The Search for the Tree of Life: A Location Problem in the Phylogenetic Tree Space

Marco Botte and Anita Schöbel

Institut für Numerische und Angewandte Mathematik
Georg-August Universität Göttingen

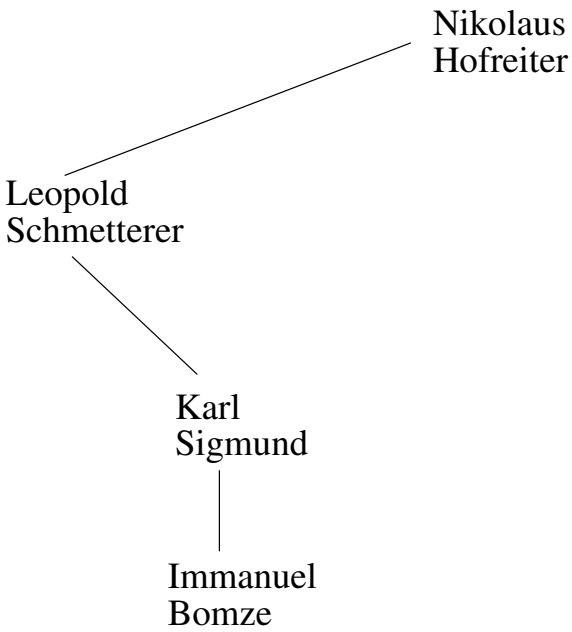
20. December 2018

The trees of this talk: phylogenetic trees

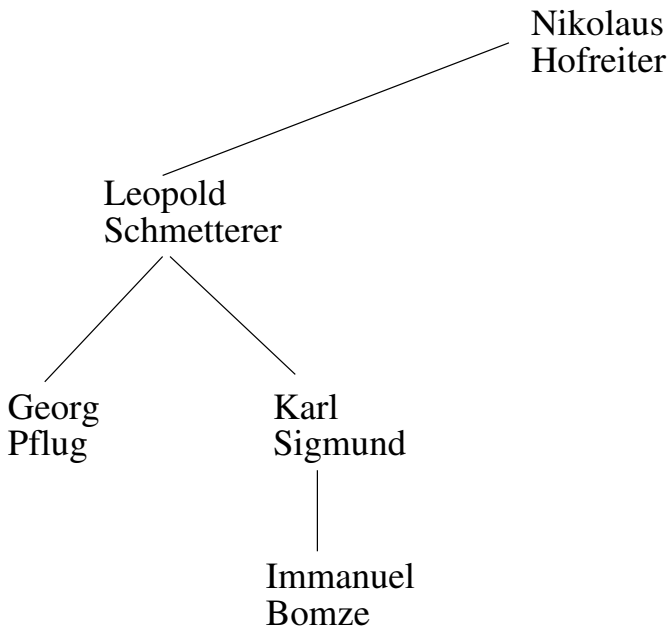
More interesting: Math genealogy

Immanuel
Bomze

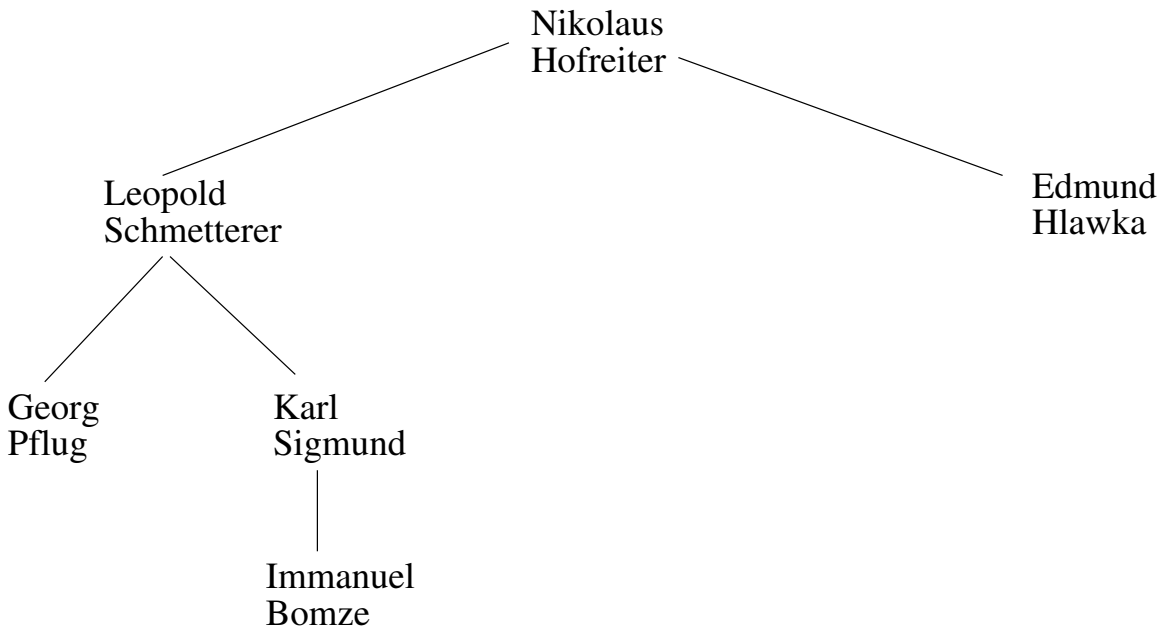
More interesting: Math genealogy



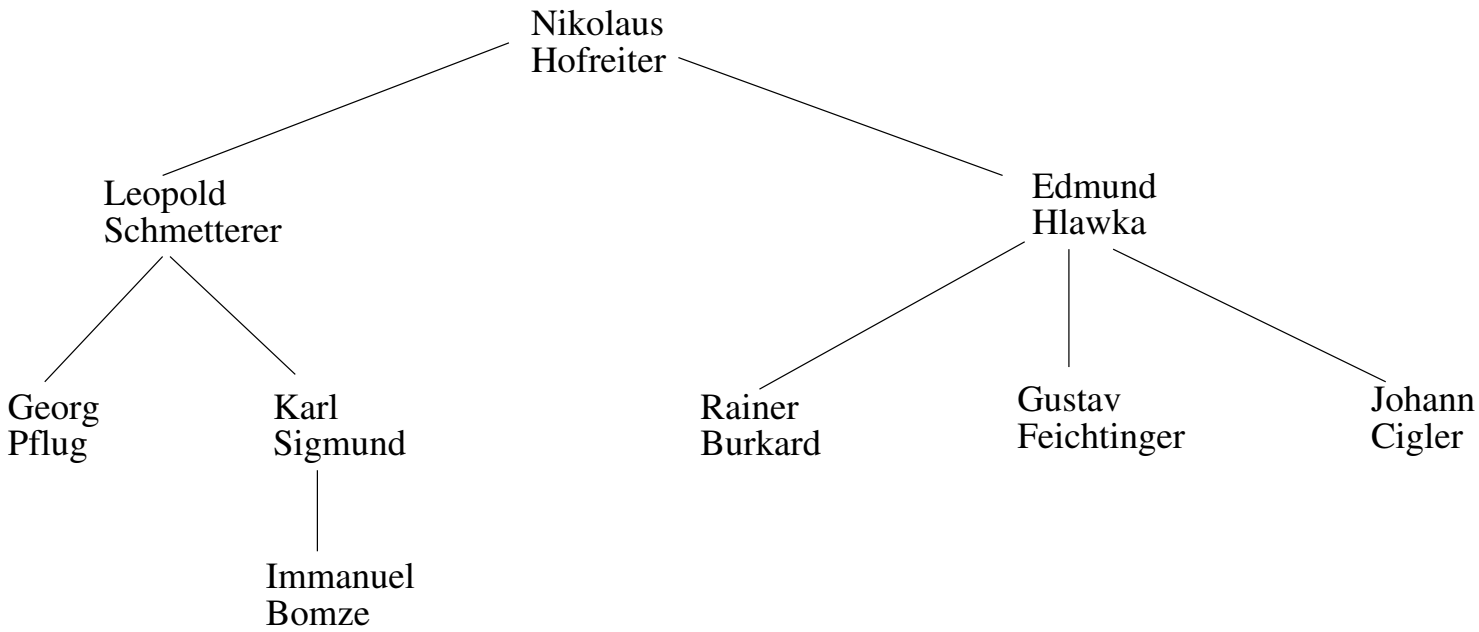
More interesting: Math genealogy



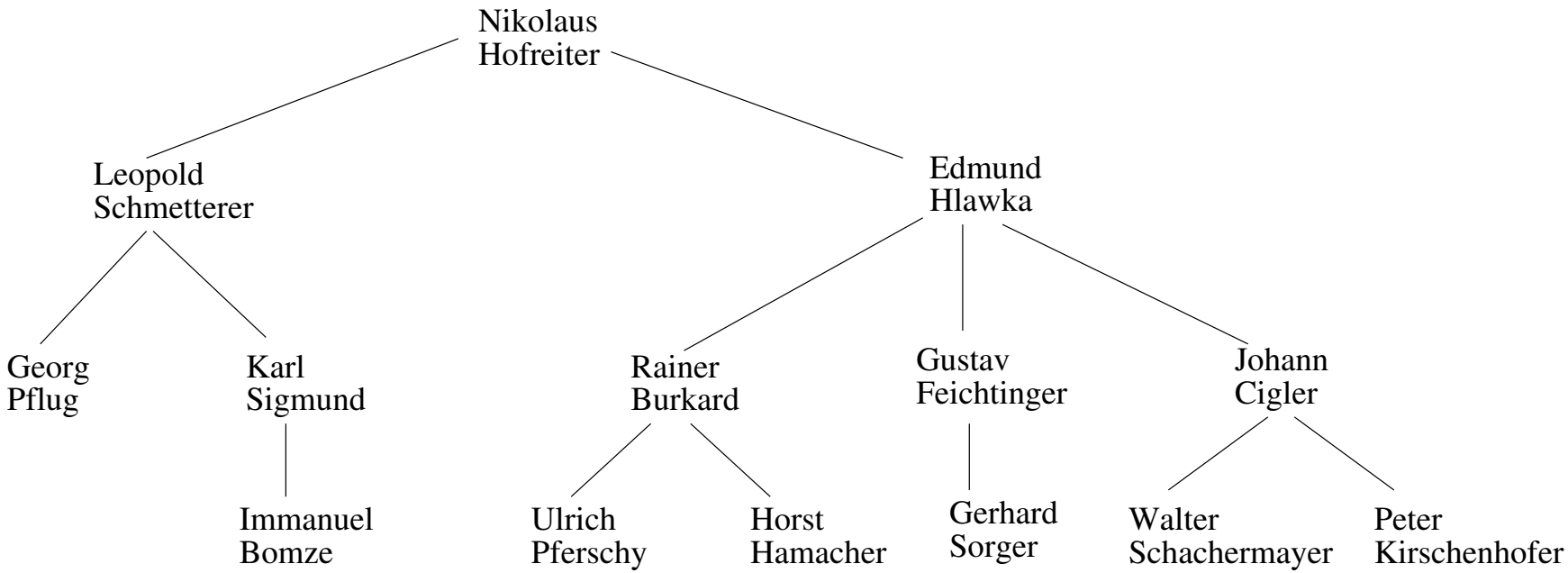
More interesting: Math genealogy



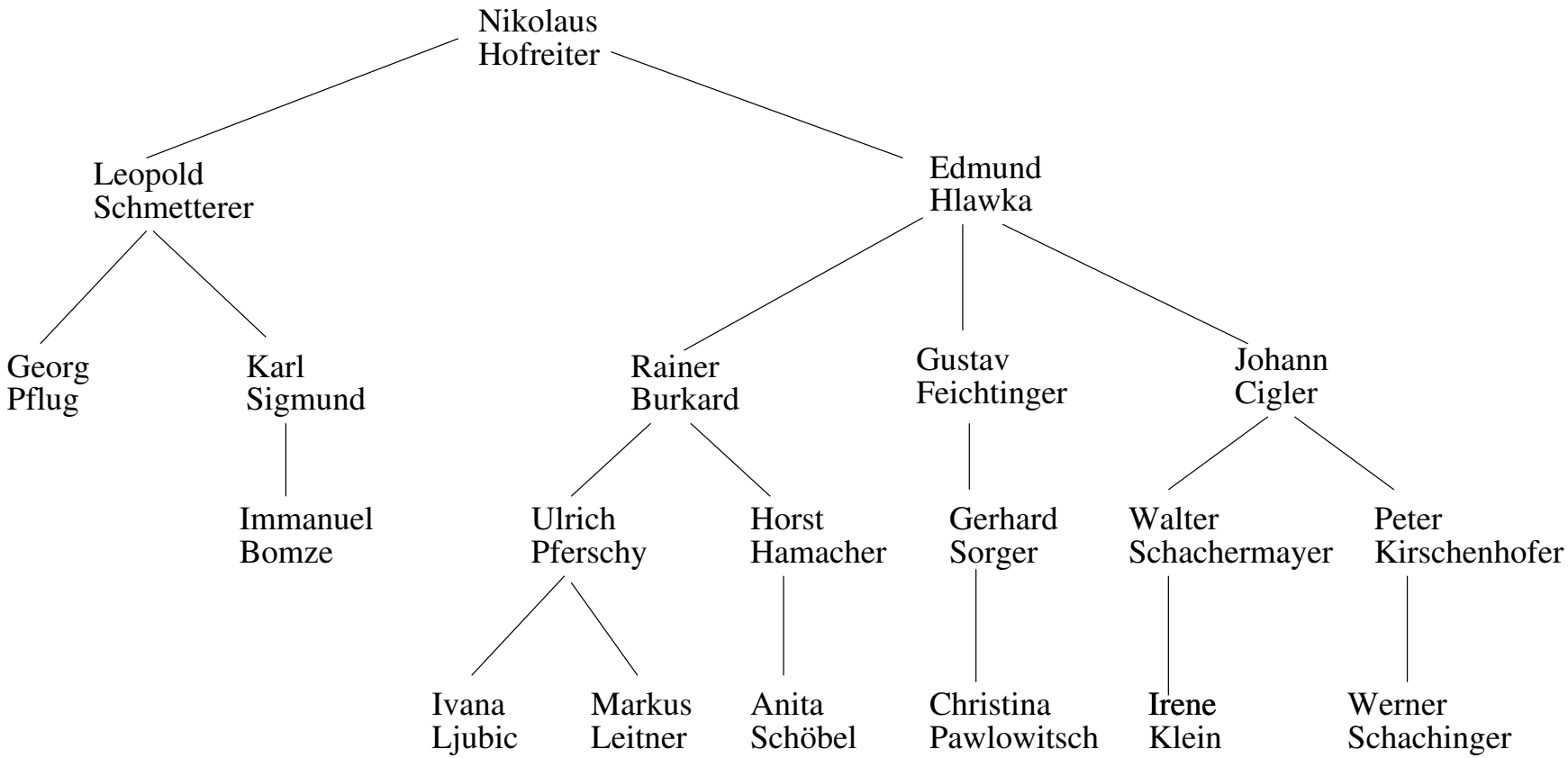
More interesting: Math genealogy



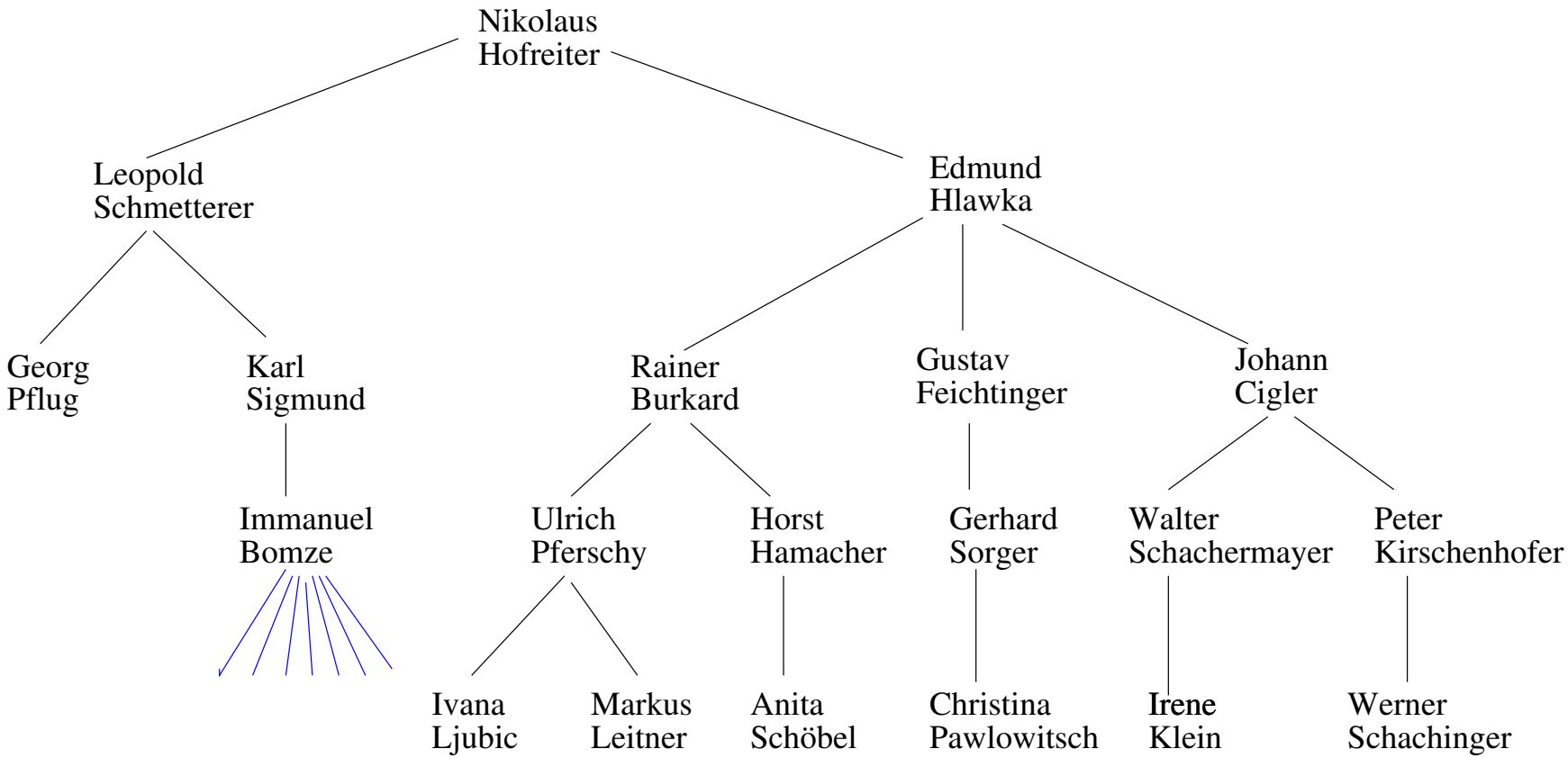
More interesting: Math genealogy



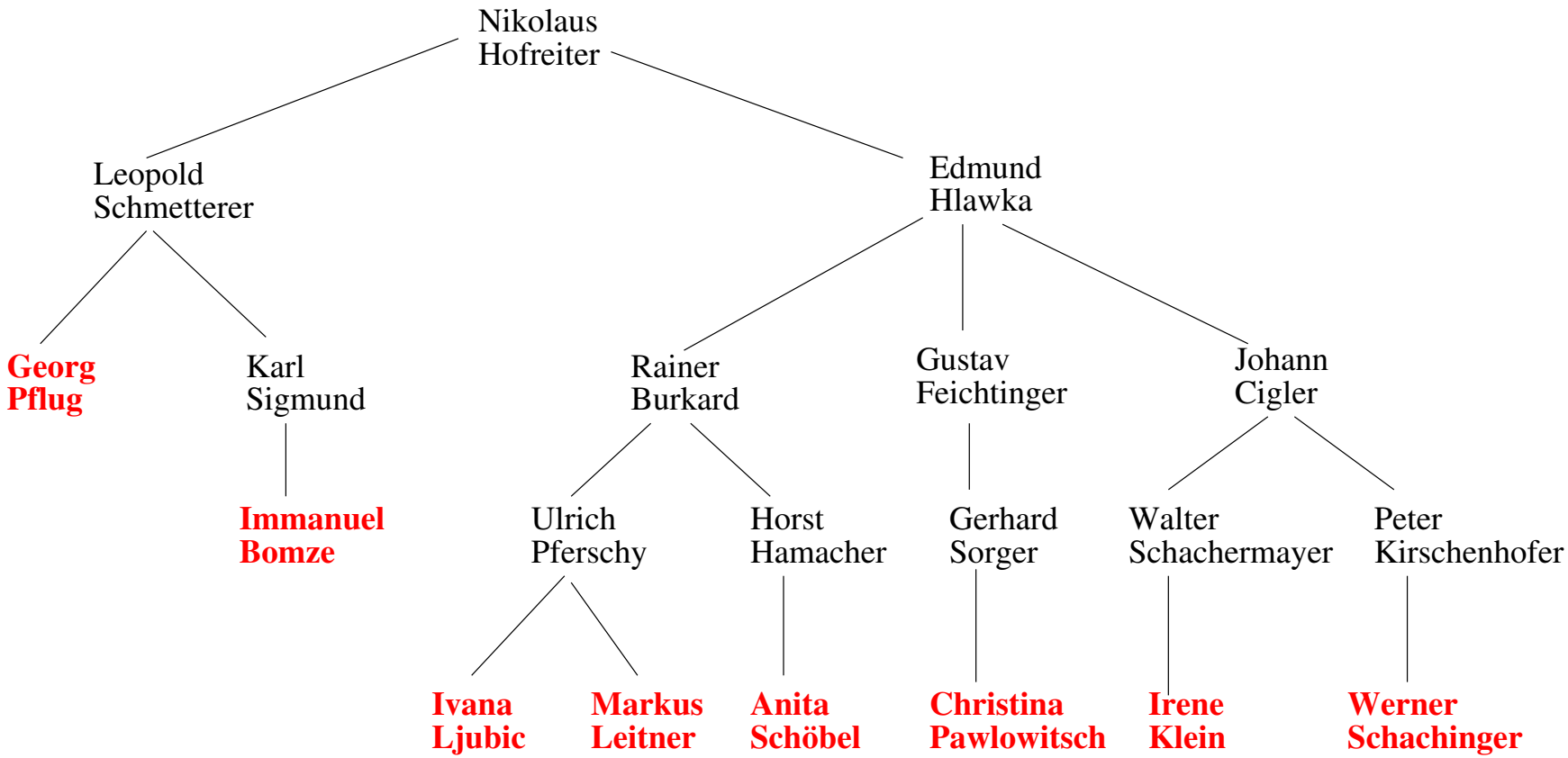
More interesting: Math genealogy



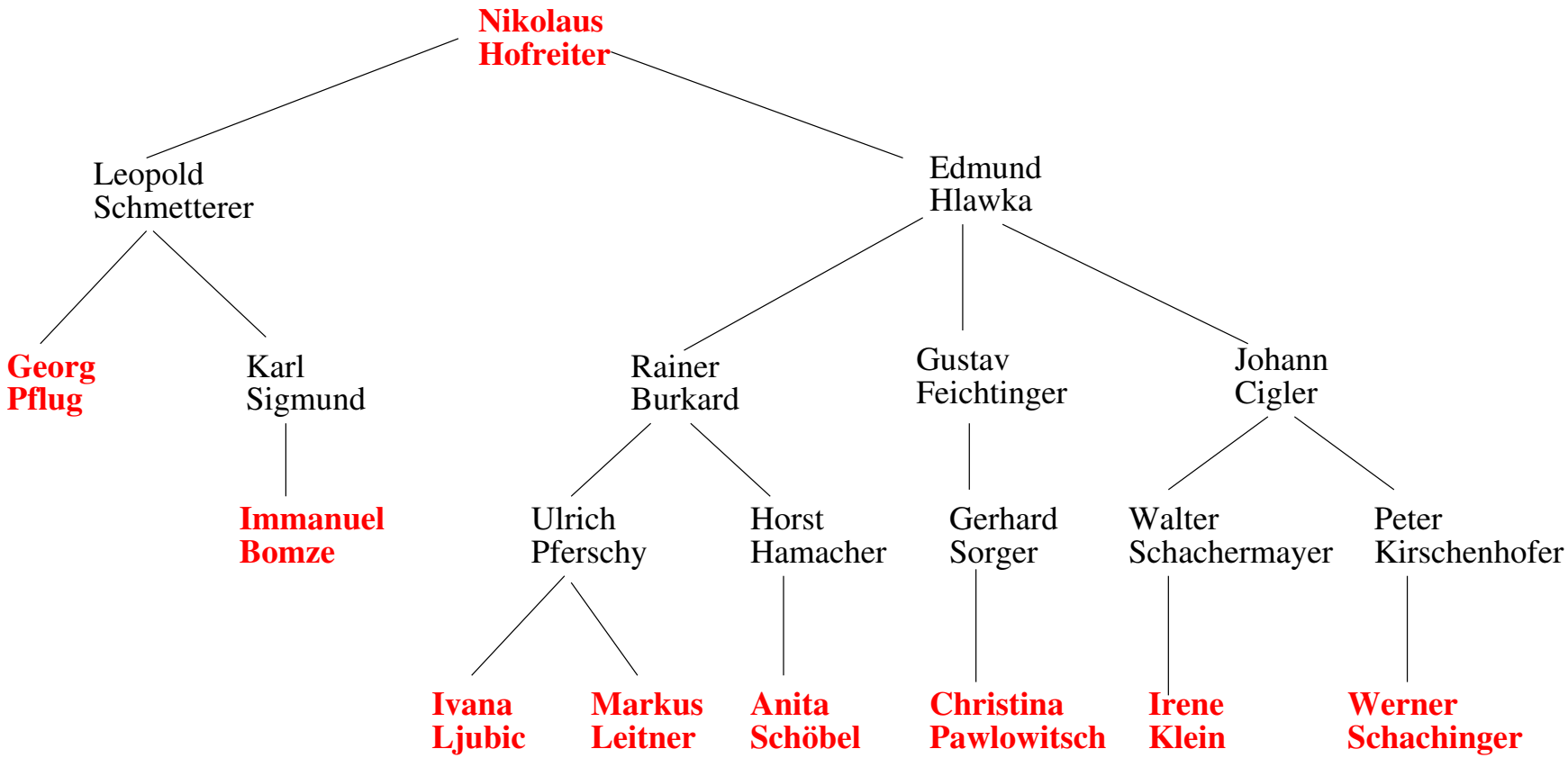
More interesting: Math genealogy



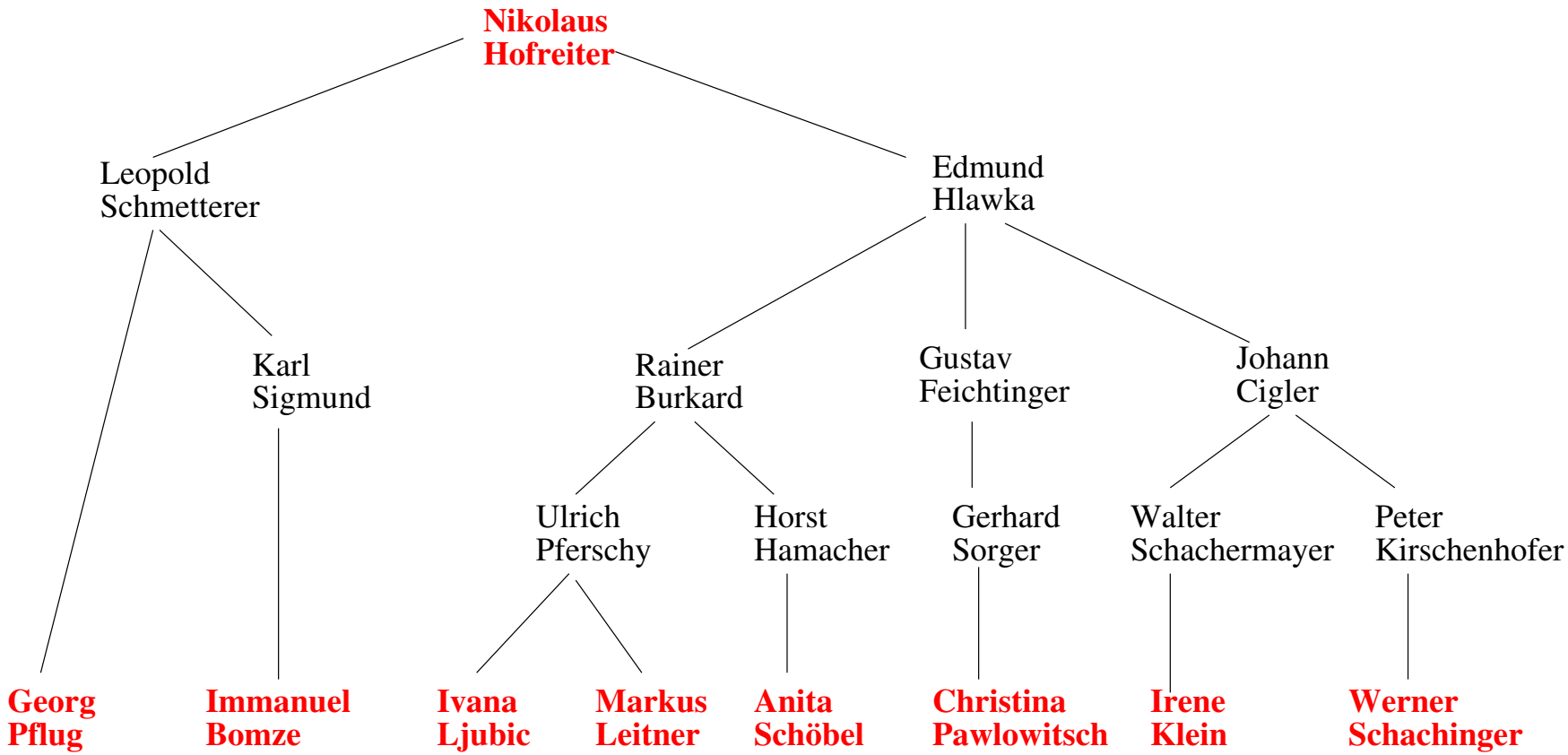
More interesting: Math genealogy



More interesting: Math genealogy



More interesting: Math genealogy

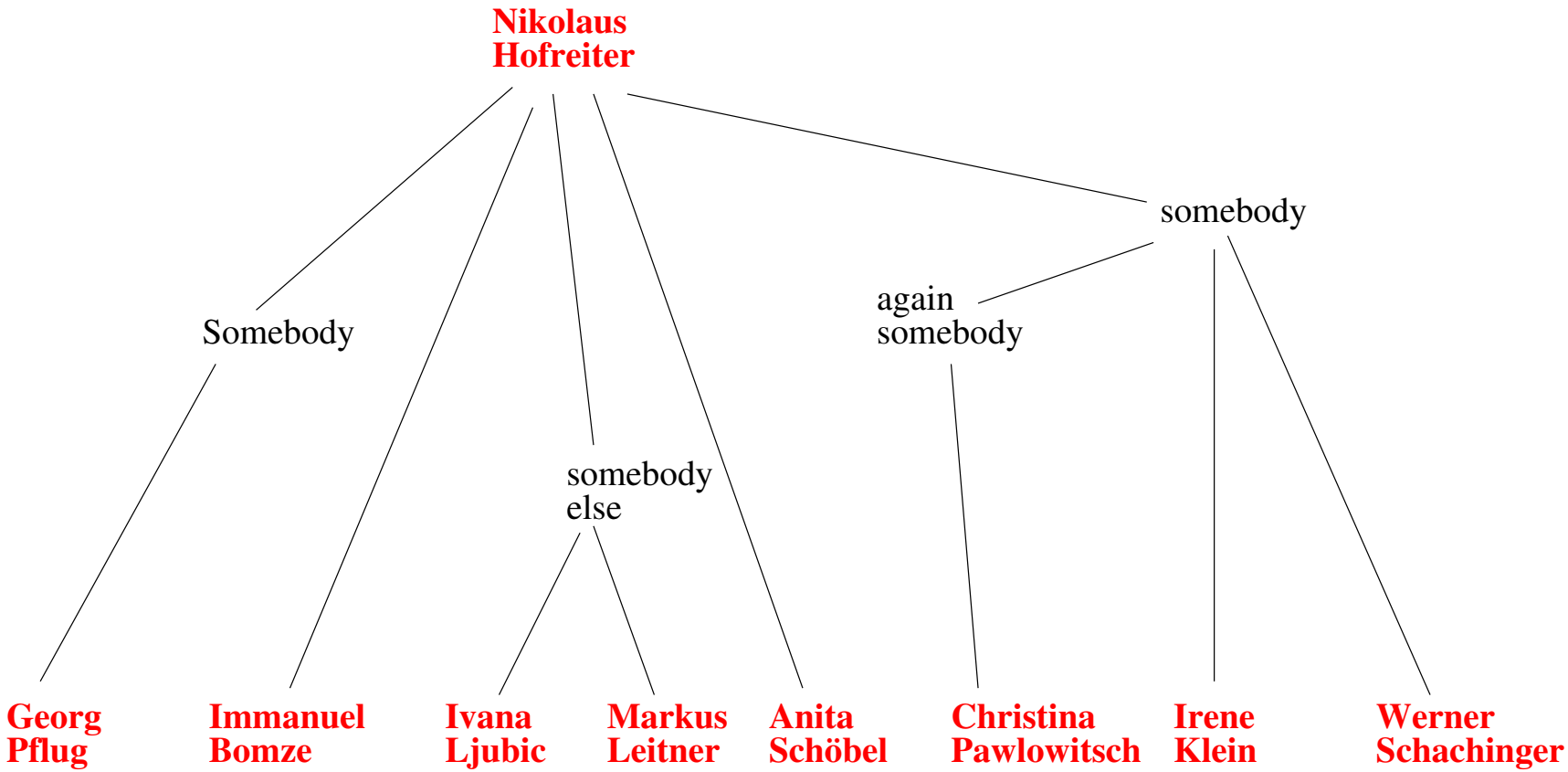


More interesting: Math genealogy

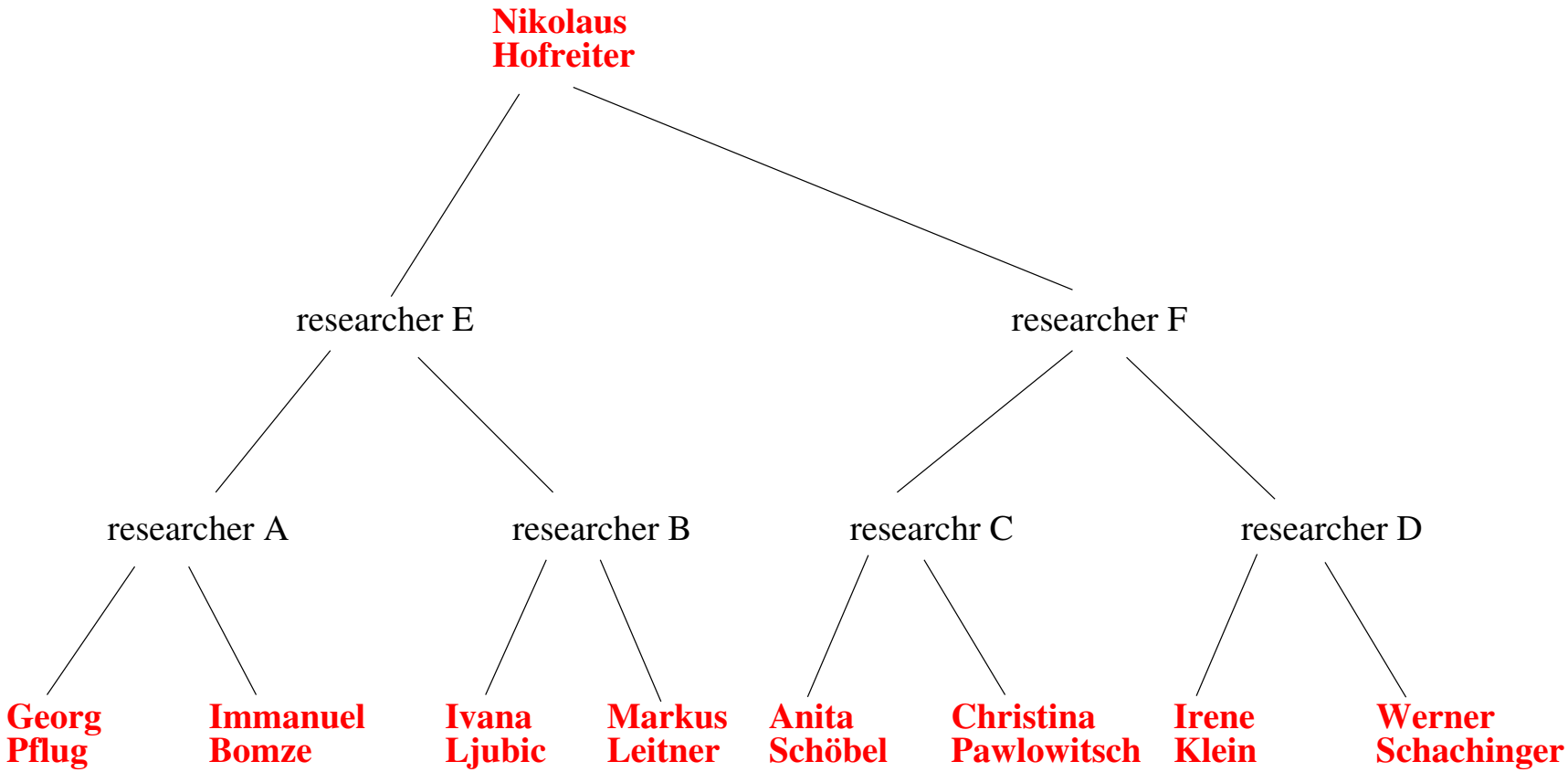
Nikolaus Hofreiter

Georg Pflug | **Immanuel Bomze** | **Ivana Ljubic** | **Markus Leitner** | **Anita Schöbel** | **Christina Pawlowitsch** | **Irene Klein** | **Werner Schachinger**

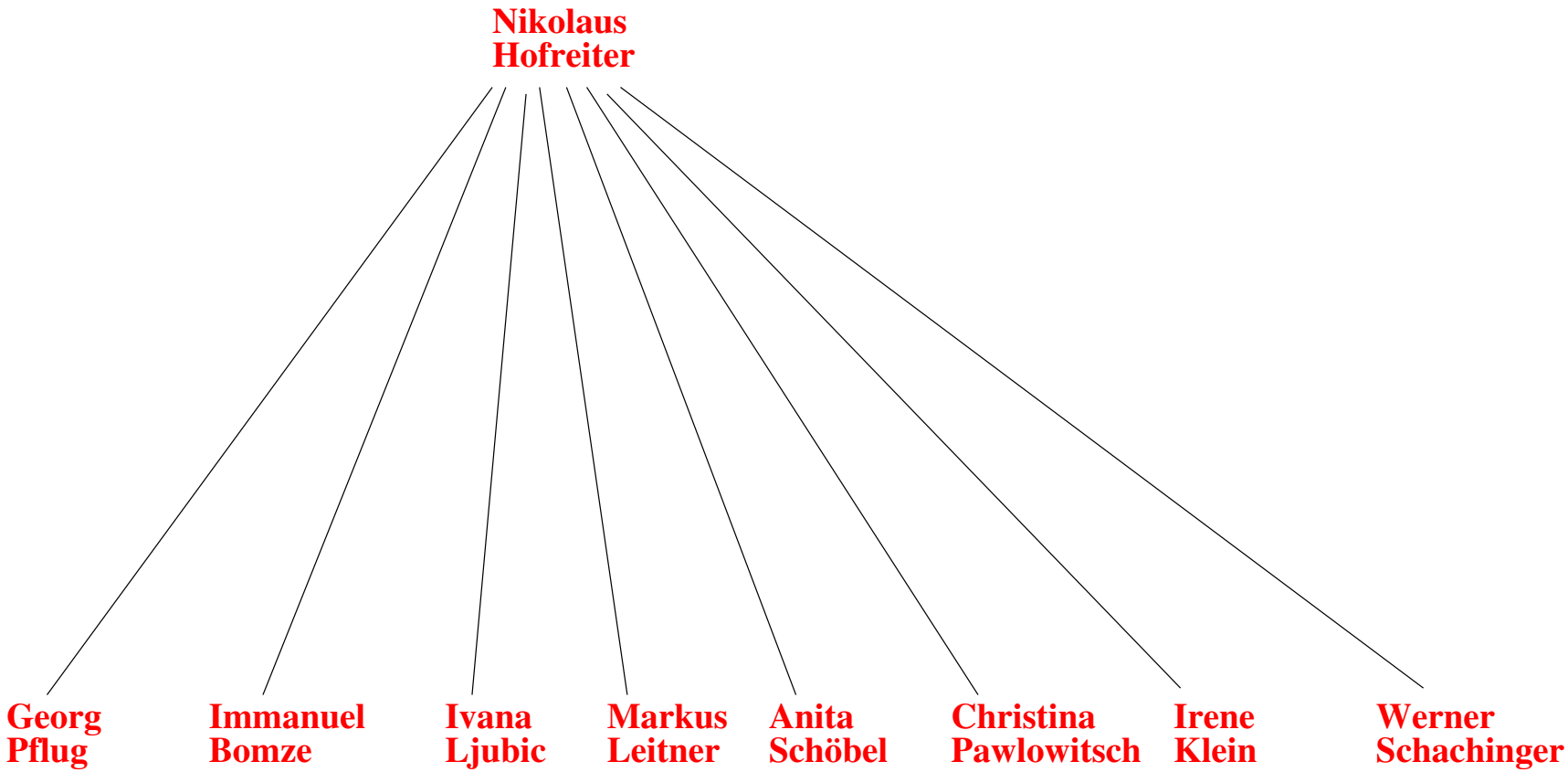
More interesting: Math genealogy



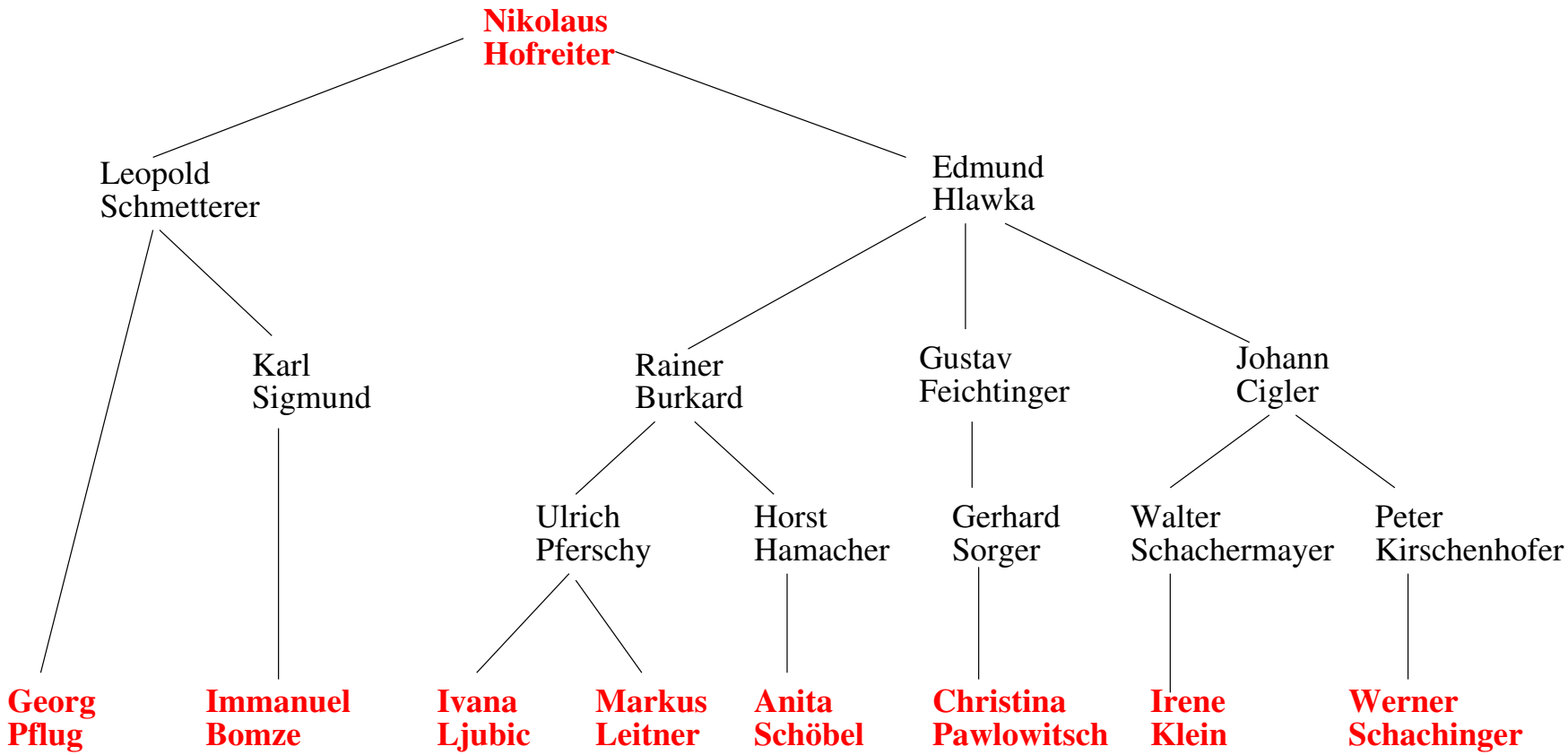
More interesting: Math genealogy



More interesting: Math genealogy



More interesting: Math genealogy

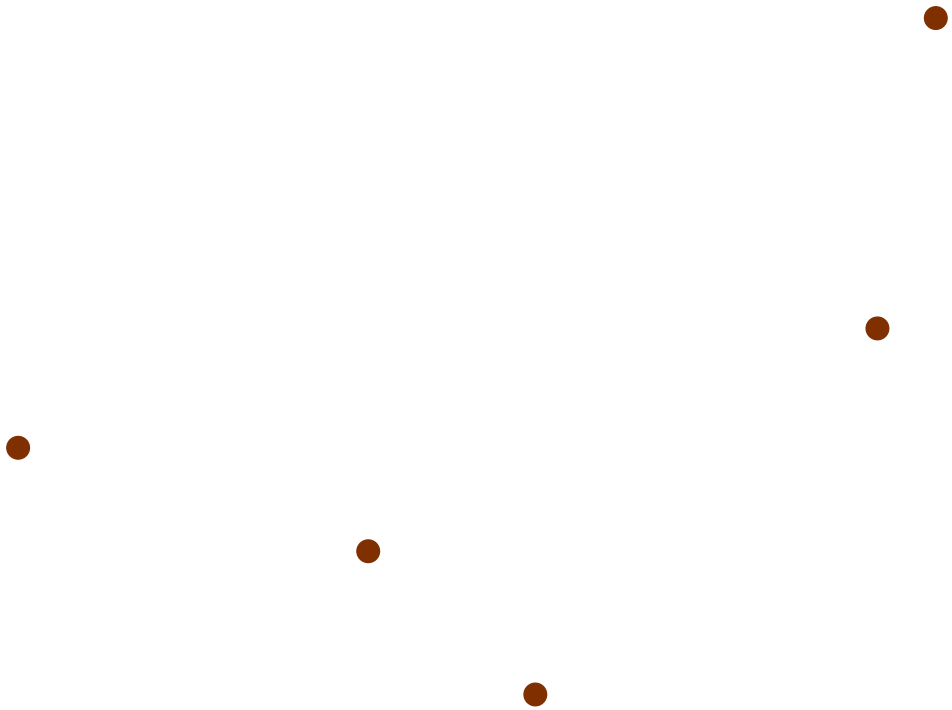


Contents

Contents

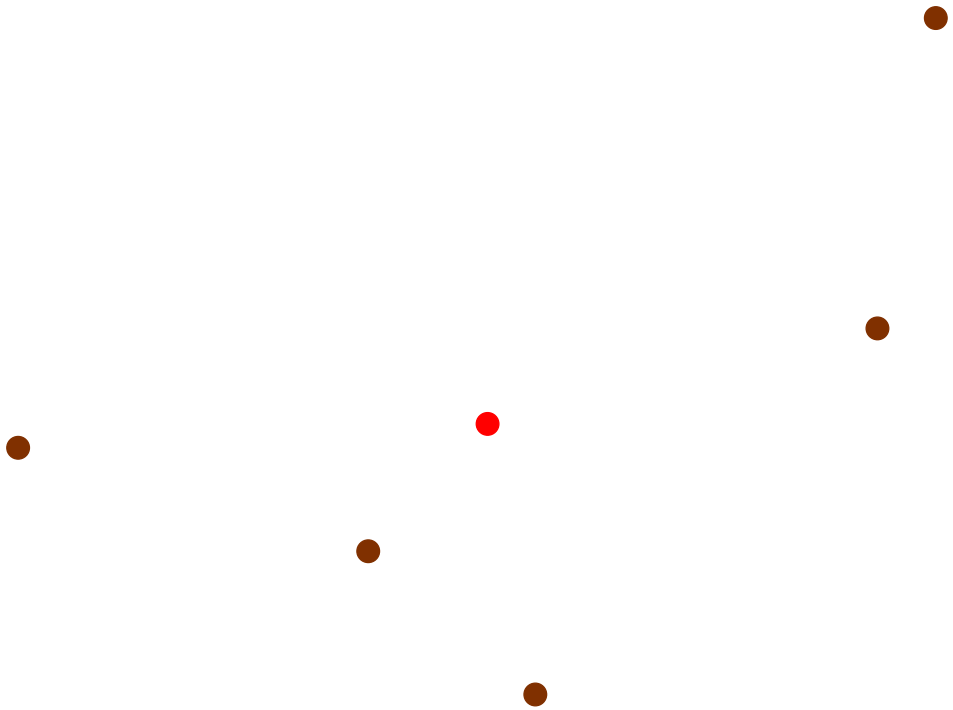
Euclidean location problems

Question: Given a set of L points in the plane, find the point that minimizes the sum of Euclidean distances to them.



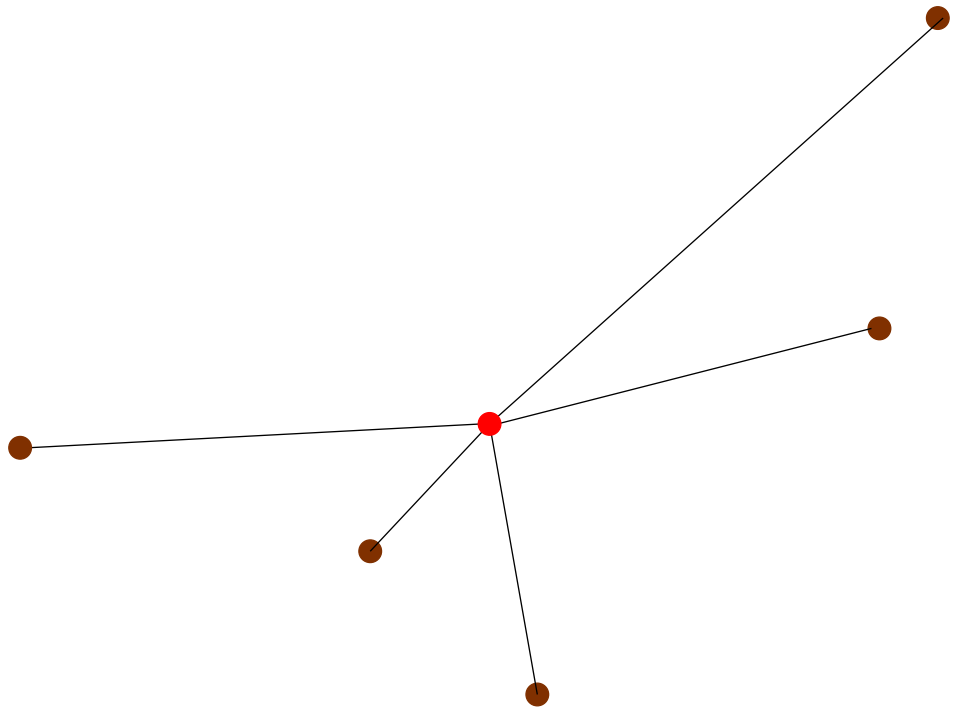
Euclidean location problems

Question: Given a set of L points in the plane, find the point that minimizes the sum of Euclidean distances to them.



Euclidean location problems

Question: Given a set of L points in the plane, find the point that minimizes the sum of Euclidean distances to them.



1-median Euclidean location problem
(Weber problem/ Fermat-Torricelli-problem)

The Tree Space

Is a metric space.

The Tree Space

Is a metric space.

- Elements of the space: All trees such that n given species are leaves of the tree.
- Distance: geodesic distance describing how quickly a tree can be transformed into another one.

The Tree Space

Is a metric space.

- Elements of the space: All trees such that n given species are leaves of the tree.
- Distance: geodesic distance describing how quickly a tree can be transformed into another one.

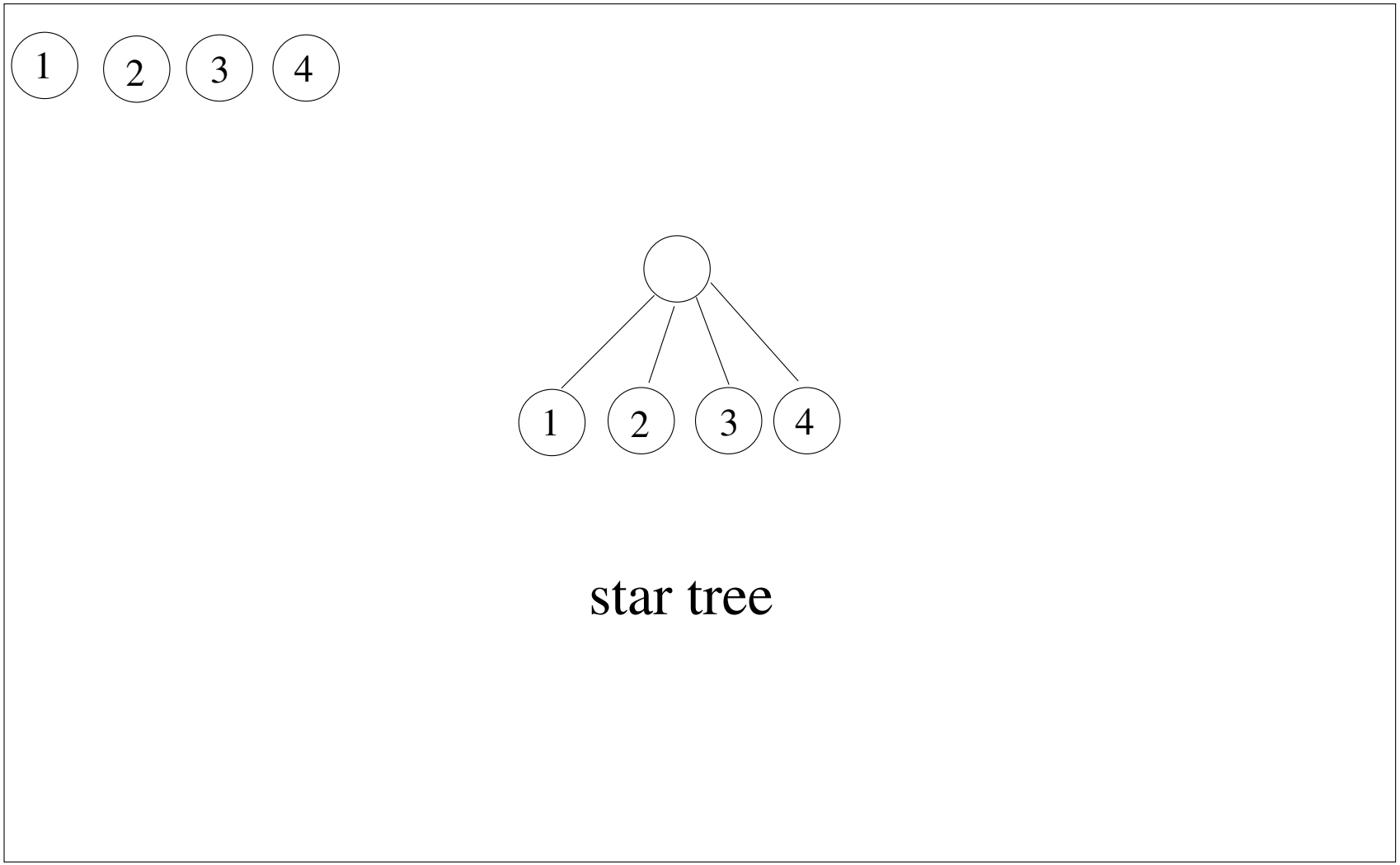
Question: Given a set of L such trees (estimated by biologists), find the *real* one.

Idea: The real one minimizes the geodesic distances to the estimated trees. We receive a

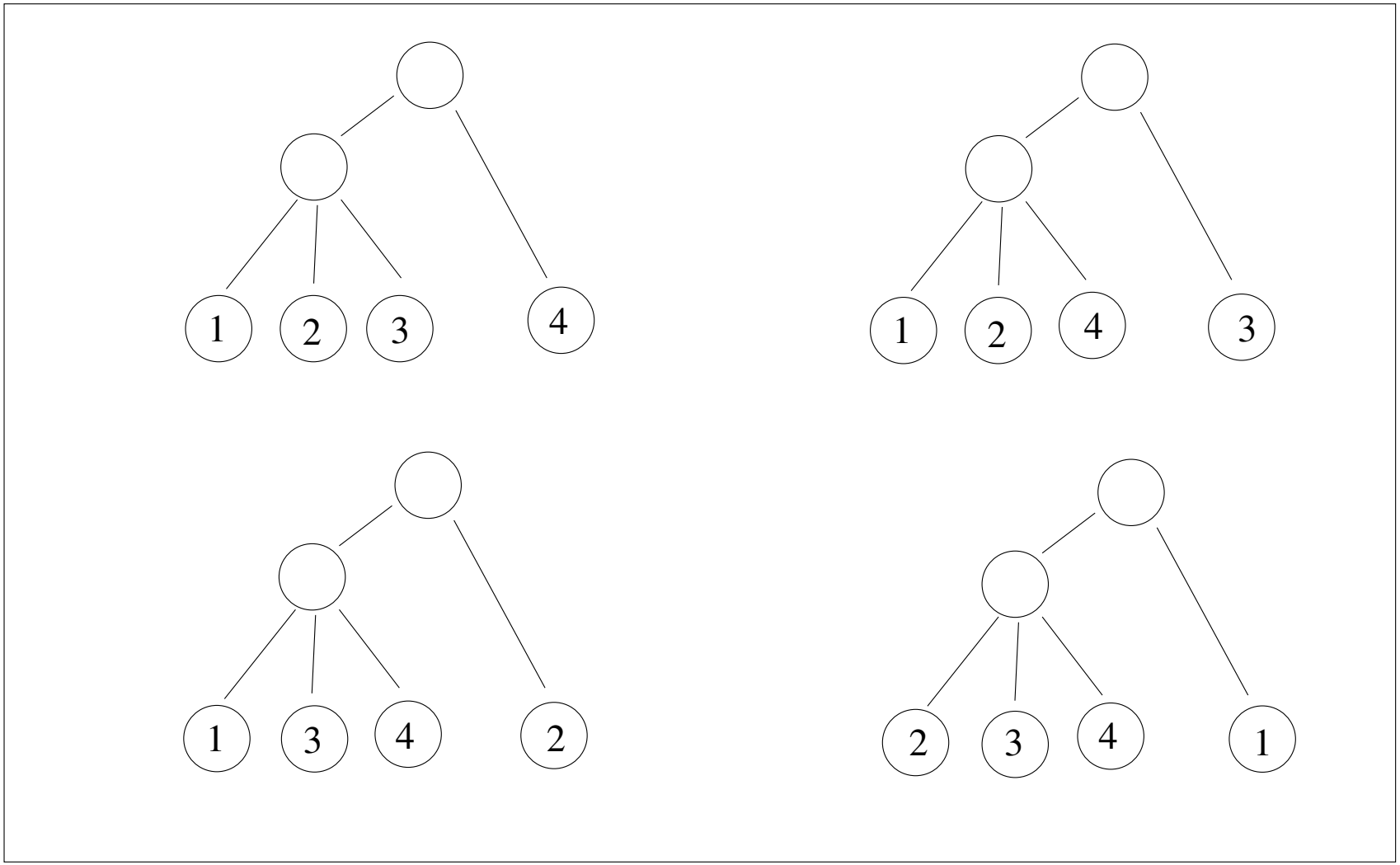
1-median location problem in the tree space.

Contents

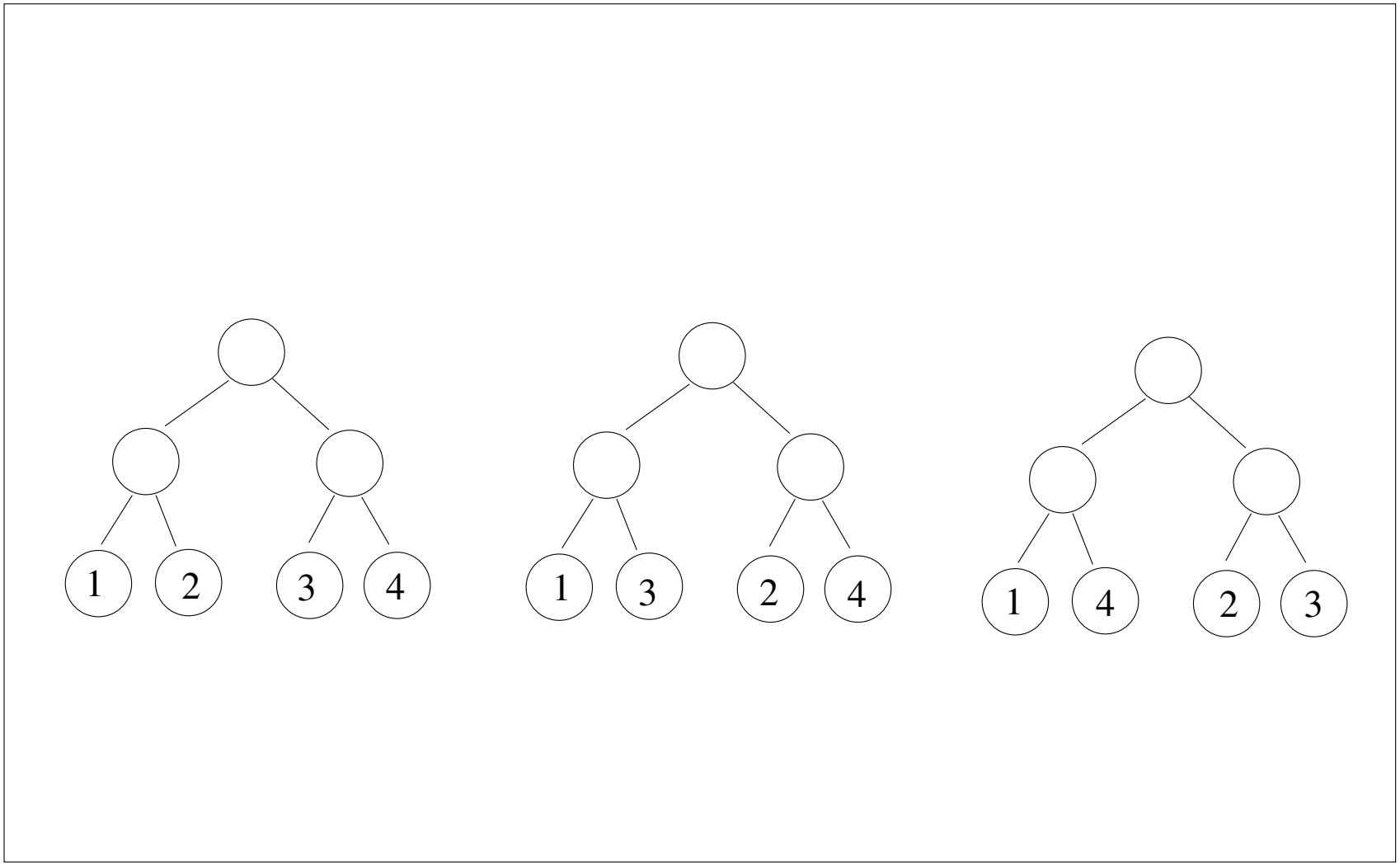
Example: Phylogenetic Trees with four species



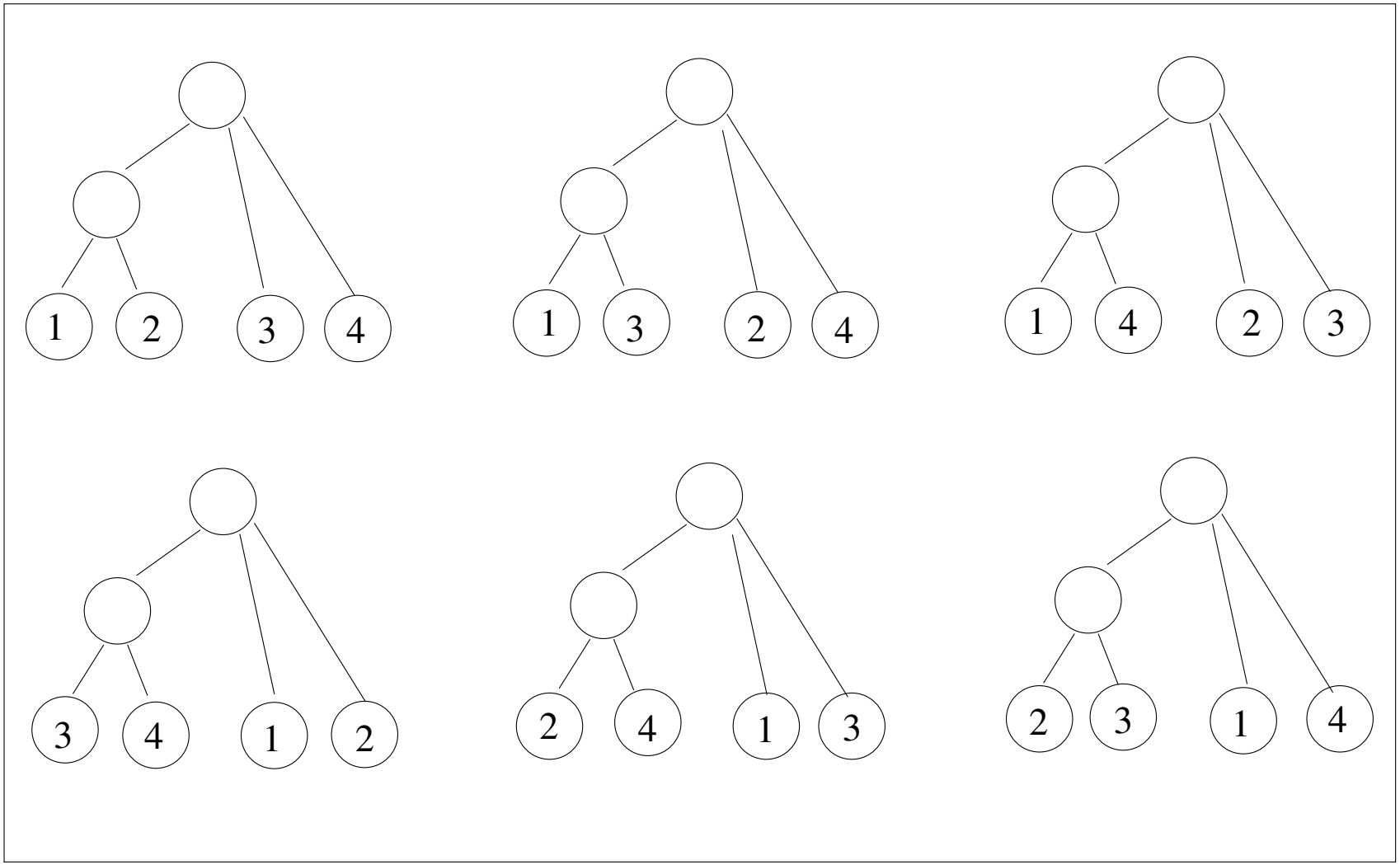
Example: Phylogenetic Trees with four species



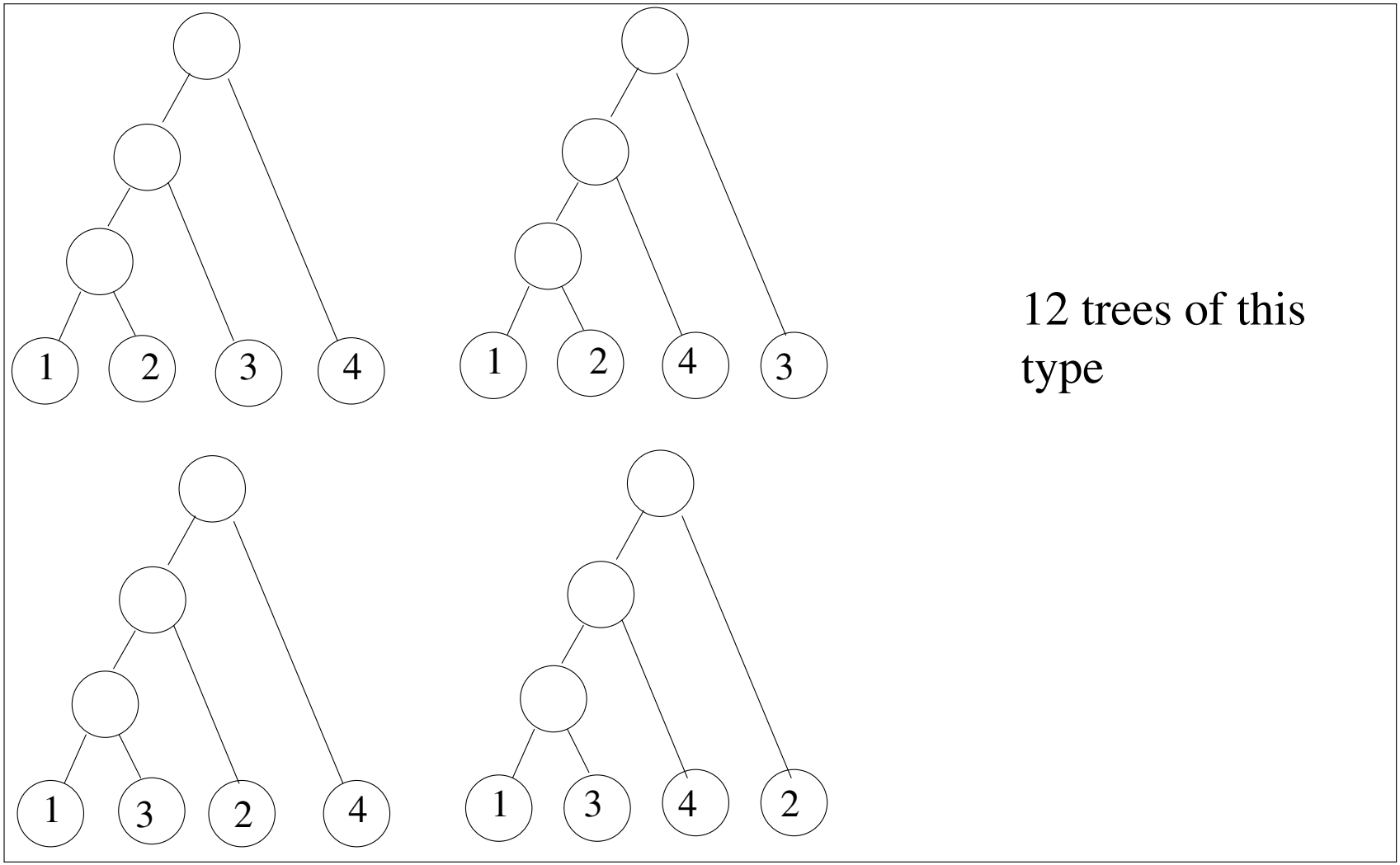
Example: Phylogenetic Trees with four species



Example: Phylogenetic Trees with four species



Example: Phylogenetic Trees with four species



Example: Phylogenetic Trees with four species

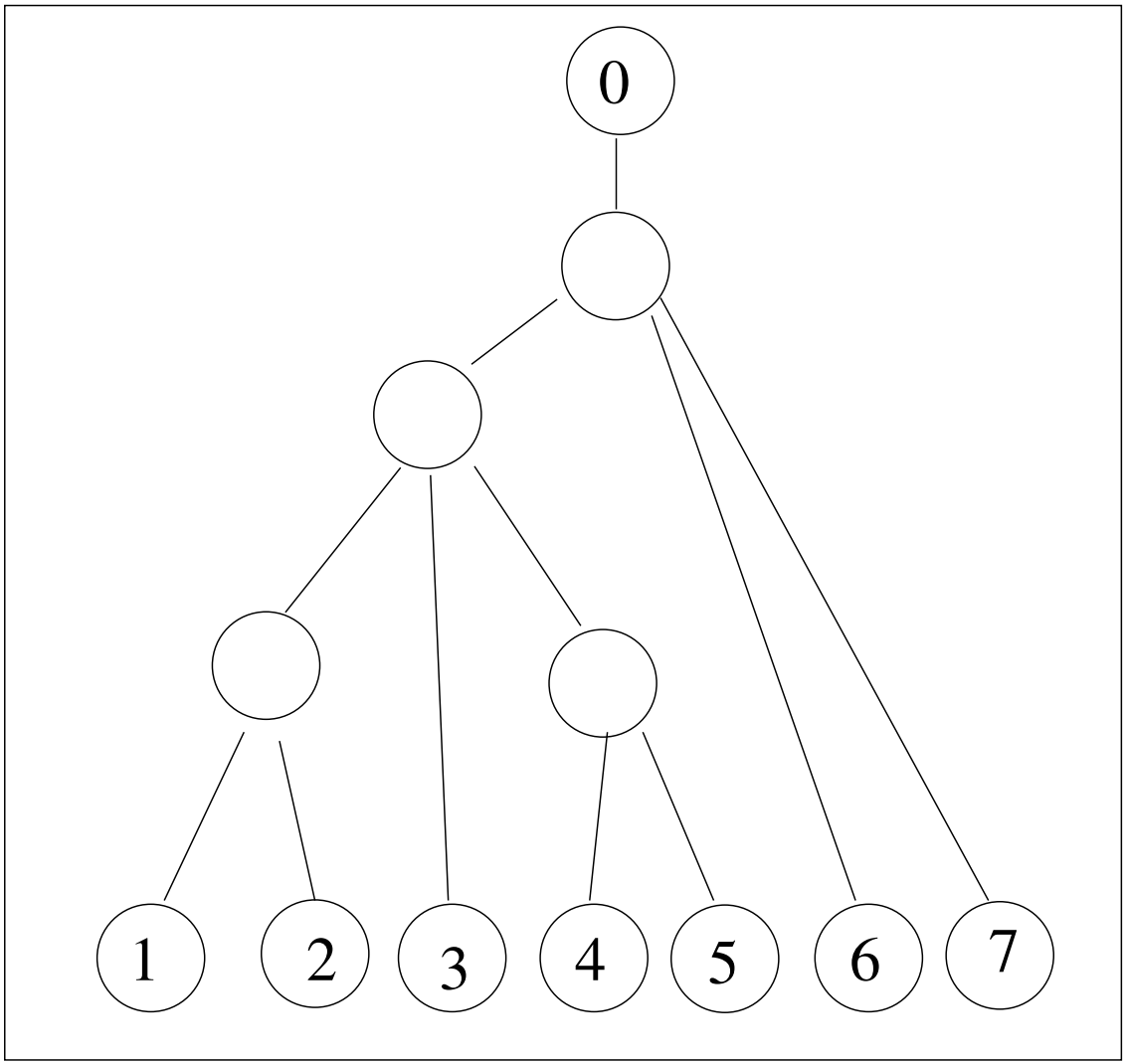
The tree space for 4 species $\mathcal{T}_4$ consists of all these trees with all positive edge weights.

Example: Phylogenetic Trees with four species

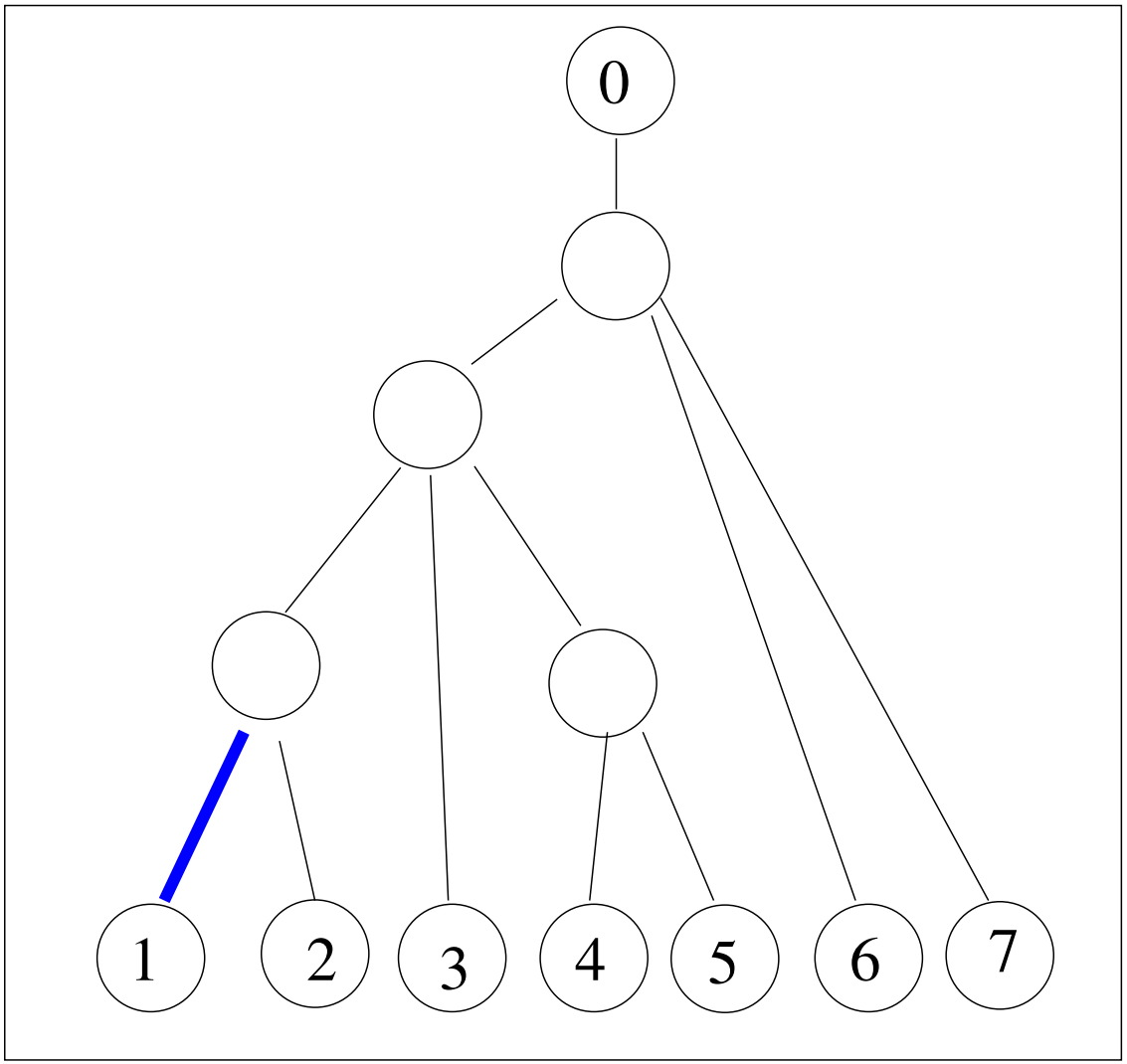
The tree space for 4 species $\mathcal{T}_4$ consists of all these trees with all positive edge weights.

What is a mathematically sound description of these trees?

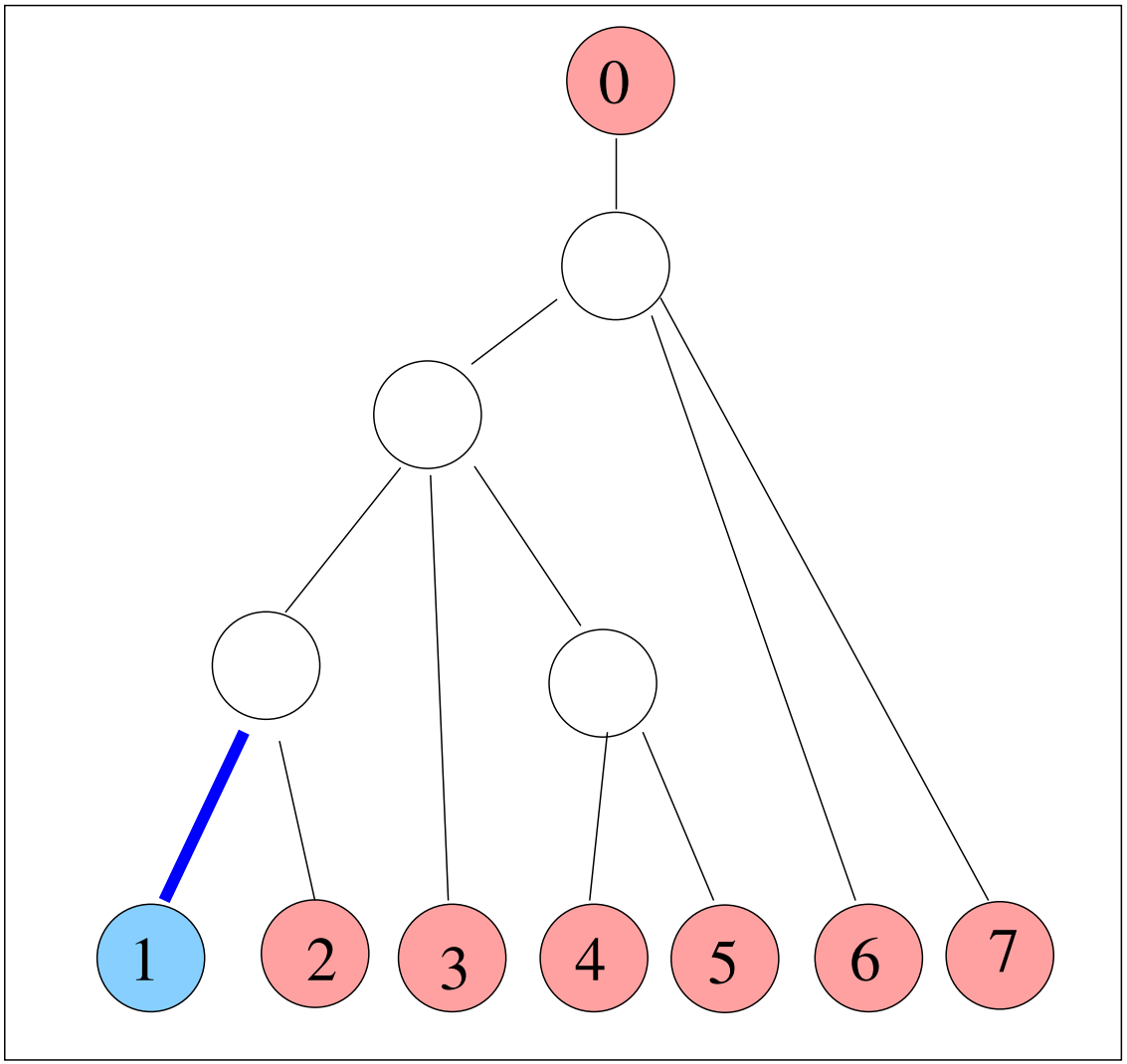
Each edge makes a partition of the leaves



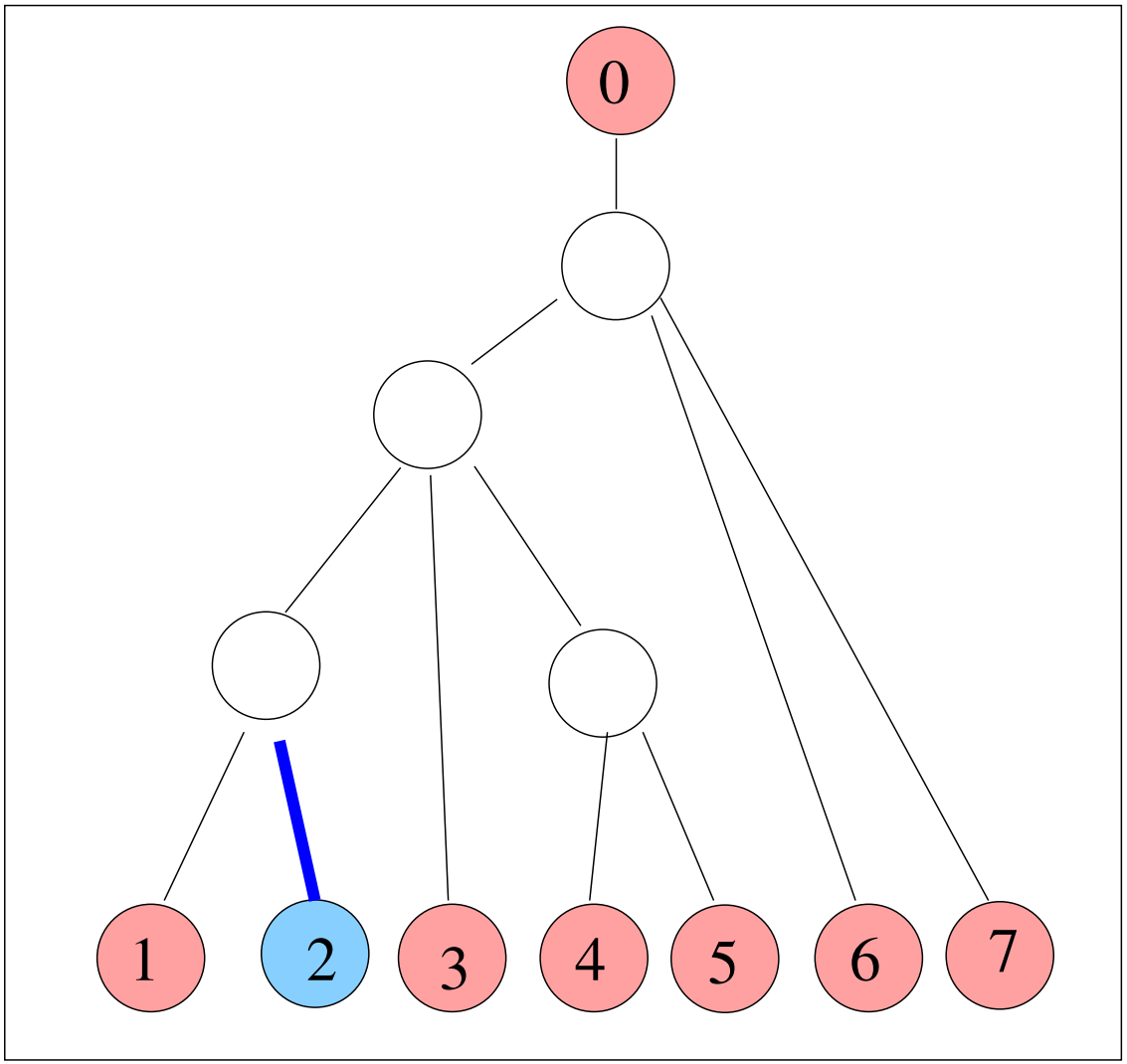
Each edge makes a partition of the leaves



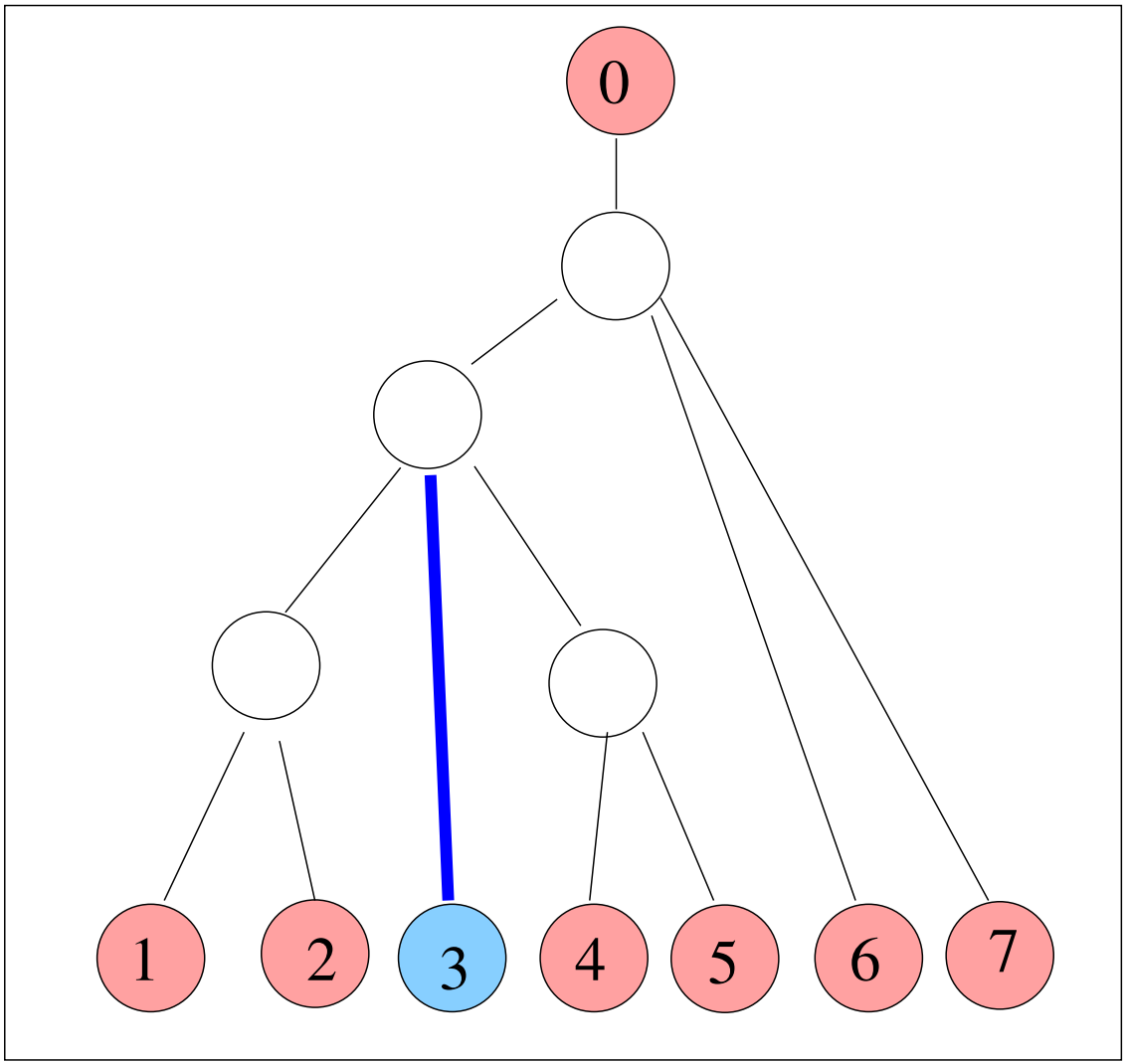
Each edge makes a partition of the leaves



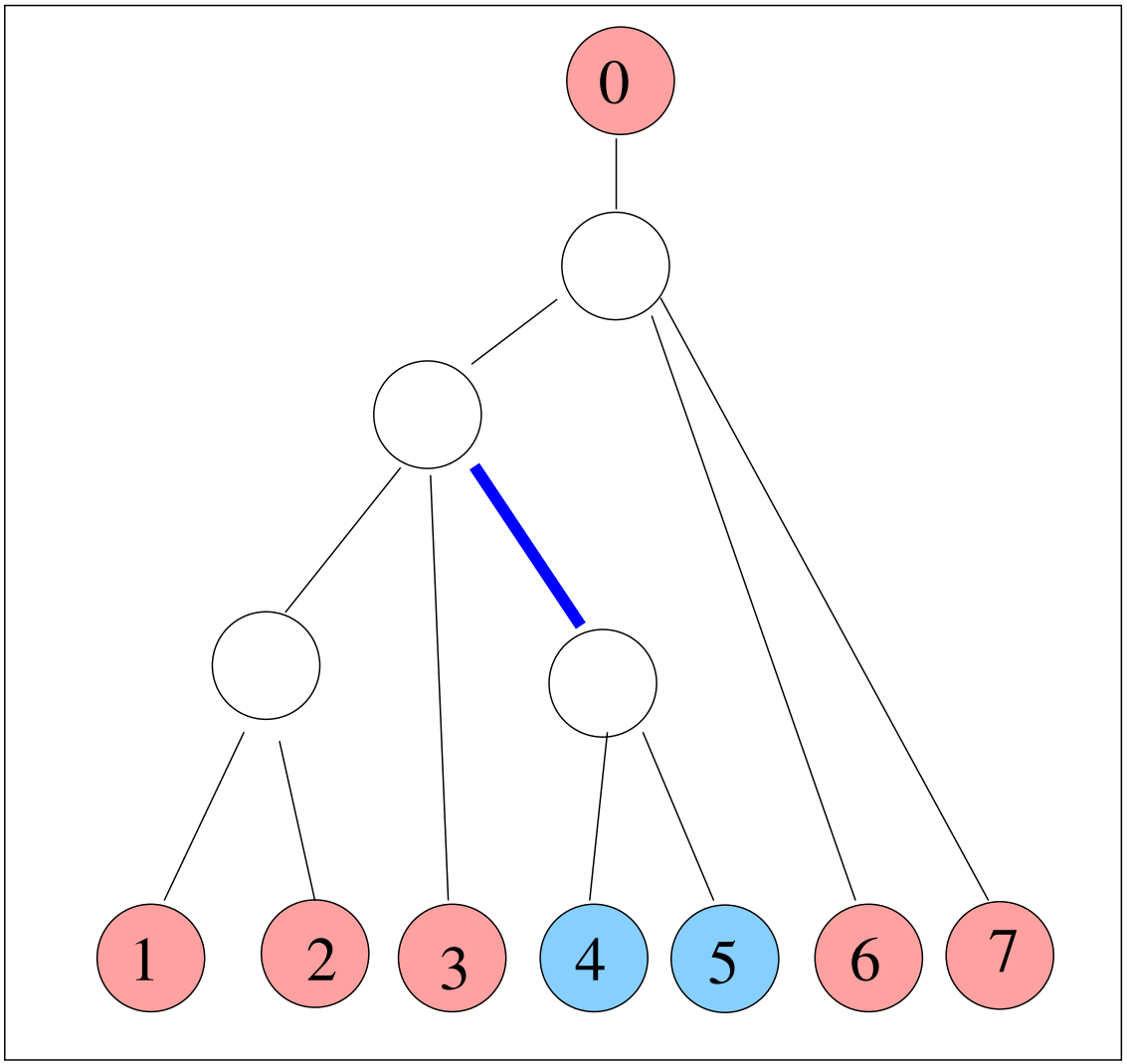
Each edge makes a partition of the leaves



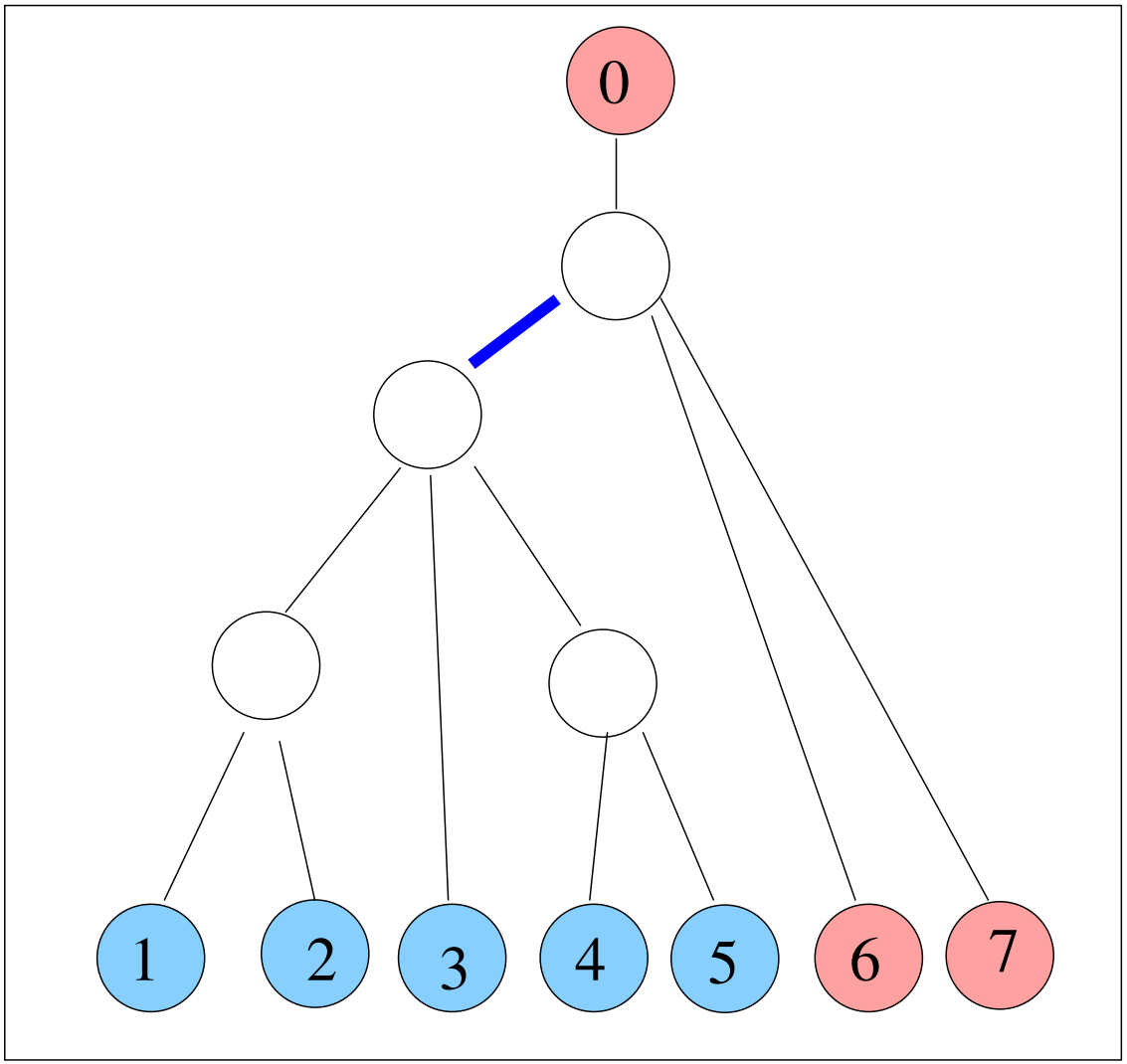
Each edge makes a partition of the leaves



Each edge makes a partition of the leaves

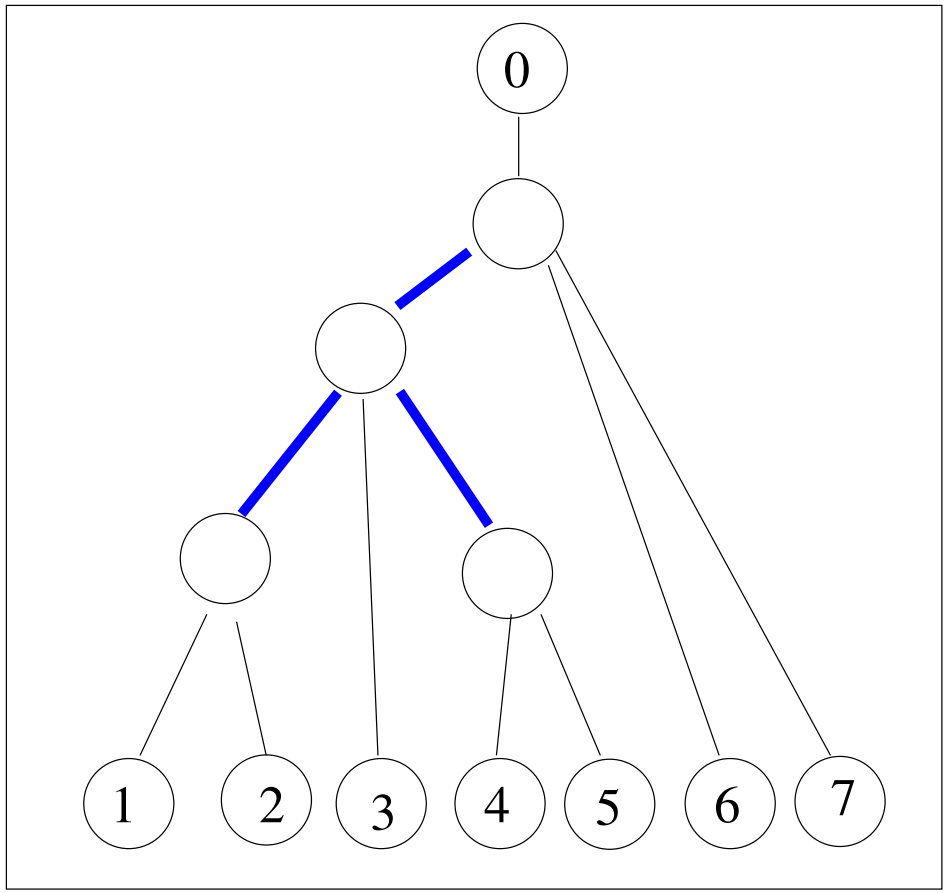


Each edge makes a partition of the leaves



Describing the topology of a weighted tree

- Each edge gives a partition of the leaves.
- Only the *inner edges* (not connected to a leaf) are interesting.



Splits

Definition (Split)

A split of a set A of leaves is a partition $(A|A^C)$ of the leaves with $|A| \geq 2, |A^C| \geq 2$.

For n species there exist $N := 2^n - (n+2)$ possible splits.

Notation: $Split(T)$ denotes the splits of T defined by its inner edges.

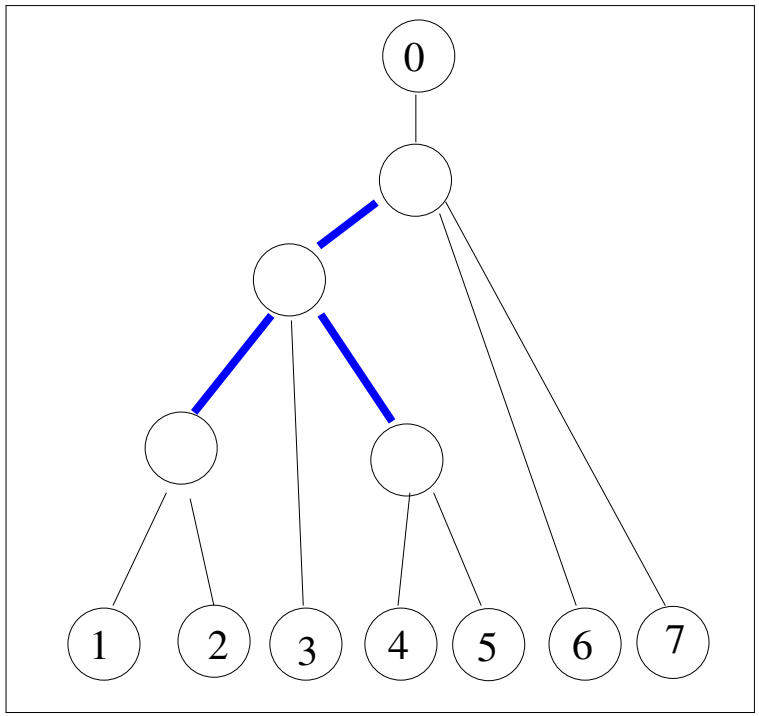
Theorem (Billera, Holmes and Vogtman (2001))

A tree is uniquely characterized by its splits and the weights on them.

Orthants

Definition

Two weighted trees T_1 and T_2 are in the same orthant if they have the same inner edges, i.e., the same topology.



For location theory: Embedding of the tree space

A tree can be represented by a vector in $x \in \mathbb{R}_+^N$ where x_s is length of the inner edge which belongs to split s.

$$\mathcal{T}_n \subseteq \mathbb{R}^N$$

For location theory: Embedding of the tree space

A tree can be represented by a vector in $x \in \mathbb{R}_+^N$ where x_s is length of the inner edge which belongs to split s.

$$\mathcal{T}_n \subseteq \mathbb{R}^N$$

Question: Does every vector in $\mathbb{R}_+^N$ represent a tree?

For location theory: Embedding of the tree space

A tree can be represented by a vector in $x \in \mathbb{R}_+^N$ where x_s is length of the inner edge which belongs to split s.

$$\mathcal{T}_n \subseteq \mathbb{R}^N$$

Question: Does every vector in $\mathbb{R}_+^N$ represent a tree?
No!
E.g.: $(\{1,2\}|\{3,4\})$ and $(\{1,3\}|\{2,4\})$ can not exist in the same tree.

Hence the tree space is only a (small) part of $\mathbb{R}^N$, consisting of many (low-dimensional) orthants.

E.g. $\mathcal{T}_4$

- is embedded in $\mathbb{R}^{10}$,
- consisting of a union of $\mathbb{R}^3$, $\mathbb{R}^2$ and $\mathbb{R}$-orthants.
- All orthants are connected through 0.

Contents

Case 1: Distance between trees in the same orthant

Definition

Let T_a, T_b be two trees in the same orthant. Then

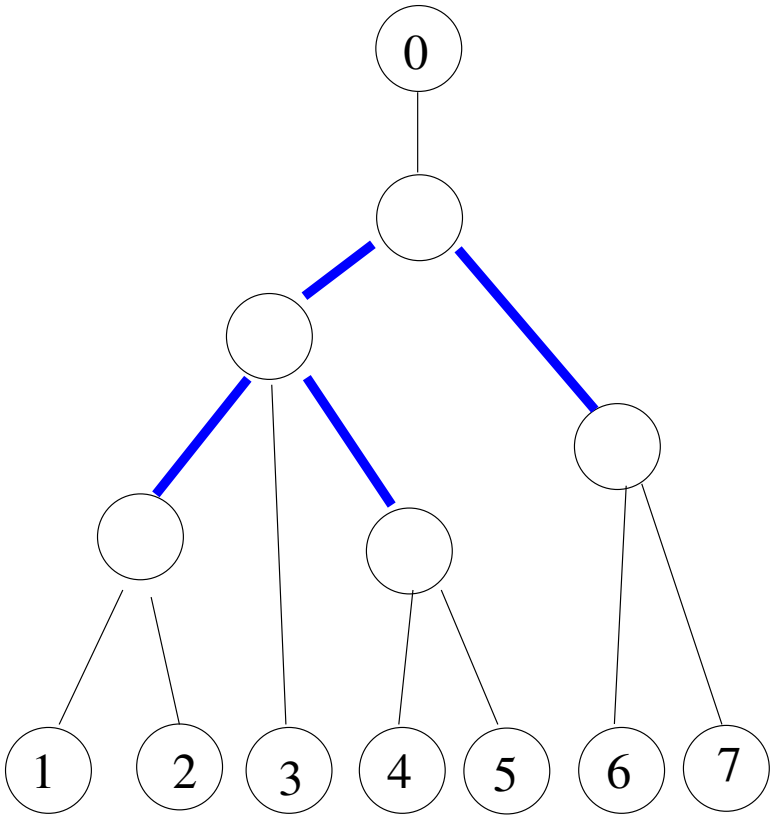
$$d(T_a, T_b) = \|T_a - T_b\|_2$$

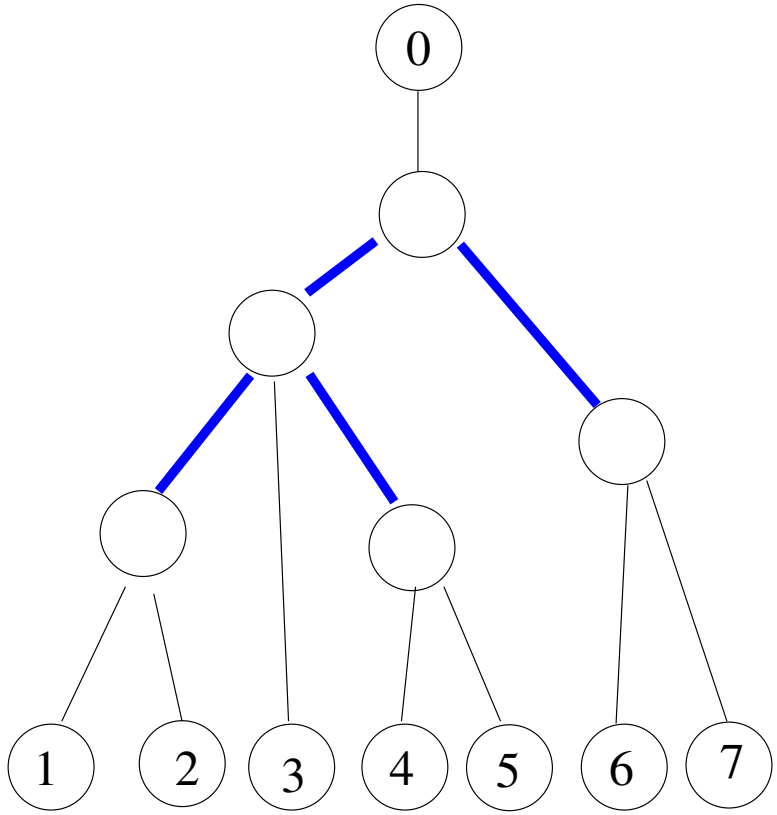
Case 2: Distance between trees in different orthants?

Trick: Edges with length zero make a tree belonging to several orthants!

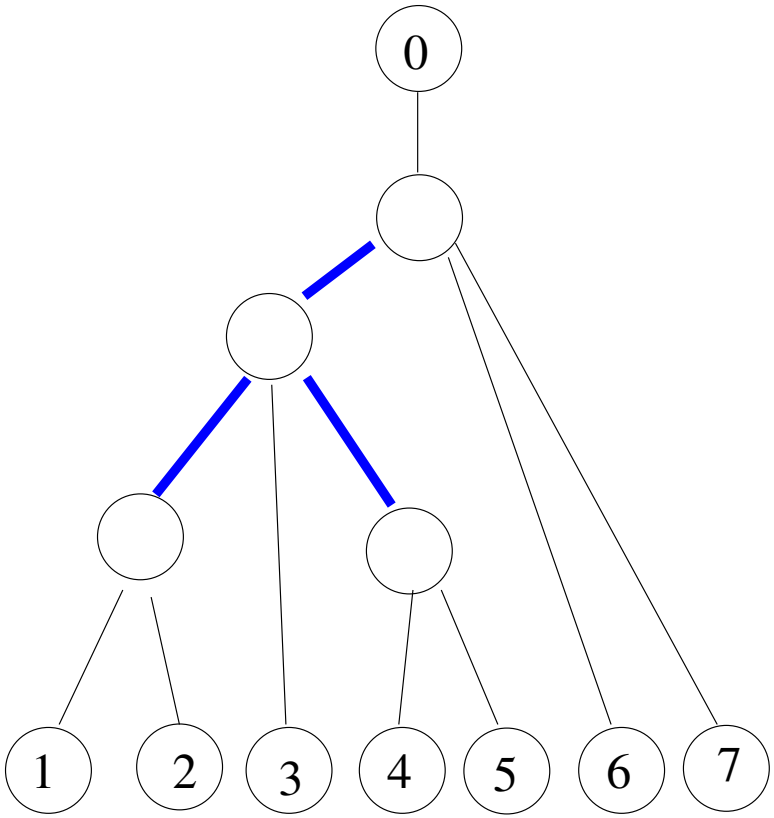
Consequence: Paths in the tree space connecting different orthants.

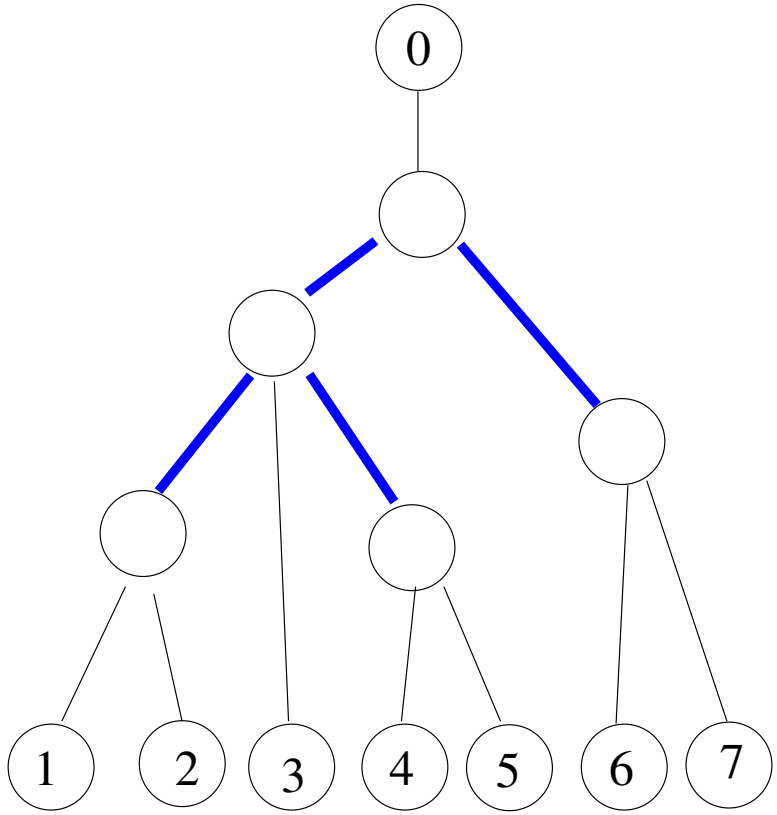
Understanding topology and orthants



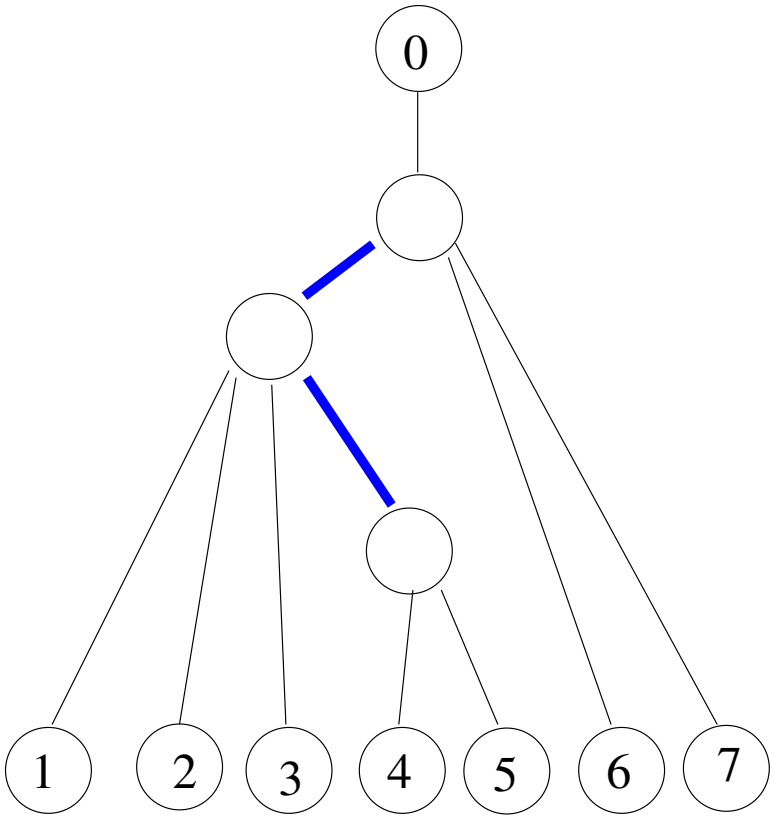


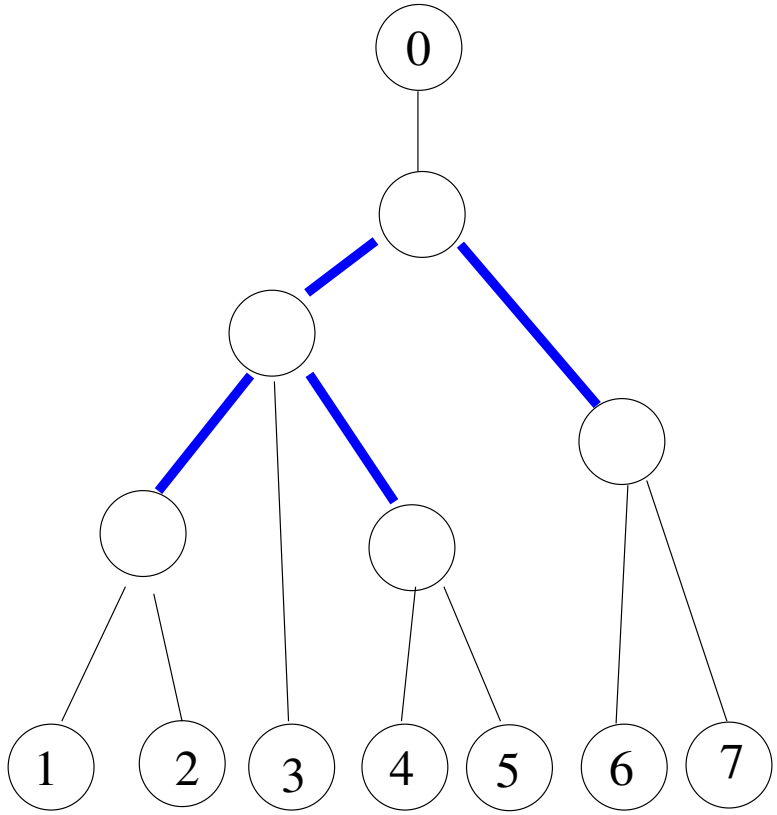
Understanding topology and orthants



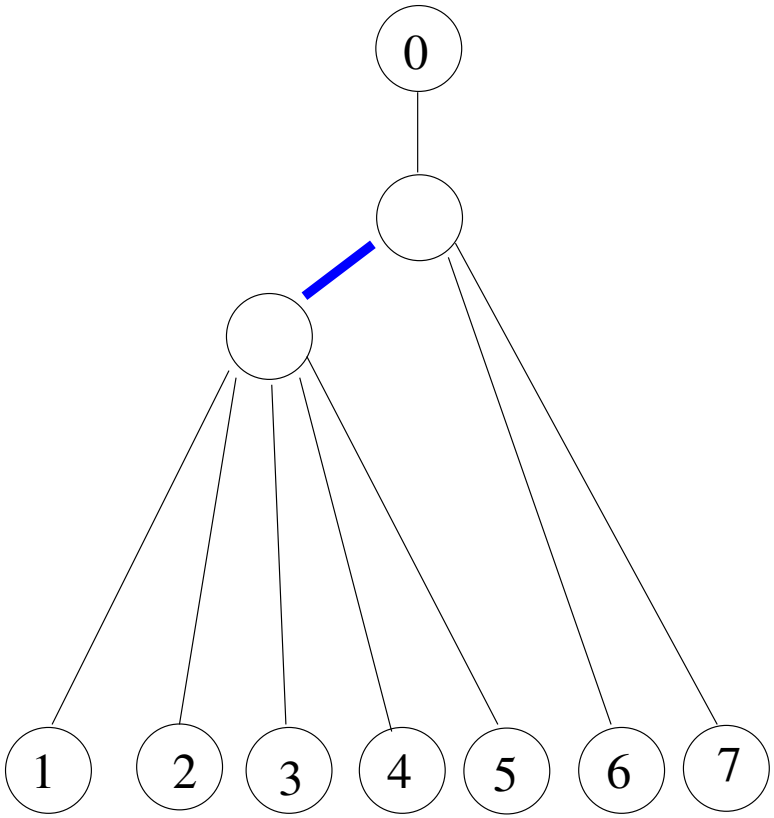


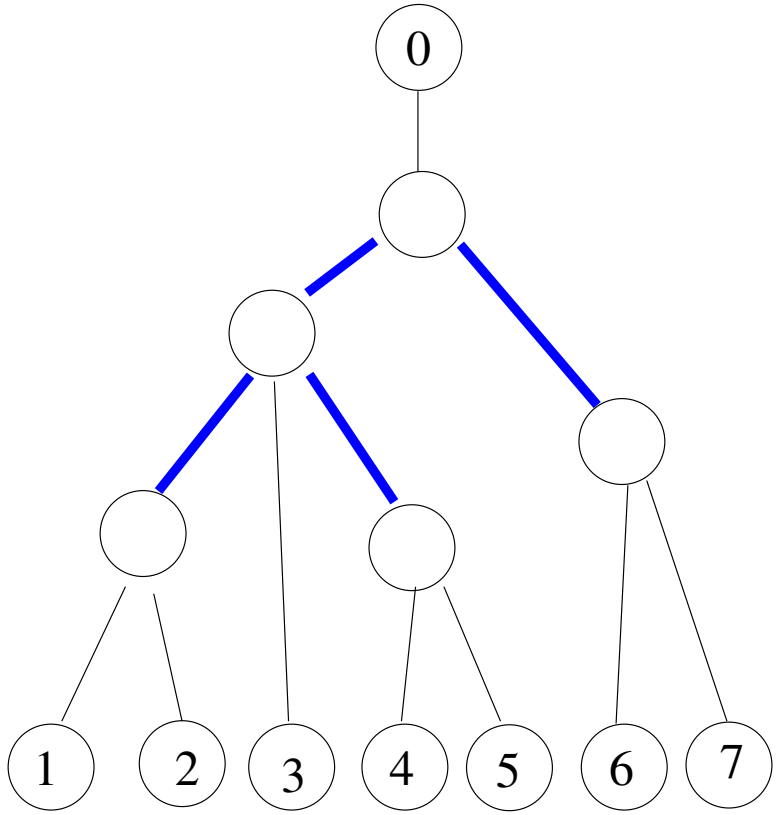
Understanding topology and orthants



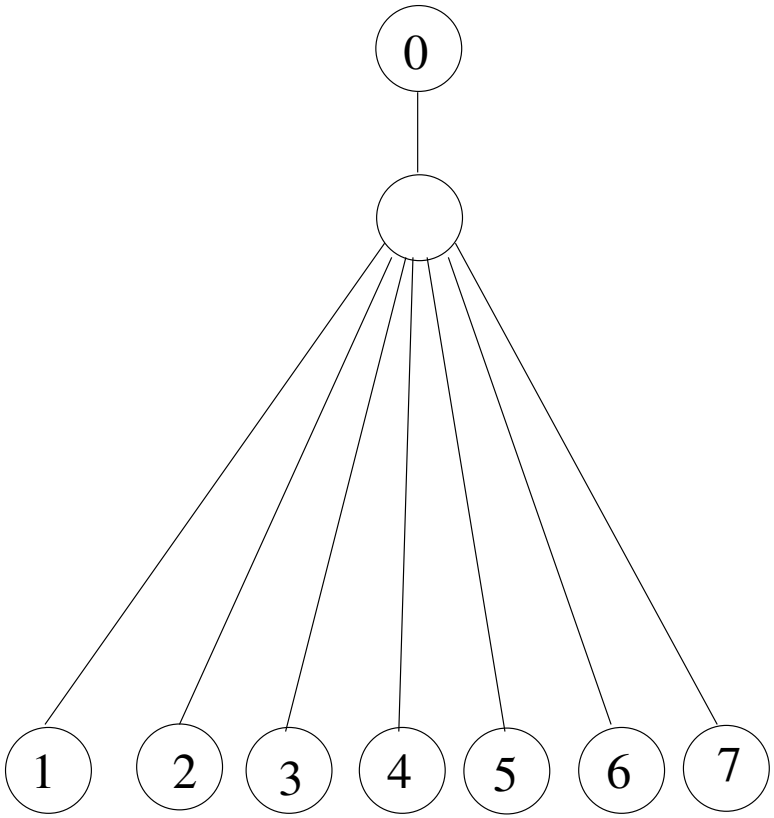


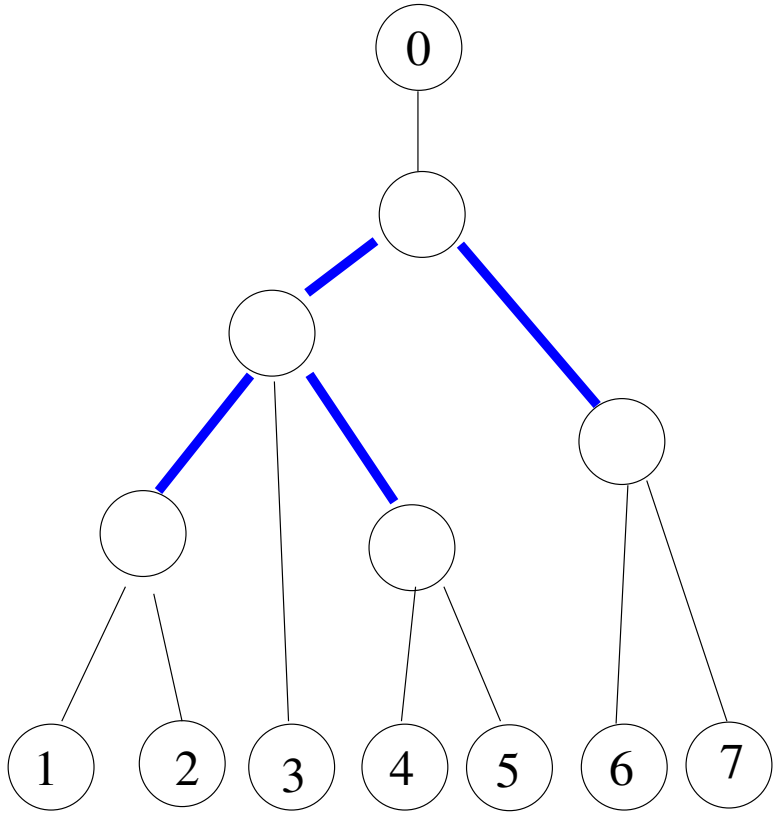
Understanding topology and orthants





Understanding topology and orthants





The geodesic distance

Definition

A sequence of trees $(T_1, T_2, \ldots, T_k)$ is a *path* in the tree space from T_1 to T_k if for each $i = 1, \ldots, k-1$ there exists an orthant $\mathcal{O}_i$ with $T_i, T_i + 1 \in \mathcal{O}_i$.

The geodesic distance

Definition

A sequence of trees $(T_1, T_2, \ldots, T_k)$ is a *path* in the tree space from T_1 to T_k if for each $i = 1, \ldots, k-1$ there exists an orthant $\mathcal{O}_i$ with $T_i, T_i + 1 \in \mathcal{O}_i$.

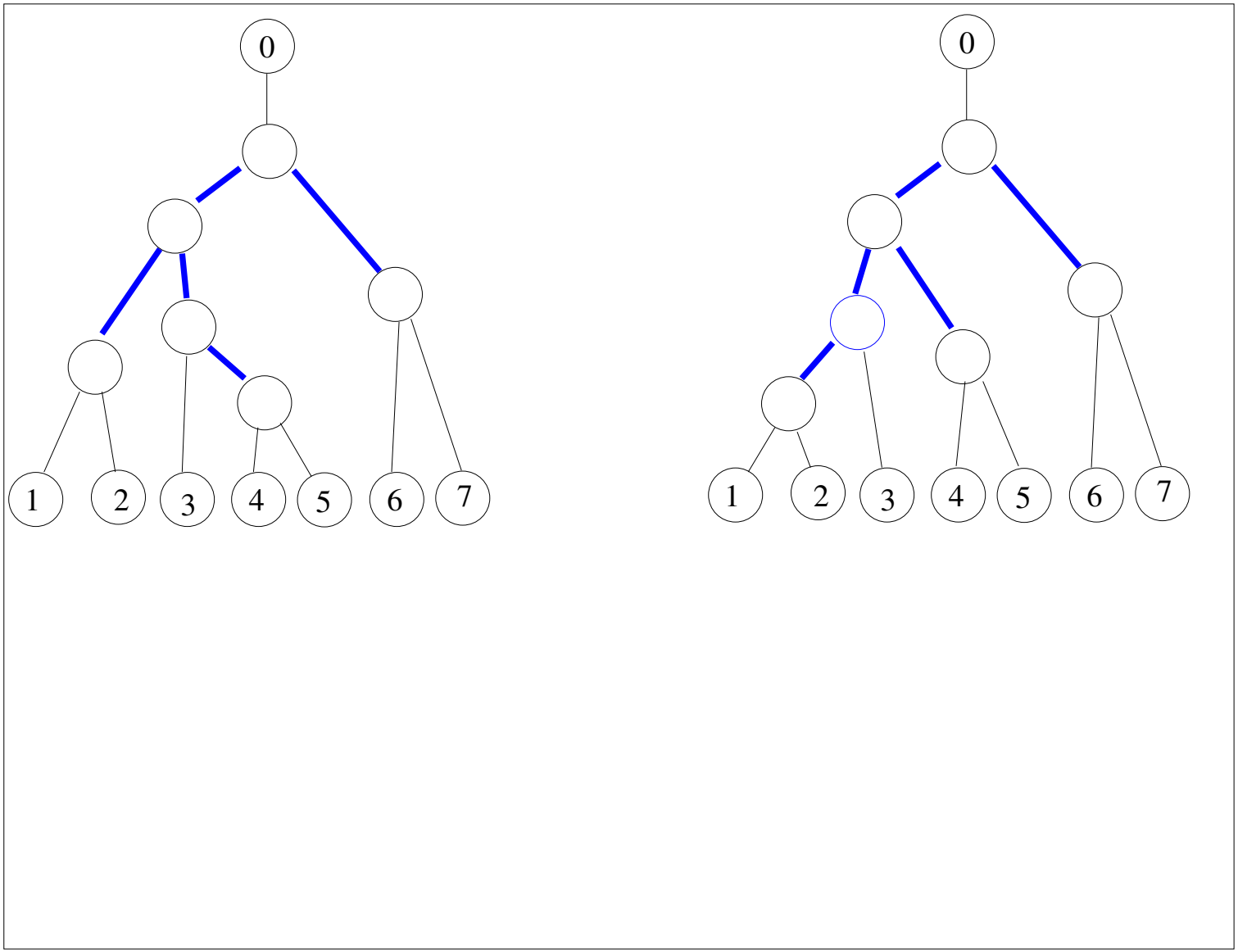
We finally get:

Definition (Geodesic distance)

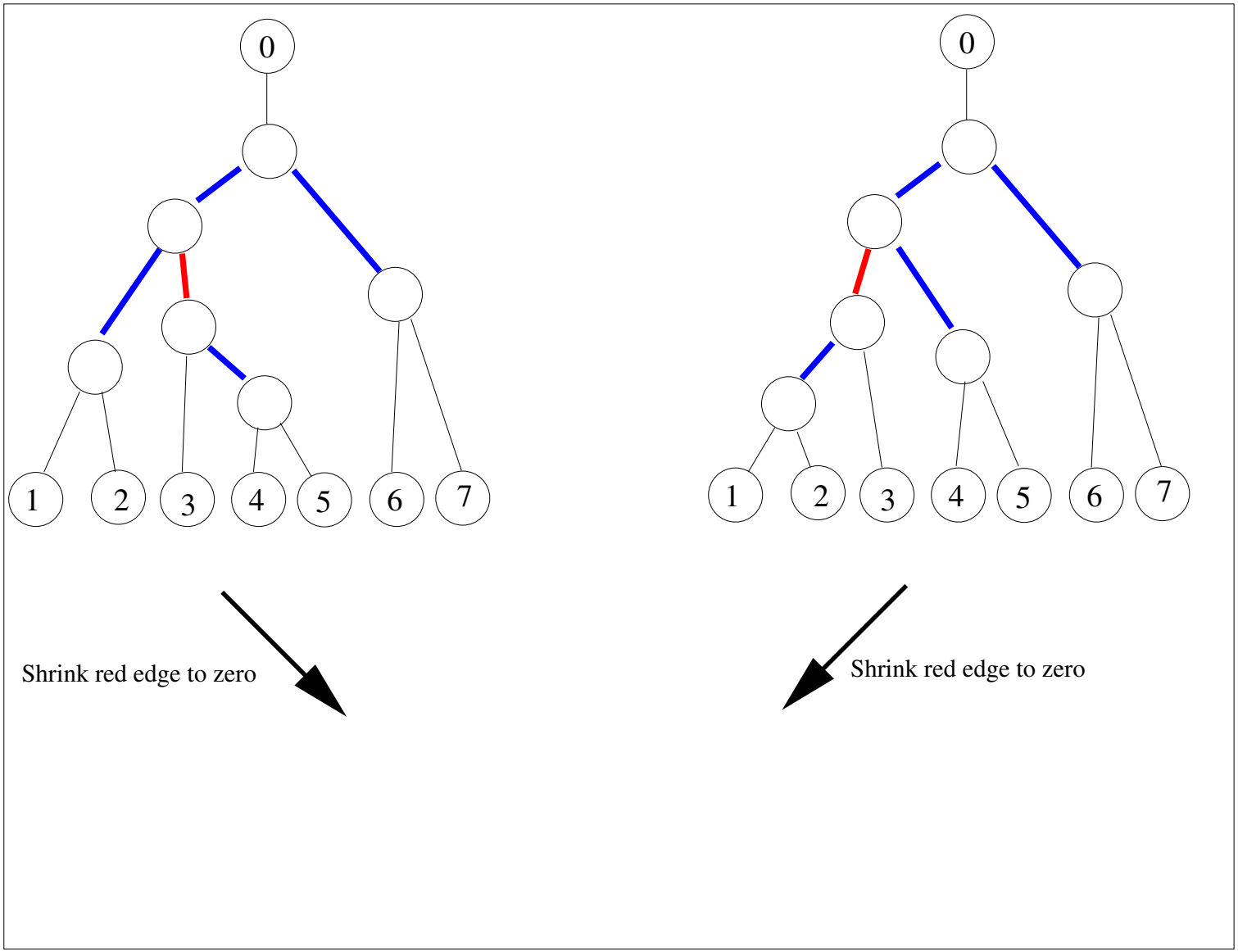
$$d(T_a, T_b) = \inf \left\{ \sum_{i=1}^{k} \| T_i - T_{i+1} \|_2 : (T_1, \ldots, T_k) \text{ is a path from } T_a \text{ to } T_b \right\}$$

Existence guaranteed since the *star tree* belongs to all orthants.

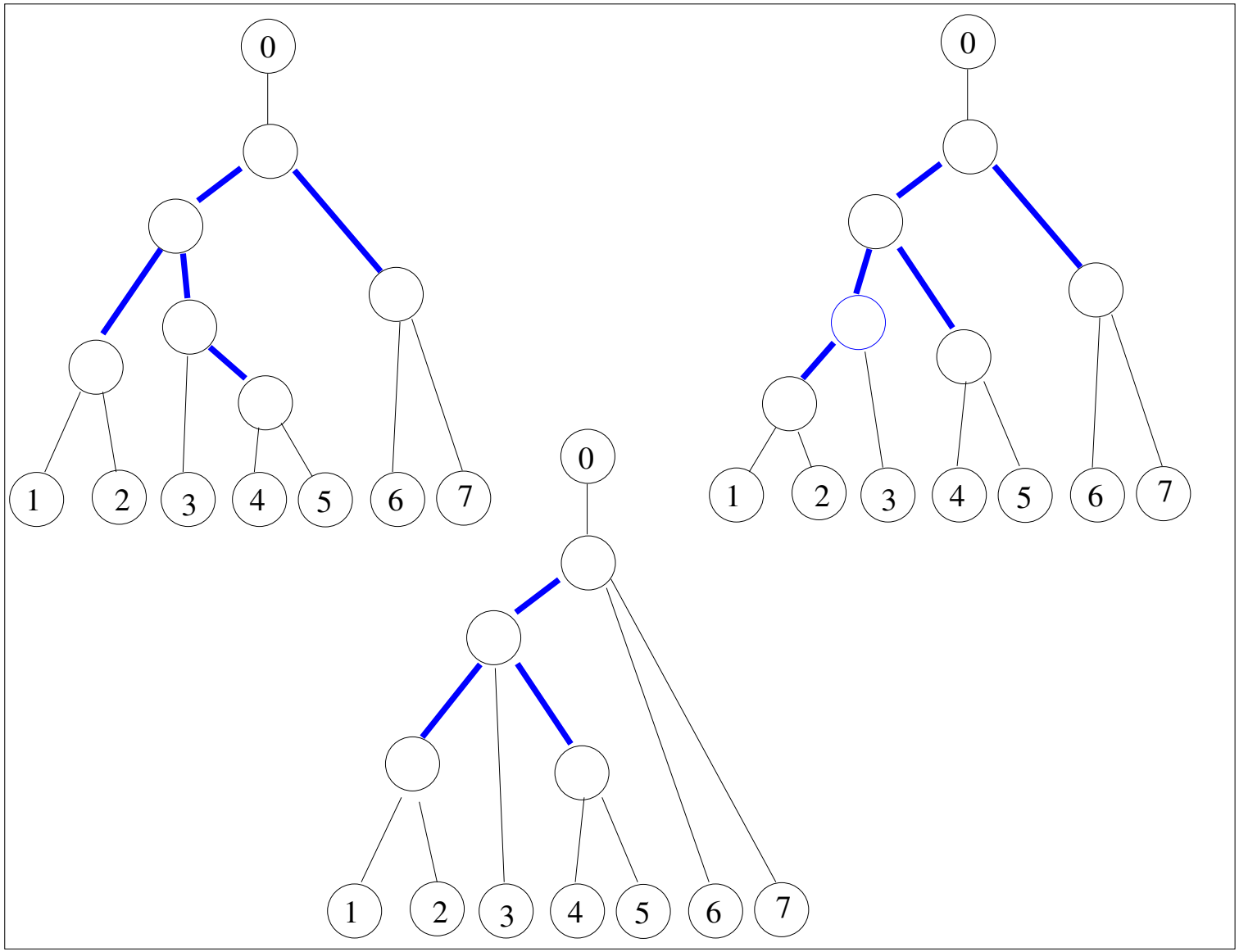
Example



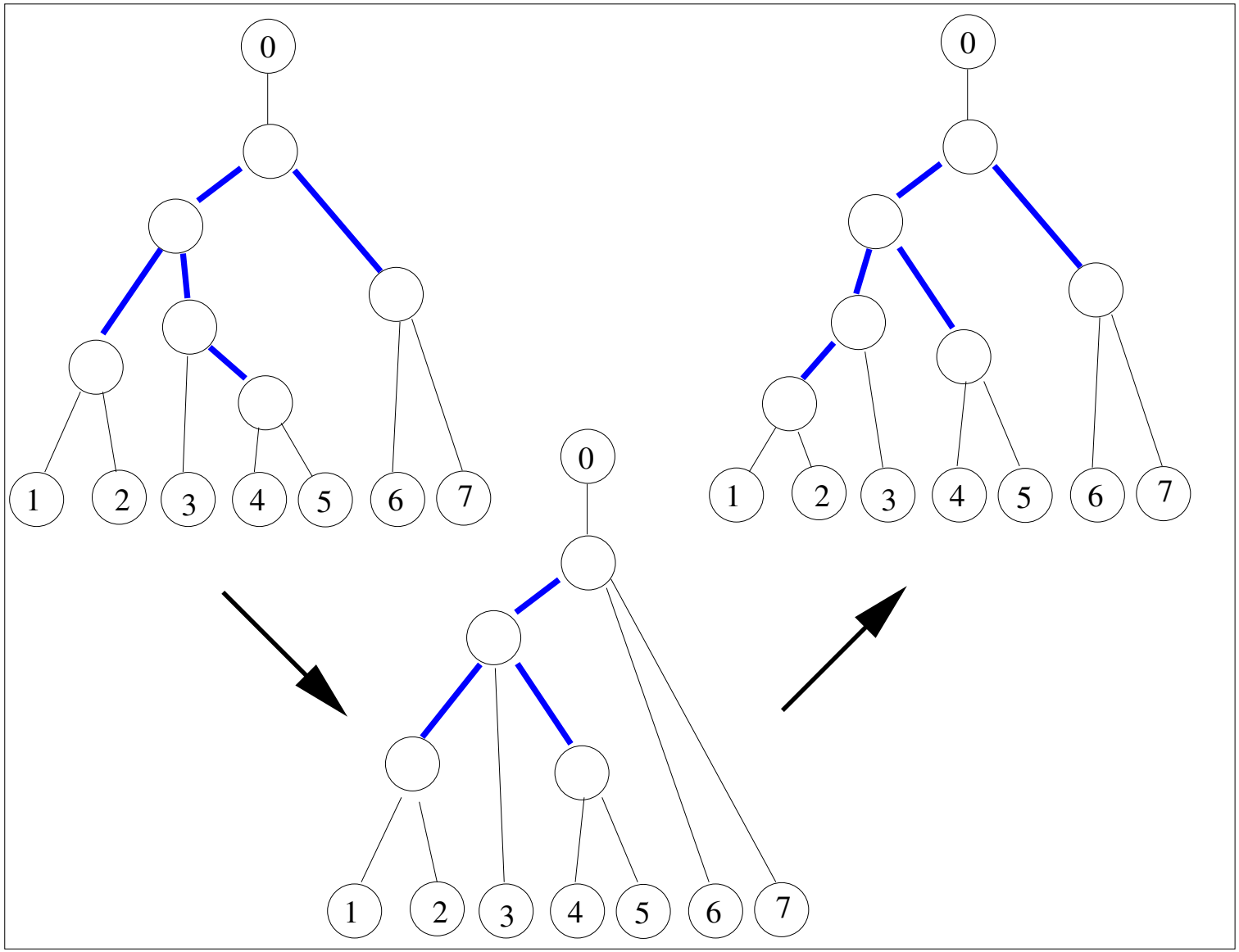
Example


0
1
2
3
4
5
6
7
Shrink red edge to zero
0
1
2
3
4
5
6
7
Shrink red edge to zero

Example



Example



The geodesic distance

Theorem (Billera, Holmes and Vogtman (2001))

The minimal path between two trees always exists and is unique.

$\Longrightarrow$ The geodesic distance is always defined.

The geodesic distance

Theorem (Billera, Holmes and Vogtman (2001))

The minimal path between two trees always exists and is unique.

$\Longrightarrow$ The geodesic distance is always defined.

Theorem (Owen (2011))

The geodesic distance can be computed in polynomial time.

Contents

A median location problem in the tree space

Given a set of phylogenetic trees, find one tree which minimizes the sum of geodesic distances to these trees.

We hope:

This is a good approximation of the tree we want.

Idea

In every orthant the geodesic distance reduces to the Euclidean distance.

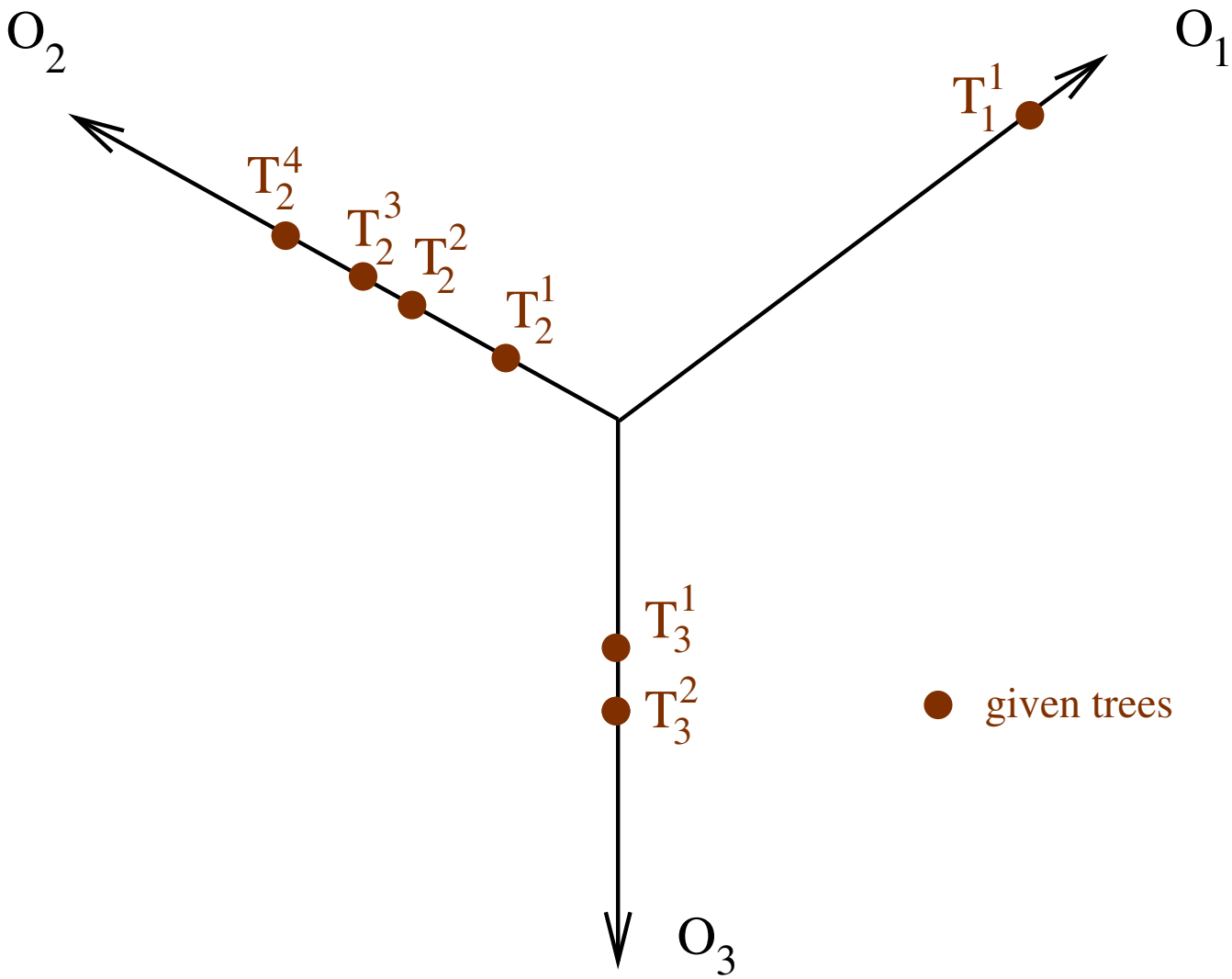
Can we use this?

Contents

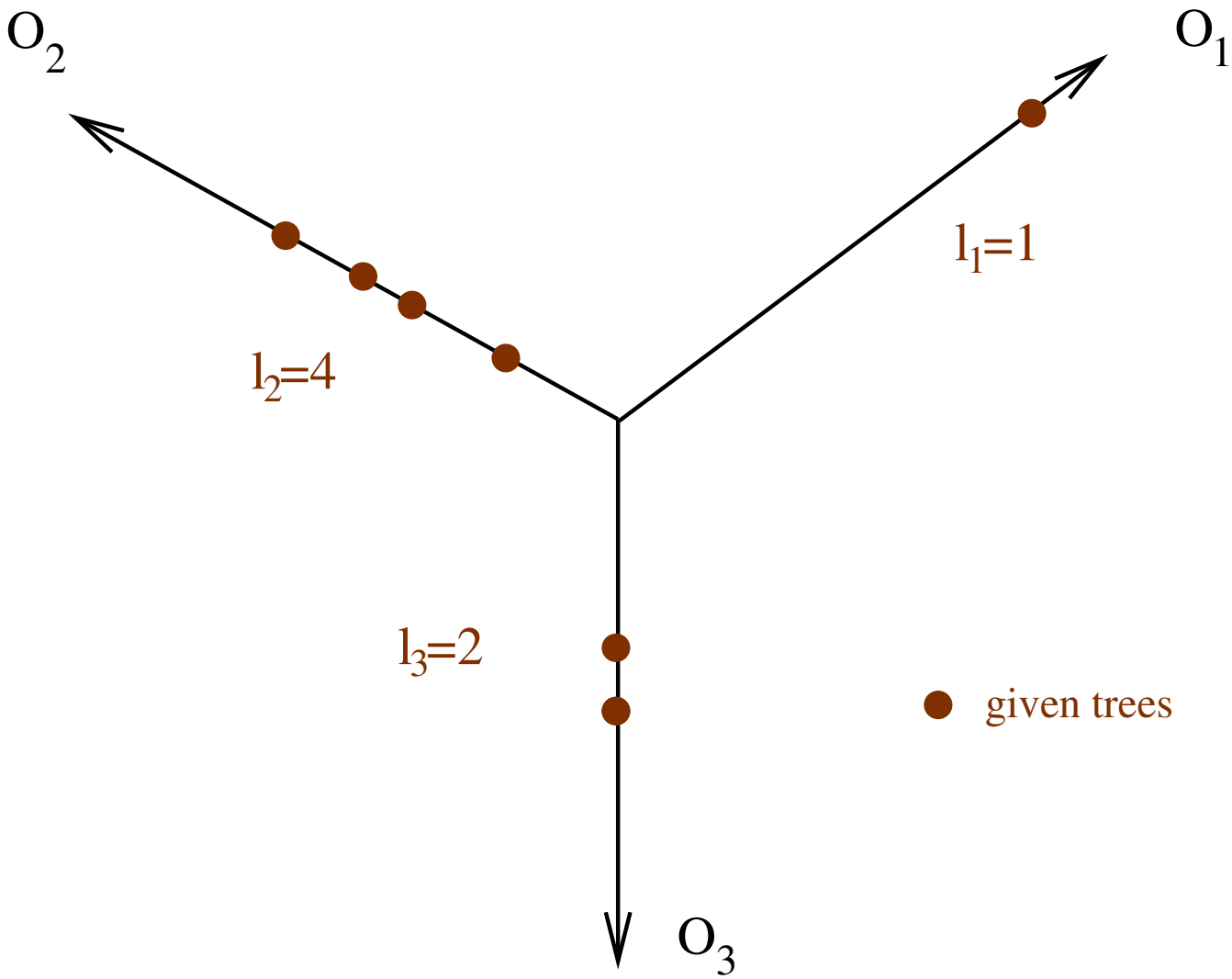
Transformation to Euclidean location problems

- Problems in $\mathcal{T}_3$
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants

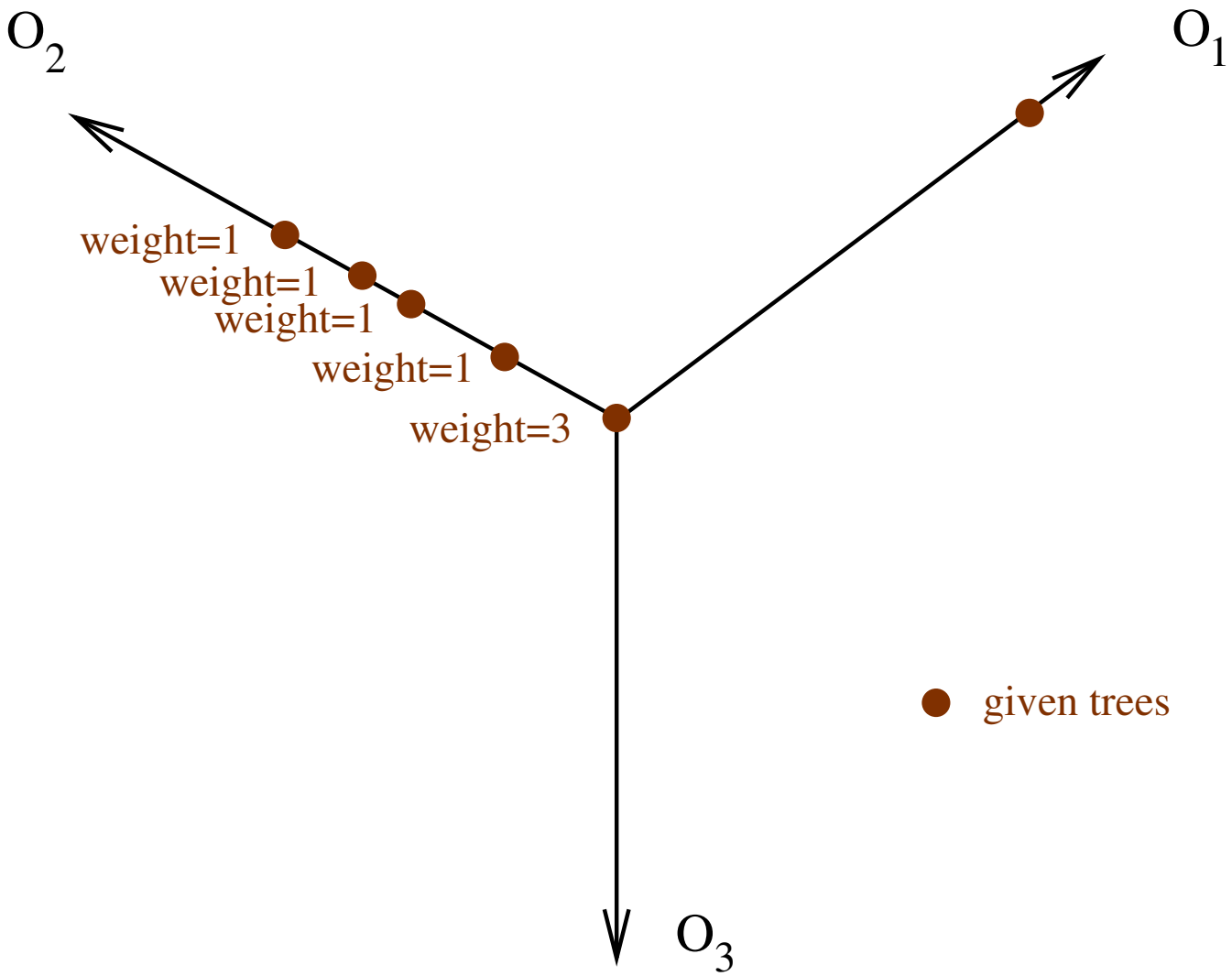
Median Location problems in $\mathcal{T}_3$ (4 orthants, 3 splits)



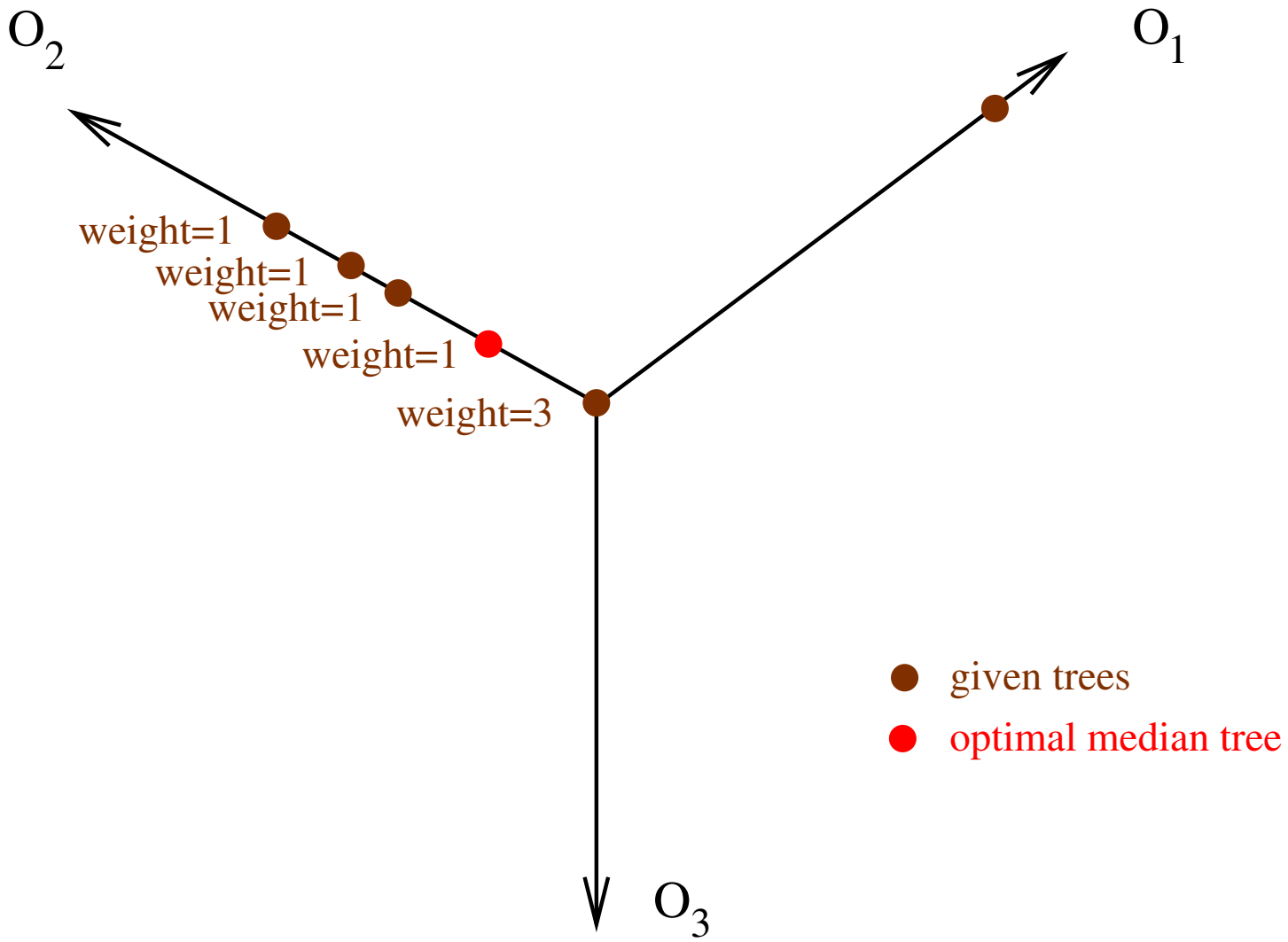
Median Location problems in $\mathcal{T}_3$ (4 orthants, 3 splits)



Median Location problems in $\mathcal{T}_3$ (4 orthants, 3 splits)



Median Location problems in $\mathcal{T}_3$ (4 orthants, 3 splits)



Transformation to Euclidean location problems

- Problems in $\mathcal{T}_3$
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants

Transformation to Euclidean location problems

- Problems in $\mathcal{T}_3$
  $\rightarrow$ Network median location problem in small tree
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants

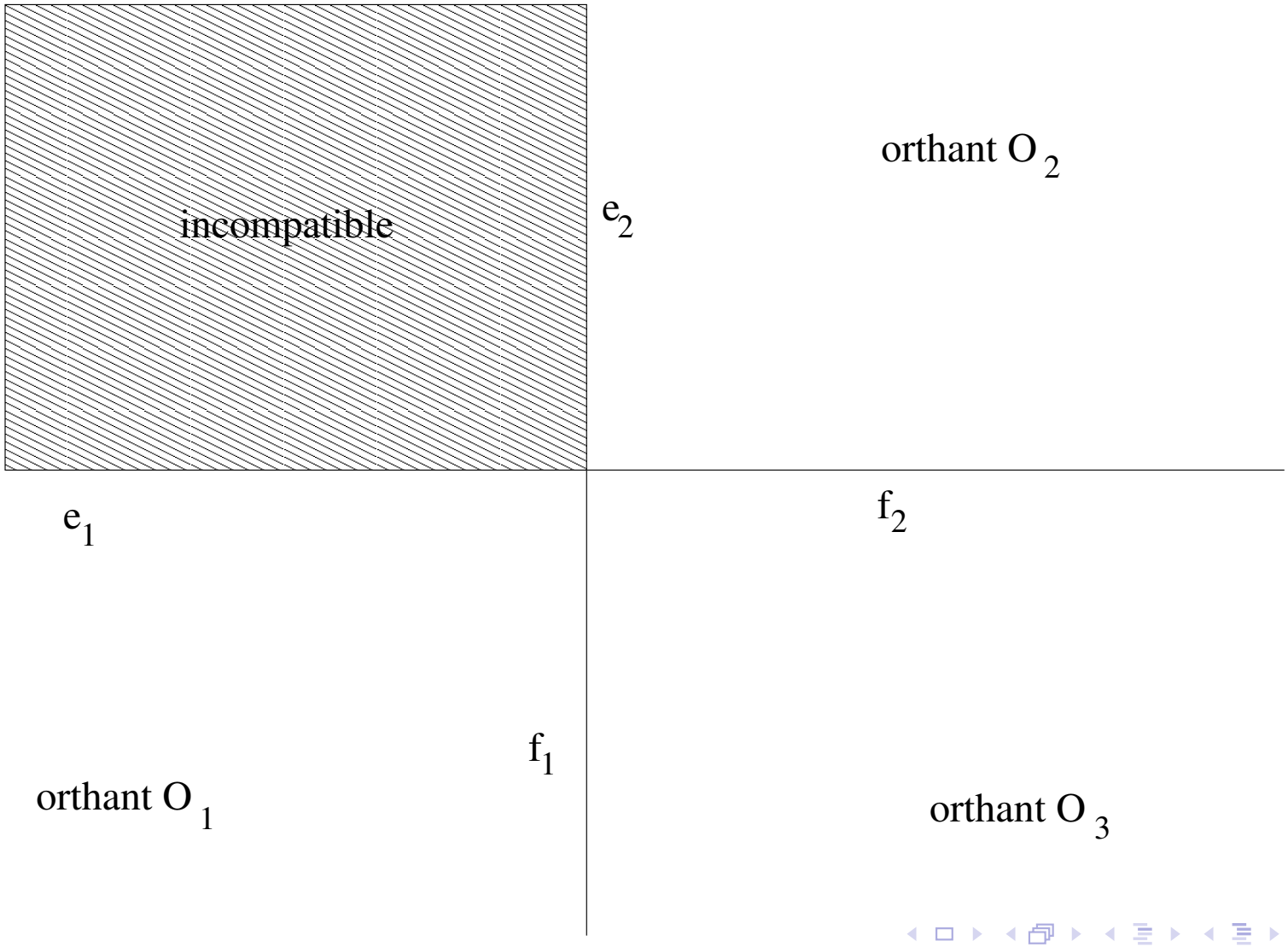
$\mathcal{T}_4$ with trees in maximal two orthants (26 orthants, 10 splits)

1. One common split: neighboring orthants
2. One pair of compatible splits
3. Completely incompatible orthants

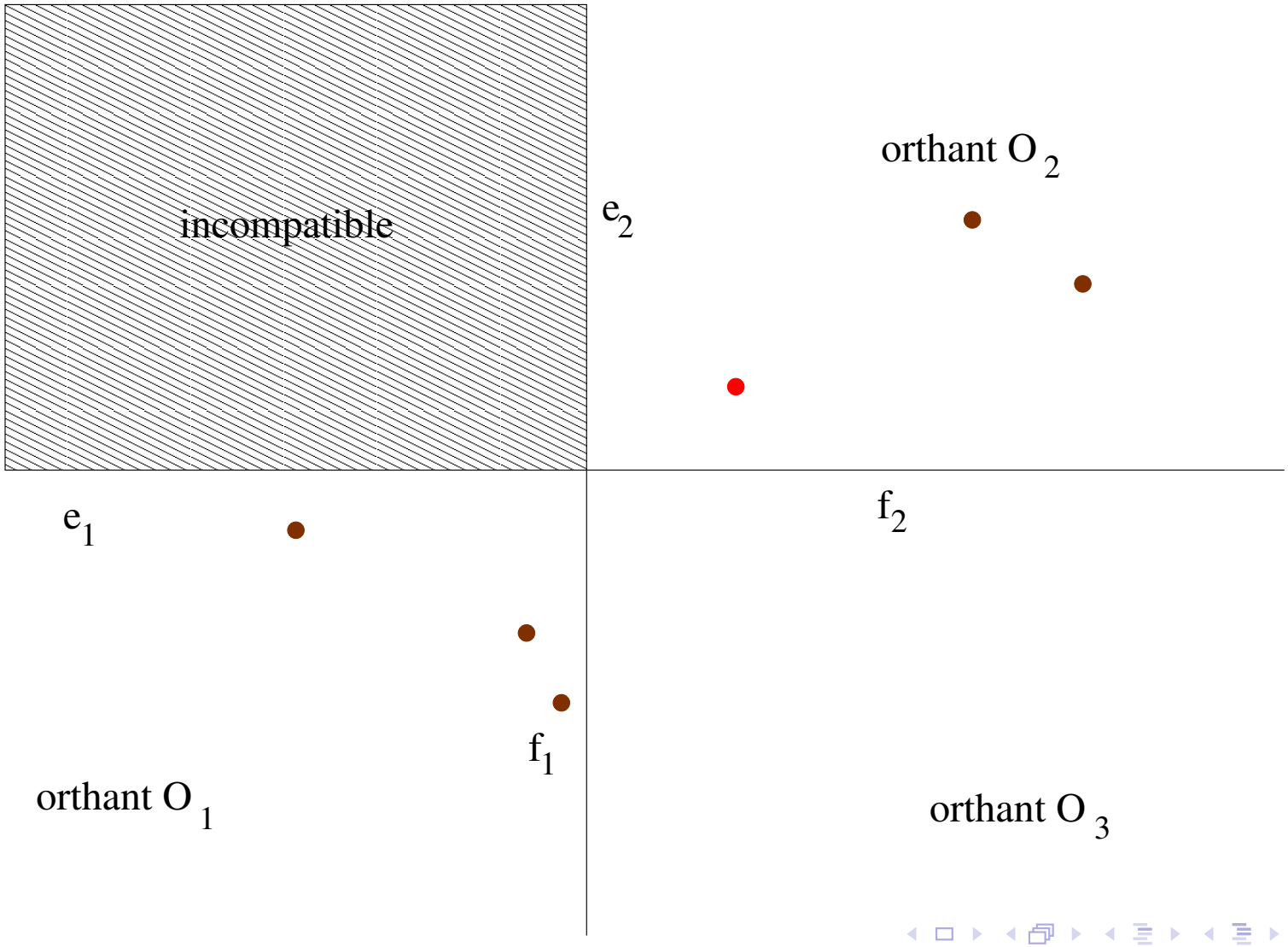
Theorem

All three cases can be modeled and solved as planar location problems.

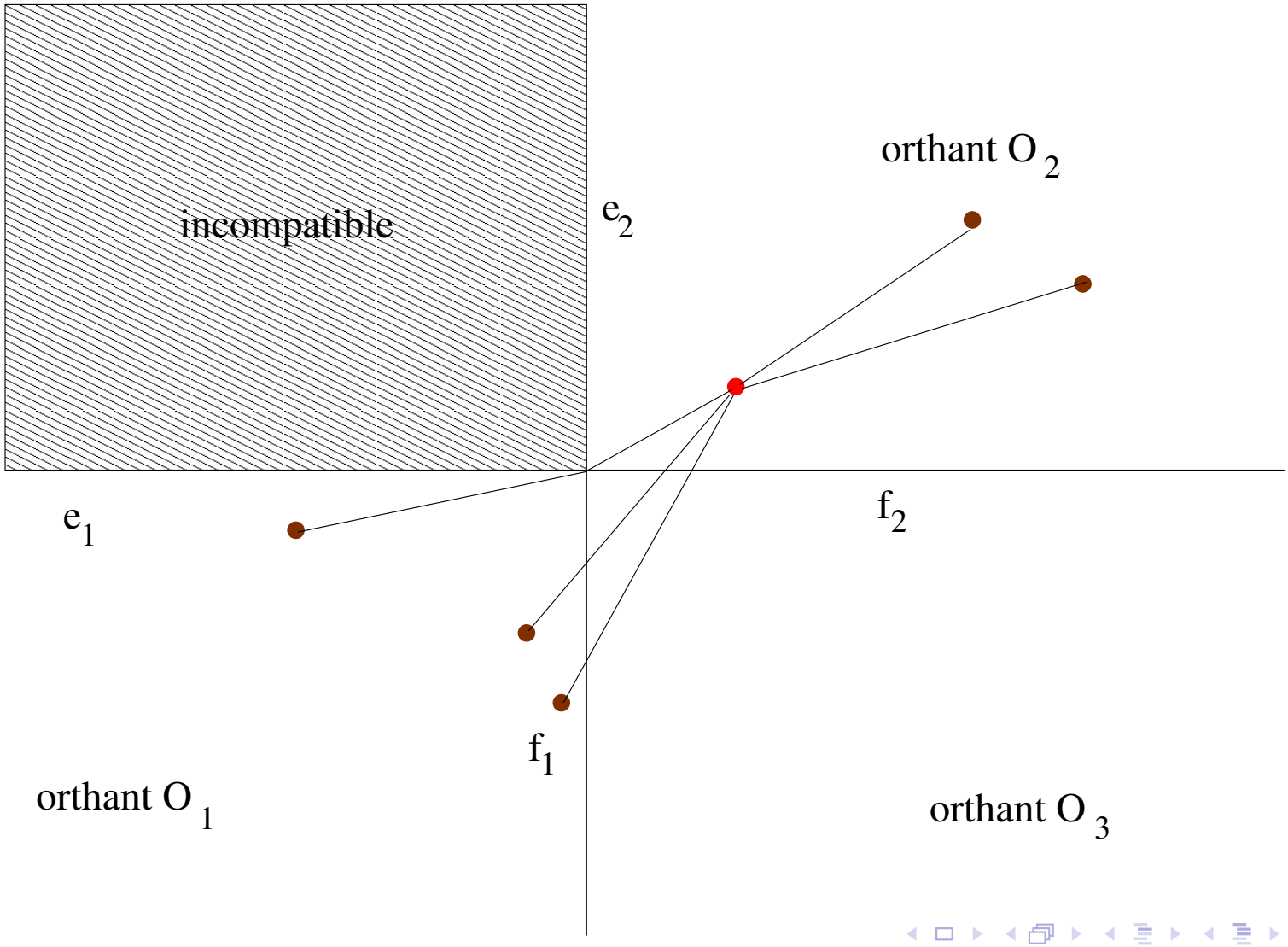
$\mathcal{T}_4$ with given trees in two orthants with exactly one compatible pair of splits



$\mathcal{T}_4$ with given trees in two orthants with exactly one compatible pair of splits



$\mathcal{T}_4$ with given trees in two orthants with exactly one compatible pair of splits



Transformation to Euclidean location problems

- Problems in $\mathcal{T}_3$
  $\rightarrow$ Network median location problem in small tree
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants

Transformation to Euclidean location problems

- Problems in $\mathcal{T}_3$
  $\rightarrow$ Network median location problem in small tree
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
  $\rightarrow$ Euclidean planar median location problem with barriers
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants

Transformation to Euclidean location problems

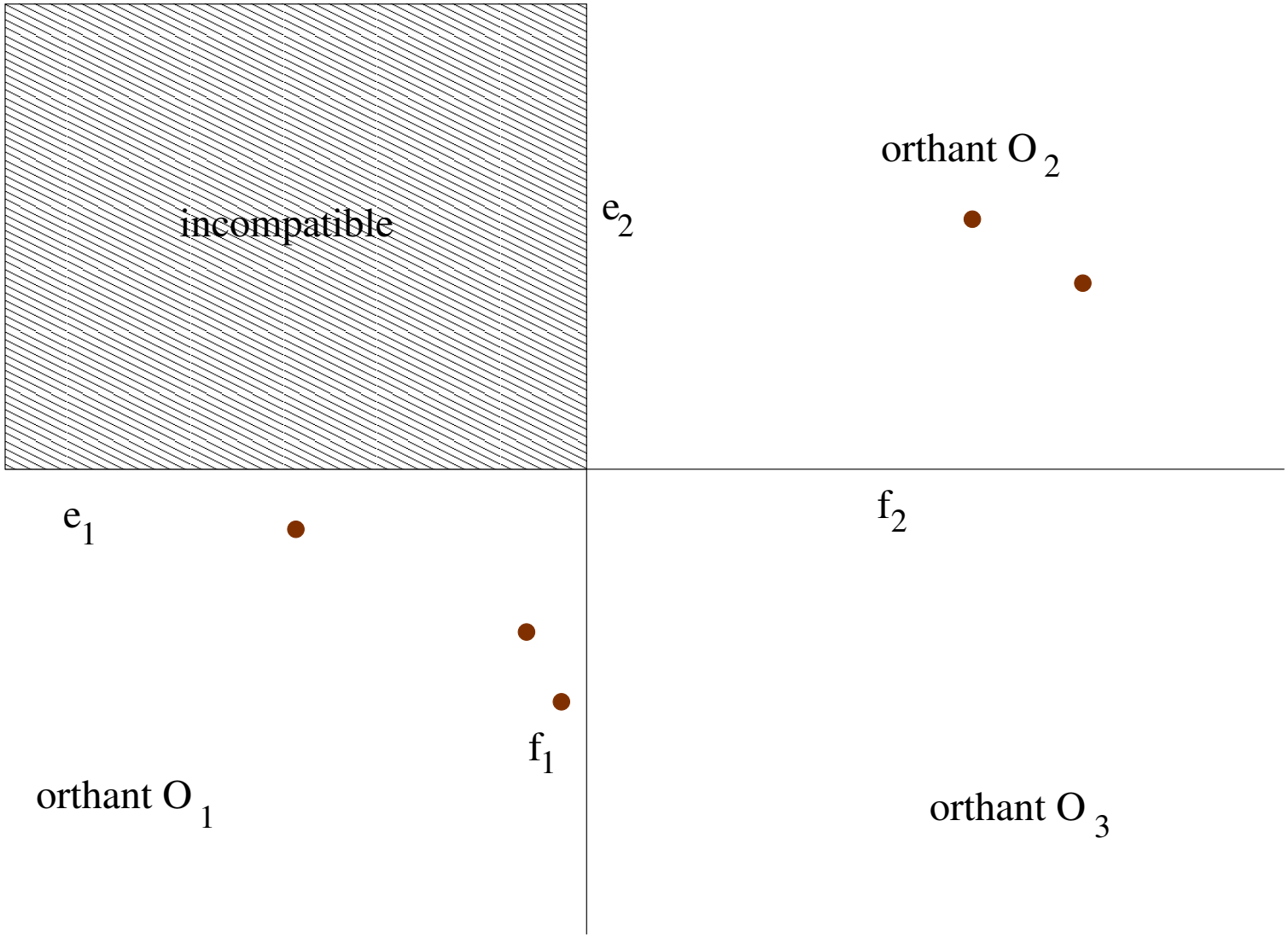
- Problems in $\mathcal{T}_3$
  $\rightarrow$ Network median location problem in small tree
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
  $\rightarrow$ Euclidean planar median location problem with barriers
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
  $\rightarrow$ Euclidean median location problem with barriers
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants

Transformation to Euclidean location problems

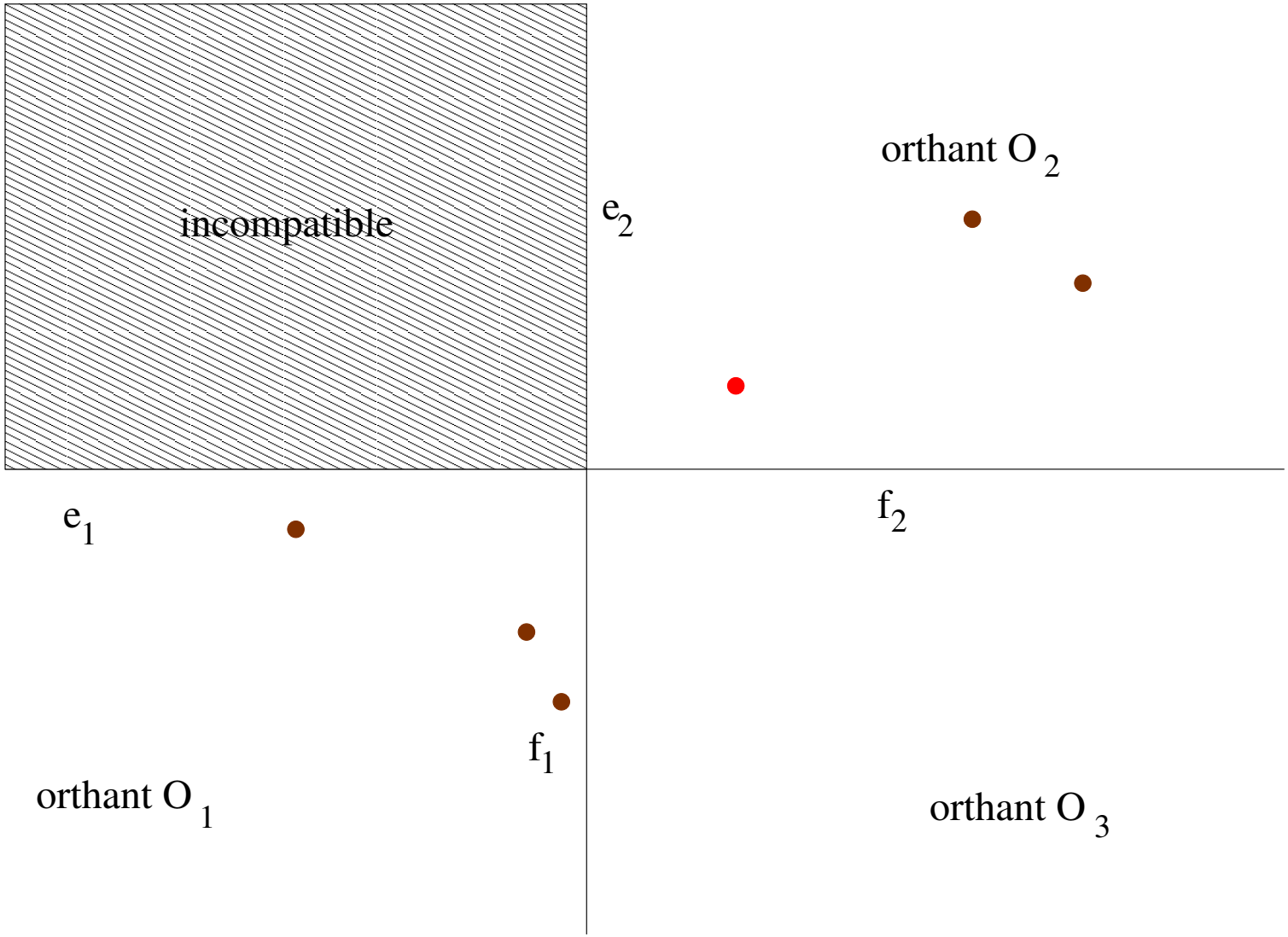
- Problems in $\mathcal{T}_3$
  $\rightarrow$ Network median location problem in small tree
- Problems in $\mathcal{T}_4$ if existing trees are in maximal two orthants
  $\rightarrow$ Euclidean planar median location problem with barriers
- Problems in $\mathcal{T}_n$ if existing trees are in maximal two orthants with special structure
  $\rightarrow$ Euclidean median location problem with barriers
- Problems in $\mathcal{T}_n$ if all existing trees are in completely incompatible orthants
  $\rightarrow$ Euclidean median location problem with a gate point

Contents

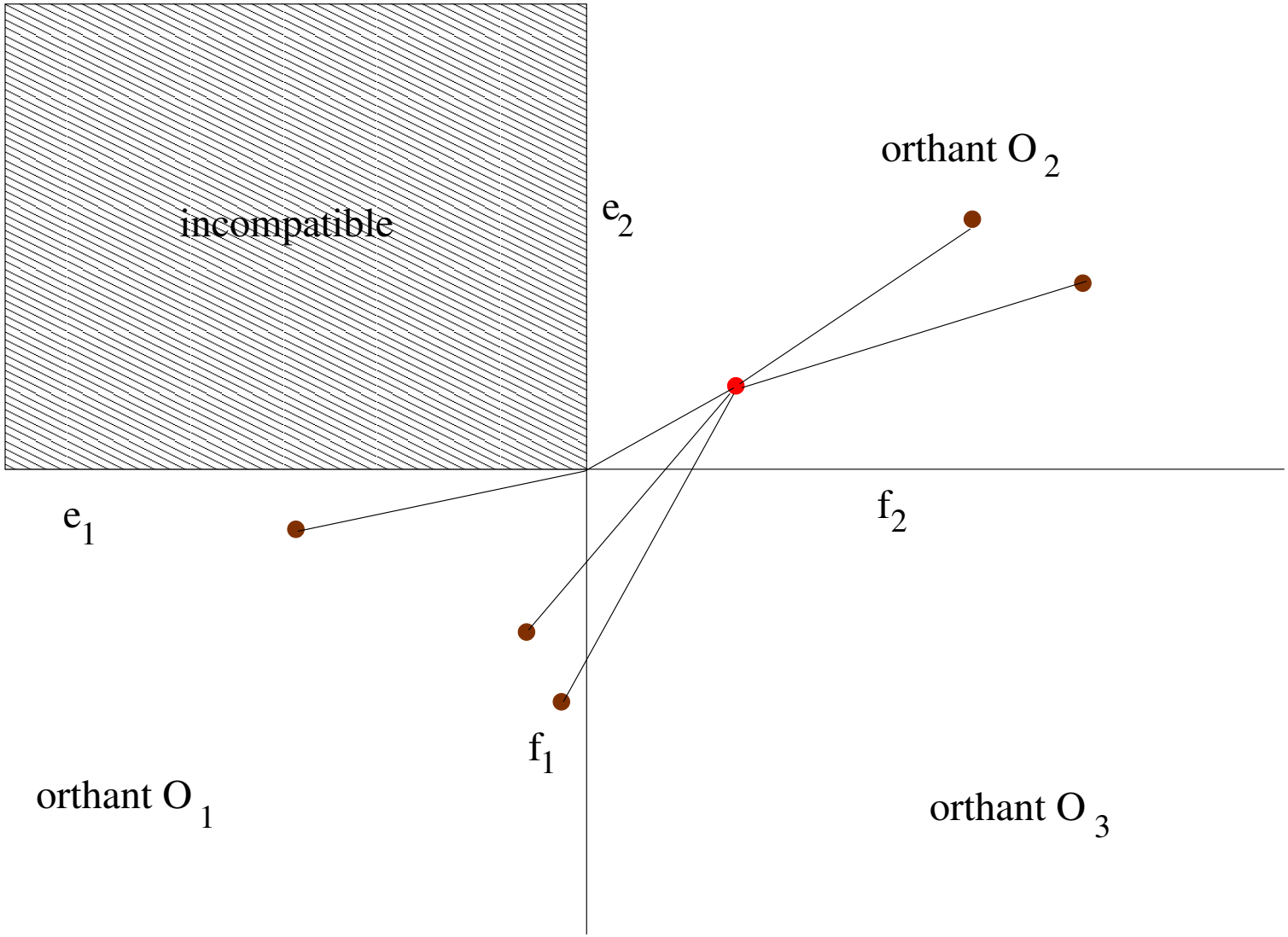
Balance Point Algorithm



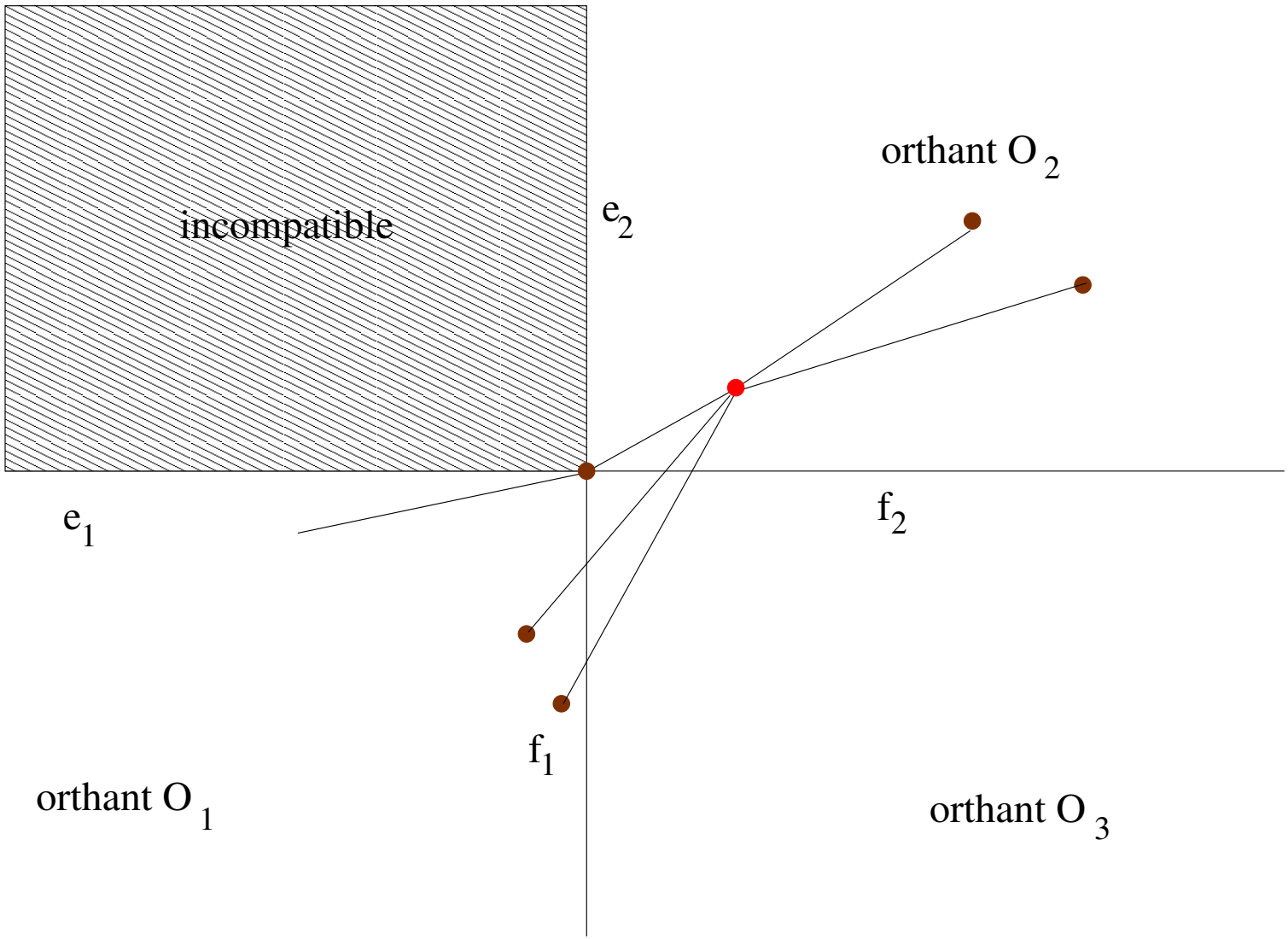
Balance Point Algorithm



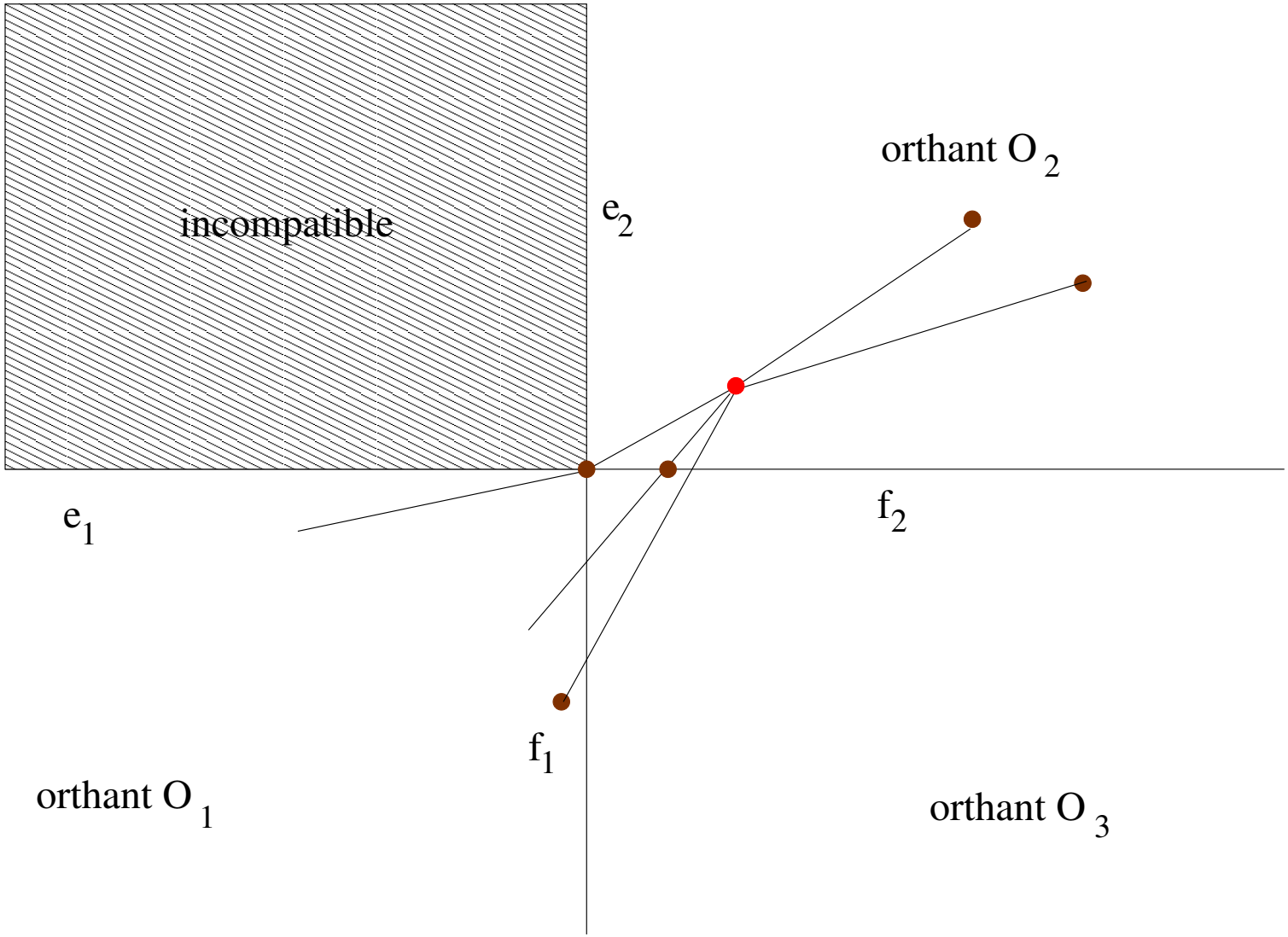
Balance Point Algorithm



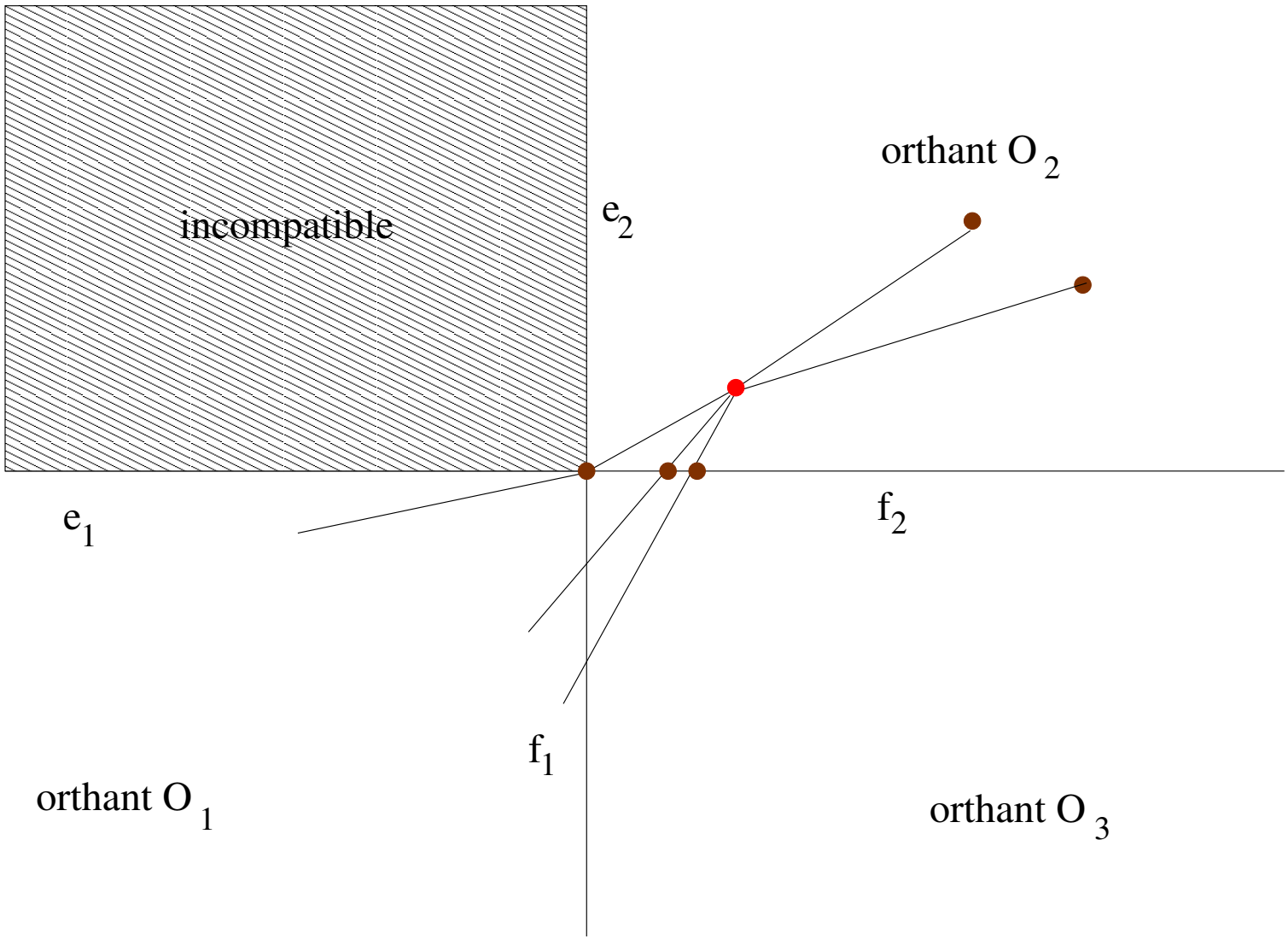
Balance Point Algorithm



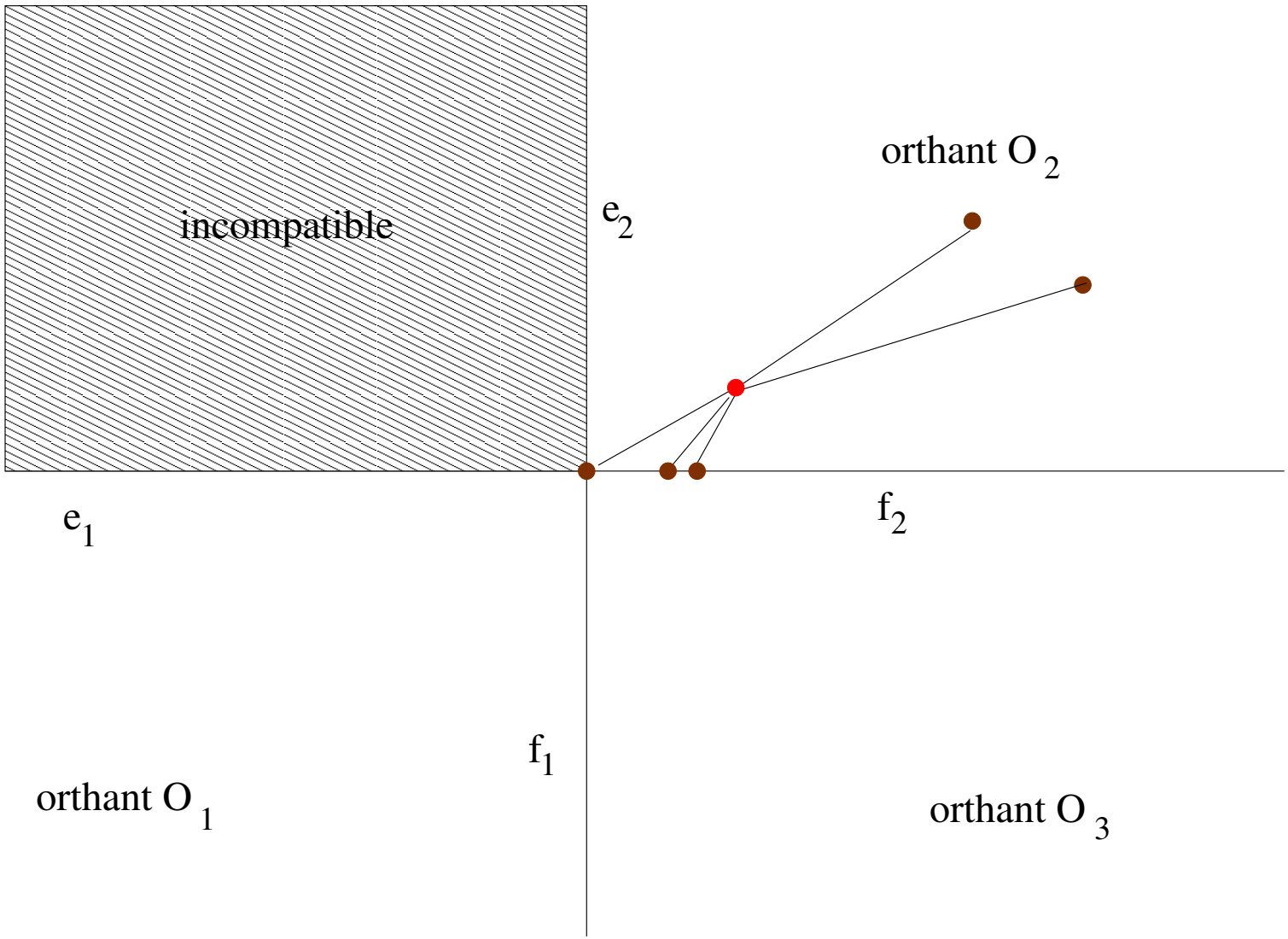
Balance Point Algorithm



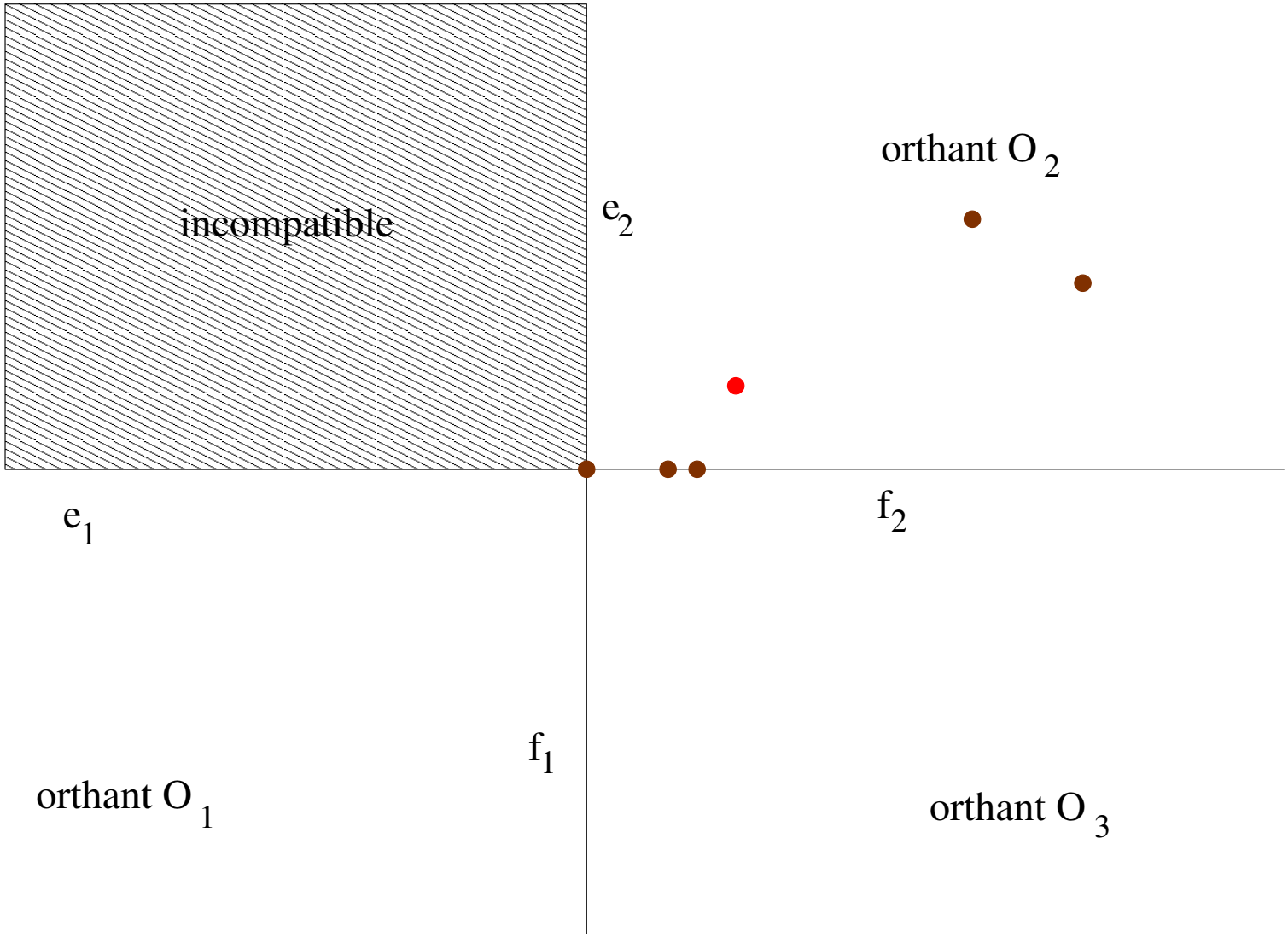
Balance Point Algorithm



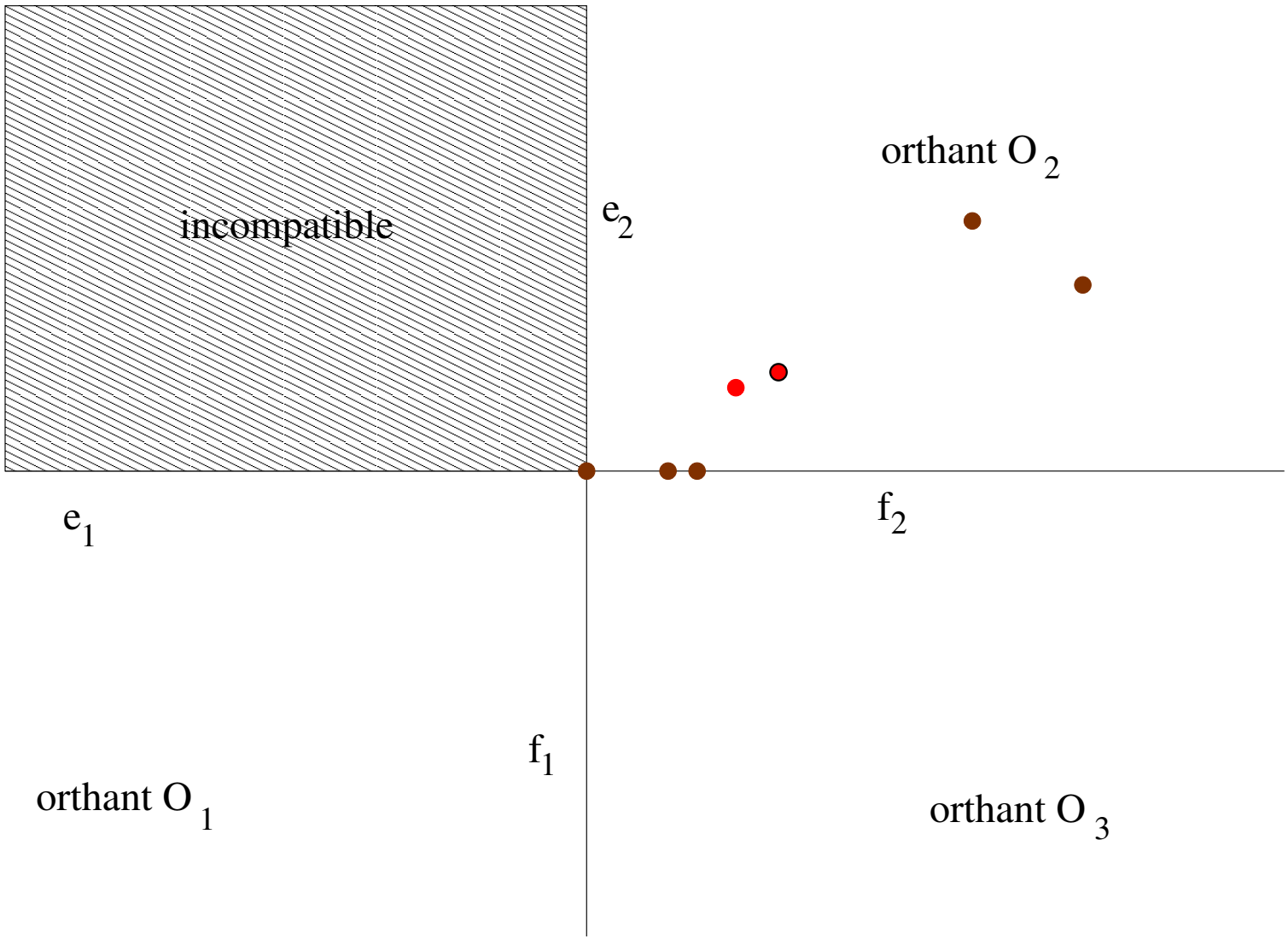
Balance Point Algorithm



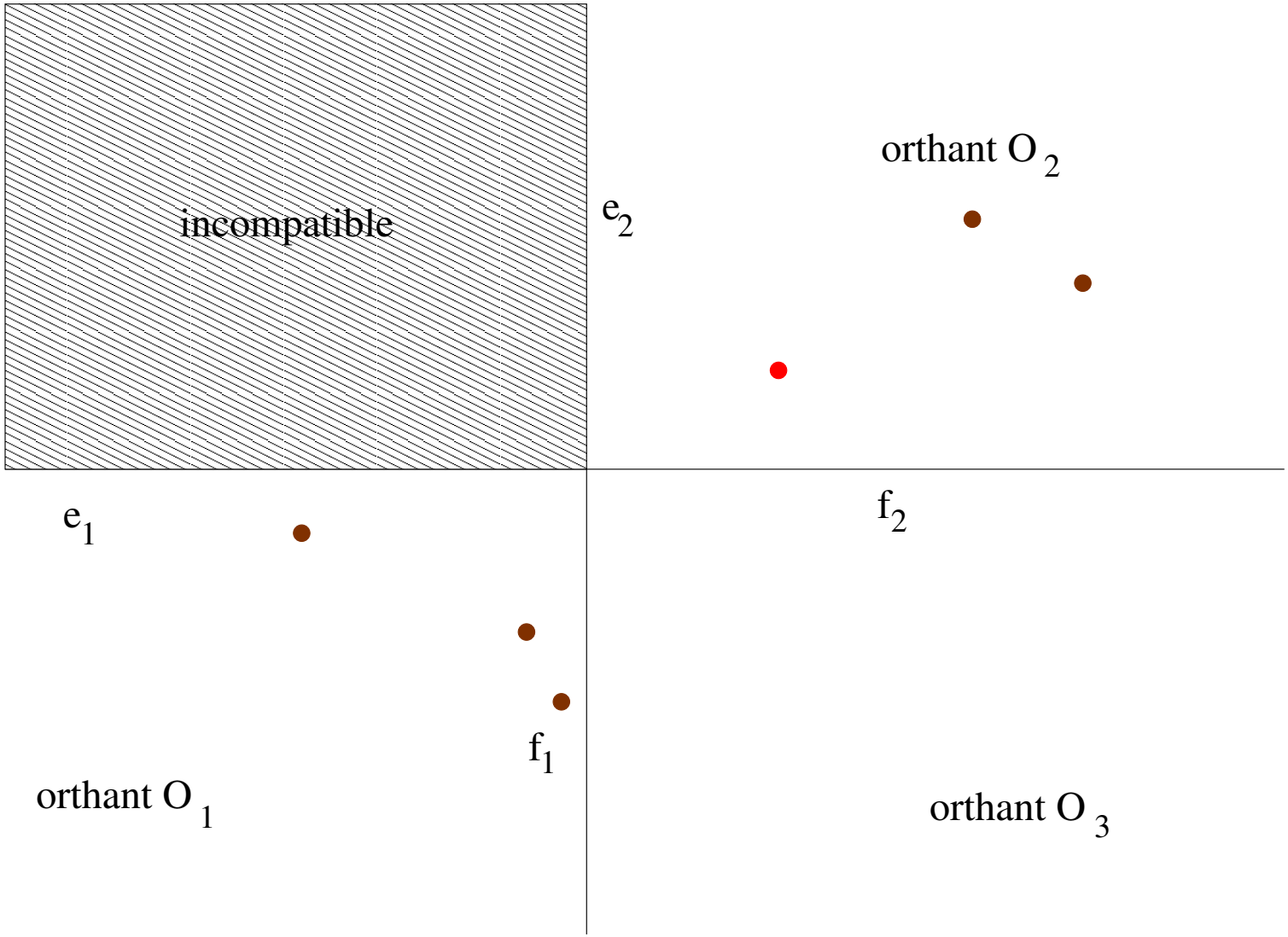
Balance Point Algorithm



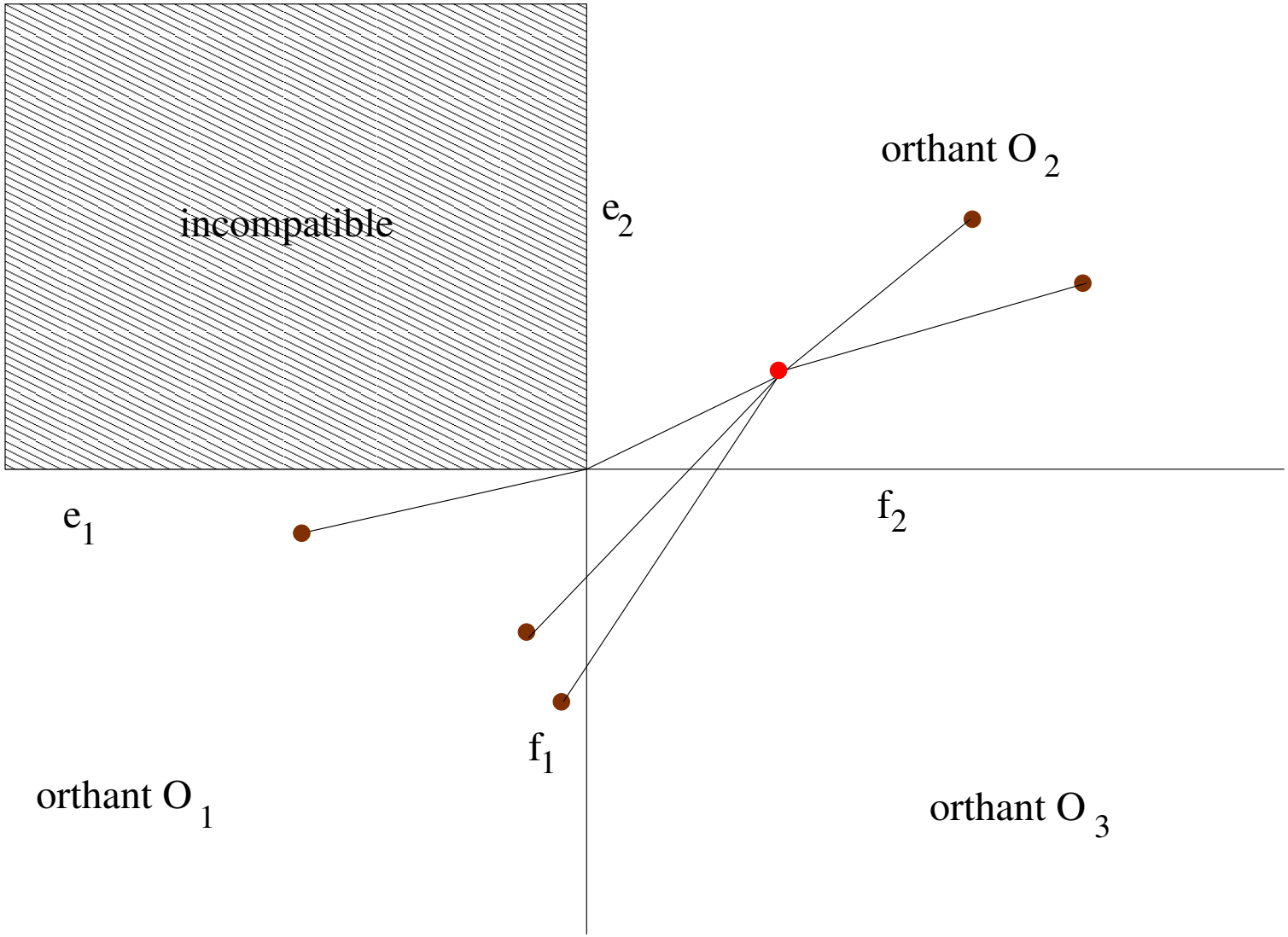
Balance Point Algorithm



Balance Point Algorithm



Balance Point Algorithm



Balance Point Algorithm

2. For each orthant $\mathcal{O}$ do
3. *Inner Loop*
   1. Start with some tree T in the interior of $\mathcal{O}$.
   2. Compute geodesics from T to all other trees together with entry points into orthant $\mathcal{O}$.
   3. Solve location problem within the orthant and get new tree T.
   4. Goto step 2 until no improvement.
4. Take best solution.

Balance Point Algorithm

1. Remove orthants by bounds
2. For each remaining orthant $\mathcal{O}$ do
3. *Inner Loop*
   1. Start with some tree T in the interior of $\mathcal{O}$.
   2. Compute geodesics from T to all other trees together with entry points into orthant $\mathcal{O}$.
   3. Solve location problem within the orthant and get new tree T.
   4. Goto step 2 until no improvement.
4. Take best solution.

Bounds

$$d(T_a, T_b) = \inf \left\{ \sum_{i=1}^{k} \| T_i - T_{i+1} \|_2 : (T_1, \ldots, T_k) \text{ is a path from } T_a \text{ to } T_b \right\}$$

Lower bound for every orthant:

- Objective of Euclidean median problem in $\mathbb{R}^N$ is a lower bound
- . . . can be improved by using structure of orthants.

Bounds

$$d(T_a, T_b) = \inf \left\{ \sum_{i=1}^{k} \| T_i - T_{i+1} \|_2 : (T_1, \ldots, T_k) \text{ is a path from } T_a \text{ to } T_b \right\}$$

Lower bound for every orthant:

- Objective of Euclidean median problem in $\mathbb{R}^N$ is a lower bound
- . . . can be improved by using structure of orthants.

Global upper bound: Take the star tree

$$UB_{star} = \sum_{\text{existing trees } T} \| T \|$$

or any other feasible solution

Theoretical results

The balance point algorithm

- is a blockwise coordinate descent method
- improves the objective value in every step (or stays the same)
- both steps are convex
  (but convexity is tricky in the tree space!)
- both steps need not find unique solutions and may be non-differentiable
- STILL: convergence to an optimal solution is provable (under a few conditions)!
  $\rightarrow$ dissertation of Marco Botte, 2019

Numerical results

- The balance point algorithm works (many experiments in low dimensions) and a real-data case study (Apicomplexa, 252 existing trees, 8 species)
- The identified trees are what biologist expect.

Comparison with the proximal point method by Baćak (2014):

- The balance point algorithm converges much quicker when close to optimal solution. (Proximal point method is not monotone.)
- The proximal point method can identify a good orthant.
- Best: starting with the proximal point method and then switching to the new approach.

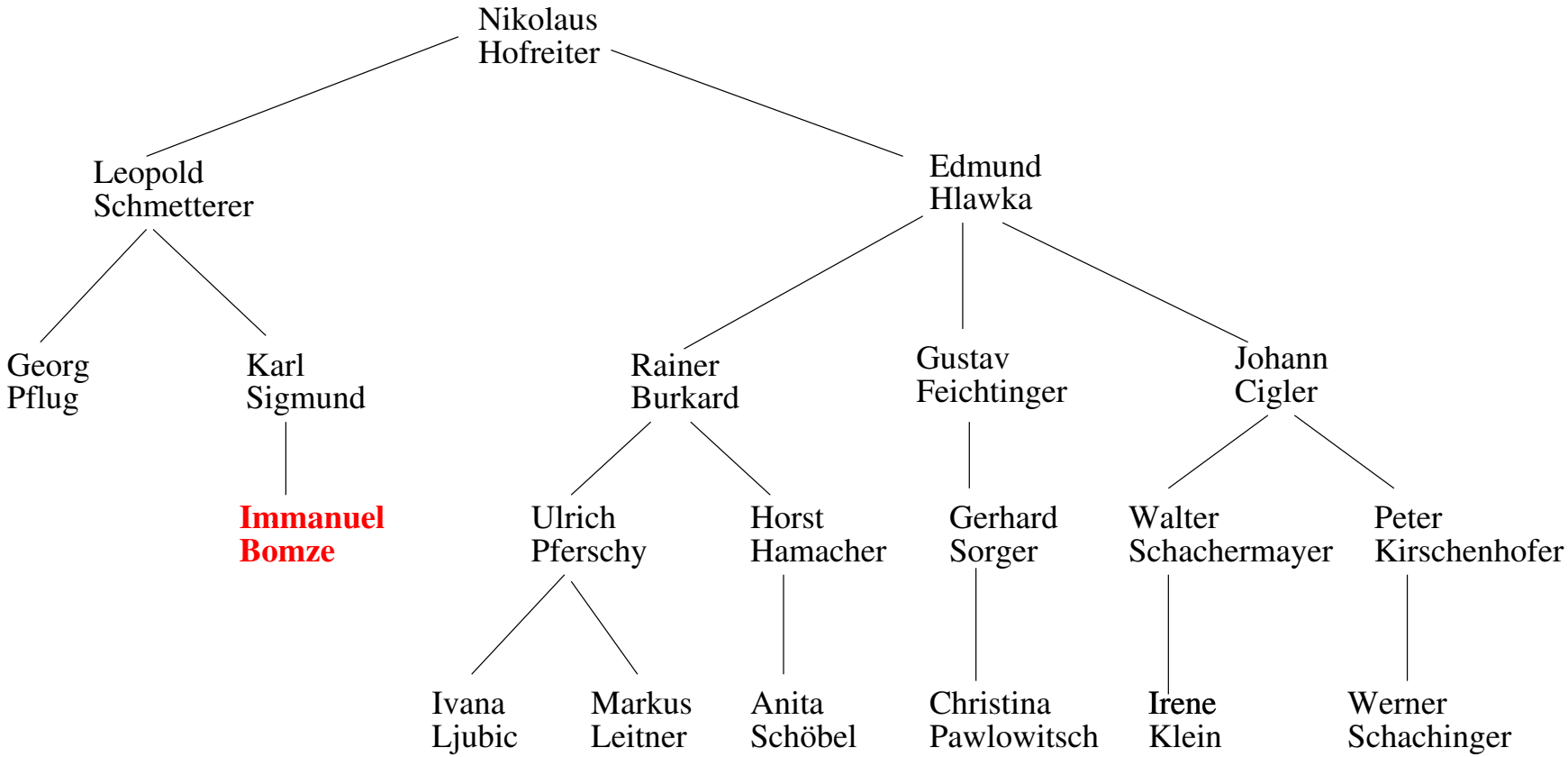
$\rightarrow$ dissertation of Marco Botte, 2019

Contents

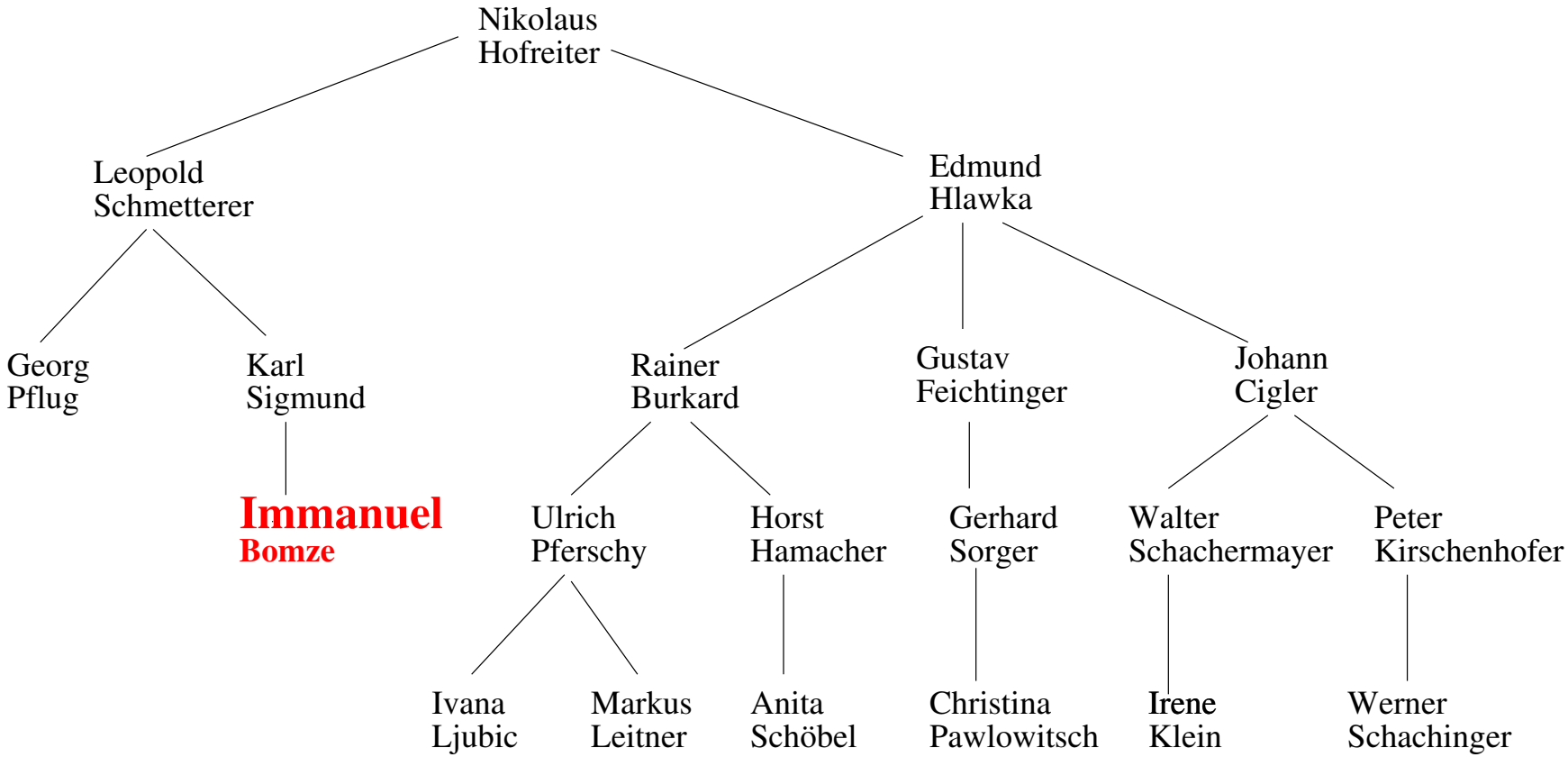
Conclusion and Outlook

- What you could have seen:
  - Finding phylogenetic trees can be modeled as a location problem in an interesting metric space
  - Methods of location theory seem appropriate for solving the problem.
- What is still to come:
  - identify more cases which can be transferred to Euclidean location problems
  - work on (high-dimensional) Euclidean location problems with barriers
  - improve the balance point algorithm, e.g., better bounds
  - more case studies

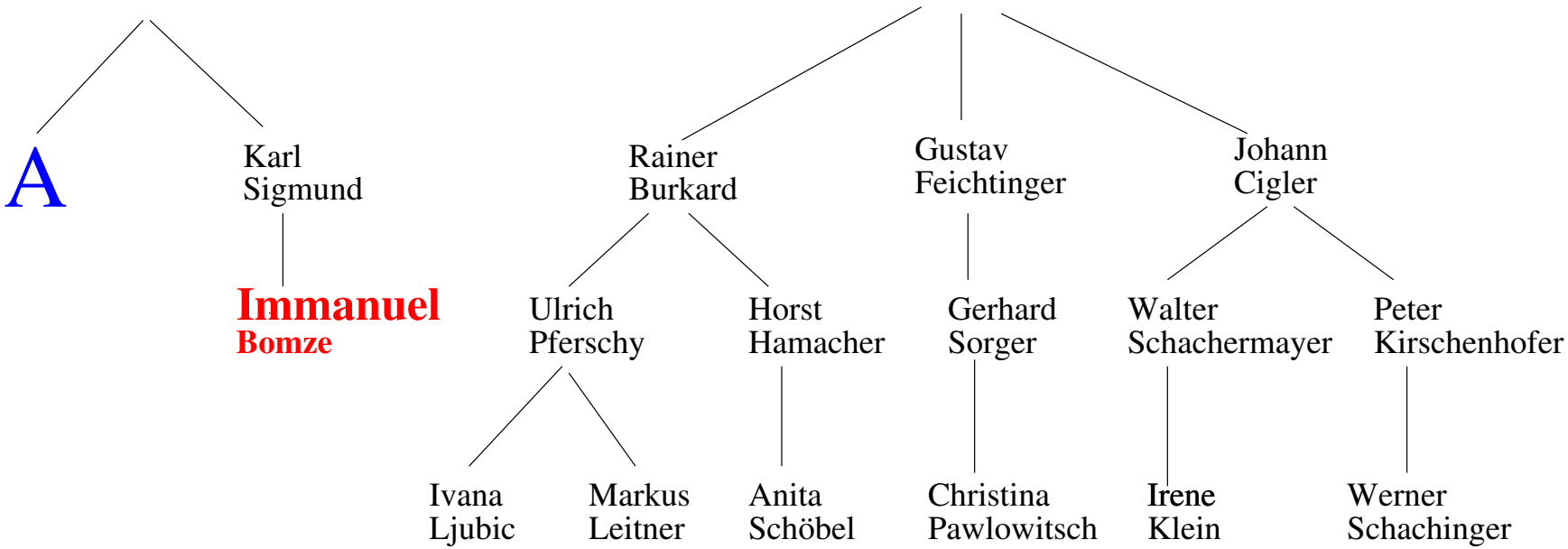
Conclusion and Outlook



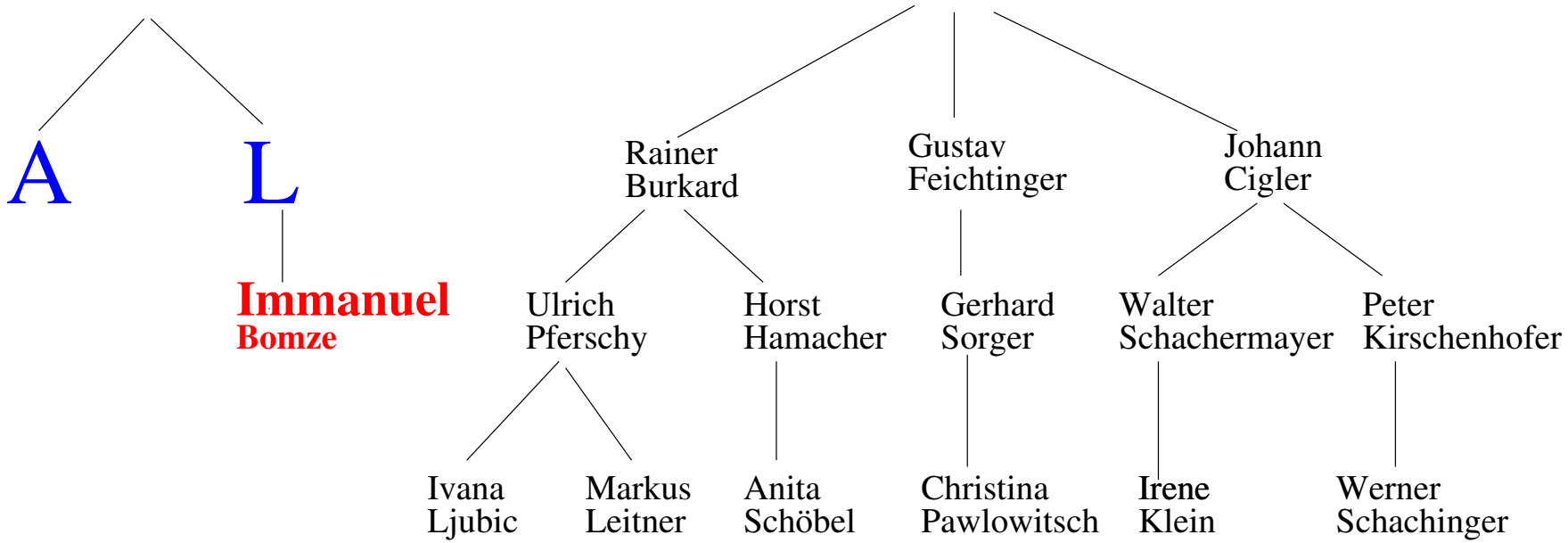
Conclusion and Outlook



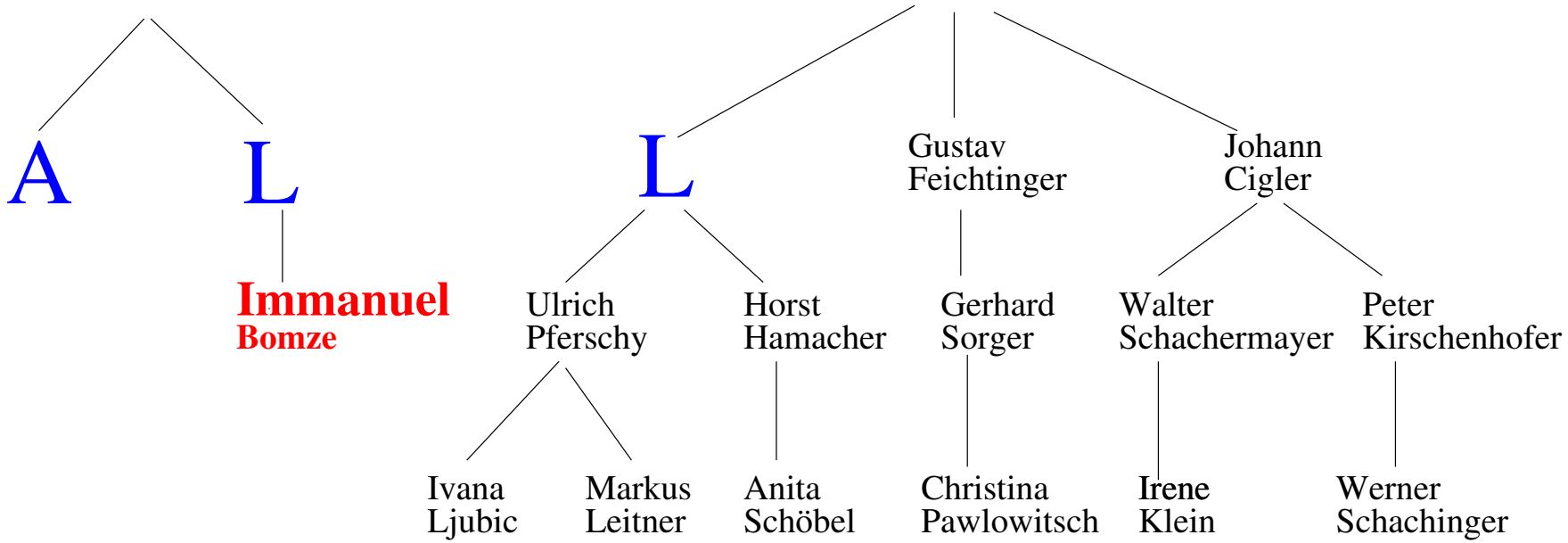
Conclusion and Outlook



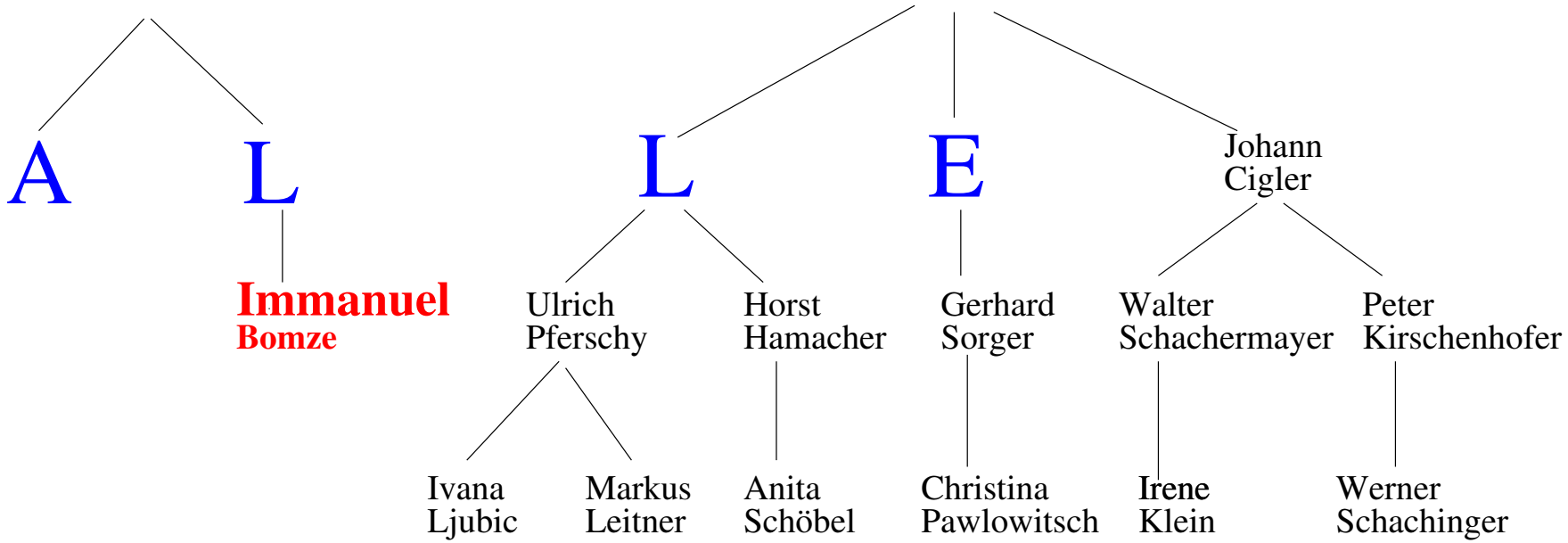
Conclusion and Outlook



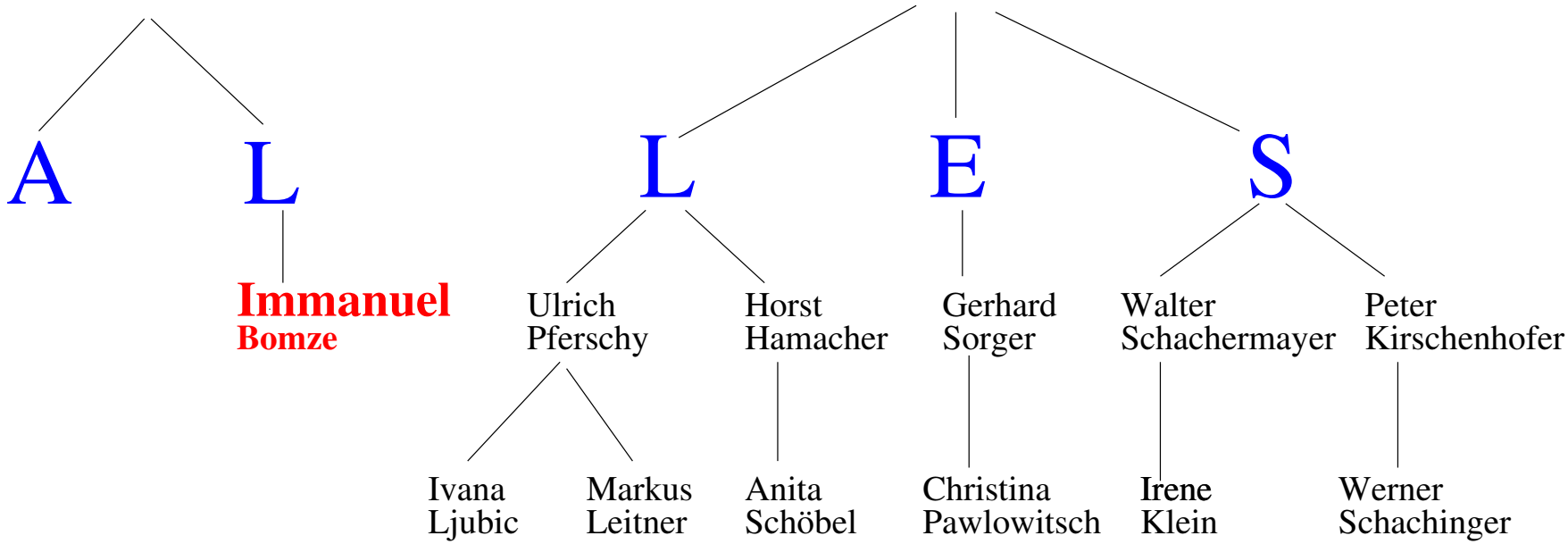
Conclusion and Outlook



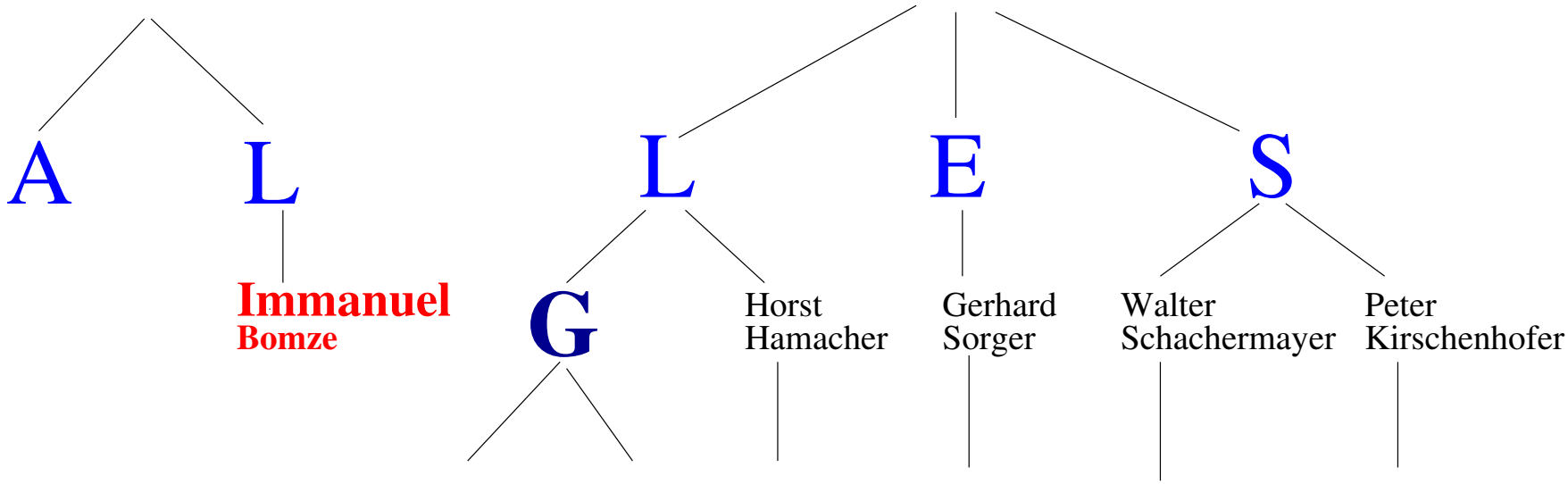
Conclusion and Outlook



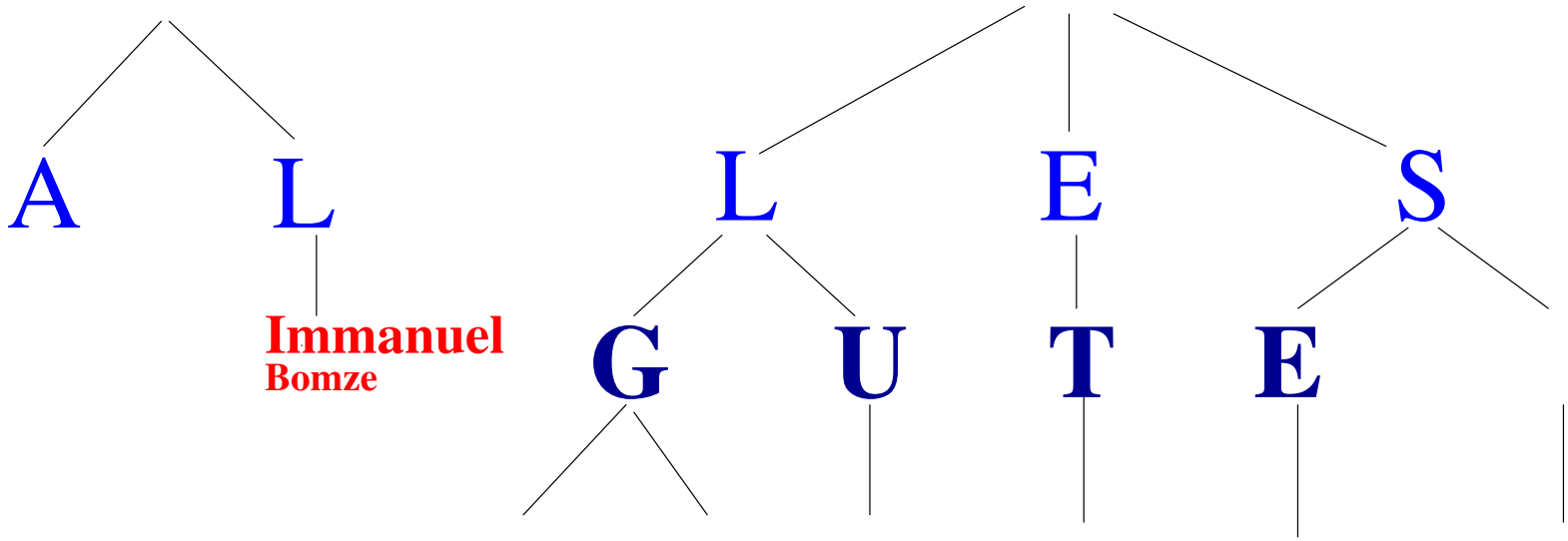
Conclusion and Outlook



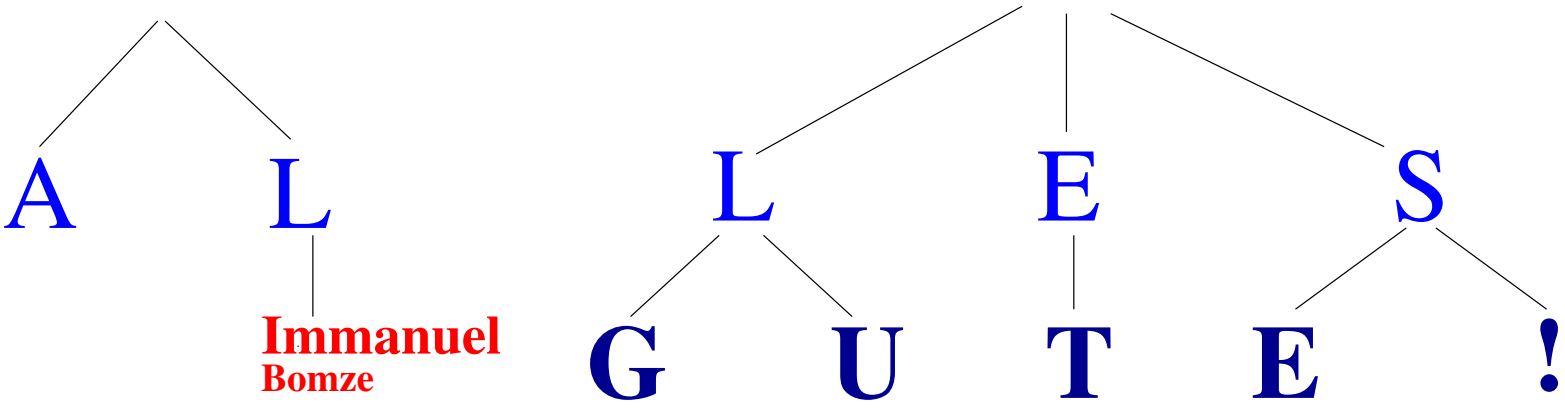
Conclusion and Outlook



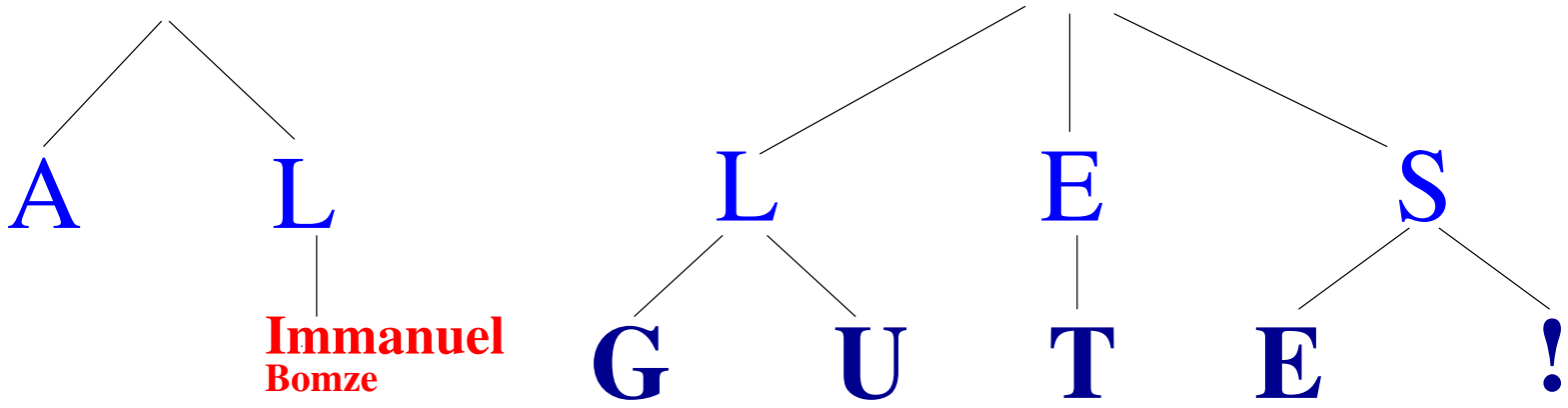
Conclusion and Outlook



Conclusion and Outlook



Conclusion and Outlook



Thank you! Questions?