University of Zurich

Prof. Dan Olteanu,
Dr. Ahmet Kara, Dr. Nils Vortmeier,
Haozhe Zhang

DaST
Data • (Systems+Theory)

Spring 2021 | Exercise Sheet 2 | 17.05.2021

- The solutions will be discussed on Monday 31.05.2021, 16:15-17:45 on Zoom
- Solutions will be posted on OLAT after the exercise session

Exercise 2.1 [Yannakakis]
Consider the query

$$\Phi(x_1, x_2, x_3) = \sum_{x_4,x_5,x_6,x_7} \psi_{124}(x_1, x_2, x_4) \cdot \psi_{2345}(x_2, x_3, x_4, x_5) \cdot \psi_{3456}(x_3, x_4, x_5, x_6) \cdot \psi_{2347}(x_2, x_3, x_4, x_7)$$

over the natural sum-product semiring. Consider the following input factors, where n and m are arbitrary natural numbers and all tuples are mapped to the domain value 1.

ψ_{124}

X_1	X_2	X_4	
a_0	b_3	d_2	
a_i	b_1	d_2	for all $i \in [n]$
a_1	b_1	d_3	

ψ_{2345}

X_2	X_3	X_4	X_5	
b_1	c_3	d_i	e_0	for all $i \in [m]$
b_1	c_1	d_3	e_0	
b_3	c_i	d_2	e_0	for all $i \in [n]$

ψ_{3456}

X_3	X_4	X_5	X_6	
c_i	d_2	e_1	f_4	for all $i \in [n]$
c_1	d_3	e_0	f_i	for all $i \in [m]$
c_3	d_2	e_0	f_1	

ψ_{2347}

X_2	X_3	X_4	X_7	
b_1	c_1	d_3	g_i	for all $i \in [m]$
b_1	c_3	d_2	g_2	
b_3	c_1	d_2	g_4	

(a) Construct a join tree for Φ.

(b) Apply Yannakakis' algorithm. Give the fully reduced input factors, the intermediate views and the final result.

Exercise 2.2 [Bunnies Playing Leapfrog]

Watch online : `https://youtu.be/jkHyYsRk0oo`
Consider three sorted unary relations R, S, and T such that $R = \{0, \ldots, 2n-1\}$, $S = \{n, \ldots, 3n-1\}$, and $T = \{0, \ldots, n-1, 2n, \ldots, 3n-1\}$ for a natural number n.

(a) Compute the result of the join of the three relations.

(b) How many linear iterator steps are needed by the LeapFrog TrieJoin algorithm to compute the result assuming we first iterate over R, then S, and finally T? Explain your answer.

(c) Now assume we use the sort-merge algorithm (recall that the three relations are already sorted) and join the three relations all at the same time (so-called multi-way join). In the sort-merge algorithm the iterator of each relation visits one value after the previous one, without jumping as done by leapfrogging. How many steps of the relation iterators would we need to compute the result of the join? Explain your answer.

(d) Assume that we decide instead of the multi-way join to use binary sort-merge join, that is, we join two of the three relations and later join their intermediate result with the third relation. Explain whether this approach would need fewer or more iterator steps than the multi-way sort-merge approach.

(e) Consider the following three join orders: $Q(x) = (R(x) \otimes S(x)) \otimes T(x)$, $Q(x) = (R(x) \otimes T(x)) \otimes S(x)$, and $Q(x) = (S(x) \otimes T(x)) \otimes R(x)$. Which join order leads to the smallest intermediate result? For instance, in case of $Q(x) = (R(x) \otimes S(x)) \otimes T(x)$, the intermediate result is the result of joining R and S.

Exercise 2.3 [SAT]

(a) For $m \in \mathbb{N}$, consider the hypergraph $\mathcal{H}_m = (\mathcal{V}_m, \mathcal{E}_m)$ with
- $\mathcal{V}_m = \{0\} \cup [2m]$
- $\mathcal{E}_m = \{\{2i+2\}, \{2i+1, 2i+2\}, \{0, 2i+1, 2i+2\} \mid i \in \{0, \ldots, m-1\}\}$

Let $\mathbf{H} = \{\mathcal{H}_m \mid m \in \mathbb{N}\}$.

(i) Draw the hypergraphs $\mathcal{H}_2$ and $\mathcal{H}_3$.

(ii) Are the hypergraphs in $\mathbf{H}$ cyclic, α-acyclic, or β-acyclic?

(iii) Given any $m \in \mathbb{N}$, can you find a marginalisation order such that for any CNF formula F with hypergraph $\mathcal{H}_m$, the DP procedure with this order determines satisfiability for F in polynomial time?

(iv) Consider the following CNF formula with hypergraph $\mathcal{H}_2$:

$$F = (x_0 \vee \neg x_1 \vee x_2) \wedge (\neg x_0 \vee x_3 \vee x_4) \wedge (x_1 \vee \neg x_2) \wedge (\neg x_3 \vee x_4) \wedge (x_2) \wedge (\neg x_4)$$

Using the marginalisation order proposed under (2.3.a.iii), apply the DP procedure on F to determine whether F is satisfiable. Give after each marginalisation step the hypergraph of the resulting CNF formula.

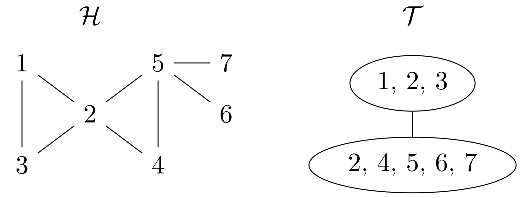
(b) Given any class of CNF formulas, assume that for every formula F of this class the DP procedure correctly determines satisfiability of F in polynomial time. What would this imply?

Exercise 2.4 [Solving FAQs]

(a) Consider the following FAQ-SS over the sum-product semiring (for brevity, input variables of factors are skipped)

$$\Phi(x_1, x_2, x_3) = \sum_{x_4} \sum_{x_5} \sum_{x_6} \sum_{x_7} \psi_{12} \cdot \psi_{23} \cdot \psi_{13} \cdot \psi_{24} \cdot \psi_{25} \cdot \psi_{45} \cdot \psi_{56} \cdot \psi_{57}$$

with hypergraph $\mathcal{H}$ and hypertree decomposition $\mathcal{T}$:



(i) Solve Φ using the hypertree decomposition $\mathcal{T}$. Apply all steps introduced in the lecture for solving FAQ-SS. Give the time complexity of your approach.

(b) Consider the multi-semiring FAQ

$$\Phi(x_1, x_3, x_7, x_8) = \prod_{x_4} \max_{x_6} \sum_{x_5} \max_{x_2} \psi_{12} \cdot \psi_{13} \cdot \psi_{34} \cdot \psi_{24} \cdot \psi_{25} \cdot \psi_{26} \cdot \psi_{37} \cdot \psi_{38}$$

using the natural max-product semiring $(\mathbb{N}, \max, \cdot, 0, 1)$ and the natural sum-product semiring $(\mathbb{N}, +, \cdot, 0, 1)$. Solve Φ using the InsideOut algorithm with marginalisation order $\sigma = (4, 6, 5, 2)$ and an optimal hypertree decomposition for the query after all bound variables are marginalised. Give the computation time at each marginalisation step. Give the computation time needed to materialise the bags of the hypertree decomposition. Give the overall computation time.

(c) Now consider the multi-semiring FAQ

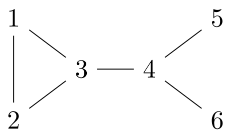
$$\Phi(x_1, x_4, x_7, x_8) = \prod_{x_3} \prod_{x_2} \sum_{x_5} \max_{x_6} \psi_{12} \cdot \psi_{13} \cdot \psi_{34} \cdot \psi_{24} \cdot \psi_{25} \cdot \psi_{26} \cdot \psi_{37} \cdot \psi_{38}$$

Solve Φ using the InsideOut algorithm with marginalisation order $\sigma = (3, 2, 5, 6)$ and an optimal hypertree decomposition for the query after all bound variables are marginalised. Give the computation time at each marginalisation step. Give the computation time to materialise the bags of the hypertree decomposition. Give the overall computation time.

Exercise 2.5 [Entropies and Size Bounds]
Consider the join

$$\Phi(x_1, \ldots, x_6) = \psi_{12}(x_1, x_2) \cdot \psi_{23}(x_2, x_3) \cdot \psi_{13}(x_1, x_3) \cdot \psi_{34}(x_3, x_4) \cdot \psi_{45}(x_4, x_5) \cdot \psi_{46}(x_4, x_6)$$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$:



(a) Assume that the output of Φ consists of the following tuples:

X_1	X_2	X_3	X_4	X_5	X_6
1	2	3	1	2	3
2	2	3	1	2	3
3	1	2	2	3	4
4	2	1	4	1	2
1	2	3	4	1	2
2	2	3	4	1	2
4	2	3	1	2	3

Assuming that the output tuples are uniformly distributed, the entropy of $\mathbf{X}_{[6]} = (X_1, \ldots, X_6)$ is $H(\mathbf{X}_{[6]}) = \log 7$. Compute the entropies $H(X_1, X_2)$, $H(X_1, X_3)$, $H(X_2, X_3)$, $H(X_3, X_4)$, $H(X_4, X_5)$, and $H(X_4, X_6)$.

(b) Give $\rho^*(\mathcal{H})$.

(c) Assume that each input factor of Φ has size N. Using Shearer's Lemma, show that $|\Phi| \leq N^{\rho^*(\mathcal{H})}$. You can do this by following the general upper-bound proof on Slide 23 of the slide deck on Worst-Case Optimal Size Bounds for Joins. How can you choose $(p_S)_{S\in\mathcal{E}}$, q, and $\mathcal{J}$ so that the proof works?

(d) Assume $N = 2^3$. Following the lower bound proof on Slides 30 and 31 of the slide deck on Worst-Case Optimal Size Bounds for Joins, explain how to construct input factors of size N^q such that $|\phi| \geq (N^q)^{\rho^*(\mathcal{H})}$. In particular, give the values for $(p_i)_{i\in\mathcal{V}}$ and q.