Department of Computer Science

Prof. Dr. Nikolaus Augsten
Jakob-Haringer-Str. 2
5020 Salzburg, Austria
Phone: +43 662 8044 6347
E-Mail: nikolaus.augsten@plus.ac.at

PARIS LODRON UNIVERSITÄT SALZBURG

Similarity Search in Large Databases
Winter Semester 2024/2025

Midterm Exam II
29.01.2025

Name: _____ **Student ID:** _____

Hints

- Check whether you received all pages of the exam (6 pages).
- Write your name or your student ID on each sheet of the exam and hand in all pages.
- All answers are expected to be written on the exam sheets.
- Clearly highlight and enumerate additional pages that are used for longer answers. Match your text with the according exercise.
- Only use pencils that are permanent and non-red colored.
- Use the notation and techniques discussed in the lecture.
- Exercises with more than one solution are not graded.
- You are allowed to use one A4 sheet with your personal notes (single-sided, hand written or printed).
- Exam duration: 75 minutes

Signature _____

Grading — Filled by the examiner

Exercise	1	2	3	4	5	Sum
Total points	7	7	7	7	7	35
Points reached						

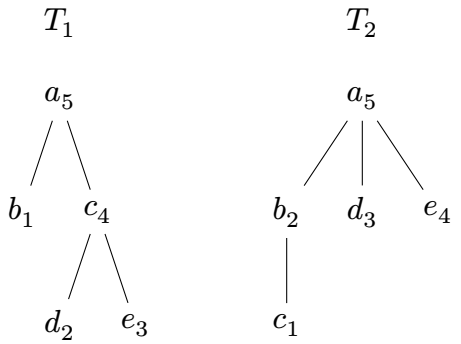
Exercise 1 - *Forest Distance Matrix.* 7 Points

Consider trees T_1 and T_2 in Figure 1a and the tree distance matrix td in Figure 1b. $td[i, j]$ stores the distance between the subtrees rooted in nodes i of T_1 and j in T_2.

a) Compute the left-most leaf descendant arrays l_1, l_2 and the key root arrays kr_1, kr_2 for trees T_1, T_2, respectively.

b) Fill the missing values into the forest distance matrix fd for the key root nodes $i = 4$ and $j = 5$.

fd	1	2	3	4	5
2	1	2	2		
3	2	3	3		
4	2	3	4		



(a) Two example trees T_1 and T_2.

td	1	2	3	4	5
1					
2			0	1	
3	1	2	1	0	4
4			2	2	
5					

(b) Tree distance matrix td for trees T_1 and T_2.

Exercise 2 - *Traversal String Lower Bound.* 7 Points

Compute the traversal string lower bounds between all pairs of trees in the collection $C = \{T_1, T_2, T_3, T_4\}$ in Figure 2.

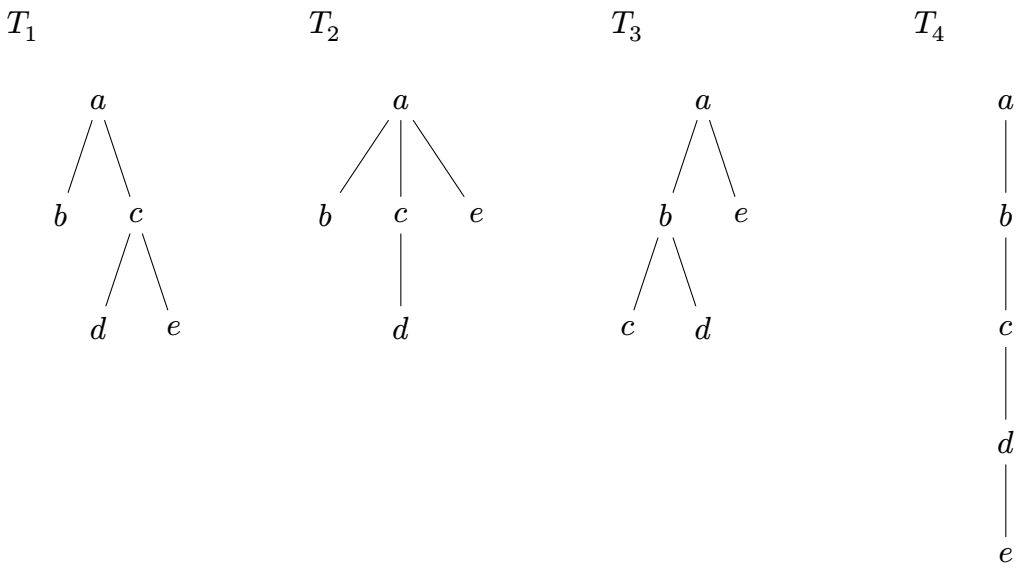


Figure 2: Collection of trees: $C = \{T_1, T_2, T_3, T_4\}$

Exercise 3 — 7 Points

Compute the binary branch distance between trees T_1 and T_2 in Figure 3. Show how the binary branch distance provides a lower bound for the tree edit distance.

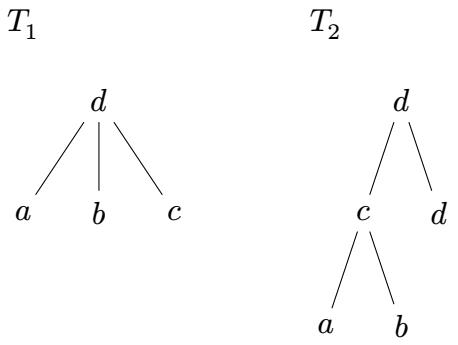


Figure 3: Two example trees T_1 and T_2.

Exercise 4 — 7 Points

Consider the collection $R = \{s_1, s_2, s_3, s_4\}$ of sets in Figure 4.

$$\begin{aligned} s_1 &= \{A, B, C, D, F\} \\ s_2 &= \{A, C, D, F, G\} \\ s_3 &= \{A, C, G, H\} \\ s_4 &= \{D, G, H\} \end{aligned}$$

Figure 4: Set collection $R = \{s_1, s_2, s_3, s_4\}$.

Construct an inverted list index for the partition signatures using Hamming threshold $h = 2$ and an appropriate partitioning. Provide the set of candidates C and the join result M.

Exercise 5 — 7 Points

Consider the following definition of SuperTED extending TED by an **additional condition**.

A SuperTED mapping M between trees T_1 and T_2 is a set of node pairs that satisfies the following conditions:

1. $(a, b) \in M \Rightarrow a \in N(T_1), b \in N(T_2)$
2. for any two pairs (a, b) and (x, y) of M:
   a) $a = x \Leftrightarrow b = y$ (one-to-one condition)
   b) a is to the left of x [1] $\Leftrightarrow$ b is to the left of y (order condition)
   c) a is an ancestor of $x \Leftrightarrow b$ is an ancestor of y (ancestor condition)
   d) If a is in a k-neighborhood of x, then b is in a k-neighborhood of y. A k-neighborhood of a node are all nodes reachable in at most k steps, i.e., by traversing at most k undirected edges in the tree. **(additional condition)**

The SuperTED of two trees is the cost of a minimal cost SuperTED mapping.

Prove either that

- SuperTED is an upper bound for TED,
- TED is an upper bound for SuperTED, or
- TED and SuperTED are incomparable.

[1] i.e., a precedes x in both preorder and postorder