Exam Re-Take

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

March 13th, 2018

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 4 independent exercises: exercises 1 and 2 are about uncertain data management (Antoine's part), and exercises 3 and 4 are about social data management (Silviu's part). **You must write your answer to exercises 3 and 4 on a *separate sheet* of paper.**

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: Knights of the Round Table (4 points)

Consider the following BID table representing the uncertain location of knights of the round table. The key attribute of the BID table is **knight**.

T

knight	**location**	
Lancelot	Camelot	0.4
Lancelot	Brocéliande	0.4
Galahad	Camelot	0.6

Question 1 (0.5 point). Which of the following two tables is a possible world of T? (No justification is expected.)

T_1

knight	**location**
Lancelot	Camelot
Galahad	Camelot

T_2

knight	**location**
Lancelot	Camelot
Lancelot	Brocéliande

Answer. Only T_1 is a possible world; T_2 violates the semantics of BID tables because it contains two tuples from the first block.

Question 2 (0.5 point). Consider the query Q_1 $\pi_{\textbf{location}}(\sigma_{\textbf{knight}=\text{“Galahad”}}(T))$ asking for the location of Galahad. Give a TID instance representing the result of evaluating Q_1 on the table T.

Answer.

location	
Camelot	0.6

Question 3 (1 point). Consider the query Q_2 asking for the knights located at Camelot. The query Q_2 should return a table with only one attribute, named **knight**. Write down the query Q_2 in the relational algebra.

Answer. The query Q_2 is defined by:

$$\pi_{\textbf{knight}}(\sigma_{\textbf{location}=\text{“Camelot”}}(T))$$

Question 4 (1 point). Consider the evaluation of Q_2 on the table T. Can the result be represented as a TID instance? If yes, give a suitable TID instance and justify; if not, prove that no TID instance can represent it.

Answer. The query result can be correctly represented as a TID instance as follows:

knight	
Lancelot	0.4
Galahad	0.6

The reason why this is correct is because the two tuples that match the selection of Q_2 are in different blocks, so they are independent.

Question 5 (1 point). Is there a query Q_3 such that the result of evaluating Q_3 on the table T cannot be represented as a pc-instance? If yes, give such a query Q_3 and justify; if not, prove that no such query Q_3 exists.

Answer. No such query Q_3 exists. Indeed, for any query Q_3, the probabilistic instance obtained by evaluating Q_3 on T can be represented as a pc-instance, because we have seen in class that any probabilistic instance can be represented as a pc-instance.

Exercise 2: Numbers of Possible Worlds (6$^+$ points)

In this exercise, recall that a *row* of a table is simply a line of the table, e.g., the BID instance T given in Exercise 1 has 3 rows.

Question 1 (1 point). Give an example of a TID instance that has at least 42 possible worlds.

Answer.

x	
a	0.5
b	0.5
c	0.5
d	0.5
e	0.5
f	0.5

Question 2 (1 point). If a TID instance has n rows, how many possible worlds can it have at most? Give the best possible bound, and justify why it is correct.

Answer. For each row, we know that either the row is present or it is absent, so we have at most 2^n possible worlds, and this bound can be achieved with a table like the one used in Question 1.

Question 3 (1 point). If a BID instance has n rows, how many possible worlds can it have at most? Give the best possible bound, and justify why it is correct.

Answer. For each row, we know that either the row is present or it is absent, so we have at most 2^n possible worlds. This bound can indeed be achieved with a TID instance (which is a special case of a BID instance) where we ensure that each tuple is in its own block. (In other words, having blocks with more than one tuple is never helpful to achieve a large number of possible worlds.)

Question 4 (1 point). Is there a TID instance having *exactly* 42 possible worlds? If yes, give an example; if not, explain why.

Answer. The number of possible worlds of a TID instance is always a power of two, and 42 is not a power of two, so there is no such TID instance.

Question 5 (2 points). Is there a BID instance having *exactly* 42 possible worlds? If yes, give an example; if not, explain why.

Answer.

x	y	
u	a	1/2
u	b	1/2
v	a	1/3
v	b	1/3
v	c	1/3
w	a	1/7
w	b	1/7
w	c	1/7
w	d	1/7
w	e	1/7
w	f	1/7
w	g	1/7

The number of possible worlds is $2 \times 3 \times 7 = 42$.

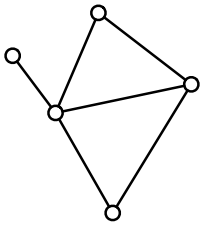
Question 6 (bonus). We consider a positive integer n, and search for a BID instance having exactly n possible worlds (if one exists) and whose number of rows is as small as possible. We call $r(n)$ the minimal number of rows of a BID instance having n possible worlds. Explain how to compute $r(n)$ as a function of n.

Answer. The number of possible worlds of a BID instance is the product of the number of possible worlds for each block, and the number of possible worlds of a block containing d rows is at most $d+1$, as can be obtained by setting the probabilities so that we keep either one of the tuples of the block or no tuple for this block. Hence, if we decompose n into prime factors $n = p_1 \times \cdots \times p_k$, then we can construct a BID instance having exactly n possible worlds and having a number of rows equal to $\sum_i (p_i - 1)$.

We argue that the bound is the best possible, so that $r(n)$ is equal to this value. The reason is that any BID instance having exactly n possible worlds implies a decomposition of n into factors, as given by the blocks. Hence, any BID instance with n possible worlds can be seen as obtained from the BID instance that we described, by merging blocks. Now it is obvious that merging blocks can never reduce the number of rows, because for any integers $p > 0$ and $q > 0$ it is clear that $(p-1) + (q-1)$ is no greater than $pq - 1$. This concludes the proof.

Exercise 3: Graph Measures (6 points)

In this exercise, we will work with some graph analysis measures, as defined in the course. Consider the following *undirected* graph G:



Question 1 (0.5 point). Write down the degree distribution of the graph G.

Answer. To compute the degree distribution, we have to compute, for each degree d, its probability $P(d)$ as the fraction between the number of nodes having degree d and the total nodes in the graph.

In this case:

$$P(1) = \frac{1}{5} = 0.2,$$
$$P(2) = \frac{2}{5} = 0.4,$$
$$P(3) = \frac{1}{5} = 0.2,$$
$$P(4) = \frac{1}{5} = 0.2.$$

Notice that is is indeed a distribution, that is $\sum_d P(d) = 1$.

Question 2 (1 point). What is the average degree $\langle k \rangle$ in the graph G? Explain how it can be computed from the degree distribution.

Answer. The average degree is equal to the expectation of the degree distribution, computed as:

$$\langle k \rangle = \sum_d d \times P(d),$$

where $P(d)$ is the probability of a node having degree d in the graph, computed above.

In our case:

$$\langle k \rangle = 1 \times 0.2 + 2 \times 0.4 + 3 \times 0.2 + 4 \times 0.2 = 2.4$$

Question 3 (0.5 point). How many triangles are there in the graph G? *Reminder*: a triangle is a subgraph of size 3 that is complete.

Answer. There are only 2 triangles in this graph, easily visible in the figure.

Question 4 (1 point). We work now with the Erdős-Rényi random graph model, and we wish to get random graphs having the same average degree as G. What is the resulting parameter p – the probability of an edge existing – in such a graph?

Answer. From the course, we know that in a random graph $\langle k \rangle = p \times (|V| - 1)$.

Then, in our case:

$$p = \frac{\langle k \rangle}{|V| - 1} = \frac{2.4}{4} = 0.6.$$

Question 5 (3 points). Compute the expected number of triangles in the random graph model, as a function of the number of nodes $|V| = n$ and of the parameter p. Now compute this value using the value of p obtained in Question 4, and using the same number of vertices as G, i.e., 5. How does it compare to the number of triangles in the example graph above?

Hint: you will need to use the formula for the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Answer. For computing the expected number of triangles in a random graph of n nodes and parameter p, we need two ingredients:

1. the number of possible triangles in the graph: $\binom{n}{3}$, and
2. the probability of a given triangle existing: this is $p \times p \times p$.

This results in:

$$\langle \Delta \rangle = \binom{n}{3} \times p^3.$$

Notice this works because of the linearity of expectation and the fact that the triangles in the graph are identically distributed.

In our case:

$$\langle \Delta \rangle = \binom{5}{3} \times p^3 = 10 \times 0.6^3 = 2.16.$$

This is comparable to what we counted in Question 3: 2 triangles.

Exercise 4: Random Graphs Modeled as Uncertain Graphs (4 points)

In this exercise, we will work with the Erdős-Rényi random graph model again. A random graph with parameter p can also be thought of as an uncertain graph $\mathcal{G} = (V, E, p)$, where the underlying graph is complete, i.e., there exists an edge between every node in both directions, and each edge has the same probability of existing, equal to p.

Question 1 (2 points). Take the random graph having $|V| = 3$ and $p = 0.2$, and take two distinct nodes s and t of this graph. Compute the reachability probability between s and t. *Hint*: the probability is the same no matter our choice for s and t.

Answer. There are two cases:

1. either the direct edge exists between s and t, with probability $p = 0.2$; or
2. the edge does not exist with probability $1 - p = 0.8$ and the other two edges in the graph exist with probability $p \times p = 0.04$.

Then the final probability is $P = p + (1 - p) \times p \times p = 0.2 + 0.8 \times 0.04 = 0.232$.

Question 2 (2 points). Now consider the general case, where we write $n := |V|$ and write p the probability parameter. We assume that $n > 1$, and choose two distinct nodes s and t. Compute the probability that s and t are at distance *exactly* 2: that is, to reach t from s, the shortest path passes through one intermediary node x (different from s and t).

Answer. Here, we have a conjunction of two events. First, the direct edge between s and t must not exist (because then the distance would be 1) – this occurs with probability $1 - p$. Second, at least one of the other $n - 2$ nodes x must act as the intermediary node on the path $s \to x \to t$; this occurs, for each possible node x, with probability $p \times p$.

Then the final probability $P(d_{st} = 2)$ that s and t are at distance 2 of each other is:

$$P(d_{st} = 2) = (1 - p)(1 - \prod_{n-2}(1 - p^2)) = (1 - p)(1 - (1 - p^2)^{n-2}).$$