Exercices

Databases 2 tutorial, M2 Data Science

Mikaël Monet

February 12th, 2024

1 Exercice 1: query evaluation over tuple-independent databases

Consider the following tuple-independent database (TID) $T = (D, \pi)$, with the first table recording who teaches which course and the second table recording which courses are part of which programs:

Teaches		π
Mary	Web technologies	0.3
Charles	Databases	0.8
Mikaël	Databases	0.5
Jean	Databases	0.7
Jean	Logics	0.2
Sylvain	Logics	1

Courses		π
Databases	Data Science master	0.3
Databases	ML master	0.6
Logics	L3 math	0.5
Web technologies	L2 CS	0.8

Q1. What is the probability of the following possible world?

Teaches	
Mikaël	Databases
Sylvain	Logics
Jean	Logics

Courses	
Web technologies	L2 CS

Q2. Let q_1 be the Boolean query asking whether there are two people that teach the Databases course; formally $q_1 = \exists x, y : \ x \neq y \wedge \text{Teaches}(x, \text{Databases}) \wedge \text{Teaches}(y, \text{Databases})$. What is the probability of q_1 on T?

Q3. More generally, give a formula to compute the probability of q_1 over an arbitrary TID T'.

Q4. Let q_2 be the Boolean query asking whether there are two people that teach the same course; formally $q_2 = \exists x, y, z :\ x \neq y \wedge \text{Teaches}(x, z) \wedge \text{Teaches}(y, z)$. What is the probability of that query on T?

Q5. More generally, give a formula to compute the probability of q_2 over an arbitrary TID T'.

Q6. Let q_3 be the Boolean query asking whether there exists someone teaching a course and there exists a course that is part of some program (i.e., the two tables are not empty); formally q_3 can be expressed as the SJFBCQ $q_3 = \exists x, y, z, t :\ \text{Teaches}(x, y) \wedge \text{Courses}(z, t)$. Is q_3 hierarchical?

Q7. What is the probability of q_3 over T?.

Q8. Propose an algorithm to compute the probability of q_3 over an arbitrary TID.

Q9. Let q_4 be the Boolean query asking whether there exists someone teaching a course that is part of some program; formally q_3 can be expressed as the SJFBCQ $q_3 = \exists x, y, z :\ \text{Teaches}(x, y) \wedge \text{Courses}(y, z)$. Is q_3 hierarchical?

Q10. Propose an algorithm to compute the probability of q_4 over an arbitrary TID.

Q11. We now extend the schema to contain an additional unary table WearsGlasses indicating which teachers wear glasses, and we consider the query q_5 asking whether there exists a teacher that wears glasses and teaches a course that is part of some program, i.e., q_5 is the SJFBCQ $\exists x, y, z :\ \text{WearsGlasses}(x) \wedge \text{Teaches}(x, y) \wedge \text{Courses}(y, z)$. Is q_4 hierarchical?

Q12. Prove that $\mathbf{PQE}_{\text{TID}}(q_4)$ is #P-hard by a reduction from #PP2DNF.

2 Block-independent databases

One shortcoming of tuple-independent databases is that they are not able to represent arbitrary probabilistic databases (as we defined them in the course). Consider the probabilistic database $\mathcal{D} = (W, \Pr)$ with $W = \{D_1, D_2\}$, $\Pr(D_1) = \Pr(D_1) = 0.5$ and D_1 contains only the fact P(c) and D_2 only the fact P(r).

Q1. Show that there is no TID that can represent $\mathcal{D}$.

Another way to represent probabilistic databases is using *Block-independent databases* (BIDs), which extend TIDs with the possibility of expressing that some tuples are mutually exclusive. Formally, a BID B consists of a DB D and a probability function $\pi : D \to [0, 1]$, and D is partitioned into so-called disjoint "blocks" $D_1, \ldots, D_m$ that together form D (i.e., $D = \bigcup_{i=1}^{m} D_i$ with the union being disjoint). The function π satisfies that for every block D_i we have that $\sum_{f \in D_i} \pi(f) \leq 1$. The semantics is that the blocks are independent of each others, and inside one block the facts are mutually exclusive with their given probabilities. The probabilities inside a block do not necessarily sum-up to 1, to account for the possibility that no tuple in that block is present.

For instance the following BID forecasts who will become president of what in 2024; it has only two blocks (delimited with the horizontal line).

President		
Régis	Univ. Lille	0.9
Charles	Univ. Lille	0.1
Thomas	Centrale Lille	0.3
Jean	Centrale Lille	0.1
Mary	Centrale Lille	0.2

Possible worlds are again defined simply to be subsets of D. Then for instance, the probability of obtaining the following possible world is 0.1×0.2:

President	
Charles	Univ. Lille
Mary	Centrale Lille

Q2. Formalize what is the probabilistic database represented by an arbitrary BID B.

Q3. For the PDB $\mathcal{D}$ from **Q1**, is there a BID that represents it?

Q4. Show that any PDB represented by a TID can also be represented by a BID.

Q5. Is there a PDB that cannot be represented by a BID? (justify your answer)