# Analysis of ranked gene tree probability distributions under the coalescent process for detecting anomaly zones

Anastasiia Kim

Department of Mathematics and Statistics, University of New Mexico, Albuquerque

Understanding phylogenies

- A species tree represents the evolutionary relationships among various species.
- Gene trees represent the genealogical relationships among the gene sequences sampled from the species.
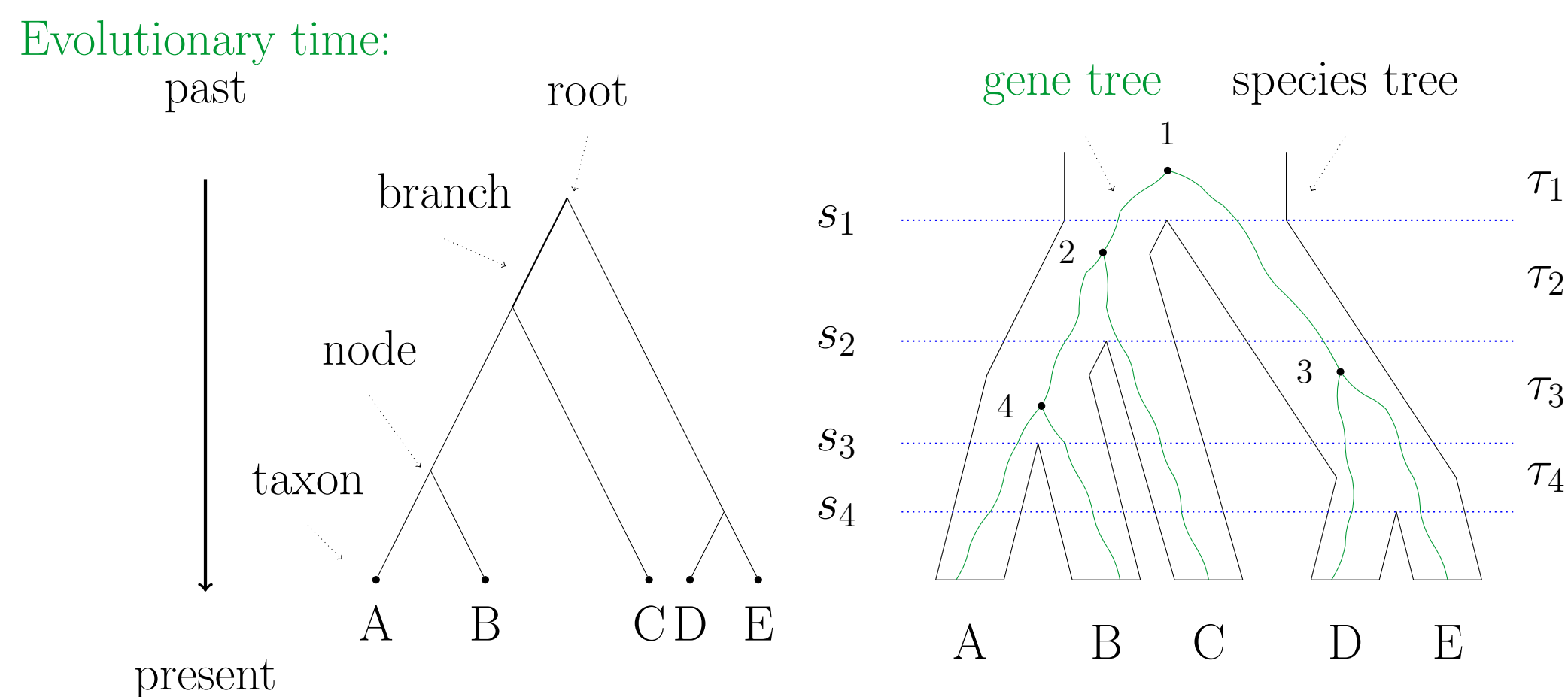
A rooted 5 taxon phylogeny where s_i is the time of the interior node of rank i

Ranked vs Unranked gene trees

- Unranked trees depict the topological relationships among gene lineages.
- Ranked trees also depict the sequence in which the lineages coalesce (join).

Calculating the probability of a ranked gene tree topology $\mathcal{G}$ given a species tree $\mathcal{T}$

$$P(\mathcal{G}|\mathcal{T}) = \sum_{x \in \mathcal{Y}} H_{\ell_1}(x) \underbrace{\prod_{i=2}^{n-1} P(G_{\tau_i}, x|T)}_{\text{product over speciation intervals } \tau_i}$$

$$\underbrace{\phantom{\sum_{x \in \mathcal{Y}} H_{\ell_1}(x) \prod_{i=2}^{n-1} P(G_{\tau_i}, x|T)}}_{\text{sum over all ranked histories}}$$

- $H_{\ell_1}(x)$ is the probability for the coalescence above the root appearing in the right order[2].
- $P(\mathcal{G}_{\tau_i}, x|\mathcal{T})$ is the probability in interval τ_i for ranked history x.

Matching and non-matching gene trees evolving on the species tree

- Gene trees that have different ranked topologies but share the same unranked topology $((AB)C)(DE)$.
- Both gene trees have the ranked history of $(1, 2, 2, 2)$.
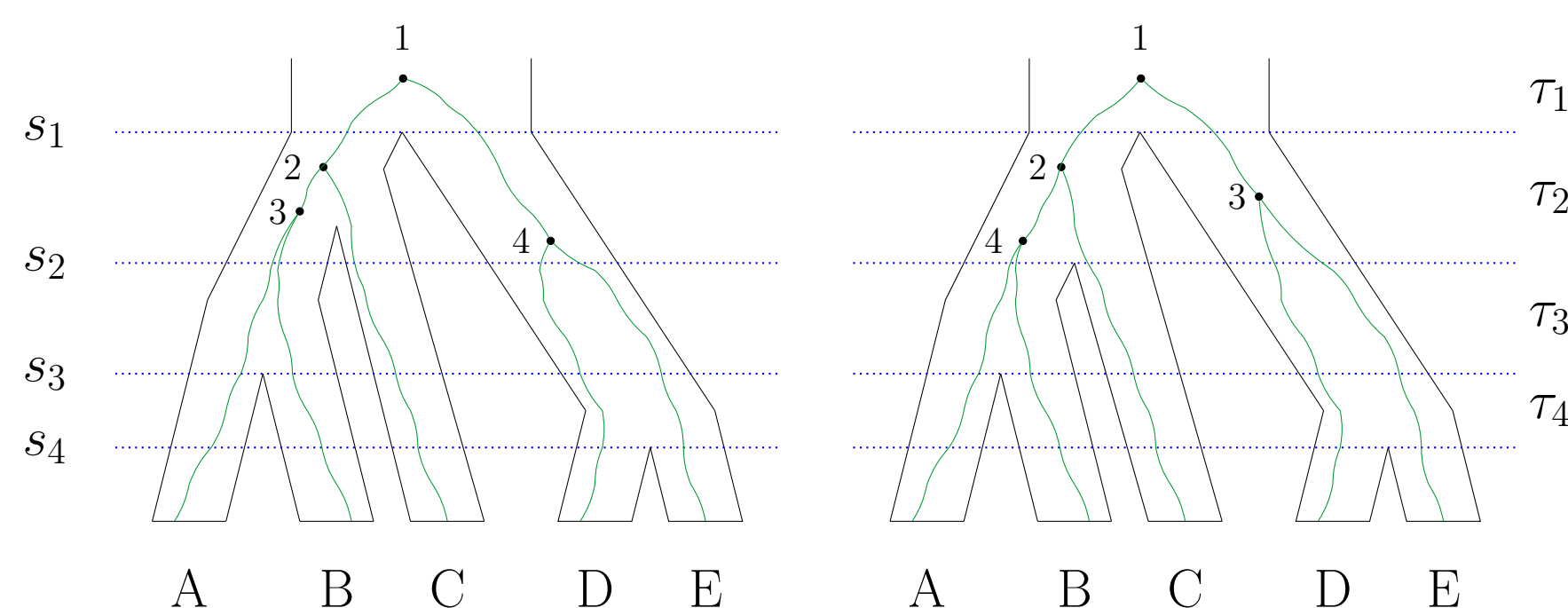
References

[1] J. H. Degnan and N.A. Rosenberg. Discordance of species trees with their most likely gene trees. PLoS Genet., 2006.

[2] J. H. Degnan, N.A. Rosenberg, and T. Stadler. The probability distribution of ranked gene trees on a species tree. Math. Biosci., 2012.

Anomalous gene trees

- The incorrect gene tree topology (one that does not match the species tree) that is more probable than the correct one is termed anomalous gene tree[1].
- Species trees that can generate anomalous gene trees are said to be in the anomaly zone.
- The method of choosing the most common gene tree as the estimate of the species tree in the anomaly zone will be statistically inconsistent.
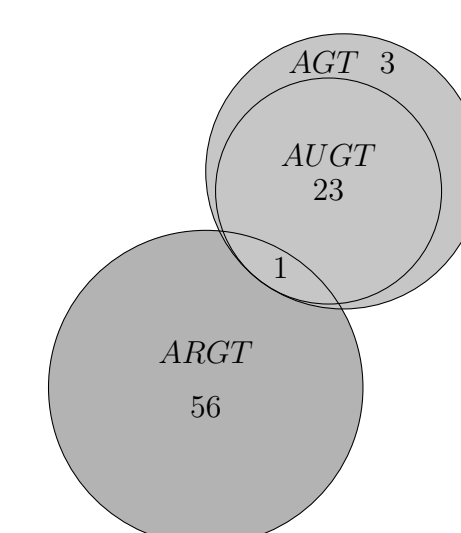
Anomaly zones
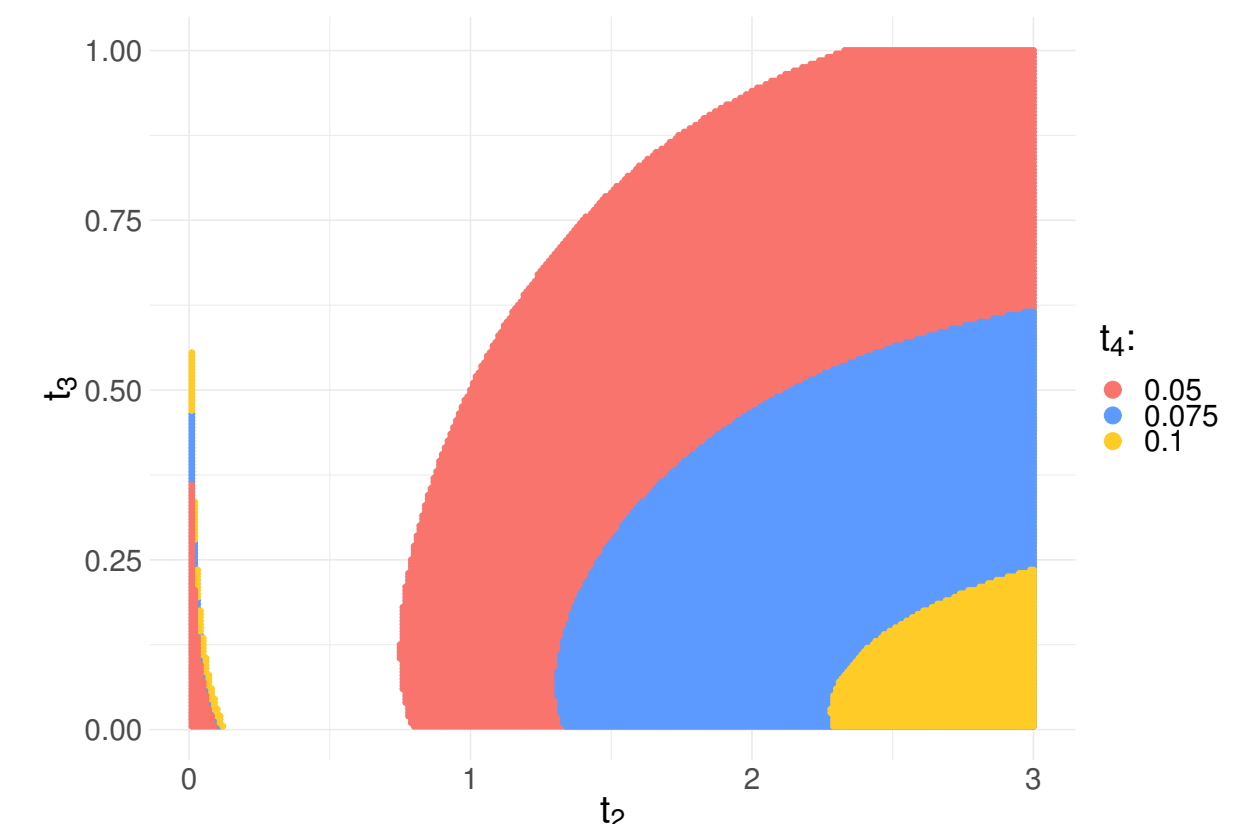
Slices of the unranked and ranked anomaly zones for the 5-taxon species-tree topology
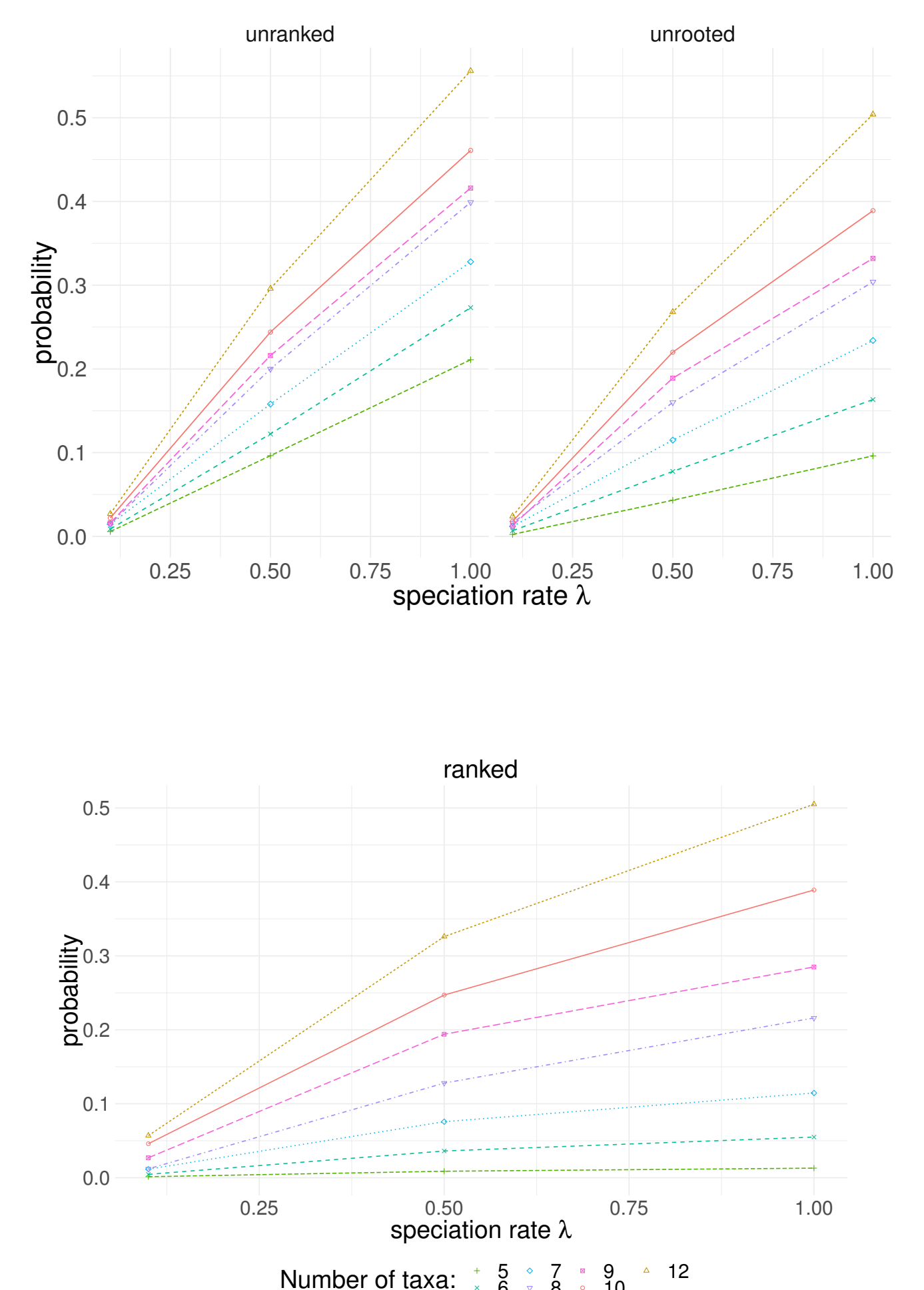
The impact of the speciation rate parameter on the existence of anomaly zones

Heuristics for larger trees

12 taxa, $\lambda = 0.1$

12 taxa, $\lambda = 1$