Exam

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

February 6th, 2018

This is the final exam for the Uncertain Data Management class, which will determine your grade for this class in conjunction with the class project. The duration of the exam is 2 hours. The exam consists of 4 independent exercises: exercises 1, 2, and 4 are about uncertain data management (Antoine's part), and exercise 3 is about social data management (Silviu's part). **You must write your answer to exercise 3 on a *separate sheet* of paper.**

Write your name clearly on the top right of every sheet used for your exam answers, and number every page.

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.

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.

Exercise 1: From c-tables to Boolean c-tables (10 points)

Recall that a *c-table* is a relational table where tuples carry a *condition*. The *condition* is a Boolean formula on equalities and inequalities that involve constant values and variables (called *named nulls*). In this exercise, we consider c-tables where *nulls may only occur in conditions*, i.e., they do not occur elsewhere in the table.

Recall that a *valuation* of a c-table T is a function that maps each named null of T to a value. Given a valuation ν of T, we can apply it to T by replacing each named null with its value in ν, evaluating the equalities and inequalities, evaluating the Boolean formulae, and keeping the tuples of T whose condition evaluates to true. The result of this process is a relational table, called a *possible world* of T, and we say that it is *obtained* from ν. Remember that a given possible world may be obtained by several different valuations.

Consider the following c-table, which describes upcoming events in the distant future in some major world cities:

Events		
city	**year**	
London	2042	$\texttt{NULL}_1 = \text{“GB”} \lor \texttt{NULL}_2 = \text{“Europe”}$
Mexico	2042	$\texttt{NULL}_1 = \text{“MX”}$
Singapore	2044	$\texttt{NULL}_1 = \text{“SG”} \land \texttt{NULL}_2 = \text{“Asia”}$

One example of a valuation is the function mapping $\texttt{NULL}_1$ to “GB” and $\texttt{NULL}_2$ to “Africa”. The possible world obtained by this valuation is the following relational table:

city	**year**
London	2042

Question 1 (0.5 point). Consider the valuation mapping $\texttt{NULL}_1$ to “MX” and $\texttt{NULL}_2$ to “Europe”. Write down the possible world obtained by this valuation. (No justification is expected.)

Question 2 (1 point). Is there a valuation that yields the following possible world? If there is one, give an example valuation; if there is none, explain why.

city	**year**
London	2042
Singapore	2044

Question 3 (1 point). How many different possible worlds are there? (No justification is expected.)

Question 4 (0.5 point). Consider the relational algebra query $Q := \Pi_{\mathbf{year}}(\textsf{Events})$ that projects the table Events on the **year** attribute. Write down the result of evaluating this query on the possible world obtained at question 1.

Question 5 (1 point). Is the following table a possible result of the query Q on table Events?

year
2042
2044

Question 6 (2 points). Construct a c-table T_6 with nulls appearing only in conditions that represents the result of evaluating the query Q on the c-table Events. In other words, for any valuation ν of $\texttt{NULL}_1$ and $\texttt{NULL}_2$, the possible world of T_6 obtained for ν should be the result of evaluating Q on the possible world of Events obtained for ν. (No justification is expected.)

Question 7 (3 points). Recall that a *Boolean c-table* is a c-table whose nulls can only be replaced by two values, True and False; in other words, we only consider valuations whose domain is $\{\textsf{True}, \textsf{False}\}$, unlike the c-tables considered so far where nulls can be replaced by arbitrary values.

Using the construction shown in class, construct a Boolean c-table T_7 whose possible worlds are exactly the possible worlds of the Events c-table. Explain briefly each step of your construction, and try to simplify the conditions that occur in T_7.

Question 8 (1 point). Construct a Boolean c-table T_8 that represents the result of evaluating the query Q on the c-table Events. (You do not need to use a minimal number of variables.)

Exercise 2: Probabilistic modeling (5 points)

You are part of a secret service which monitors which countries buy and sell which types of radioactive materials. As you do not expect to be sure of which country is doing what, you would like to represent the information as two TID tables, Buy and Sell, that would look like the following, with some probabilities p_1, p_2, p_3, p_4:

Buy

country	material	
Syldavia	uranium	p_1
Syldavia	plutonium	p_2
Borduria	uranium	p_3

Sell

country	material	
Borduria	uranium	p_4

You do not know what the probabilities p_1, p_2, p_3, p_4 should be. However, your field agents have been able to figure out the probability of some query results. Your job is to choose the right probabilities for p_1, p_2, p_3, p_4 so that the query results have the right probability.

For instance, consider the Boolean query $Q_0 := \Pi_\emptyset(\sigma_{\textbf{country}=\text{"Syldavia"}\wedge\textbf{material}=\text{"uranium"}}(\mathsf{Buy}))$. This query asks whether Syldavia is buying uranium. Your field agents have determined that this query has 50% probability of being true. To represent accurately this information in your TID tables, you will simply set $p_1 := 0.5$. This ensures that Q_0 has 50% probability of being true. In the rest of the exercise, we will determine p_2, p_3, p_4 in a similar way.

Question 1 (1 point). A new report from your field agents indicates that the Boolean query $Q_1 := \Pi_\emptyset(\sigma_{\textbf{country}=\text{"Syldavia"}}(\mathsf{Buy}))$ has 80% probability of being true. What is the value of p_2? (Remember that $p_1 = 0.5$.)

Question 2 (1 point). Your field agents now report that there is probability $\frac{5}{8}$ that Borduria is transacting uranium, i.e., it is either buying or selling it (or both). Write down the relational algebra query Q_2 corresponding to this field report.

Question 3 (1 point). Is the field report of Question 2 sufficient to determine the probabilities p_3 and p_4? Explain why, or why not.

Question 4 (2 points). A new field report arrives: there is, in addition, a $\frac{1}{8}$ probability that Borduria is *both* selling *and* buying uranium. Give all possible solutions for p_3 and p_4.

Hint: remember that quadratic equations of the form $ax^2 + bx + c = 0$ can be solved with the quadratic formula: compute $\Delta := b^2 - 4ac$, and then there are three cases:

- *If $\Delta < 0$ then there are no solutions;*
- *If $\Delta > 0$ there are two solutions: $x_1 = \frac{-b+\sqrt{\Delta}}{2a}$ and $x_2 = \frac{-b-\sqrt{\Delta}}{2a}$;*
- *If $\Delta = 0$ there is one solution $x_1 = x_2$ given by the formulas above.*

Exercise 3: Uncertain Graphs (3 points)

Consider an uncertain graph $\mathcal{G} = (V, E)$ as defined in class, where each edge $e \in E$ has attached a probability $p_e \in (0, 1)$, encoding the probability of e existing in the graph, and where all p_e's are considered independent.

Question 1 (1 point). How many *possible worlds* does the graph have (give the formula)? State which graph variables it depends on (number of nodes, number of edges, etc.). If we were to store this graph in a probabilistic relational database, which of the block-independent or tuple-independent models would we need?

Question 2 (2 points). Consider now an uncertain graph where each edge e is annotated with a Boolean formula on named nulls (the formulas are constructed like in Exercise 1). We assume that the graph has at least 1 and at most $n = |E|$ named nulls. Give the minimal number and the maximal number of possible worlds, and state on which variable(s) it depends on.

Exercise 4: From TID to BID (2 points or more)

Note: this exercise is substantially harder than the previous ones. It is recommended not to attempt it until you have solved the other exercises.

Recall that a relational algebra query is called *monotone* if it does not use the difference operator.

Question 1. Prove that, for every monotone relational algebra query Q and TID instance I, if the probabilistic instance $Q(I)$ cannot be represented by a TID instance, then it cannot be represented by a BID instance either.

Question 2. Give an example of a relational algebra query Q and TID instance I such that the probabilistic instance $Q(I)$ cannot be represented by a TID instance but can be represented by a BID instance.