Exam Re-Take

Uncertain Data Management
Université Paris-Saclay, M2 Data&Knowledge

May 20th, 2017

This is the re-take of the final exam for the Uncertain Data Management class. The grade in this exam will replace your grade of the first session of the final exam, and will become your final grade for the class. The exam consists of two independent exercises. Each exercise *must* be answered on a separate sheet of paper. Each sheet *must* be numbered and carry your name on the top right.

No additional explanations will be given during the exam, and no questions will be answered. If you think you have found an error in the problem statement, you should report on your answer sheet what you believe to be the error, and how you chose to interpret the intent of the question to recover from the alleged error.

You are allowed up to two A4 sheets of personal notes (i.e., four page sides), printed or written by hand, with font size of 10 points at most. If you use such personal notes, you must hand them in along with your answers. You may not use any other written material.

The exam is strictly personal: any communication or influence between students, or use of outside help, is prohibited. No electronic devices such as calculators, computers, or mobile phones, are permitted. Any violation of the rules may result in a grade of 0 and/or disciplinary action.

Exercise 1: Constructing probabilistic relations (10 points)

We consider one relation $T(\mathsf{pupil}, \mathsf{teacher})$ indicating which pupil follows classes by which teacher, and one relation $M(\mathsf{manager}, \mathsf{subordinate})$ indicating which teacher is a direct manager of which teacher. The relation are given below as tuple-independent (TID) instances, with probabilities expressed as $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$:

T

pupil	teacher	
Jane	Alice	p_a
Jane	Bob	p_b
Jane	Carol	p_c

M

manager	subordinate	
Alice	Bob	q_{ab}
Alice	Carol	q_{ac}
Bob	Carol	q_{bc}

We consider the Boolean conjunctive query Q asking whether there is a pupil who is taught by some teacher and who is also taught by some manager of that teacher.

We will be interested in the probability of the query Q on T and M, as a function of the values of $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$. We denote this probability by $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc})$. Here are some examples:

- If $p_a = p_b = p_c = q_{ab} = q_{ac} = q_{bc} = 0$ then we have $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 0$.
- Whenever $q_{ac} = 0$ and $p_b = 0$ then we always have $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 0$.
- Whenever $p_a = p_b = q_{ab} = 1$ then we always have $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 1$.

- Whenever $p_c = 0$, $p_a = p_b = 1$, and $q_{ab} = 0.5$, then we always have $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 0.5$.

All questions are independent except when indicated otherwise. No justification is expected for your answers unless indicated otherwise.

Question 1 (1 point). Express the query Q in the relational algebra.

> **Answer.**
>
> $$Q : \pi_{\emptyset}(T \bowtie \rho_{\mathsf{teacher} \mapsto \mathsf{teacher2}}(T) \bowtie \rho_{\mathsf{manager} \mapsto \mathsf{teacher}, \mathsf{subordinate} \mapsto \mathsf{teacher2}}(M))$$

Question 2 (1 point). Express the query Q in the relational calculus.

> **Answer.**
>
> $$\exists p\, m\, s\ T(p, m) \wedge T(p, s) \wedge M(m, s)$$

Question 3 (1 point). Give a choice of values for $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ such that $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 0.42$.

> **Answer.** Take $p_a = p_b = 1$, $p_c = 0$, and $q_{ab} = 0.42$.

Question 4 (1 point). Give a choice of values for $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ such that *all these values are in* $\{0, 0.5, 1\}$ and $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 3/4$.

> **Answer.** Take $p_c = 1$, $p_a = p_b = 0.5$, $q_{ac} = q_{bc} = 1$ and $q_{ab} = 0$.

Question 5 (1 point). The *denominator* of a rational value $0 \leq v \leq 1$ is the smallest positive integer p such that pv is an integer. In other words, p is the denominator of v when written as a fraction in irreducible terms. For instance, the denominator of 0 and 1 are 1, the denominator of 0.5 is 2, the denominator of 3/7 is 7, the denominator of 2/8 is 4, and the denominator of 0.42 is 50.

Explain briefly why, for any rational values $0 \leq v \leq 1$ and $0 \leq v' \leq 1$, if the denominator of v is d and the denominator of v' is d', then the denominator of vv' is a divisor of dd'. Give an example of *positive* v and v' where the denominator of vv' is equal to dd'. Give another example where it is strictly less than dd'.

> **Answer.** Write $v = p/d$ and $v' = p'/d'$ as irreducible fractions. We have $vv' = (pp')/(dd')$. We can simplify this fraction to an irreducible fraction by dividing the numerator and denominator by the greatest common divisor of pp' and dd', and we obtain an irreducible fraction representation of vv' whose denominator is a divisor of the original fraction, namely, dd', so the claim is proven.
>
> For $v = 1/2$ and $v' = 1/2$, we have $d = d' = 2$, and the denominator of $vv' = 1/4$ is 4 which is dd'.
>
> For $v = 1/2$ and $v' = 2/3$, we have $d = 2$, $d' = 3$, and the denominator of $vv' = 1/3$ is 3 which is less than $dd' = 6$.

Question 6 (2 point). Using the previous question, show that, for any choice of $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ where all values are in $\{0, 0.5, 1\}$, the denominator of $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc})$ is at most 64.

> **Answer.** When the probability values are all in $S = \{0, 0.5, 1\}$, the probability of any possible world can be expressed as a product of six values which are either in S or of the form $(1 - s)$ for some $s \in S$, i.e., which are in S. Now, the denominator of values in S is ≤ 2, so the denominator of the product of six values in S is a divisor of $2^6 = 64$ by the previous question. This means that the probability of each possible world can be written as a fraction whose denominator is equal to 64.
>
> Now, the probability of the query is a sum of the probability of the possible worlds where it is true, and it is clear that the sum of fractions with denominator 64 is a rational number whose denominator divides 64, which concludes.

Question 7 (2 point). Show that, for any choice of values $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ that are all in $\{0, 0.5, 1\}$, we have either $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 0$ or $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) \geq 1/8$.

Give an example of a choice of $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ in $\{0, 0.5, 1\}$, such that $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 1/8$.

> **Answer.** Consider a choice of $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ in $\{0, 0.5, 1\}$ and assume that we have $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 0$. It must then be the case that there are two rows $\{t, t'\}$ in T and one row $\{m\}$ in M such that $\{t, t', m\}$ (without probabilities) satisfies Q, and the probabilities assigned to t, t', and m are all positive. This means they must be at least equal to 0.5, so the probability that all three rows t, t', m' are present is at least $(0.5)^3 = 1/8$. Now, in all possible worlds where all these rows are present, the query holds, so indeed the probability that the query holds is at least $1/8$.
>
> One choice where the probability is equal to $1/8$ is obtained by setting $p_a = p_b = 0.5$, $p_c = 0$, and $q_{ab} = 0.5$.

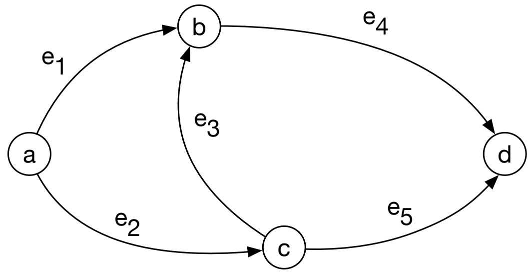
Question 8 (1 point). Give a choice of values $p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}$ that are all in $\{0, 0.5, 1\}$ such that $f(p_a, p_b, p_c, q_{ab}, q_{ac}, q_{bc}) = 7/16$.

> **Answer.** Take $p_a = p_b = 0.5$, $p_c = 1$, $q_{ab} = 0$, and $q_{ac} = q_{bc} = 0.5$. The query is now true iff Jane is taught by Alice and Alice is the manager of Carol, or Jane is taught by Bob and Bob is the manager of Carol, which are independent events. So the overall probability is $1 - (1 - (0.5)^2)^2 = 1 - (3/4)^2 = 7/16$.

Exercise 2: Reachability Queries on Probabilistic Graphs (10 points)

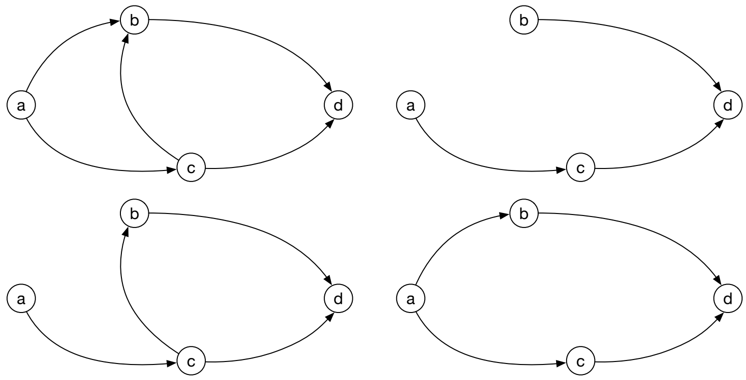
In this exercise, we will consider reachability queries on probabilistic graphs. A *probabilistic* graph is a directed graph in which edges are annotated with probabilistic events. In this exercise we consider *independent edge events*: each edge is annotated by a specific probabilistic event e, which is associated to a probability $P(e)$ that the edge exists, assuming independence between events.

For the following questions, we will work with the following probabilistic graph, where edges are annotated with events e_1, e_2, e_3, e_4, e_5:



Question 1 (2 points). Consider the following function P assigning probabilities to events: $P(e_1) = 0.5$, $P(e_2) = 1$, $P(e_3) = 0.5$, $P(e_4) = 1$, and $P(e_5) = 1$. How many *distinct possible worlds*, having non-zero probability, are there for the probabilistic graph $\mathcal{G}$? Draw each of them, and write down their probabilities. (Hint: possible worlds for a probabilistic graph have the same semantics as in TID databases.)

Answer. Only 2 edges, e_1 and e_3, are uncertain to exist. Hence, we have $2^2 = 4$ possible worlds, each with a probability of $0.5 \times 0.5 \times 1 \times 1 \times 1 = 0.25$. The corresponding 4 graphs are the following:



Question 2 (3 points). On a probabilistic graph, a reachability query $r(s,t)$ asks whether there exists at least one path in the graph between the nodes s and t. Moreover, the lineage $\Phi(s,t)$ is the logical expression (where terms are the edge events described above) which encodes the condition under which there exists at least one path between s and t.

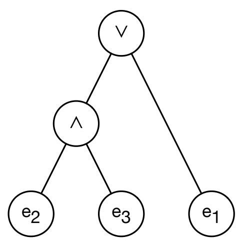
For instance, for $r(a,b)$ on $\mathcal{G}$, the corresponding lineage is:

$$\Phi(a,b) = e_1 \vee (e_2 \wedge e_3),$$

signifying that at least one of two paths, $a \to b$ or $a \to c \to b$, should exist for $r(a,b)$ to be true.

Draw the *read-once circuit* corresponding to $\Phi(a,b)$. Write down the probability formula for $r(a,b)$, as a function of edge events e_1, e_2, e_3, e_4, e_5. Compute the final probability using the probabilities given in Question 1.

Answer. The corresponding read-once circuit is the following:



By applying the formulas for independent probabilities, the probability formula is the following:

$$P(\Phi(a,b)) = 1 - (1 - P(e_1)) \times (1 - P(e_2) \times P(e_3)).$$

The actual probabilities are then:

$$\begin{aligned} P(\Phi(a,b)) &= 1 - (1 - 0.5) \times (1 - 0.5 \times 1) \\ &= 1 - (1 - 0.5) \times (1 - 0.5) \\ &= 1 - 0.25 = 0.75. \end{aligned}$$

Question 3 (3 points). Consider now $r(a,d)$ on $\mathcal{G}$. Write down the corresponding lineage, $\Phi(a,d)$. Write down its probability formula, as a function of edge events e_1, e_2, e_3, e_4, e_5, and compute the probability, using the probabilities given in Question 1.

Answer. The lineage $\phi(a,d)$ is:

$$\Phi(a,d) = (e_1 \wedge e_4) \vee (e_2 \wedge e_3 \wedge e_4) \vee (e_2 \wedge e_5).$$

To compute the probability formula, we need to do first a Shannon expansion on either e_2 or e_4. Expanding on e_4:

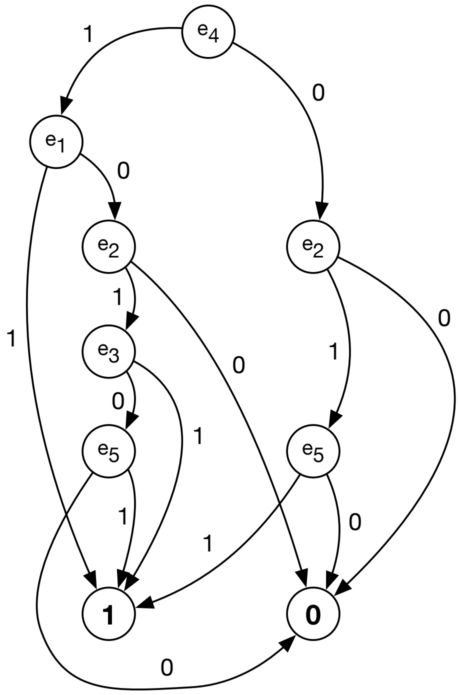
$$\begin{aligned} P(\Phi(a,d)) &= P(e_4 = \text{True})P(e_1 \vee (e_2 \wedge (e_3 \vee e_5))) + P(e_4 = \text{False})P(e_2 \wedge e_5) \\ &= P(e_4)(1 - (1 - P(e_1))(1 - P(e_2)(1 - (1 - P(e_3)(1 - P(e_5))))))+ \\ &\quad + (1 - P(e_4))P(e_2)P(e_5). \end{aligned}$$

To compute the probability, we can easily see from the answer in Question 1 that there always exists a path between a and d, in any possible world of $\mathcal{G}$, so the probability equals 1. Another way is to plug the probabilities in the above formula, and reach the same answer.

Question 4 (2 points). Can $\Phi(a,d)$ be rewritten as a *read-once* formula? Give a read-once rewriting, or explain why none exists. Draw an OBDD corresponding to $\Phi(a,d)$.

Answer. There is no rewriting of $\Phi(a,d)$ so that it is a *read-once* formula. The culprits are e_2 and e_4: both appear in different clauses and cannot be factored out without the other appearing in more than one place.

A possible OBDD is:



e4
1
0
e1
0
e2
e2
1
0
e3
1
1
0
0
1
e5
e5
1
0
1
1
0
0