Max Planck Institute for Intelligent Systems – Autonomous Motion Department

# Navigation in the Space of Hierarchies

Ömür Arslan

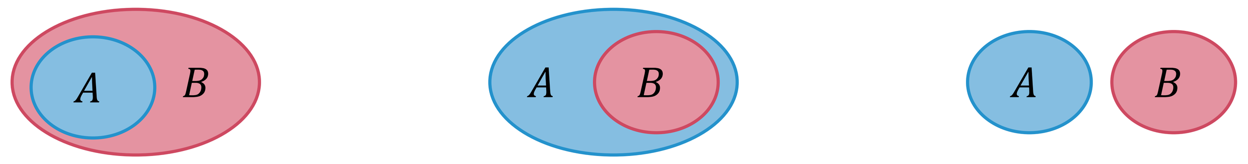
## Motivation

- Efficient and informative comparison of hierarchical structures used in bioinformatics, pattern recognition and data mining [1]
- Adaptive (randomized) restructuring of hierarchical clustering models for dynamically evolving (big) data [2]
- Analysis and (high-level) planning of structural transitions in grouping of multiagent systems [3]
- Adaptive dependency trees for approximating probability distributions [4]
- Structural anomaly detection and context-aware pattern recognition [5]

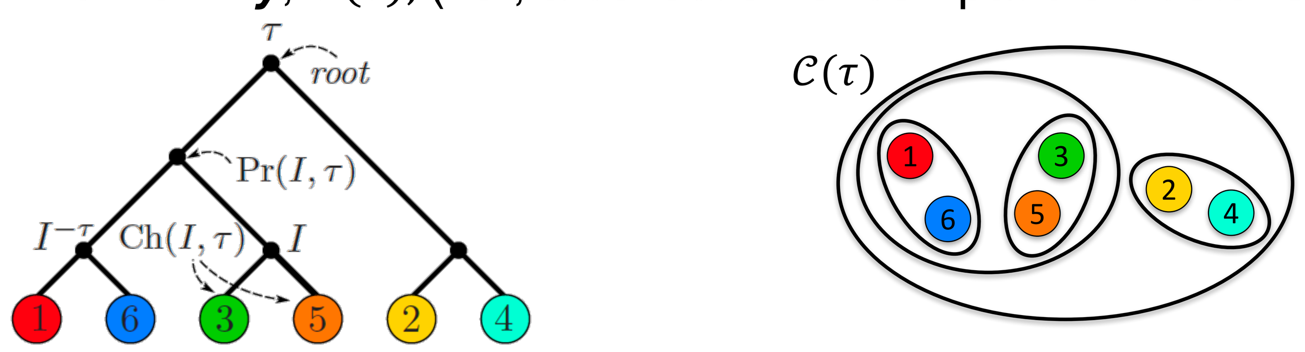
## A Set Theoretic View of Hierarchies

**Definition:** Two sets, $A$ and $B$, are said to be ***compatible*** if one is a subset of the other or they are disjoint, i.e.,

$A \subseteq B$ or $A \supseteq B$ or $A \cap B = \emptyset$




**Definition:** A ***hierarchy*** is equivalently represented, in graph theory, by a ***rooted tree***, $\tau$,(i.e., a connected directed acyclic graph) and, in set theory, by a ***laminar family***, $\mathcal{C}(\tau)$, (i.e., a collection of compatible cluster sets).

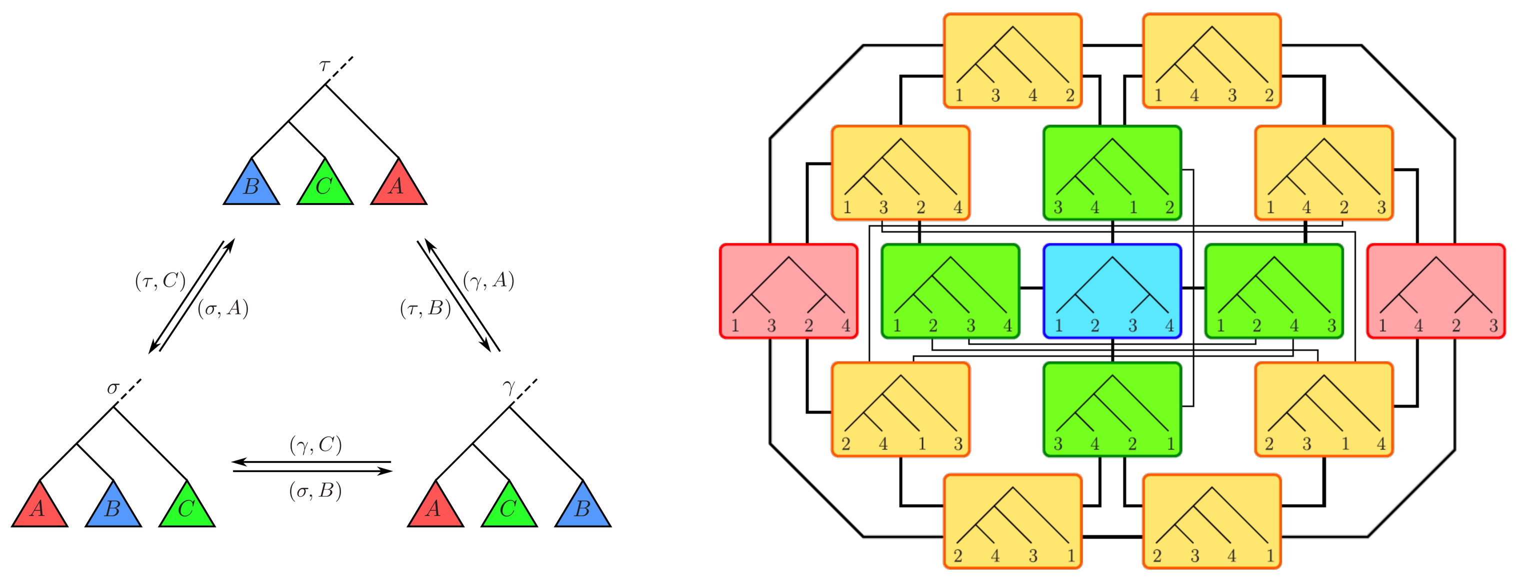



**Remark:** A binary tree, $\tau \in \mathcal{BT}_n$, is a maximal collection of compatible subsets of its leaf set, $\{1, 2, \ldots, n\}$.

## Nearest Neighbor Interchange Moves

**Definition:** A ***nearest neighbor interchange*** (NNI) move on a hierarchy, $\tau \in \mathcal{BT}_n$, swaps a cluster, $G \in \mathcal{C}(\tau)$, with its parent's sibling, $\mathrm{Pr}(G,\tau)^{-\tau}$.

Accordingly, the ***NNI graph*** is formed over the vertex set of binary trees, $\mathcal{BT}_n$, by declaring two trees to be connected by an edge if and only if one can be obtained from the other by a single NNI move.

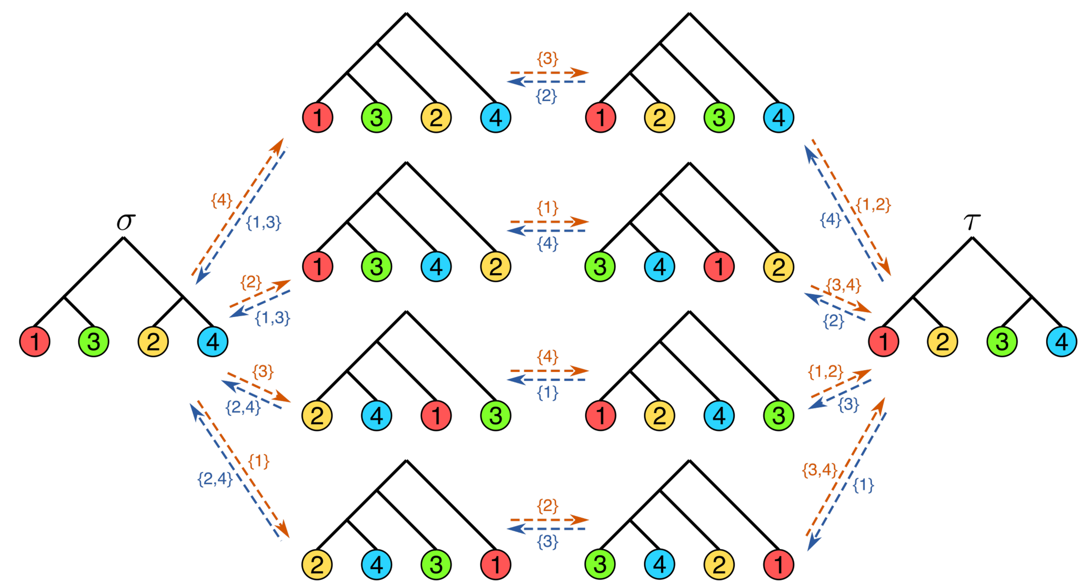


## NNI Navigation Algorithm

To navigate from any given initial hierarchy $\sigma \in \mathcal{BT}_n$ towards a desired goal hierarchy $\tau \in \mathcal{BT}_n$, one can find an NNI move on $\sigma$ at cluster $G \in \mathcal{C}(\sigma)$ as follows:

1) If $\sigma = \tau$, then return the identify move, $G \leftarrow \emptyset$.
2) Otherwise,
   a) Find a common cluster $K$ of $\sigma$ and $\tau$ with incompatible children.
   b) Find a descendant $I$ of $K$ in tree $\sigma$ which is incompatible with $Ch(K,\tau)$ and whose children $Ch(I,\sigma)$ are compatible with $Ch(K,\tau)$.
   c) Return a proper NNI navigation move on $\sigma$ at a child cluster in $Ch(I,\sigma)$:
      i. If $G^{-\sigma} \cup I^{-\sigma}$ is compatible with $Ch(K,\tau)$ for some $G \in Ch(I,\sigma)$, then return $G$.
      ii. Otherwise, return an arbitrary NNI move at a child $G \in Ch(I,\sigma)$.

**Example:**




**Proposition:** All NNI navigation paths between a pair of binary trees have the same length.

**Proposition:** An NNI navigation move over $\mathcal{BT}_n$ can be computed in $O(n)$ time with the number of leaves $n$.

## NNI Navigation Dissimilarity

**Definition:** The NNI navigation dissimilarity, $d_{nav}(\sigma,\tau)$, on $\mathcal{BT}_n$ is the count of NNI moves along an NNI navigation path joining a pair of trees, $\sigma$ and $\tau$.

**Important Properties:**

- $d_{nav}$ has a closed form formula that is a weighted count of pairwise incompatibilities of clusters of trees.
- $d_{nav}$ is positive definite, symmetric, but it is not a metric (because it fails to satisfy the triangle inequality).
- $d_{nav}$ on $\mathcal{BT}_n$ can be computed in $O(n^2)$ time.
- $\mathrm{diam}(\mathcal{BT}_n, d_{nav}) = \frac{1}{2}(n-1)(n-2)$.

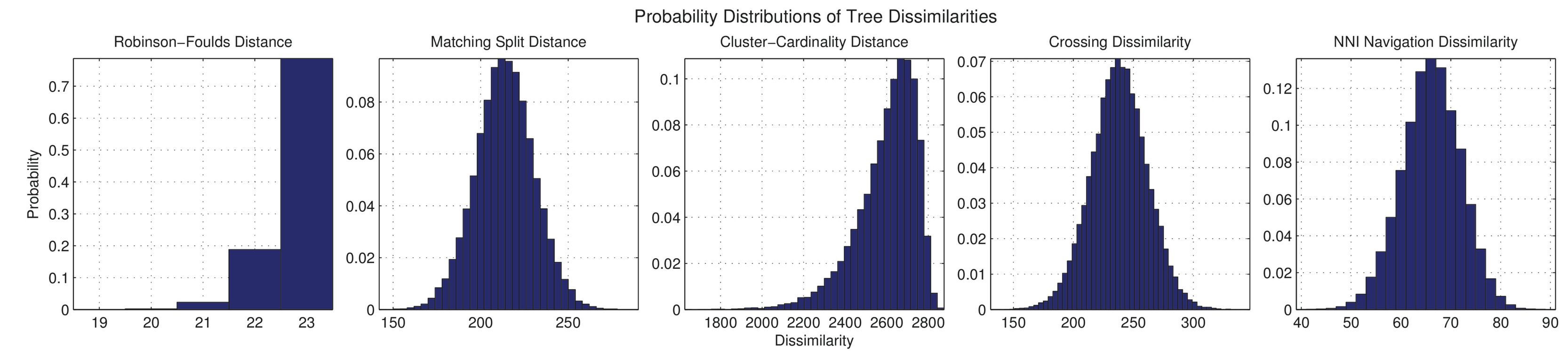
## References

[1] O. Arslan, D.P. Guralnik and D.E. Koditschek, "Discriminative Measures for Comparison of Phylogenetic Trees," *Discrete Applied Mathematics*, vol. 217, pp. 405-426, 2017.
[2] O. Arslan and D.E. Koditschek, "Anytime Hierarchical Clustering", *arXiv:1404.3439*, 2014.
[3] O. Arslan, D.P. Guralnik and D.E. Koditschek, "Coordinated Robot Navigation via Hierarchical Clustering," *IEEE Transactions on Robotics*, vol. 32, no. 2, pp. 352-371, 2016.
[4] C. Chow and C. Liu. "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462-467, 1968.
[5] M.J. Choi, J.J. Lim, A. Torralba and A.S. Willsky, "Exploiting Hierarchical Context on a Large Database of Object Categories," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 129-136.

## Relations with Other Tree Measures

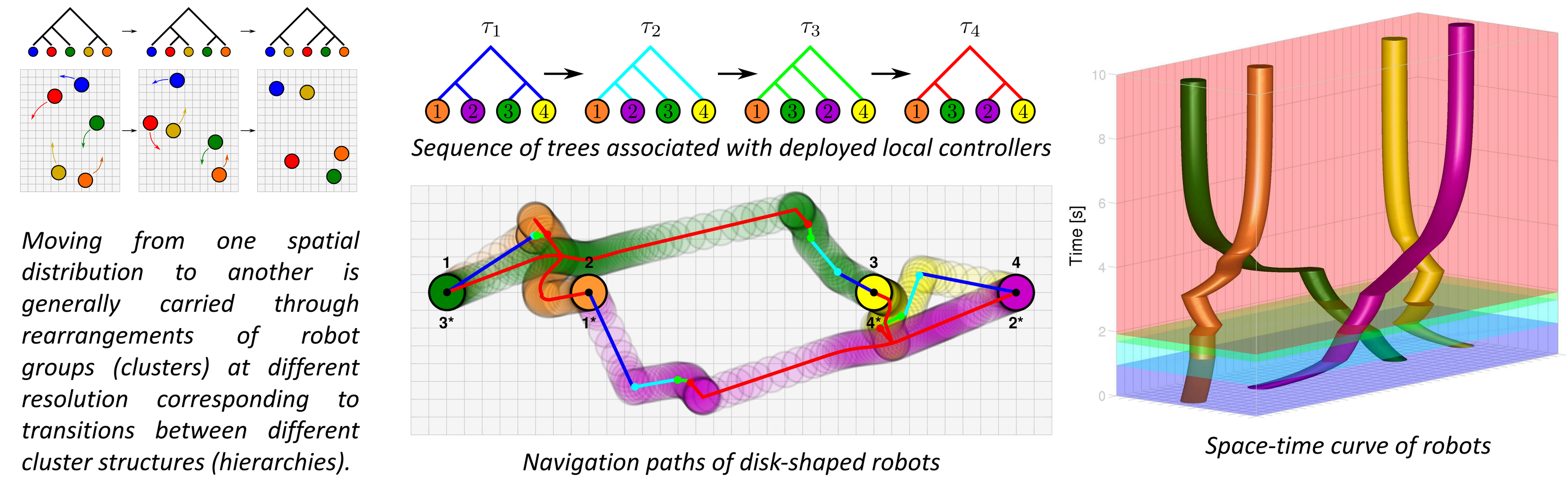
**Theorem:** $\frac{2}{3} d_{RF} \leq \frac{2}{3} d_{NNI} \leq \frac{2}{3} d_{nav} \leq d_{CM} \leq d_{CC}$

- The Robinson-Foulds distance, $d_{RF}$, is the count of the disparate edges of trees ($\mathrm{diam}(\mathcal{BT}_n, d_{RF}) = n-2$, Time Complexity: $O(n)$).
- The NNI distance, $d_{NNI}$, is the shortest path distance in the NNI graph ($\mathrm{diam}(\mathcal{BT}_n, d_{NNI}) = O(n \log n)$, Time Complexity: NP-hard).
- The crossing dissimilarity, $d_{CM}$, is the count of pairwise incompatible clusters of trees ($\mathrm{diam}(\mathcal{BT}_n, d_{CM}) = (n-2)^2$, Time Complexity: $O(n^2)$).
- The cluster-cardinality distance, $d_{CC}$, is a pullback of the matrix norm of an ultrametric embedding of hierarchies ( $\mathrm{diam}(\mathcal{BT}_n, d_{CC}) = \mathrm{O}(n^3)$ , Time Complexity: $O(n^2)$).

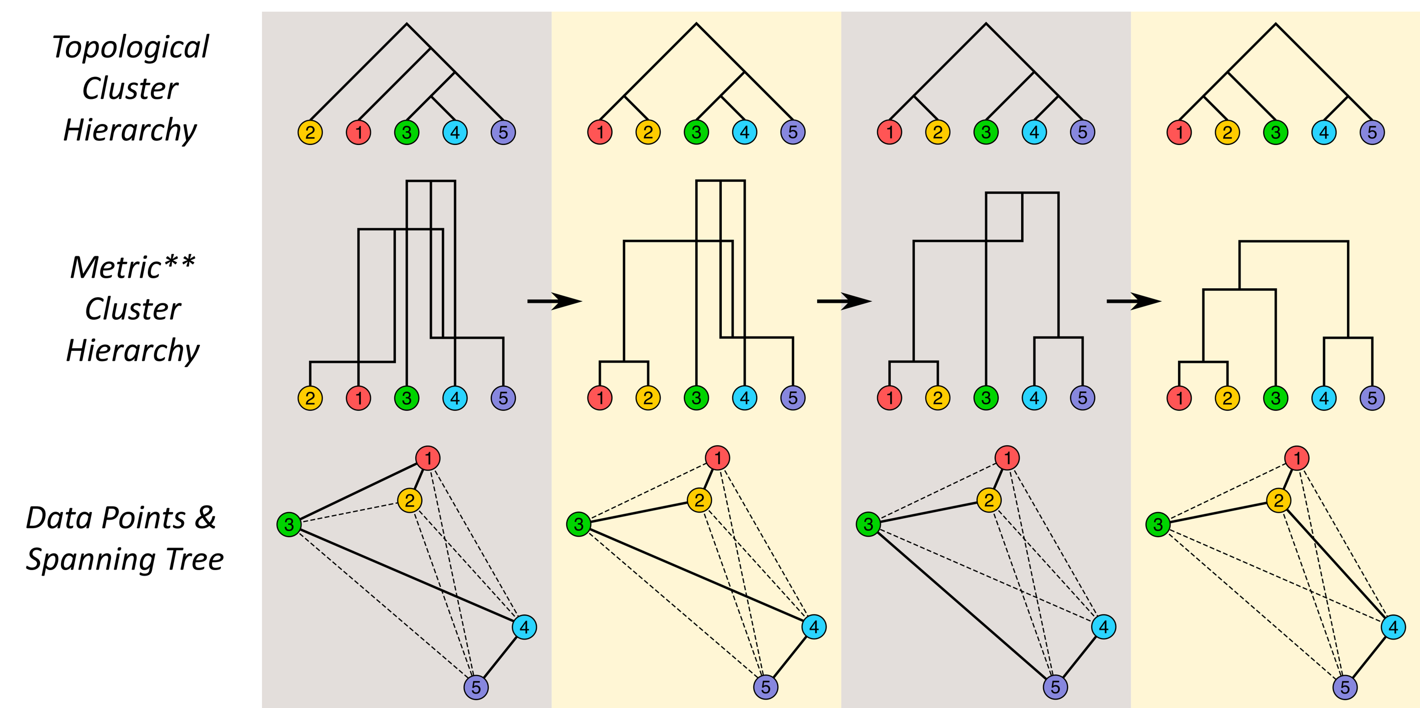



## Applications

- Coordinated Multirobot Navigation via Hierarchical Clustering [3]




- Anytime Hierarchical Linkage Clustering* [2]




* Instead of cluster compatibility, each NNI move aims to increase cluster homogeneity, i.e., $l(x; I, I^{-\tau}) \leq min\left(l(x; I, Pr(I,\tau)^{-\tau}), l(x; I, Pr(I,\tau)^{-\tau})\right)$.
** Here, the single linkage function is used to measure cluster dissimilarity.