1 = 2$
⟨5, mouse, grey⟩	$f_2 = 1 \wedge s = 0$
⟨5, rat, grey⟩	$f_2 = 2 \wedge s = 0$

dictionary

φ	**prob**	**description**
$s = 0$	0.8	Tom and Tommy not same kid
$s = 1$	0.2	Tom and Tommy same kid
$i = 1$	0.4	Inconsistency: Tom is right name
$i = 2$	0.6	Inconsistency: Tommy is right name
$f_1 = 1$	0.8	Tom's grey furry friend is a mouse
$f_1 = 2$	0.2	Tom's grey furry friend is a rat
$f_2 = 1$	0.8	Tommy's grey furry friend is a mouse
$f_2 = 2$	0.2	Tommy's grey furry friend is a rat
$l = 1$	0.4	Lindsay has a bear
$l = 2$	0.3	Lindsay has a monkey
$l = 3$	0.3	Lindsay has a dog
$b = 1$	0.5	Betty has the first of the other two
$b = 2$	0.5	Betty has the second of the other two

Figure 1: Example probabilistic data on kids and their furry friends.

Solution: The record exists if either "Lindsay" or "Betty" have a brown bear, so $l = 1 \lor b = 1$ (or any sentence equivalent with this one).

(c) [2%] In case "Tom" and "Tommy" and different kids, is it possible according to the database that they have the same furry friend (i.e., both a mouse or both a rat)? Explain your answer.

Solution: Yes, because in possible worlds $s = 0 \land f_1 = 1 \land f_2 = 1$ they both have a mouse, and in possible worlds $s = 0 \land f_1 = 2 \land f_2 = 2$ they both have a rat.

(d) [3%] Calculate the probability of $(l = 3 \lor b = 2) \land (f_2 = 2 \land s = 0)$. Explain your answer by giving the complete calculation of the answer.

Solution:

$$
\begin{aligned}
& P((l = 3 \lor b = 2) \land (f_2 = 2 \land s = 0)) = \\
&\quad = P(l = 3 \lor b = 2) \times P(f_2 = 2 \land s = 0) = \\
&\quad = (P(l = 3) + P(b = 2) - P(l = 3) \times P(b = 2)) \times P(f_2 = 2) \times P(s = 0) = \\
&\quad = (0.3 + 0.5 - 0.3 \times 0.5) \times 0.2 \times 0.8 = \\
&\quad = 0.65 \times 0.2 \times 0.8 = \\
&\quad = 0.104
\end{aligned}
$$

(e) [3%] Given the following DuBio query (the attribute containing the sentences is denoted with φ in the query)

```
SELECT f.animal, round(prob(d.dict,k.φ & f.φ)::numeric,4) AS probability
FROM kids k, furryfriends f, dictionary d
WHERE k.name='tommy' AND k.ff=f.fid;
```

You are allowed to consult `https://github.com/utwente-db/DuBio/wiki` for the documentation of DuBio.

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

◯ The query returns two rows: one for "mouse" and one for "rat"

√ **The query constructs the correct sentence for a join between "kids" and "furryfriends".**

◯ The query produces an uncertain table.

√ **The type of $\texttt{k.}\varphi$ needs to be `Bdd` in DuBio.**

Solution:

- It returns four rows, because from "kids" it selects two rows that each join with two other rows of "furryfriends". It is of course true that the resulting rows have either "mouse" or "rat" in them, but they are each produced twice by the query.
- A join calls for an 'and' of the sentences, so $\texttt{k.}\varphi$ `&` $\texttt{f.}\varphi$ is correct.
- An uncertain table is a table with sentences in it. This table produces certain data: records with a string and a number and no sentences. Although there are probabilities in the resulting table, these are certain probabilities.
- The type for a sentence in DuBio is indeed `Bdd`.

(f) [8%] Given the probabilistic algebra expression E below ('k.a' refers to attribute a of table "kids"; 'f.a' refers to attribute a of table "furryfriends").

$$\pi_{f.animal}(\bowtie_{\text{k.ff=f.fid}} (\sigma_{\text{k.name='lindsay'}}(\text{kids}), \text{furryfriends}))$$

Since there are 3 records with name "lindsay" and 5 records in 'furryfriends' with a matching "fid", the query can have many results. Give exactly *one result* of E.
NB: I do not ask for a derivation, nor do I ask for all results, but I do ask for one of those many results *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.

Solution: The result of a probabilistic algebra expression, is a probabilistic table itself, i.e., a set of pairs containing a record and a sentence. One result, hence, is a pair of a record with a sentence. One such results of this query is $(\langle \text{dog} \rangle, l = 3 \vee b = 2)$. Its sentence is obtained by an *and* (because of the join) between the sentence of the $\langle 1, \text{lindsay}, 3 \rangle$ record of "kids" and the sentence of the $\langle 3, \text{dog}, \text{brown} \rangle$ record of "furryfriends": $(l = 3) \wedge (l = 3 \vee b = 2$, which can be simplified to $l = 3 \vee b = 2$.

(g) [5%] In indeterministic duplicate detection, an M-graph is constructed from similarity match results of a duplicate detection tool which ran on tuples a, b, c, d. The tool determines the following similarities: $s(a - b) = 0.5$, $s(a - c) = 0.95$, $s(a - d) = 0.8$, $s(b - c) = 0.5$, $s(b - d) = 0.3$, $s(c - d) = 0.8$. We set the upper threshold to 0.9 and the lower threshold to 0.4.

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.

Solution: Enforcing the thresholds has as a consequence that we consider $a - c$ as a certain edge, no edge between b and d, and all other edges as uncertain.
With four uncertain edges, there are $2^4 = 16$ possible W-graphs. A W-graph can be inconsistent because equality is transitive. Since $a - c$ is a certain edge, we know that each possible database has a merge between a and c. Since $b - d$ has no edge, we also know that each possible database has them seperately. This leaves the following possibilities.

- $\{acd, b\}$
- $\{ac, b, d\}$
- $\{abc, d\}$