Exam Re-Take

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

June 6th, 2016

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 four independent exercises.

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.

Write your name clearly on the top right of every sheet used for your exam answers. Number every page. It is highly recommended to answer the exercises on separate sheets.

The exam is strictly personal: any communication or influence between students, or use of outside help, is prohibited. Any violation of the rules may result in a grade of 0 and/or disciplinary action.

Exercise 1: Unions of TID (3 points)

Consider the three TID instances R, S, and T defined as follows:

R

attr1	**attr2**	
a	b	0.5

S

attr1	**attr2**	
c	d	0.8

T

attr1	**attr2**	
a	b	0.2

Question 1 (0.5 point). Let U_1 be the probabilistic instance defined by the query $R \cup S$ (in relational algebra). Write a representation of U_1 as a TID instance.

Question 2 (0.5 point). Define likewise U_2 as $R \cup T$. Write a representation of U_2 as a TID instance.

Question 3 (2 points). Show that the union $R_1 \cup R_2$ of two arbitrary TID instances R_1 and R_2 can always be represented as a TID instance U. Describe how U is constructed from R_1 and R_2.

Exercise 2: Unions of BID (7 points)

Consider the three BID instances R, S, and T defined as follows:

R

attr1	**attr2**	
a	b	0.3
a	c	0.4

S

attr1	**attr2**	
d	e	0.1
d	f	0.1

T

attr1	**attr2**	
a	g	0.1

Question 1 (0.5 point). Define U_1 as $R \cup S$. Write a representation of U_1 as a BID instance. How many blocks does U_1 contain?

Question 2 (1.5 points). Let U_2 be $R \cup T$. Prove that U_2 cannot be represented as a BID instance with key **attr1**.

Question 3 (1 point). Can U_2 be represented as a BID instance over **attr1**, **attr2** but with some other choice of key attributes? If yes, write such a representation; if not, prove that there is no such representation.

Question 4 (1 point). Give two *different* BID instances R' and S' over the schema **attr1**, **attr2** and with key **attr1** such that R' and S' both contain some common tuple t with some probability $0 < p < 1$, and yet $R' \cup S'$ can be represented as a BID instance. Specifically, write a suitable choice of BID R' and S' and write the representation of their union $R' \cup S'$ as a BID.

Question 5 (1 point). Let R_1 and R_2 be two arbitrary BID instances over the schema **attr1**, **attr2** with key **attr1**. Give a characterization of when $R_1 \cup R_2$ cannot be represented as a BID instance over this schema with this key. In other words, write a necessary and sufficient condition on R_1 and R_2 that holds if and only if $R_1 \cup R_2$ cannot be represented as a BID instance with key **attr1**. You are not required to prove that your proposed condition is correct. (Hint: use the examples of questions 1, 2, and 4 to verify that your condition correctly classifies them.)

Question 6 (1 point) Given a BID instance W over **attr1**, **attr2** with key **attr1**, we denote by $\overline{W}$ the result of keeping the same tuples with the same probabilities, but changing the key attribute to be **attr2** instead of **attr1**. For instance, remembering S from the beginning of the exercise, we define $\overline{S}$ as:

$\overline{S}$

attr1	**attr2**	
d	e	0.1
d	f	0.1

However, this operation is not always well-defined: i.e., it is not always possible to interpret its result as a BID instance. To illustrate this, give an example of a BID instance W over **attr1**, **attr2** with key **attr1** such that $\overline{W}$ cannot be interpreted as a BID instance, and quickly explain why.

Question 7 (1 point). Remember the BID instance R defined at the beginning of the exercise. Design a BID instance W' over **attr1**, **attr2** with key **attr1** such that $R \cup W'$ can be represented as a BID instance, $\overline{W'}$ is well-defined (i.e., it is a BID instance), but $\overline{R} \cup \overline{W'}$ cannot be represented as a BID instance. Use Question 5 to justify that your choice of W' satisfies these conditions.

Exercise 3: Probabilistic Query Processing (6 points)

Consider the following TID instances:

A

s	**m**	
a	b	p_1
a	c	p_2
b	c	p_3
b	d	p_4
b	e	p_5
c	d	p_6
c	e	p_7
e	d	p_8

B

m	**t**	
b	c	r_1
c	b	r_2
c	d	r_3
c	e	r_4

Question 1 (1 point). Consider the following query in relational calculus:

$$Q(s,t) := \exists x, y\ A(s,x) \land B(x,y) \land A(y,t).$$

Write an SQL query that evaluates $Q(a,e)$.

Question 2 (2 points). We assign probabilistic events to the tuples in A and B as follows:

A

s	**m**	
a	b	X_1
a	c	X_2
b	c	X_3
b	d	X_4
b	e	X_5
c	d	X_6
c	e	X_7
e	d	X_8

B

m	**t**	
b	c	Y_1
c	b	Y_2
c	d	Y_3
c	e	Y_4

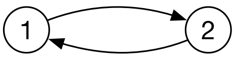
Consider again the query $Q(a,e)$. Write the lineage of the query, i.e., a Boolean formula in disjunctive normal form over the probabilistic events which evaluates to true iff the query is true when the corresponding facts are kept. Use this to write the probability $\Pr[Q(a,e)]$ that the query is true, as a function of the probability values $p_1, \ldots, p_8, r_1, \ldots, r_4$.

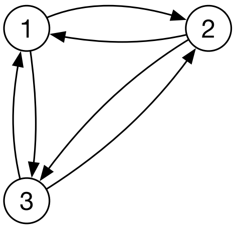
Question 3 (2 points). Consider the query $Q' := \pi_{\mathbf{s}}(\sigma_{\mathbf{s}=\text{‘a’}\lor\mathbf{s}=\text{‘b’}}(A \bowtie B))$ in relational algebra, and the following plan for Q': $\pi_{\mathbf{s}}\big(\sigma_{\mathbf{s}=\text{‘a’}\lor\mathbf{s}=\text{‘b’}}(A) \bowtie \pi_{\mathbf{m}}(B)\big)$. Compute Q' following this plan, detailing the steps on the instance represented above. Is it true that the probabilities obtained by evaluating the plan match the correct probabilities for the result of the query? Can the resulting relation be represented as a TID instance?

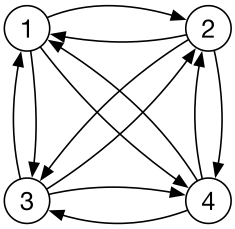
Question 4 (1 point). Consider the same plan as above in the general case, i.e., on any TID instance of the tables A and B. Is the plan safe in general? Briefly explain why or why not. What does this imply about the safeness of the query?

Exercise 4: Reachability Queries (4 points)

In this exercise, we will consider reachability queries on probabilistic complete graphs. A *complete graph* of n nodes, denoted K_n, is a directed graph where there exists an edge from each node to each node of the graph (excluding self-loops, i.e., there is never an edge from one node to itself). We represent the graphs K_2, K_3, and K_4 below:



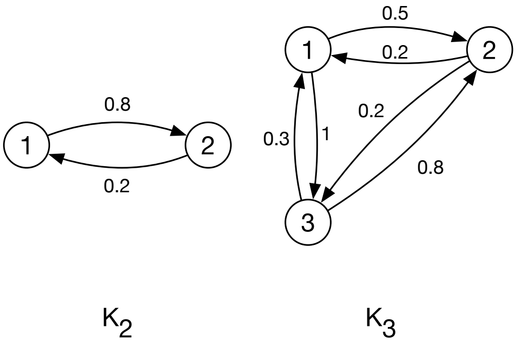




K_2 K_3 K_4

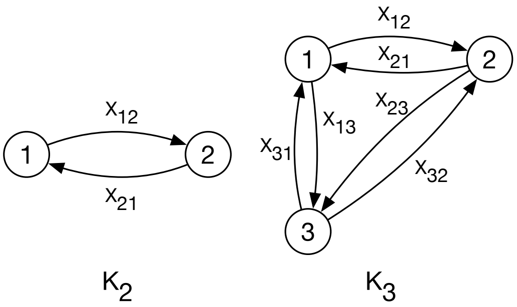
A *probabilistic complete graph* is a complete graph where each edge has a non-zero probability of existence. The *reachability query* $\mathsf{Reach}(s,t)$ on a probabilistic complete graph asks, given a source node s and a target node t with $s \neq t$, what is the probability that node t is reachable via a directed path from s.

Question 1 (2 points). Consider the following probabilistic complete graphs, along with their attached probabilities:



Compute the probability for $\mathsf{Reach}(1,2)$ in the graph K_2 above. Do the same in the graph K_3 above.

Question 2 (2 points). Consider the general case of the K_2 and K_3 graphs, where each edge is annotated with an independent probabilistic event:



Remember that a *lineage* of a query is a Boolean formula over the uncertain events of the graph which evaluates to true iff the query is true when the corresponding edges are kept; and that the lineage is *read-once* if it can be written in a form where each variable occurs at most once.

Show that, for any $n \in \{2,3\}$, for any vertices $s \neq t$ in K_n, any lineage for $\mathsf{Reach}(s,t)$ on K_n is read-once.