Approximate Cartesian Tree Matching

Bastien Auvray

April 1st, 2025 - NormaSTIC day - Caen, France

Outline

About us

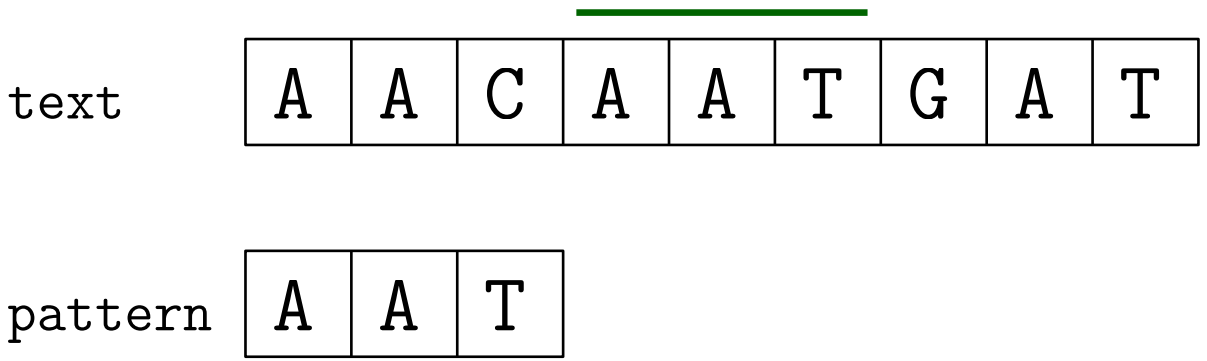
- Supervisors:
  Thierry Lecroq (LITIS), Julien David (GREYC) and Richard Groult (LITIS)
- Text **Algo**rithms + **Comb**inatorics
- Follow-up to a 2023 NormaSTIC internship

Pattern matching

Definition

In general terms, the pattern matching problem consists in finding one or all occurences of a pattern in a text.

To our disposal: 60+ years worth of research.

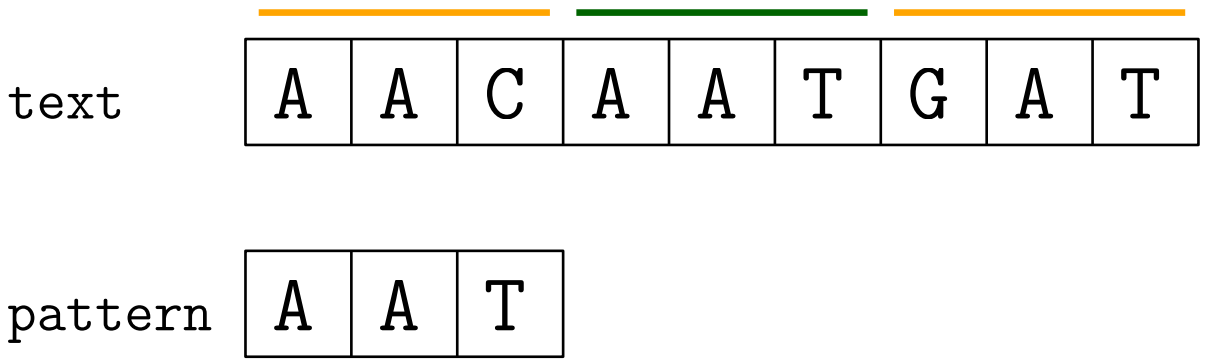


Approximate pattern matching

Idea

We want to allow for differences between the pattern and the text in the approximate pattern matching problem.

To our disposal: 50+ years of research.



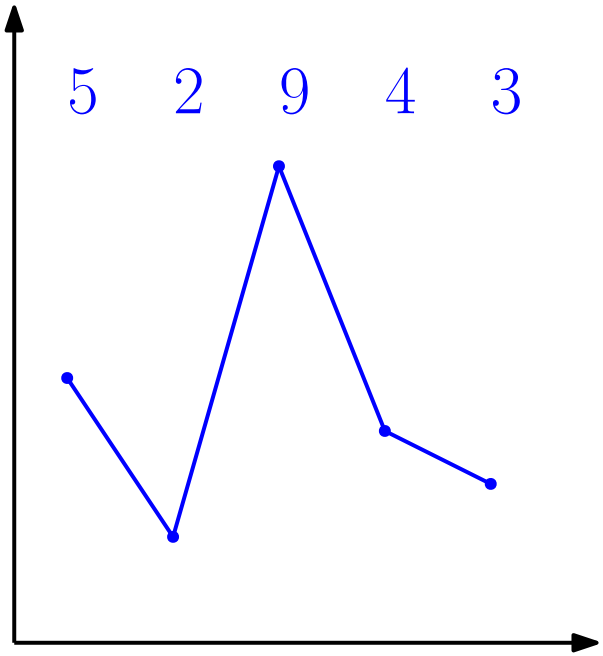
Time series

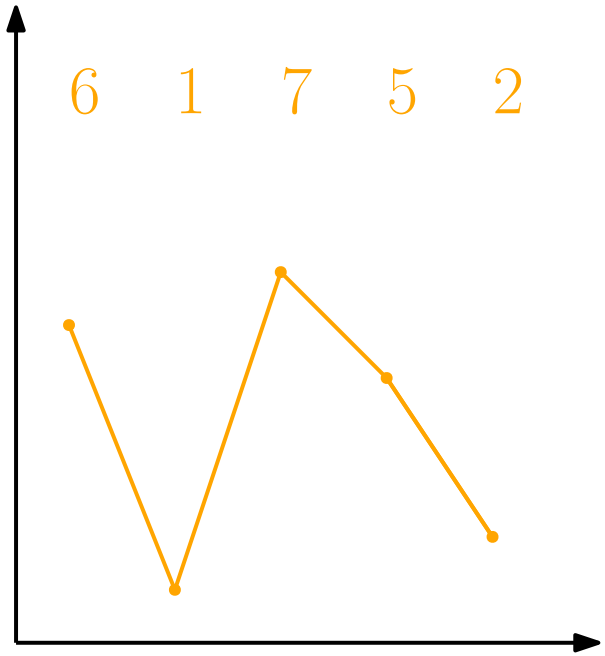
Roughly put

Time series represent "values" over time. They can be found in:

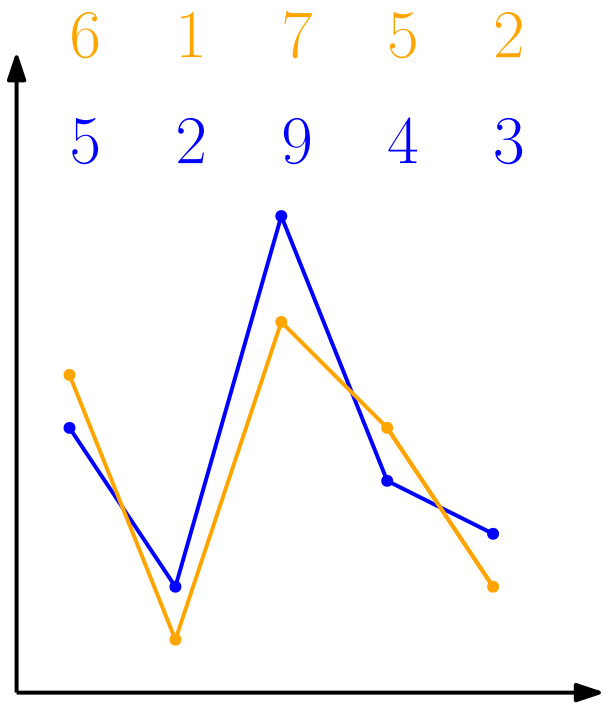
- Stock market prices
- Musicology
- Bioinformatics (Gene Sample Time data)
- And so on...

Time series

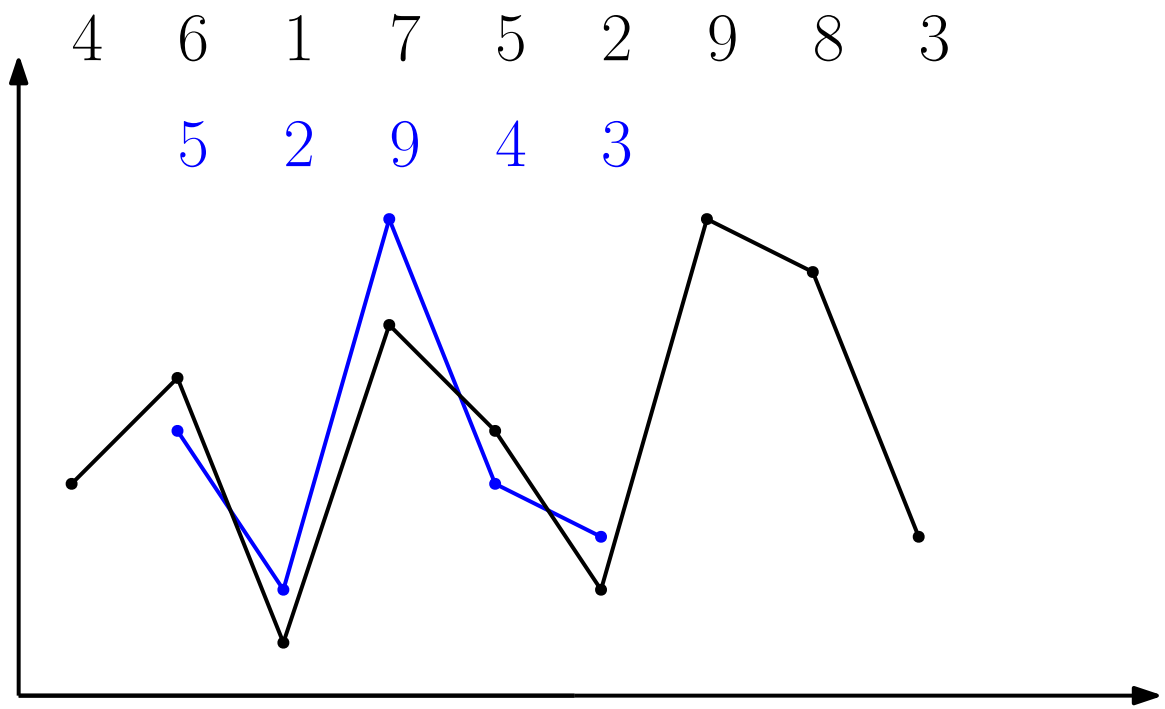




Time series



Time series



Cartesian tree

Cartesian tree [Vuillemin, 1980]

A sequence x of length m can be associated to its Cartesian tree $C(x)$ according to the following rules:

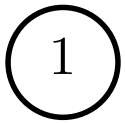
- if x is empty, then $C(x)$ is the empty tree;
- if $x[1 \dots m]$ is not empty and $x[i]$ is the smallest value of x, $C(x)$ is the Cartesian tree with:
  - the root is at position i,
  - $C(x[1 \dots i-1])$ is the left subtree,
  - $C(x[i+1 \dots m])$ is the right subtree.

Cartesian tree

x	4	5	6	2	1	7	8	3	9

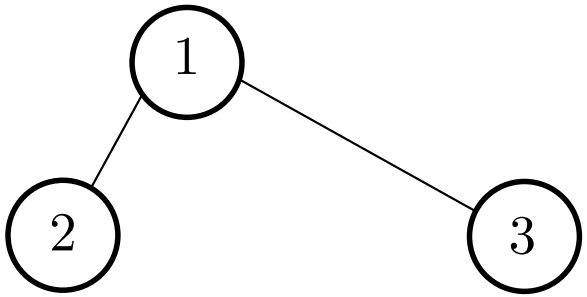
Cartesian tree

x	4	5	6	2	1	7	8	3	9



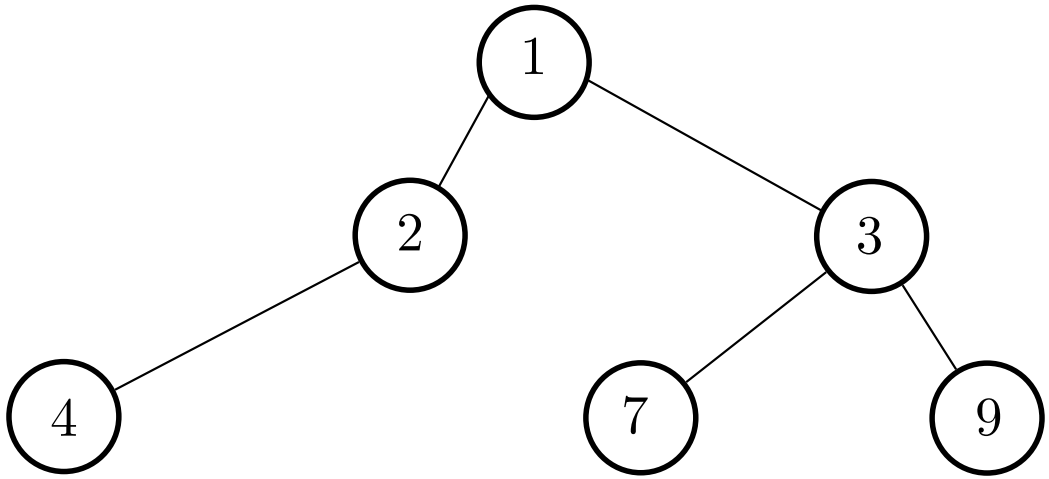
Cartesian tree

x	4	5	6	2	1	7	8	3	9



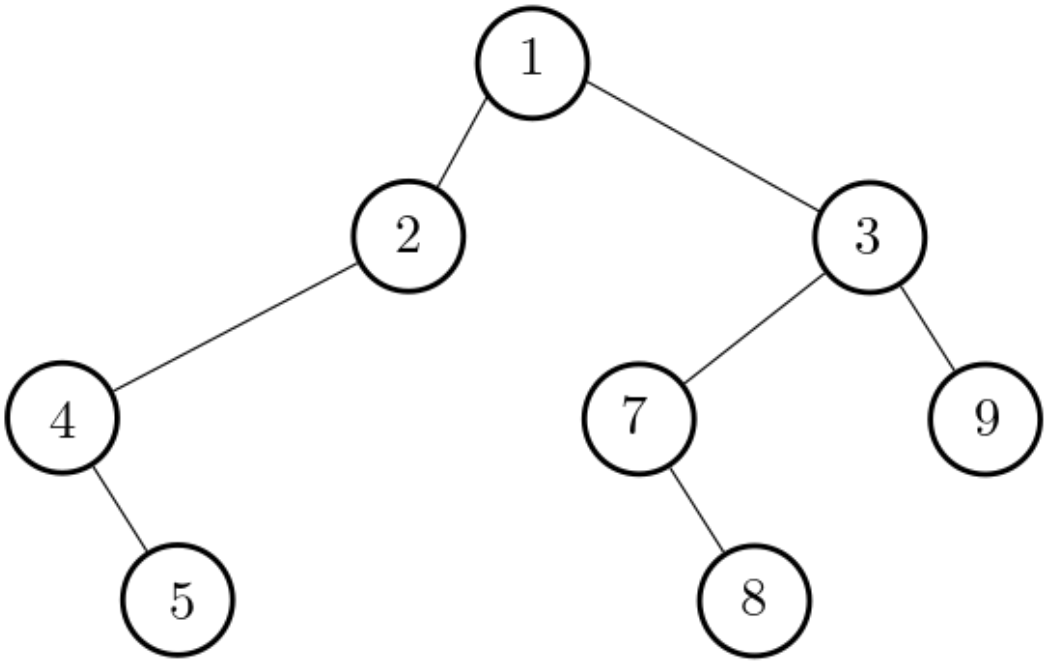
Cartesian tree

x	4	5	6	2	1	7	8	3	9



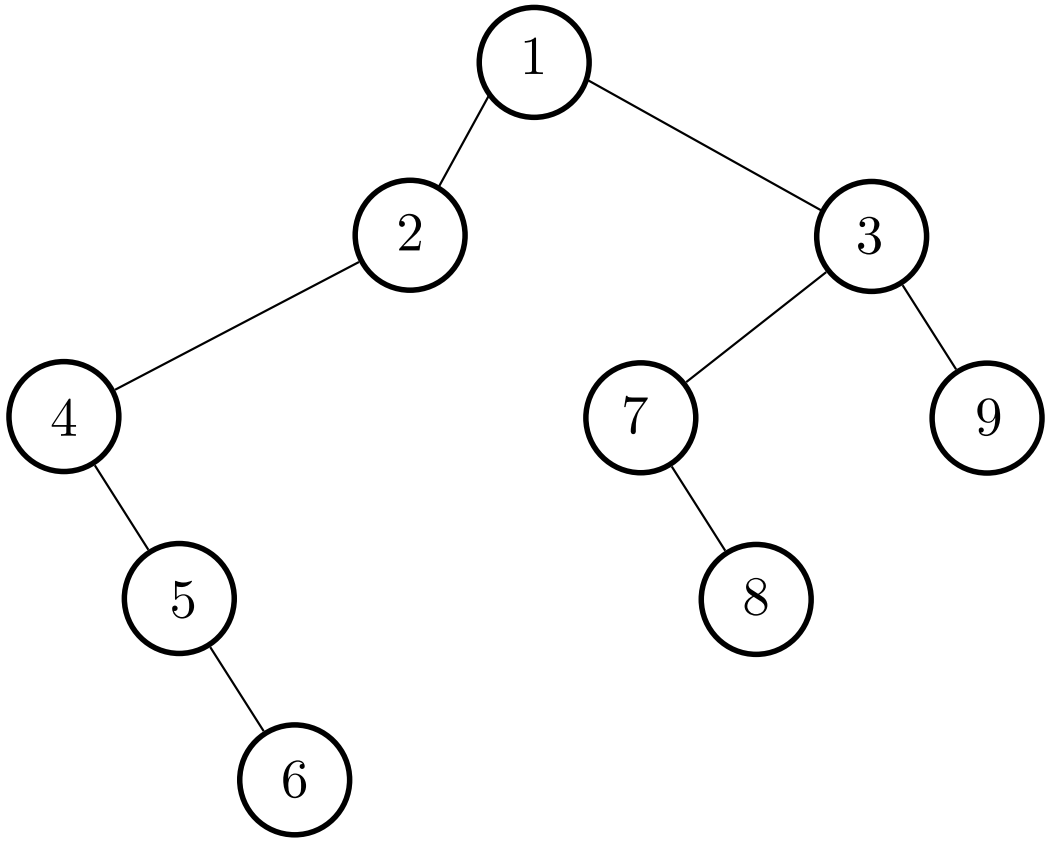
Cartesian tree

x	4	5	6	2	1	7	8	3	9



Cartesian tree

x	4	5	6	2	1	7	8	3	9

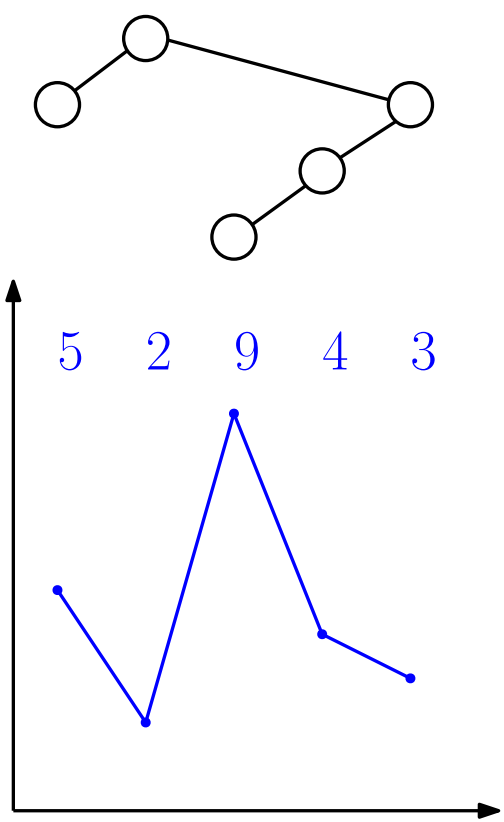


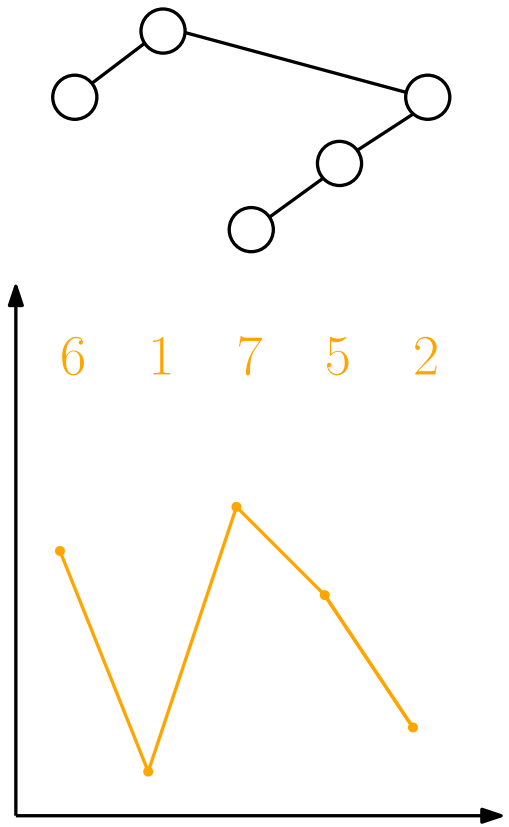
Cartesian tree matching

Similarity

Two sequences x and y are similar if they share the same Cartesian tree.

Cartesian tree matching





Cartesian tree matching

Cartesian tree matching [Park, Amir, Landau and Park, 2019]

The Cartesian tree matching (CTM) problem is the following: Given a pattern p and a text t, find every factor f of t such that f shares the same Cartesian tree as p.

First solutions for CTM

Linear time solutions for CTM

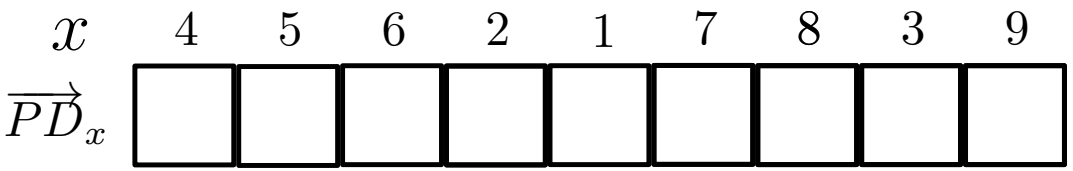
Park *et al.* adapted the KMP and Aho-Corasick to achieve single pattern matching and multiple pattern matching in linear time and space.

Parent-distance representation

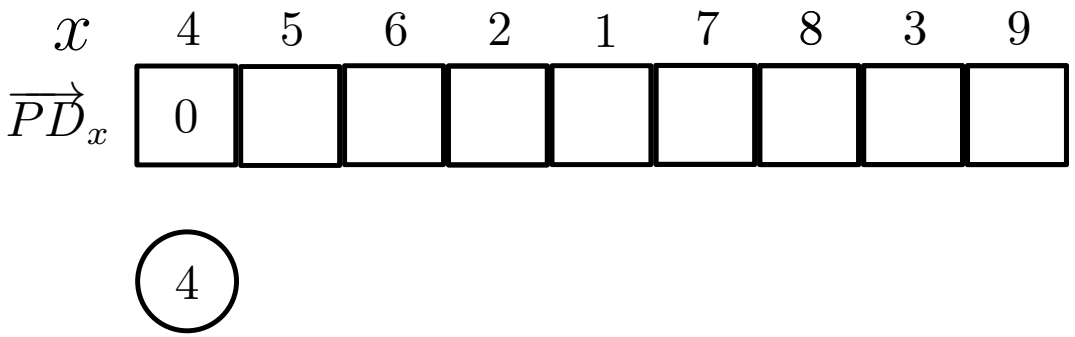
Parent-distance [PALP19]

Given a sequence $x[1 \dots m]$, the parent-distance representation of x is an integer sequence $\overrightarrow{PD}_x[1 \dots m]$, where $\overrightarrow{PD}_x[i]$ is the distance between $x[i]$ and its parent in the Cartesian tree of $x[1 \dots i]$ (if it exists).

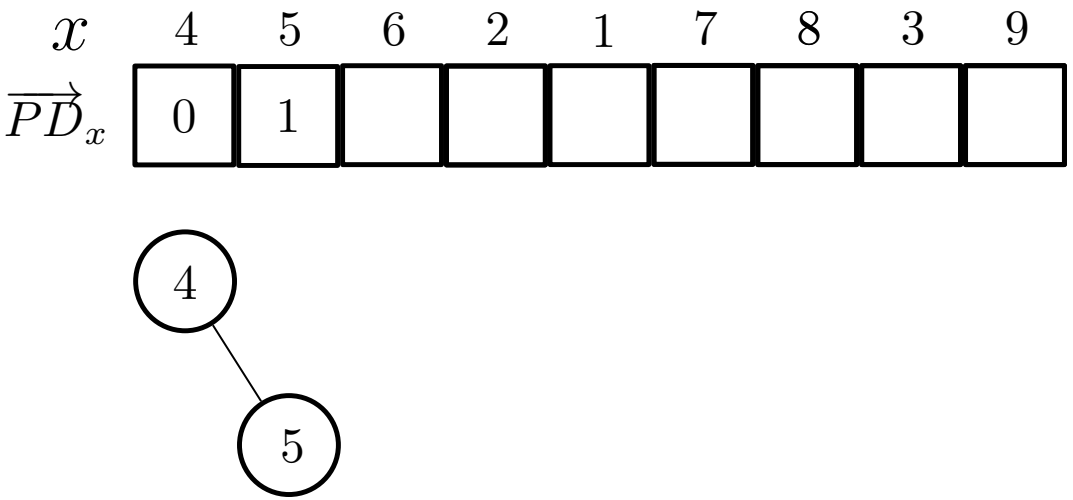
Parent-distance representation



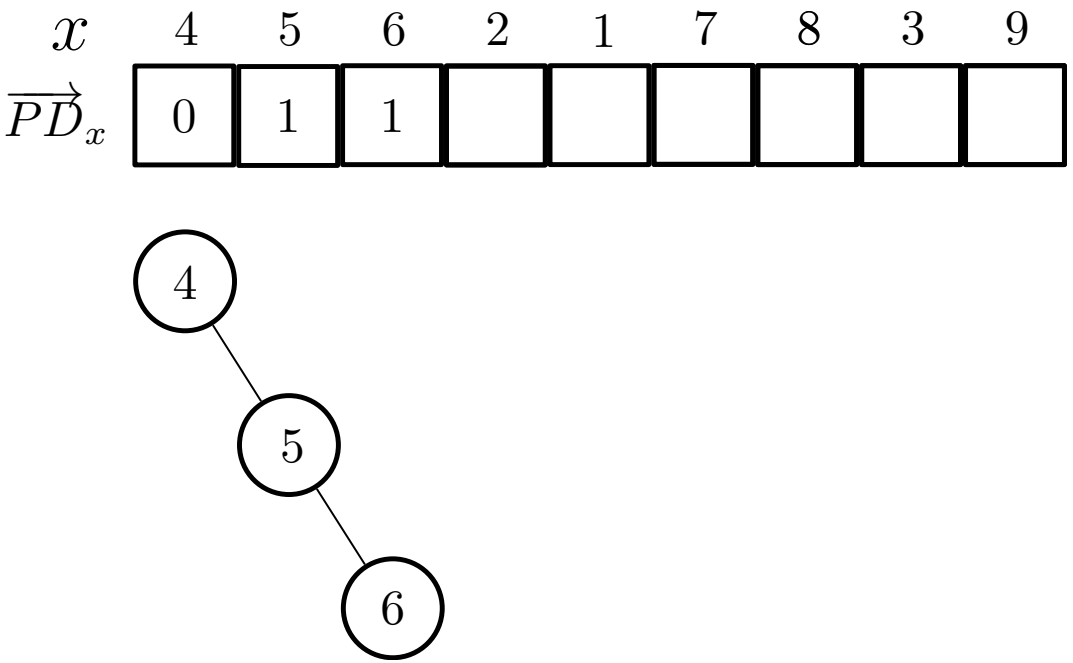
Parent-distance representation



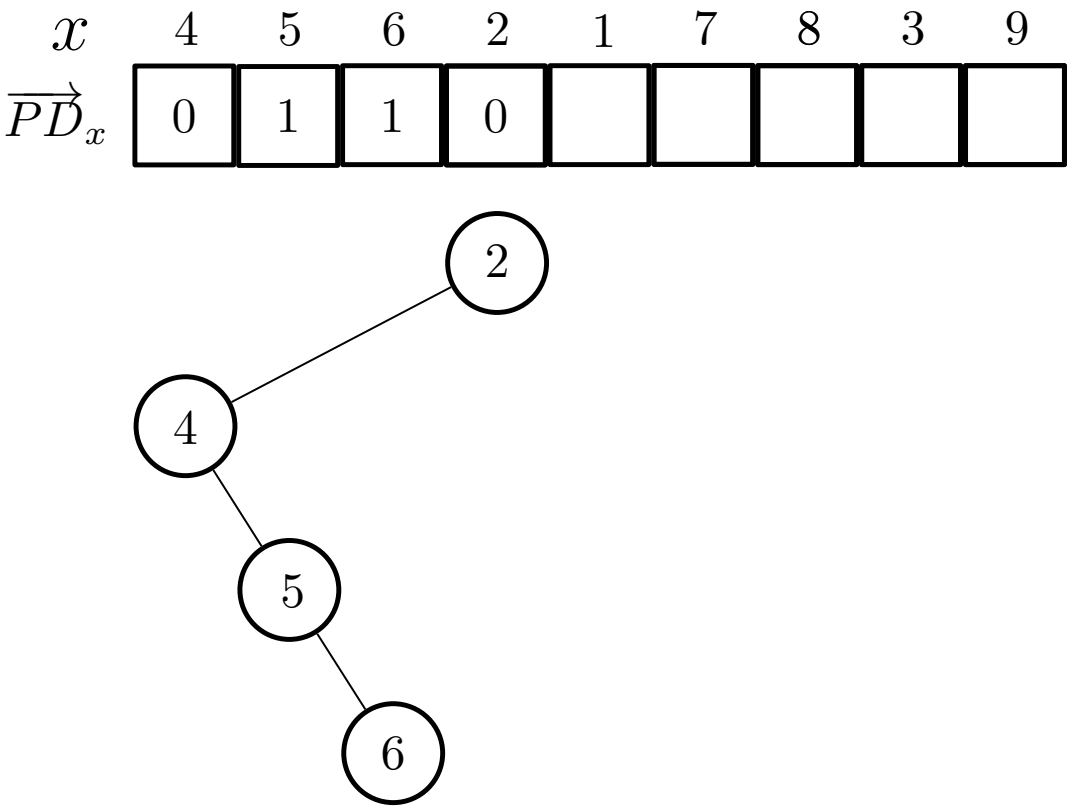
Parent-distance representation



Parent-distance representation

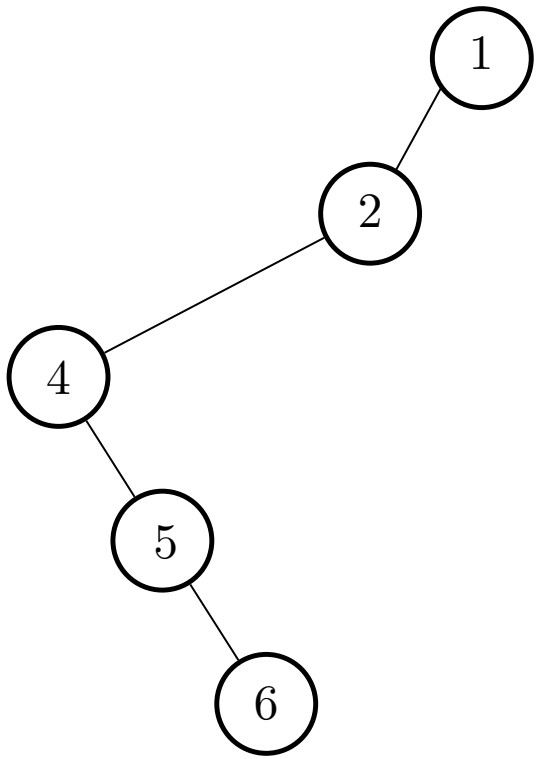


Parent-distance representation



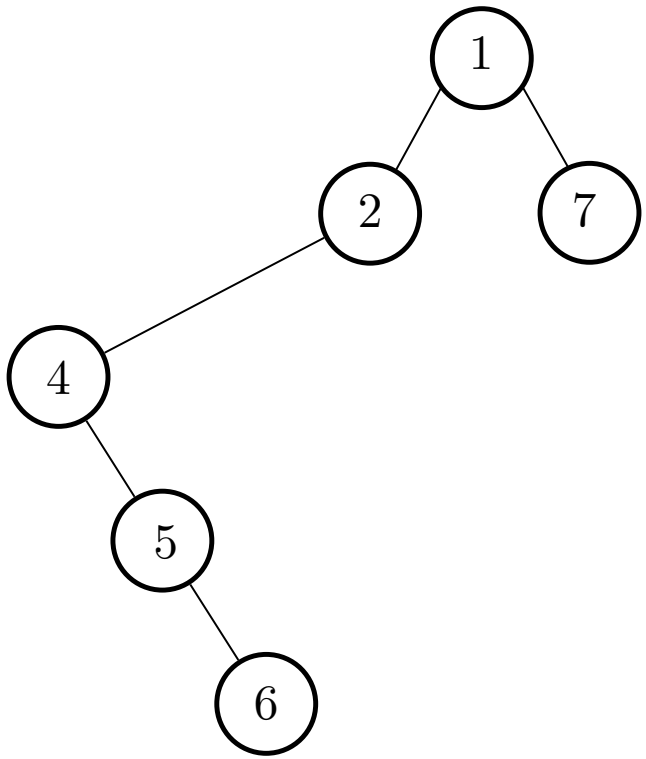
Parent-distance representation

x	4	5	6	2	1	7	8	3	9
$\overrightarrow{PD}_x$	0	1	1	0	0				



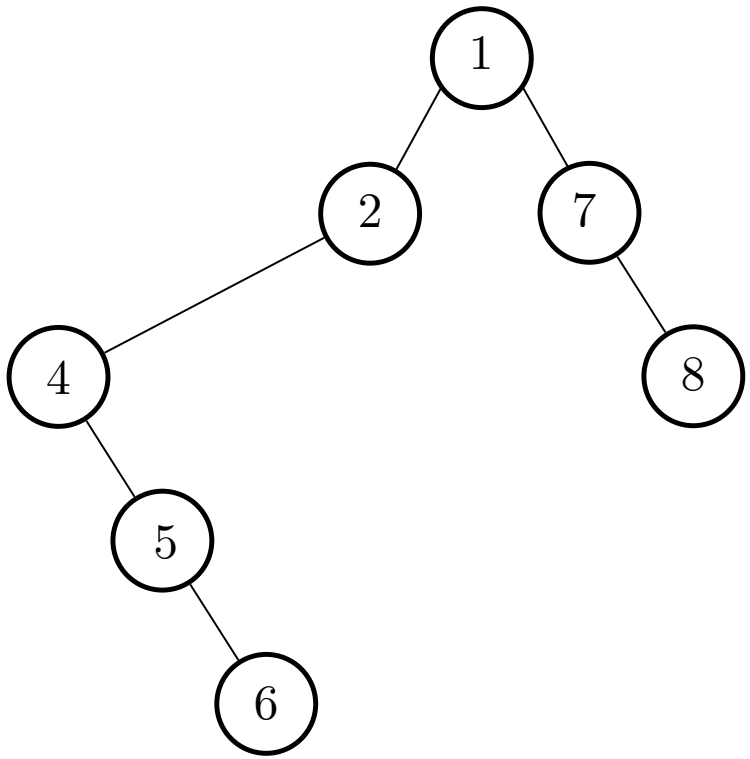
Parent-distance representation

x	4	5	6	2	1	7	8	3	9
$\overrightarrow{PD}_x$	0	1	1	0	0	1			



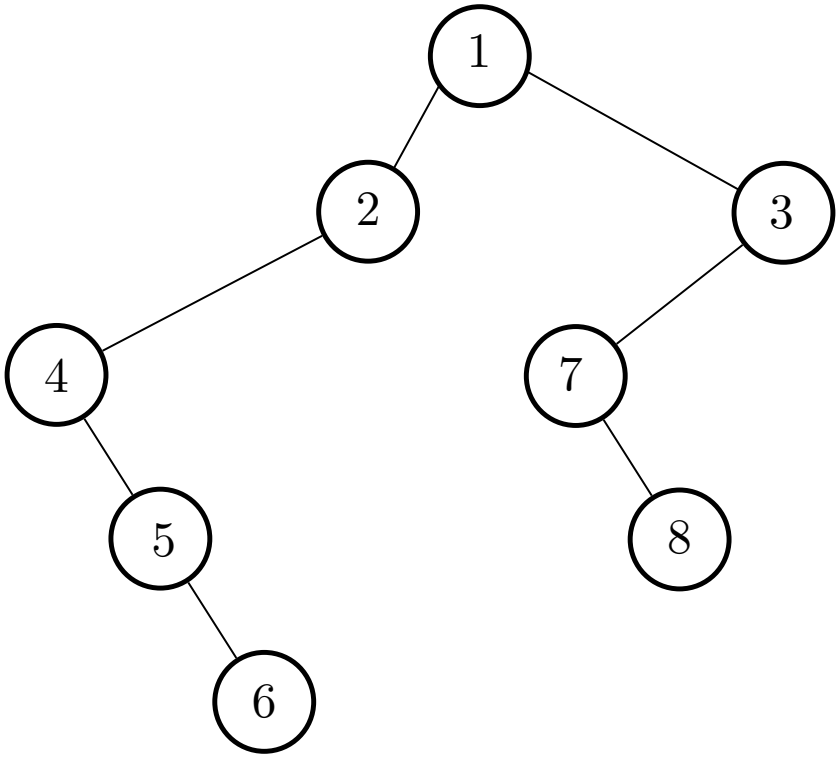
Parent-distance representation

x	4	5	6	2	1	7	8	3	9
$\overrightarrow{PD}_x$	0	1	1	0	0	1	1		



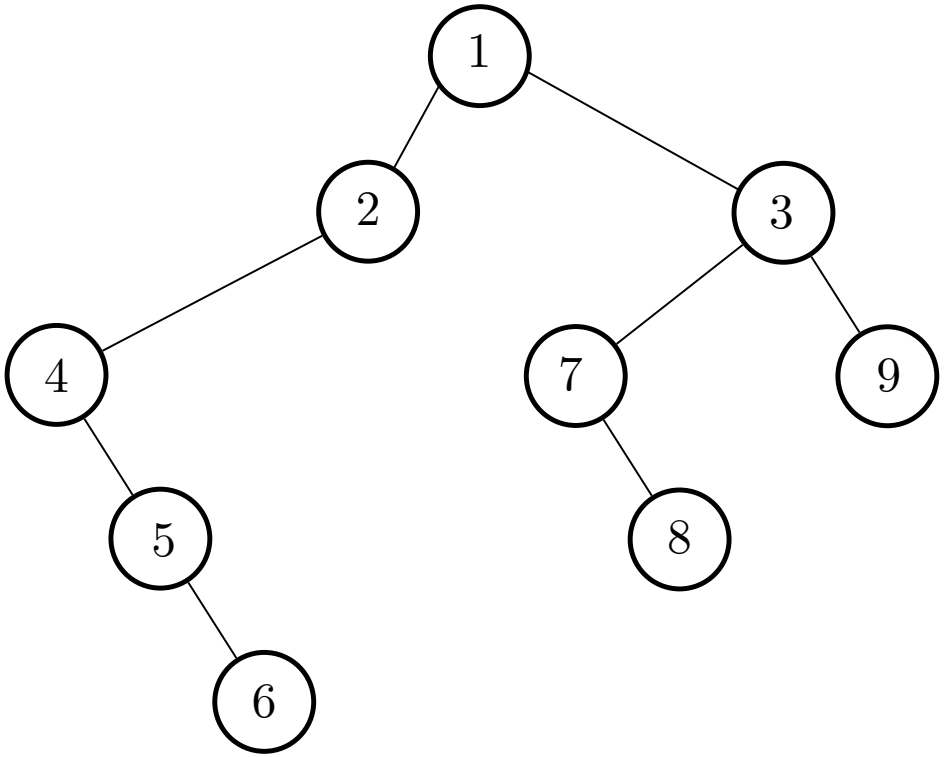
Parent-distance representation

x	4	5	6	2	1	7	8	3	9
$\overrightarrow{PD}_x$	0	1	1	0	0	1	1	3	



Parent-distance representation

x	4	5	6	2	1	7	8	3	9
$\overrightarrow{PD}_x$	0	1	1	0	0	1	1	3	1

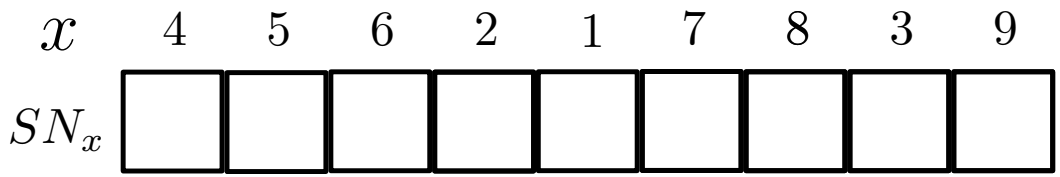


Skipped-number representation

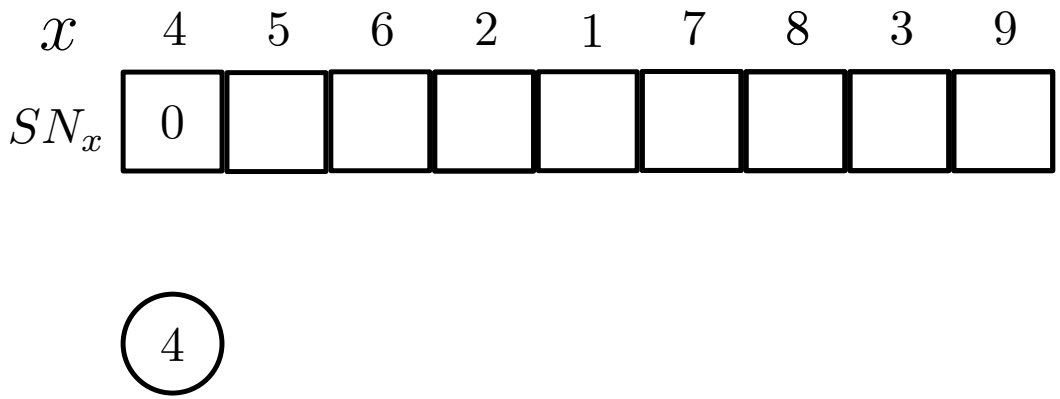
Skipped-number representation

Given a sequence $x[1 \dots m]$, the Skipped-number representation of x is an integer sequence $SN_x[1 \dots m]$, where $SN_x[i]$ is the number of nodes "eaten by i" on the right path.

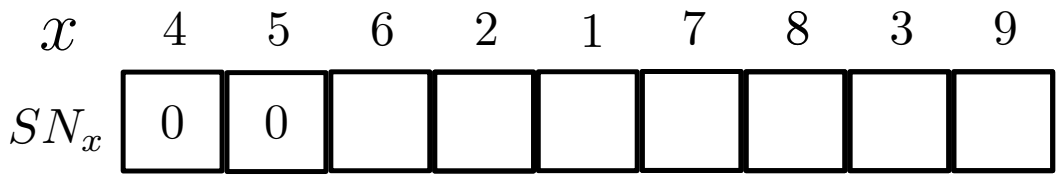
Skipped-number representation

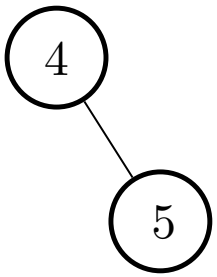


Skipped-number representation

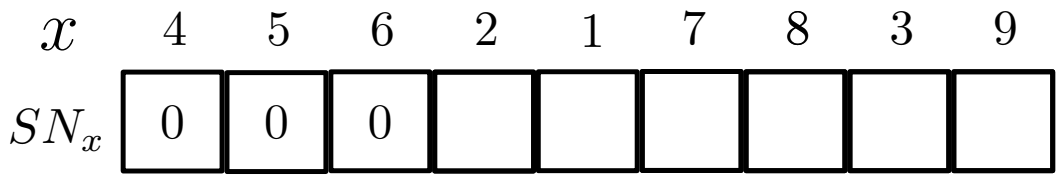


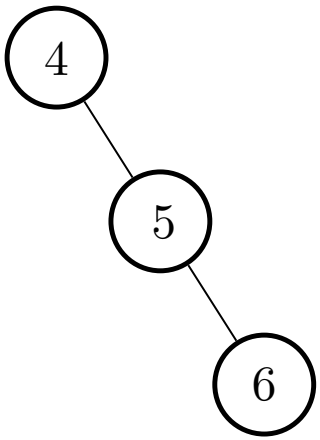
Skipped-number representation





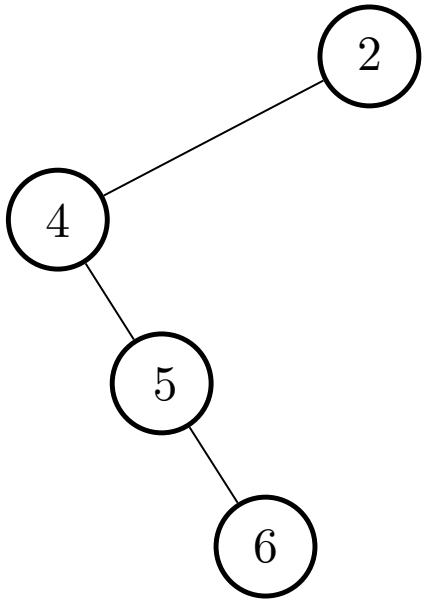
Skipped-number representation





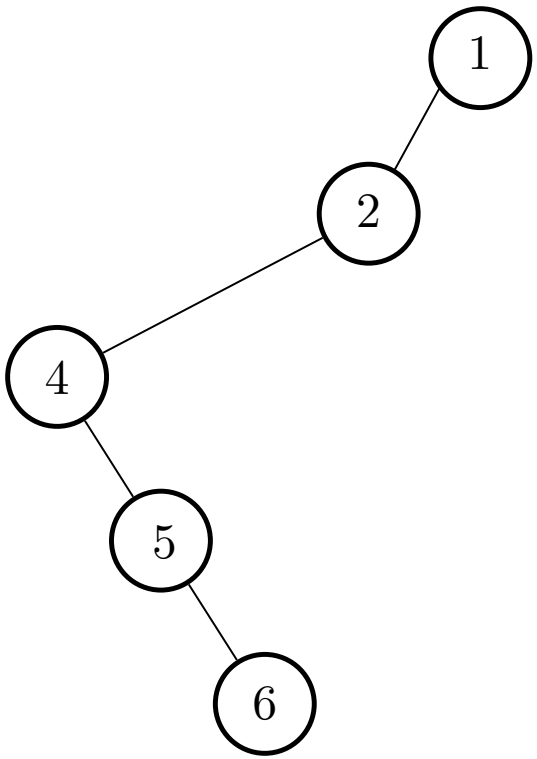
Skipped-number representation

x	4	5	6	2	1	7	8	3	9
SN_x	0	0	0	3					



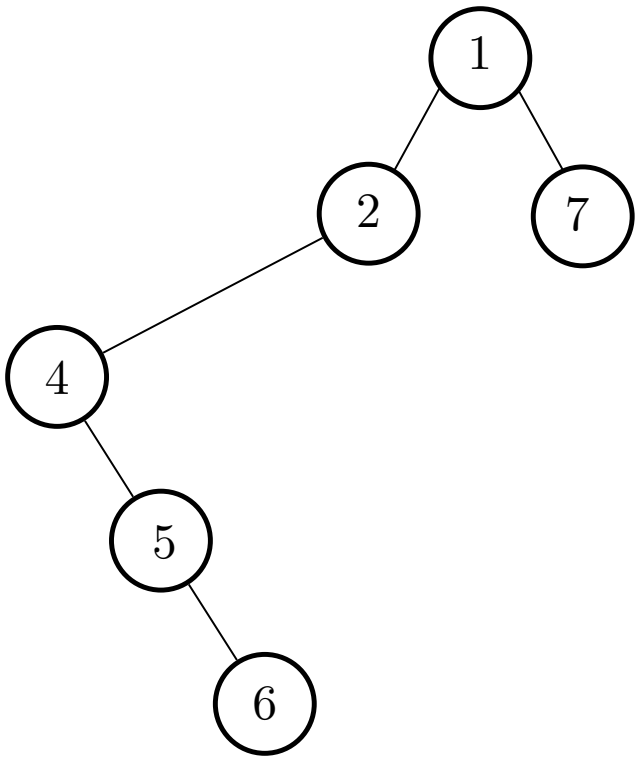
Skipped-number representation

x	4	5	6	2	1	7	8	3	9
SN_x	0	0	0	3	1				



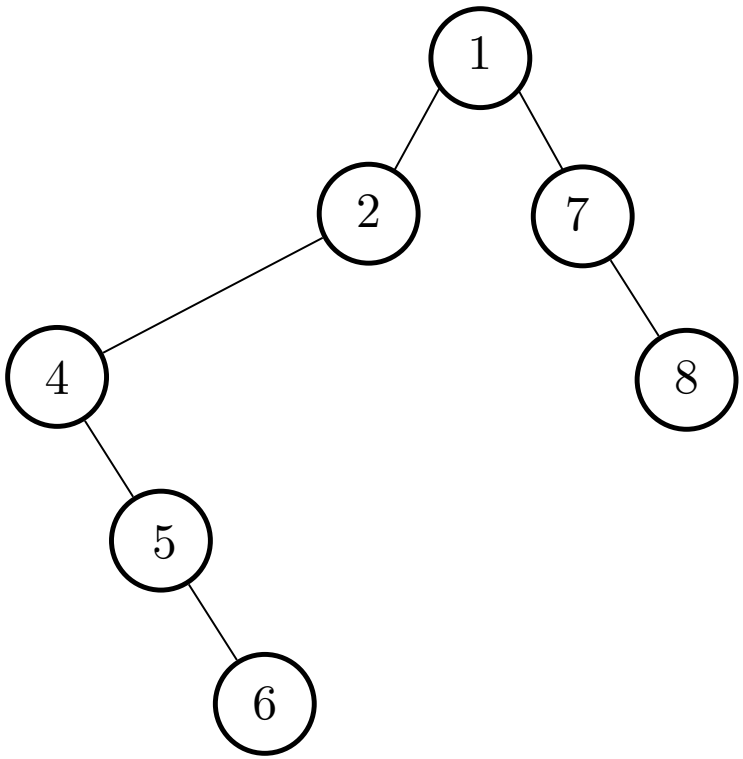
Skipped-number representation

x	4	5	6	2	1	7	8	3	9
SN_x	0	0	0	3	1	0			



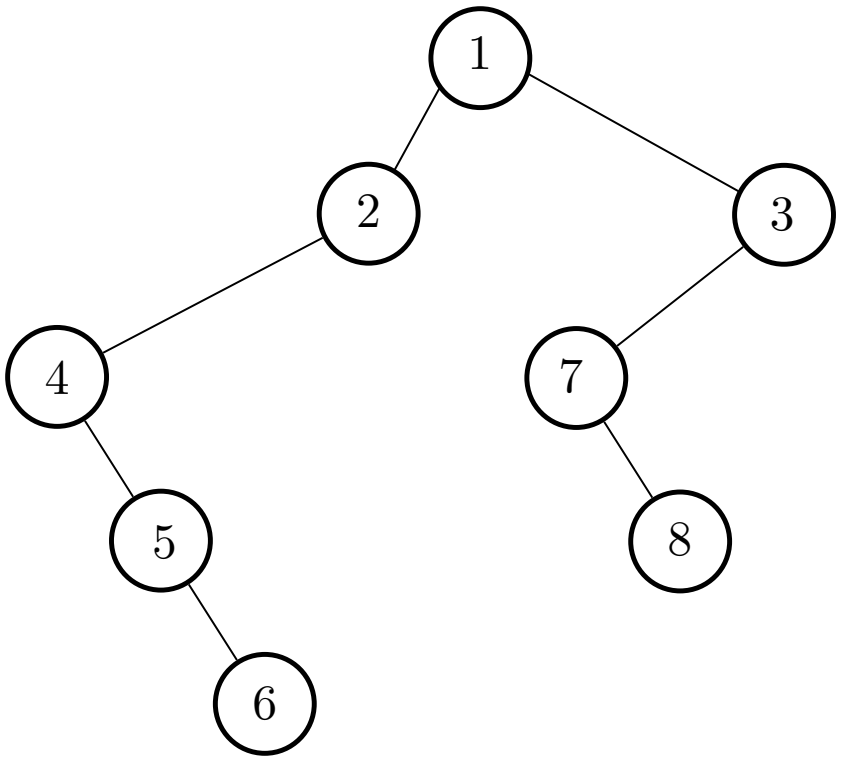
Skipped-number representation

x	4	5	6	2	1	7	8	3	9
SN_x	0	0	0	3	1	0	0		



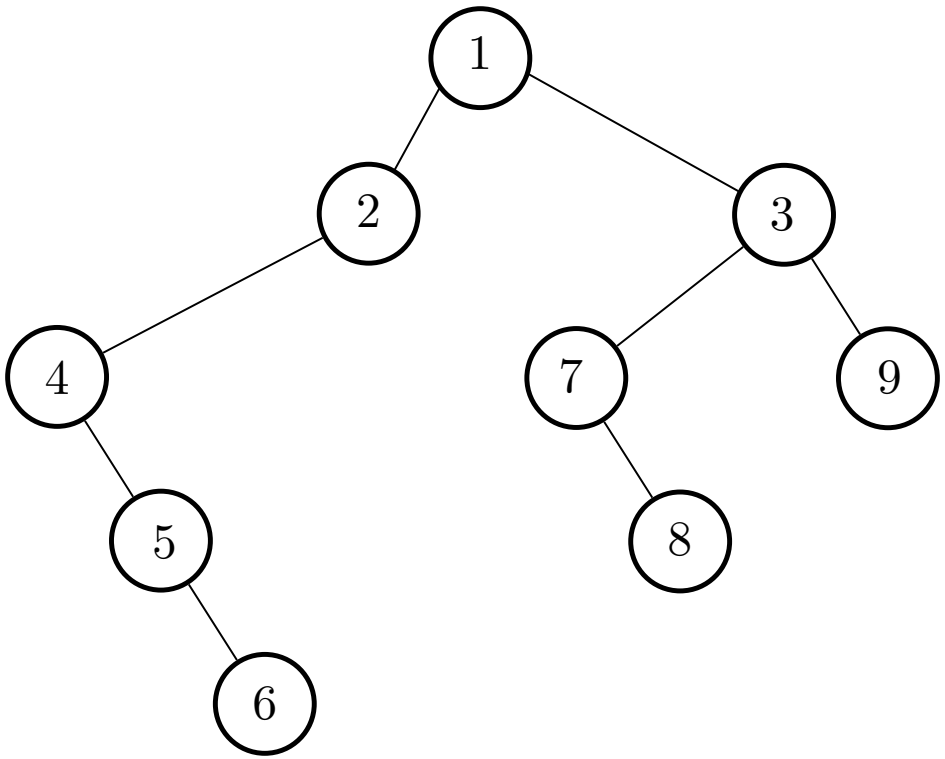
Skipped-number representation

x	4	5	6	2	1	7	8	3	9
SN_x	0	0	0	3	1	0	0	2	



Skipped-number representation

x	4	5	6	2	1	7	8	3	9
SN_x	0	0	0	3	1	0	0	2	0



Approximate Cartesian tree matching

Idea

We now want to allow for differences between the Cartesian trees. There was no approximate version of the CTM problem until recently.

Differences

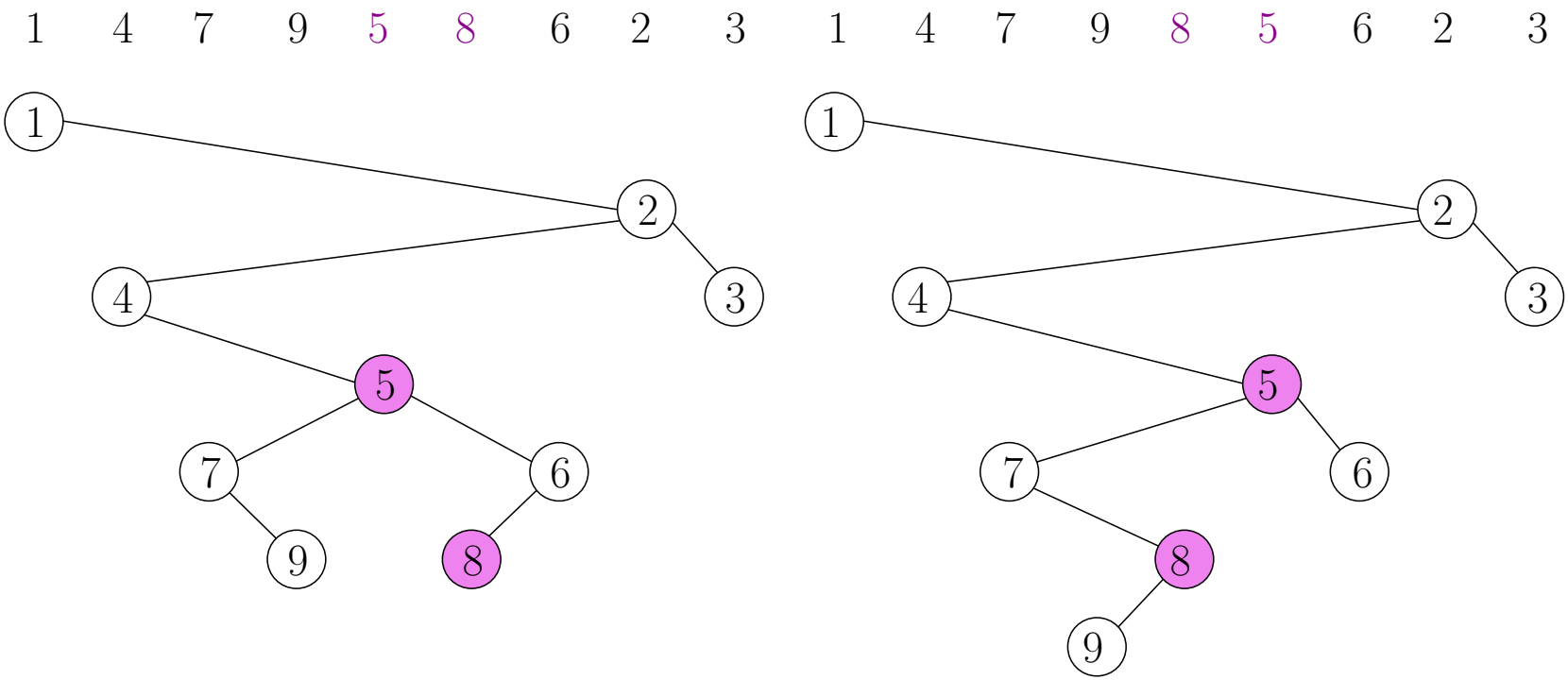
Covered ground

We will consider up to one difference for the following approximations:

- Swap
- Mismatch
- Insertion
- Deletion

Differences

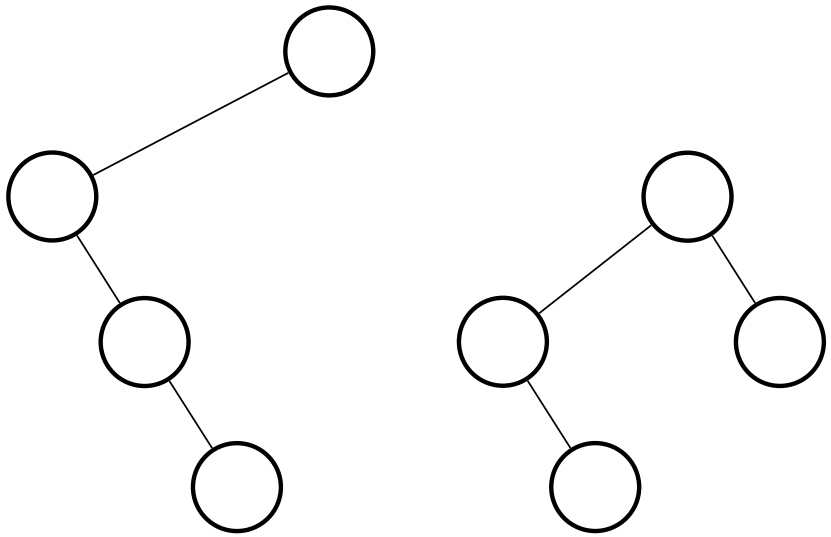
Swap



Differences

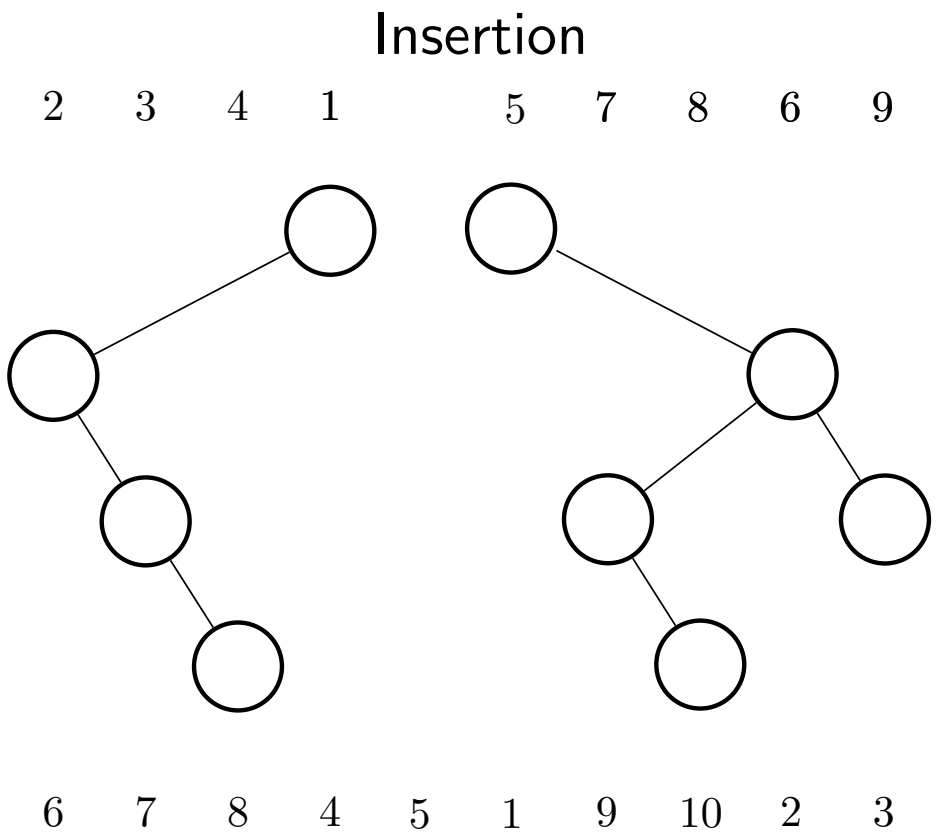
Mismatch

2 3 4 1 5 7 8 6 9



3 4 9 2 5 6 8 1 7

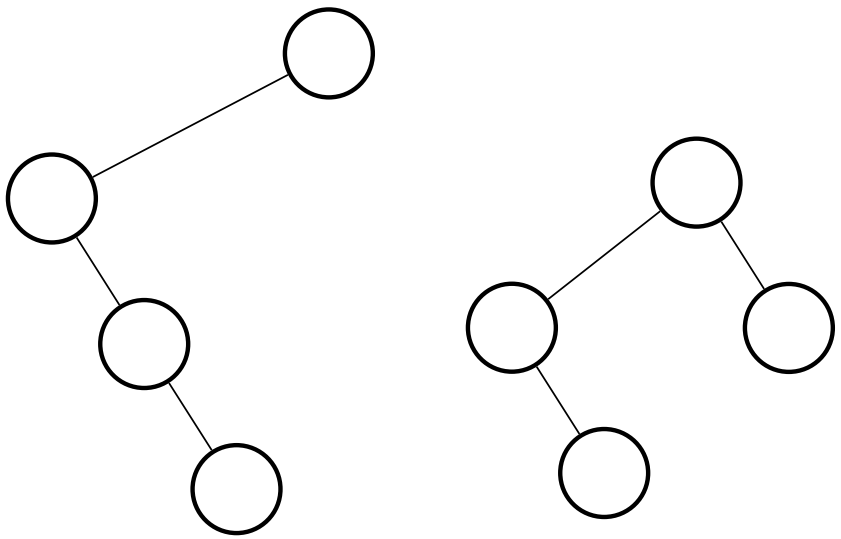
Differences



Differences

Deletion

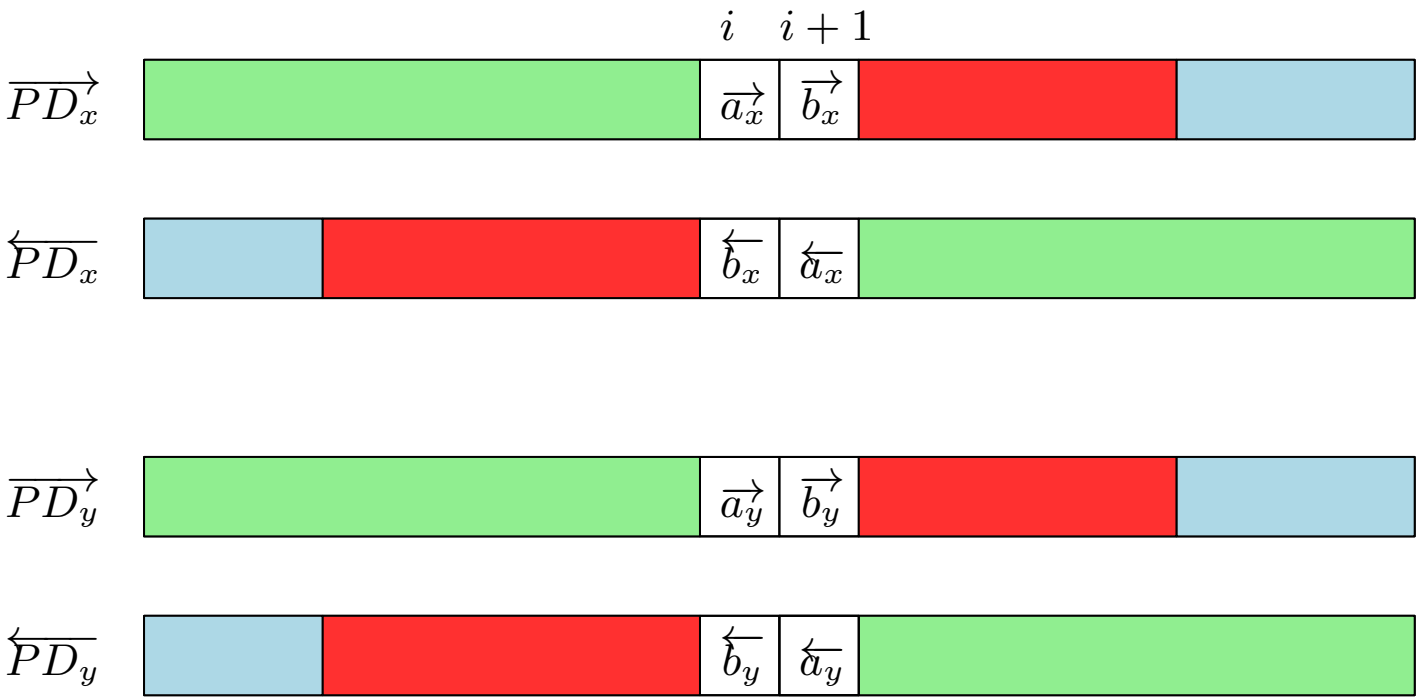
2 3 4 1 5 7 8 6 9



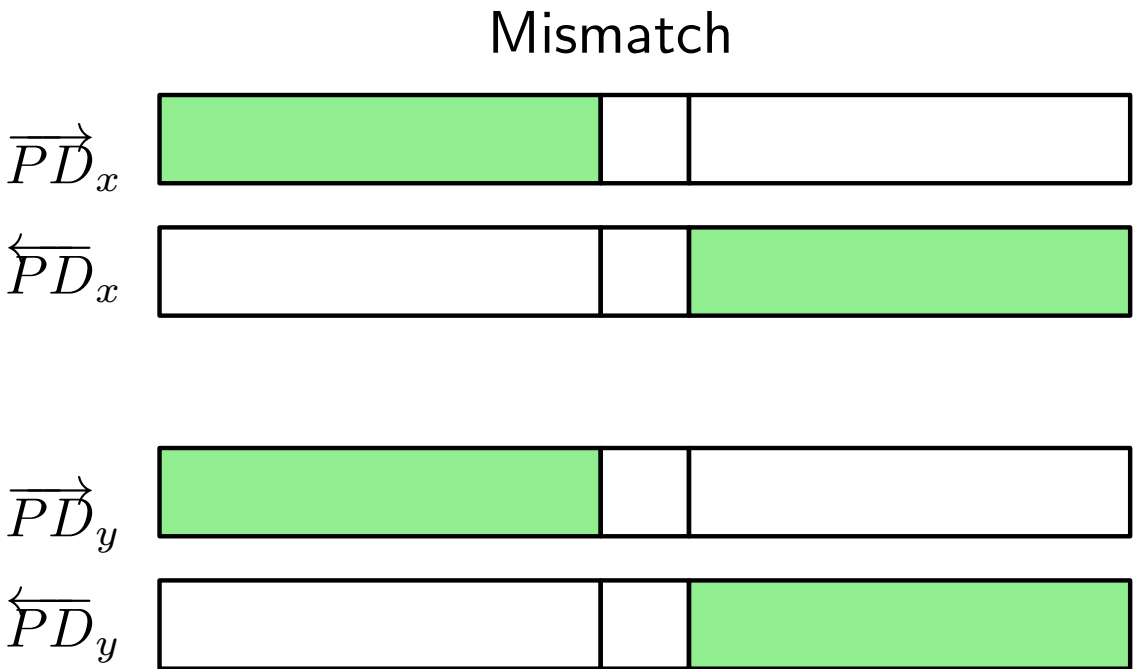
6 7 8 4 2 3 1 5

Ideas of the solutions

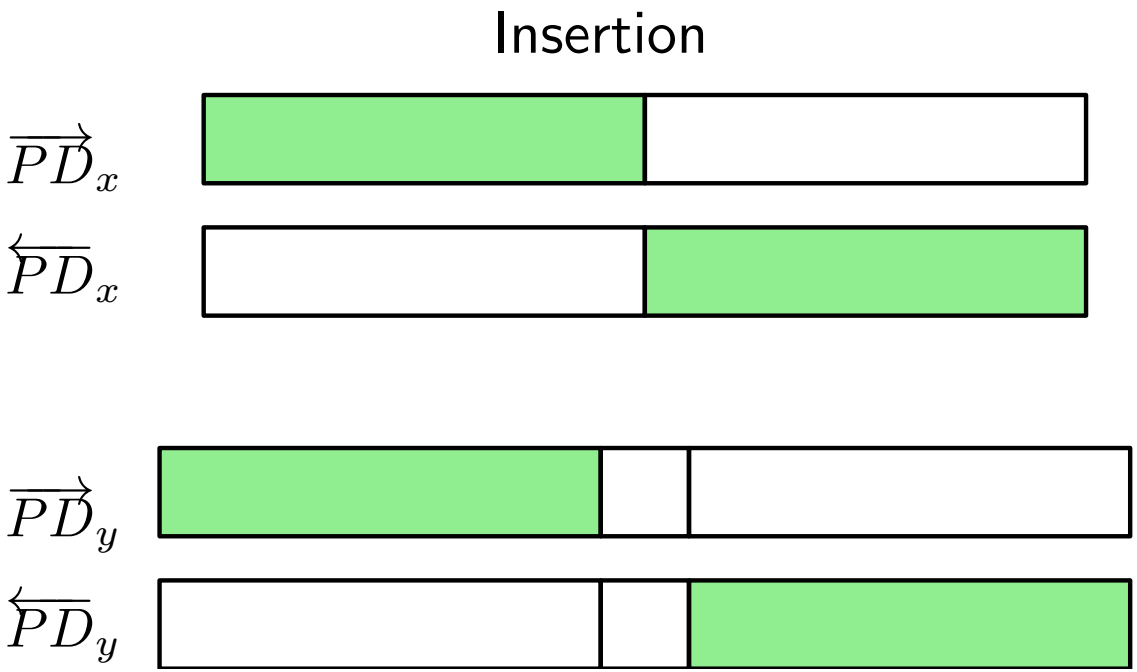
Swap



Ideas of the solutions



Ideas of the solutions



Ideas of the solutions

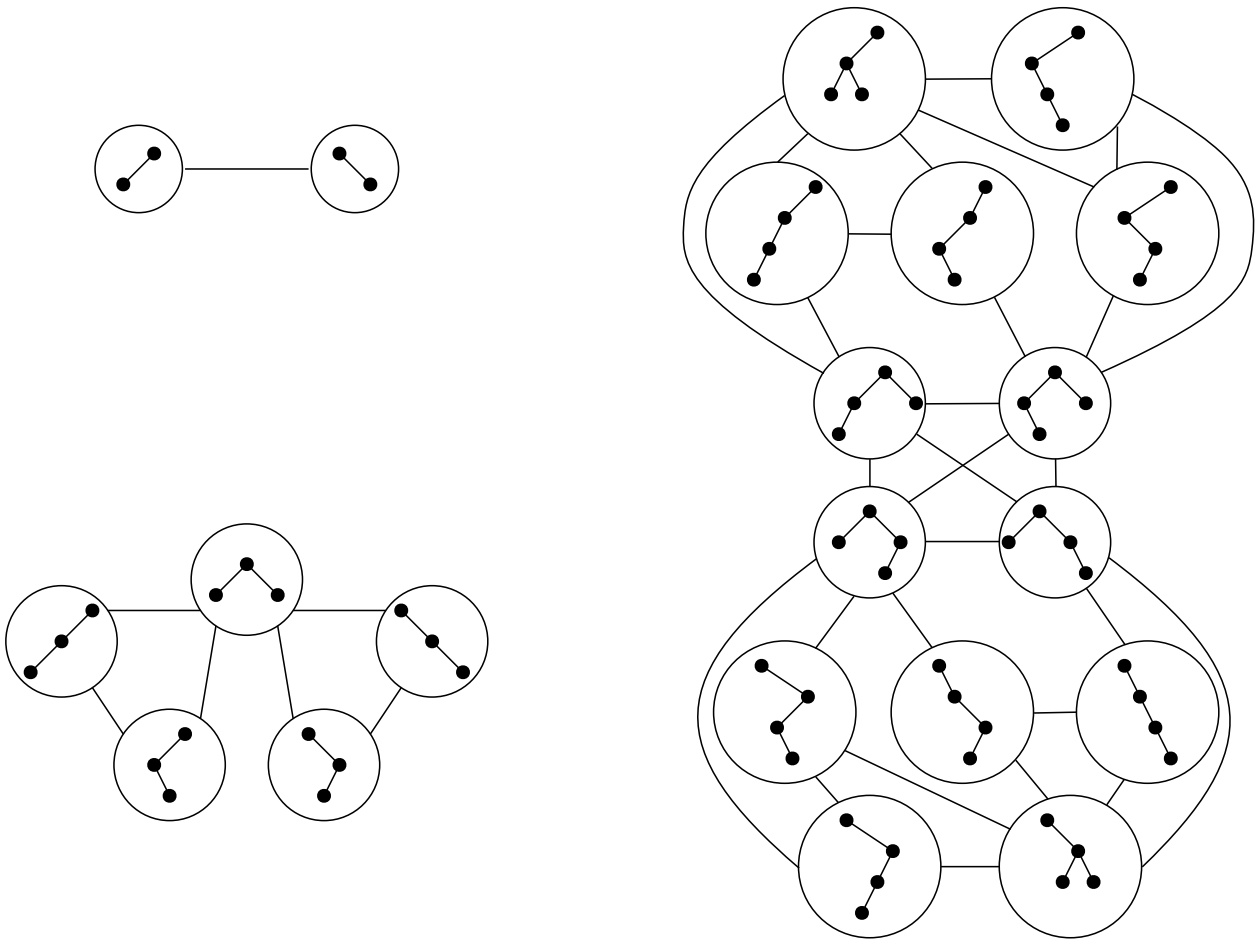
Ideas of the solutions

Skipped-number approach (swap)

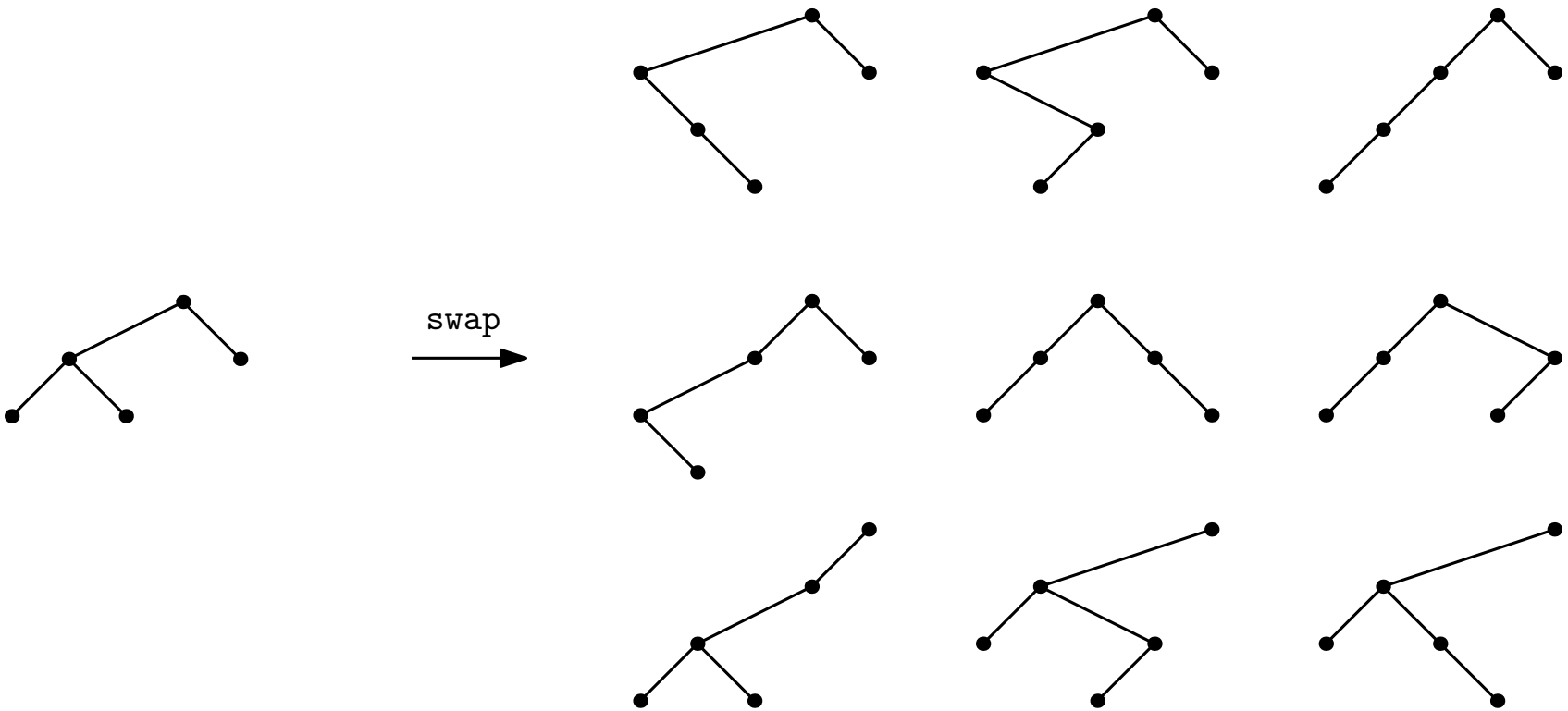
We showed that there exists at most 3 differences between the Skipped-number representations of two sequences x and y when there is a swap between the sequences.

Ideas of the solutions

Swap graph



Ideas of the solutions



Ideas of the solutions

Neighbourhood

The number of neighbours $ng(T)$ a given Cartesian tree T of size m may have in the swap graph is:

$$m - 1 \leq ng(T) \leq 3(m-2) + 1$$

Ideas of the solutions

Complexities

- Linear representations: $\Theta(n)$ time on average ($\Theta(mn)$ time in the worst case) and $\Theta(m)$ space.
- Neighbourhood: $\mathcal{O}((m^2 + n)\log(m))$ time and $\mathcal{O}(m^2)$ space.

Closing words

Perspectives

- Generalizing our results to any number of differences?
- Adapting Skipped-number and Aho-Corasick approaches?
- Searching for regularities?
- Indexing?
- Applications

- Bastien Auvray, Julien David, Richard Groult, and Thierry Lecroq. Approximate cartesian tree matching: An approach using swaps. In Proc. SPIRE, volume 14240 of LNCS, pages 49–61, 2023.
- Incoming journal version (joint work with Gad M. Landau and Samah Ghazawi from Haifa, Israel)

Thank you for your attention!