Balance indices for phylogenetic trees under well-known probability models

Universitat de les Illes Balears

Tomás M. Coronado

LINKÖPING
UNIVERSITY

1 What is a phylogenetic tree?
  - Balance

2 Probabilistic models for phylogenetic trees
  - The Yule model
  - The Uniform model
  - The α and α-γ models
  - The β-model

3 Balance indices
  - The Colless index
  - The Sackin index
  - The Cophenetic index
  - The Quadratic Colless index
  - The rooted Quartet index

4 Conclusions

5 References

1 What is a phylogenetic tree?
Balance

2 Probabilistic models for phylogenetic trees
The Yule model
The Uniform model
The α and α-γ models
The β-model

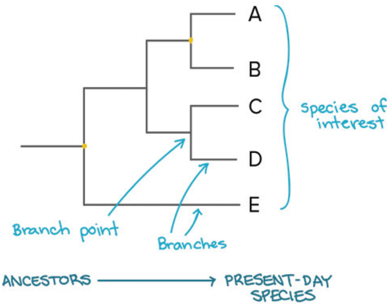
3 Balance indices
The Colless index
The Sackin index
The Cophenetic index
The Quadratic Colless index
The rooted Quartet index

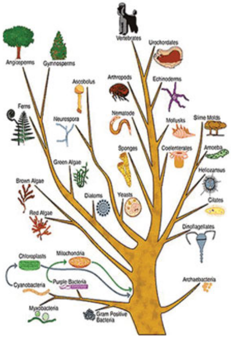
4 Conclusions

5 References

LINKÖPING
UNIVERSITY

What is a phylogenetic tree?





source:
`https://microbenotes.com/how-to-construct-a-phylogenetic-tree/`

A phylogenetic tree depicts the joint evolutionary history of a set of species.

A phylogenetic tree depicts the joint evolutionary history of a set of species.
Two main aspects are interesting to biologists:

- The length of the branches of a phylogenetic tree: the timing of speciation events.
- The *shape*, or *topology*, of the tree: differences in diversification rates among subtrees.

A phylogenetic tree depicts the joint evolutionary history of a set of species.
Two main aspects are interesting to biologists:

- The length of the branches of a phylogenetic tree: the timing of speciation events.
- The *shape*, or *topology*, of the tree: differences in diversification rates among subtrees.

Reconstructing the former is usually harder than reconstructing the latter [Drummond et al. 2006], since many reconstructing methods agree on the shape.

What is a phylogenetic tree (Mathematically)?

- Let T be a rooted tree, and understand it as a directed graph.
- Let $L(T)$ be the set of *leaves* of T; i.e., of the nodes with out-degree 0. Conversely, call $\mathring{V}(T) = V(T) \setminus L(T)$ the set of *internal nodes* of T.
- Let Λ be a set of labels, and $\lambda : L(T) \to \Lambda$ a map.

The pair (T, λ) is a *phylogenetic tree* if λ is injective.
If λ is not injective, it is a *multilabelled tree*.

Balance

- A popular way to assess the underlying shape of a phylogenetic tree is to consider a quantitative measure over it.

Balance

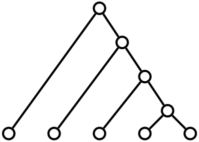
- A popular way to assess the underlying shape of a phylogenetic tree is to consider a quantitative measure over it.
- The "balance" of a phylogenetic tree is a pre-theoretic, intuitive concept reflecting its shape.
- It measures the propensity of internal nodes to have the same number of descendants.

Balance

- A popular way to assess the underlying shape of a phylogenetic tree is to consider a quantitative measure over it.
- The “balance” of a phylogenetic tree is a pre-theoretic, intuitive concept reflecting its shape.
- It measures the propensity of internal nodes to have the same number of descendants. Sort of.

Three families of trees

The caterpillar
Caterpillars are bifurcating trees all of whose internal nodes are parents to at least one leaf.



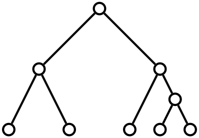
Figur: The caterpillar with five leaves

They are considered to be “the least balanced family of trees”, because

- they are completely one-sided,
- they minimize the number of automorphisms of a tree.

Three families of trees

The maximally balanced tree
Maximally balanced trees are bifurcating trees all of whose internal nodes have children whose subtrees have numbers of leaves that differ in at most 1.

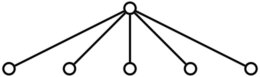


Figur: The maximally balanced tree with five leaves

They are considered to be "the most balanced family of *bifurcating* trees", because it splits "as evenly as possible" the number of descendant leaves at each step.

Three families of trees

The star
Stars are usually non bifurcating trees all of whose leaves pend from the root.



Figur: The star with five leaves

They are considered to be “the most balanced family of trees”, because

- there is only one internal node,
- they maximize the number of automorphisms of a tree.

1 What is a phylogenetic tree?

- Balance

2 Probabilistic models for phylogenetic trees

- The Yule model
- The Uniform model
- The α and α-γ models
- The β-model

3 Balance indices

- The Colless index
- The Sackin index
- The Cophenetic index
- The Quadratic Colless index
- The rooted Quartet index

4 Conclusions

5 References

LINKÖPING
UNIVERSITY

Why are we interested in probabilistic models?

- Be able to produce new trees to test evolutionary hypothesis against the trees appearing in the bibliography.

Why are we interested in probabilistic models?

- Be able to produce new trees to test evolutionary hypothesis against the trees appearing in the bibliography.
- If we know the first moments of balance indices, to test reconstructed trees against the null hypothesis “this tree is obtained under the model P_n”.

How to create a phylogenetic tree?

A probabilistic model (P_n) on phylogenetic trees is a family of functions $P_n : \mathcal{T}_n \to [0,1]$ assigning a probability $P_n(T)$ to each $T \in \mathcal{T}_n$ such that $\sum_{T\in\mathcal{T}_n} P_n(T) = 1$.

- Most models in this section only deal with bifurcating trees.
- That means that the probability of multifurcating trees is 0.

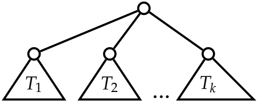
Three properties

Three properties that a probabilistic model for phylogenetic trees can have and that ease the computations are that of *Markovianity*, *shape invariance* and *sampling consistency*:

- **Markovianity** (bifurcating version): A probabilistic model (P_n) of phylogenetic trees is sampling consistent if there exists a family $q(k, n-k)$ in $[0,1]$ such that $\sum_{k=1}^{n-1} q(k, n-k) = 1$ and

  $$P_n(T_k * T_{n-k}) = q(k, n-k) P_k(T_k) P_{n-k}(T_{n-k}),$$

  where $T_k * T_{n-k}$ is the root join of $T_k \in \mathcal{T}_k$ and $T_{n-k} \in \mathcal{T}_{n-k}$: the tree whose root has T_k and T_{n-k} as children.



Figur: The tree $T_1 * \cdots * T_k$, with maximal pending subtrees $T_1, \ldots, T_k$.

Three properties

- **Shape invariance**: If T_1, T_2 have the same shape but possibly different labelling, $P_n(T_1) = P_n(T_2)$.

Three properties

- **Shape invariance**: If T_1, T_2 have the same shape but possibly different labelling, $P_n(T_1) = P_n(T_2)$.
- **Sampling consistency**: Given a tree T_{n-1} leaves, we have

$$P_{n-1}(T_{n-1}) = \sum_{T_n \in \mathcal{T}_n} \sum_{\substack{x \in L(T) \\ T_n(-x) = T_{n-1}}} P_n(T_n),$$

where $T_n(-n)$ is the tree resulting after removing the leaf labelled n from T_n.

The Yule model

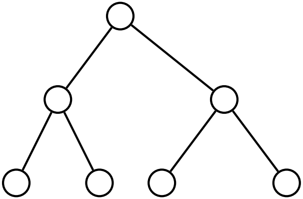
Recursive model of tree growth for bifurcating trees:

1. Start with a single node

The Yule model

Recursive model of tree growth for bifurcating trees:

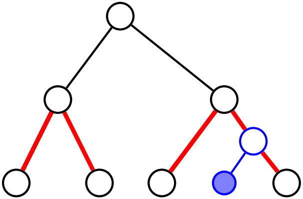
1. Start with a single node
2. For every step m, add a new leaf by choosing uniformly between pending arcs



The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step m, add a new leaf by choosing uniformly between pending arcs

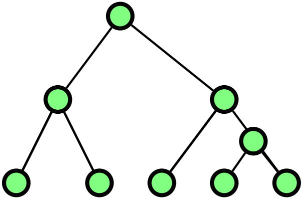


$\frac{1}{m}$

The Yule model

Recursive model of tree growth for bifurcating trees:

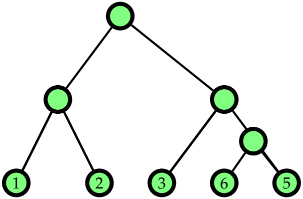
1. Start with a single node
2. For every step m, add a new leaf by choosing uniformly between pending arcs
3. Until the number of leaves n is reached



The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step *m*, add a new leaf by choosing uniformly between pending arcs
3. Until the number of leaves *n* is reached
4. Label the tree uniformly



The Yule model

The Yule model explicitly assumes that, at each speciation event, all the current species are equally likely to speciate.

The Yule model

The Yule model explicitly assumes that, at each speciation event, all the current species are equally likely to speciate.
The Yule model is

- Markovian with $q(k, n-k) = \frac{1}{n-1}$ [Semple and Steel 2003].
- Shape invariant by construction.
- Sampling consistent [Ford 2005].

The Uniform model

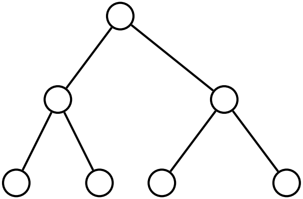
Recursive model of tree growth for bifurcating trees:

1. Start with a single node

The Uniform model

Recursive model of tree growth for bifurcating trees:

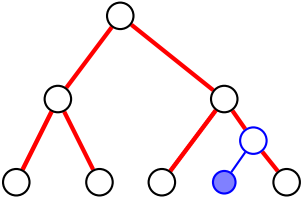
1. Start with a single node
2. For every step m, add a new leaf by choosing uniformly between *any* arc



The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step m, add a new leaf by choosing uniformly between *any* arc

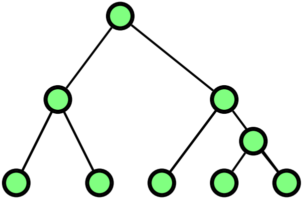


$$\frac{1}{2(m-1)}$$

The Uniform model

Recursive model of tree growth for bifurcating trees:

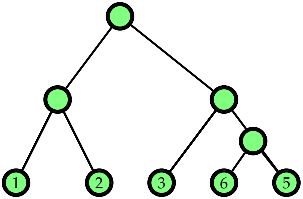
1. Start with a single node
2. For every step *m*, add a new leaf by choosing uniformly between *any* arc
3. Until the number of leaves *n* is reached



The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step m, add a new leaf by choosing uniformly between *any* arc
3. Until the number of leaves n is reached
4. Label the tree uniformly



The Uniform model

Equivalently: Uniformly choose a tree with n leaves from the set of all phylogenetic trees with n leaves.

The Uniform model

Equivalently: Uniformly choose a tree with n leaves from the set of all phylogenetic trees with n leaves.
Therefore, it assumes that all the joint evolutive histories are equally likely.

The Uniform model

Equivalently: Uniformly choose a tree with n leaves from the set of all phylogenetic trees with n leaves.
Therefore, it assumes that all the joint evolutive histories are equally likely.

- There are $(2n-3)!!$ trees with n leaves [Schröder 1870].
- Therefore, each tree has probability $\frac{1}{(2n-3)!!}$.

The Uniform model

Equivalently: Uniformly choose a tree with n leaves from the set of all phylogenetic trees with n leaves.
Therefore, it assumes that all the joint evolutive histories are equally likely.

- There are $(2n-3)!!$ trees with n leaves [Schröder 1870].
- Therefore, each tree has probability $\frac{1}{(2n-3)!!}$.

As a result, the Uniform model is

- Markovian with $q(k, n-k) = C_{k,n-k} = \frac{1}{2}\binom{n}{k}^{-1}\frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}$ [Semple and Steel 2003], where $n!! = n(n-2)(n-4)\cdots 1$ if n is odd and $n!! = n(n-2)(n-4)\cdots 2$ if it is even.
- Shape invariant by construction.
- Sampling consistent [Ford 2005].

The α-model

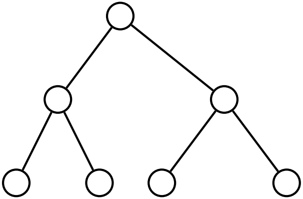
Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled

The α-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

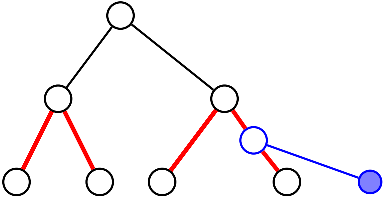
1. Start with a single node labelled
2. For every step m, add a new leaf by choosing randomly between:



The α-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step m, add a new leaf by choosing randomly between:
   - *pending arc*

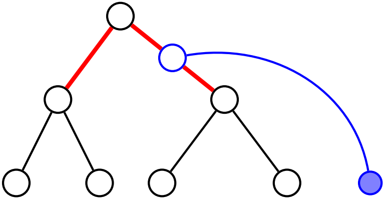


$$\frac{1-\alpha}{n-\alpha}$$

The α-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - *internal arc*

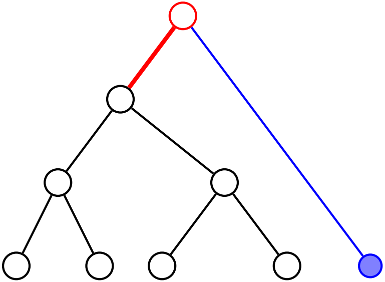


$\frac{\alpha}{n-\alpha}$

The α-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc *(including a new root)*

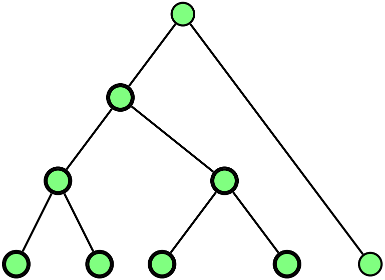


$$\frac{\alpha}{n-\alpha}$$

The α-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

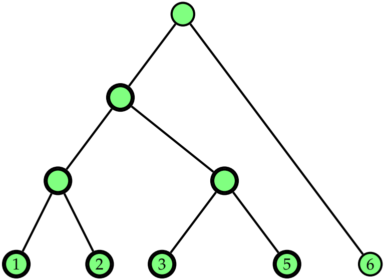
1. Start with a single node labelled
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc (including a new root)
3. Until number of leaves n is reached



The α-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc (including a new root)
3. Until number of leaves n is reached
4. Label the tree uniformly



The α-model

- Markovian [Ford 2005].
- Shape invariant by construction.
- Sampling consistent [Ford 2005].

The α-model

- Equal to the Yule model if $\alpha = 0$ [Ford 2005].
- Equal to the Uniform model if $\alpha = 1/2$ [Ford 2005].

The α-γ-model

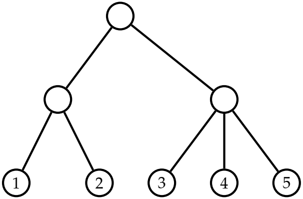
Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1

The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

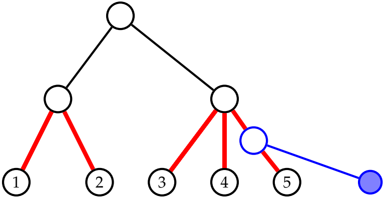
1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:



The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:
   - *pending arc*

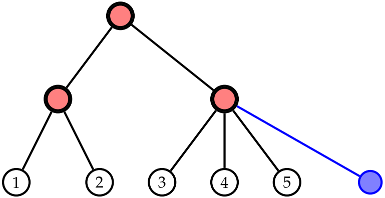


$$\frac{1-\alpha}{n-\alpha}$$

The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - *internal node*

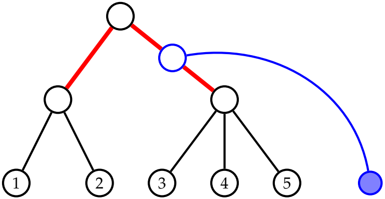


$$\frac{(deg(v)-1)\alpha-\gamma}{n-\alpha}$$

The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - *internal arc*

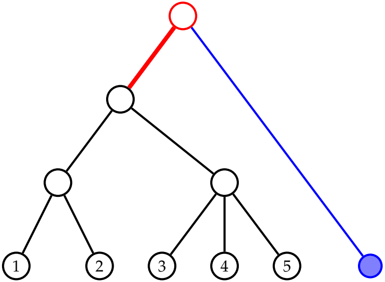


$$\frac{\gamma}{n-\alpha}$$

The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc *(including a new root)*



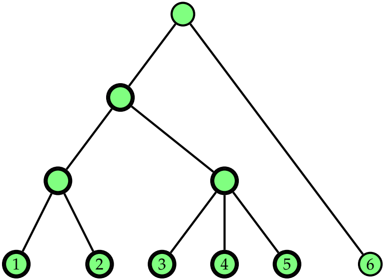
$$\frac{\gamma}{n-\alpha}$$

The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc (including a new root)

   and label it m

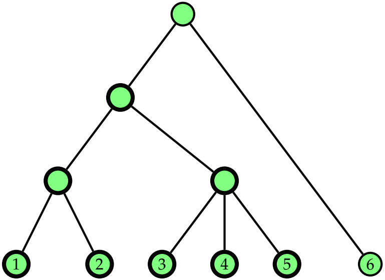


The α-γ-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step m, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc (including a new root)

   and label it m
3. Until number of leaves n is reached



The α-γ-model

The only probabilistic model presented here of multifurcating trees.

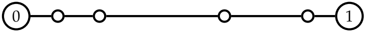
- Markovian [Chen, Ford, and Winkel 2009].
- **Not** shape invariant in general.
- Sampling consistent [Chen, Ford, and Winkel 2009].

The α-γ-model

- Equal to the α-model when $\alpha = \gamma$ if we relabel each leaf uniformly [Chen, Ford, and Winkel 2009].

The β-model

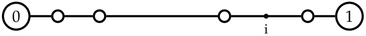
1. Start with n dots uniformly distributed over the interval $[0, 1]$



The β-model

1. Start with n dots uniformly distributed over the interval $[0,1]$
2. Choose a point in $[0,1]$ with beta density

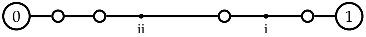
$$f(x)=\frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}x^{\beta}(1-x)^{\beta},\quad 0<x<1.$$



The β-model

1. Start with n dots uniformly distributed over the interval $[0,1]$
2. Choose a point in $[0,1]$ with beta density

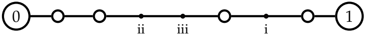
$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \quad 0 < x < 1.$$



The β-model

1. Start with n dots uniformly distributed over the interval $[0,1]$
2. Choose a point in $[0,1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \quad 0 < x < 1.$$

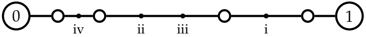


The β-model

1. Start with n dots uniformly distributed over the interval $[0,1]$
2. Choose a point in $[0,1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^{\beta}(1-x)^{\beta}, \quad 0 < x < 1.$$

3. Until each pair of leaves is separated by at least one point

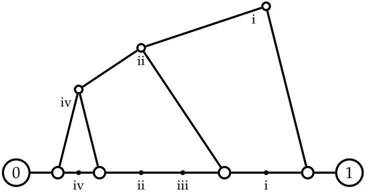


The β-model

1. Start with n dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \quad 0 < x < 1.$$

3. Until each pair of leaves is separated by at least one point
4. Construct the tree accordingly

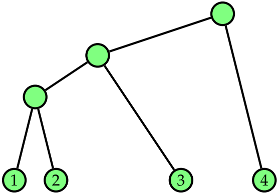


The β-model

1. Start with n dots uniformly distributed over the interval $[0,1]$
2. Choose a point in $[0,1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \quad 0 < x < 1.$$

3. Until each pair of leaves is separated by at least one point
4. Construct the tree accordingly
5. Label the tree uniformly



The β-model

- It is Markovian [Aldous 1996].
- Shape invariant by construction.
- Sampling consistent [Aldous 1996].

The β-model

- Equal to the Yule model if $\beta = 0$ [Aldous 1996].
- Equal to the Uniform model if $\beta = -3/2$ [Aldous 1996].

The β-model

- Equal to the Yule model if $\beta = 0$ [Aldous 1996].
- Equal to the Uniform model if $\beta = -3/2$ [Aldous 1996].
- Therefore, the α and β models intersect at these points...

The β-model

- Equal to the Yule model if $\beta = 0$ [Aldous 1996].
- Equal to the Uniform model if $\beta = -3/2$ [Aldous 1996].
- Therefore, the α and β models intersect at these points...
- ... and these are the only points at which them intersect (**Theorem 43** at [Ford 2005]).

1 What is a phylogenetic tree?
- Balance

2 Probabilistic models for phylogenetic trees
- The Yule model
- The Uniform model
- The α and α-γ models
- The β-model

3 Balance indices
- The Colless index
- The Sackin index
- The Cophenetic index
- The Quadratic Colless index
- The rooted Quartet index

4 Conclusions

5 References

Balance indices: What do we know?

- Most balance indices have only been studied under the models of Yule and Uniform.

Balance indices: What do we know?

- Most balance indices have only been studied under the models of Yule and Uniform.
- The only index presented here of which we know both the first and second moments under every probabilistic model presented is the rooted Quartet index.

The Colless index

- Introduced in [Colless 1982].
- Only sound for bifurcating trees.
- Let $u \in \mathring{V}(T)$, and call u_1, u_2 its two children. Let $\kappa(u_i)$ be the number of leaves of T under u_i.
- Then,

$$C(T) = \sum_{u \in \mathring{V}(T)} |\kappa(u_1) - \kappa(u_2)|.$$

The Colless index

- Introduced in [Colless 1982].
- Only sound for bifurcating trees.
- Let $u \in \mathring{V}(T)$, and call u_1, u_2 its two children. Let $\kappa(u_i)$ be the number of leaves of T under u_i.
- Then,

$$C(T) = \sum_{u \in \mathring{V}(T)} |\kappa(u_1) - \kappa(u_2)|.$$

In other words, the sum over all internal nodes of the absolute difference of numbers of leaves of each pair of subtrees rooted at the same internal node.

The Colless index

The Colless index has the undeniable quality of being intuitive, as it sums up all the “local imbalances” of a tree.

- Its maximum value for a tree with n leaves is $\binom{n-1}{2}$ and it is attained exactly by the caterpillars [Mir, Rotger, and Rosselló 2013].
- Its minimum value is $\sum_{i=0}^{\ell-1} 2^{m_i}(m_\ell - m_i - 2(\ell - i - 1))$, where $\sum_{i=0}^{\ell} 2^{m_i}$, with $m_i < m_{i+1}$, is the binary decomposition of n. It is attained by the maximally balanced trees, among other trees [Coronado, Fischer, et al. 2020].

The Colless index

The Colless index has the undeniable quality of being intuitive, as it sums up all the “local imbalances” of a tree.

- Its maximum value for a tree with n leaves is $\binom{n-1}{2}$ and it is attained exactly by the caterpillars [Mir, Rotger, and Rosselló 2013].
- Its minimum value is $\sum_{i=0}^{\ell-1} 2^{m_i}(m_\ell - m_i - 2(\ell - i - 1))$, where $\sum_{i=0}^{\ell} 2^{m_i}$, with $m_i < m_{i+1}$, is the binary decomposition of n. It is attained by the maximally balanced trees, among other trees [Coronado, Fischer, et al. 2020].
- By far, the most popular balance index in the literature.

The Colless index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_α	σ^2_α	E_β	σ^2_β
Colless	✓ [1]	✓ [2]	O [3]	×	×	×	×	×

[1] Heard 1992
[2] Cardona, Mir, and Rosselló 2013
[3] Blum, François, and Janson 1996

The Colless index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_α	σ^2_α	E_β	σ^2_β
Colless	✓ [1]	✓ [2]	O [3]	×	×	×	×	×

[1] Heard 1992

[2] Cardona, Mir, and Rosselló 2013

[3] Blum, François, and Janson 1996

- If we knew the expected value or the variance under the β or α model, we would know it under the Uniform model.

The Sackin index

- Introduced in [Sokal 1983].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

$$S(T) = \sum_{x \in L(T)} \delta(x),$$

where $\delta(x)$ is the *depth* of x; i.e., the length of the shortest path from the root to x.

The Sackin index

- Introduced in [Sokal 1983].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as
$$S(T) = \sum_{x \in L(T)} \delta(x),$$
where $\delta(x)$ is the *depth* of x; i.e., the length of the shortest path from the root to x.

In other words, the sum of the depths of all the leaves of T.

The Sackin index

Also intuitive: the caterpillar has more different depths than the maximally balanced tree does.

- Its maximum value for a tree with n leaves is $\frac{(n-1)(n+2)}{2}$ and it is attained exactly by the caterpillars [Fischer 2018].
- Its minimum value is $2^m m + 2s(m+1)$, where $n = 2^m + s$, with $s < 2^m$. It is attained exactly by the maximally balanced trees and the trees depth-equivalent to them [Fischer 2018].

The Sackin index

Also intuitive: the caterpillar has more different depths than the maximally balanced tree does.

- Its maximum value for a tree with n leaves is $\frac{(n-1)(n+2)}{2}$ and it is attained exactly by the caterpillars [Fischer 2018].
- Its minimum value is $2^m m + 2s(m+1)$, where $n = 2^m + s$, with $s < 2^m$. It is attained exactly by the maximally balanced trees and the trees depth-equivalent to them [Fischer 2018].
- The second most popular balance index in the literature.

The Sackin index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_α	σ^2_α	E_β	σ^2_β
Colless	✓	✓	O	×	×	×	×	×
Sackin	✓ [1]	✓ [2]	✓ [3]	✓ [4]	×	×	×	×

[1] Kirkpatrick and Slatkin 1993
[2] Cardona, Mir, and Rosselló 2013
[3] Mir, Rotger, and Rosselló 2013
[4] Coronado, Mir, Rosselló, and Rotger 2020

The Sackin index

This last result is known thanks to the proof in the Supplementary Material of [Coronado, Mir, Rosselló, and Rotger 2020] of **Proposition 6** thereof: the solution of the family of recurrences

$$X_n = 2\sum_{k=1}^{n-1} C_k X_k + \sum_{l=1}^{r} a_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!}\sum_{i=1}^{s} b_l \binom{n}{l},$$

with initial condition X_1 and a_l, b_l real numbers.

The Sackin index

This last result is known thanks to the proof in the Supplementary Material of [Coronado, Mir, Rosselló, and Rotger 2020] of **Proposition 6** thereof: the solution of the family of recurrences

$$X_n = 2\sum_{k=1}^{n-1} C_k X_k + \sum_{l=1}^{r} a_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{i=1}^{s} b_l \binom{n}{l},$$

with initial condition X_1 and a_l, b_l real numbers.
As a further note, the term $\frac{(2n-2)!!}{(2n-3)!!}$ appears when dealing with the expected value or the variance of recursive shape indices under the Uniform model.

The Colless and Sackin indices

In [Blum, François, and Janson 1996], we find the following results

- The Pearson correlation under the Yule model of the Sackin and Colless indices tends to
$$\mathrm{cor}_{\mathrm{Yule}}(C_n, S_n) \sim \frac{27 - 2\pi^2 - 6\log 2}{\sqrt{2(18-\pi^2-6\log 2)(21-2\pi^2)}} \sim 0.98,$$
as n goes to ∞.
- Under the Uniform model,
$$\frac{S_n - C_n}{n^{3/2}} \to 0$$
in probability as n tends to ∞.
- Let A be the Airy distribution [Flajolet and Louchard 2001]. Under the Uniform model,
$$\frac{S_n}{n^{3/2}} \to A$$
in distribution as n tends to ∞.

The Cophenetic index

- Introduced in [Mir, Rotger, and Rosselló 2013].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

$$\Phi(T) = \sum_{x,y \in L(T)} \phi(x, y),$$

where $\phi(x, y)$ is the *cophenetic value* of x and y; i.e., depth of the lowest common ancestor of both x and y.

The Cophenetic index

- Introduced in [Mir, Rotger, and Rosselló 2013].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

  $$\Phi(T) = \sum_{x,y \in L(T)} \phi(x,y),$$

  where $\phi(x,y)$ is the *cophenetic value* of x and y; i.e., depth of the lowest common ancestor of both x and y.

In other words, the sum over all pairs of leaves of the length of their shared evolutive history.

The Cophenetic index

- Its maximum value for a tree with n leaves is $\binom{n}{3}$ and it is attained exactly by the caterpillars [Mir, Rotger, and Rosselló 2013].
- Its minimum value for a multifurcating tree with n leaves is $\binom{n}{2}$ and is attained exactly at the stars.
- Its minimum value for a bifurcating tree with n leaves is

$$\binom{n}{2} - \sum_{j=1}^{s_n} 2^{m_j(n)-1}(m_j(n) + 2(s_n - j))$$

, where $\sum_{j=0}^{\ell}$ is the binary decomposition of n, $m_i < m_{i+1}$ [to be submitted]. It is attained exactly by the maximally balanced trees [Mir, Rotger, and Rosselló 2013].

The Cophenetic index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_α	σ^2_α	E_β	σ^2_β
Colless	✓	✓	O	×	×	×	×	×
Sackin	✓	✓	✓	✓	×	×	×	×
Cophenetic	✓[1]	✓[2]	✓[1]	✓[3]	×	×	×	×

[1] Mir, Rotger, and Rosselló 2013

[2] Cardona, Mir, and Rosselló 2013

[3] Coronado, Mir, Rosselló, and Rotger 2020

The Cophenetic index: limit behaviour under the Yule model

We can extend the definition of the Cophenetic index continuously taking into account edge lengths [Bartoszek 2018a], call it $\hat{\Phi}$.

- For the continuos Cophenetic index, $\binom{n}{2}^{-1}\hat{\Phi}_n$ is a positive submartingale that converges almost surely and in L^2 to a finite first and second moment random variable [Bartoszek 2018a] under the Yule model.
- For the (discrete) Cophenetic index, it can be shown that $\binom{n}{2}^{-1}\Phi_n$ is an almost surely and L^2 convergent submartingale [Bartoszek 2018a] under the Yule model.

The Sackin and Cophenetic indices

- The covariance of the Sackin and Cophenetic indices under the Uniform model is known [Coronado, Mir, Rosselló, and Rotger 2020]:

$$\mathrm{cov}_{\mathrm{unif}}(S_n, \Phi_n) = \binom{n}{2}\frac{26n^2-5n-4}{15} - \frac{3n+2}{8}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!} - \frac{n}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.$$

The Sackin and Cophenetic indices

- The covariance of the Sackin and Cophenetic indices under the Uniform model is known [Coronado, Mir, Rosselló, and Rotger 2020]:

$$\mathrm{cov}_{\mathrm{unif}}(S_n,\Phi_n) = \binom{n}{2}\frac{26n^2-5n-4}{15} - \frac{3n+2}{8}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!} - \frac{n}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.$$

- The Pearson correlation of the Sackin and Cophenetic under the Uniform model is estimated [Coronado, Mir, Rosselló, and Rotger 2020]:

$$\mathrm{cor}_{\mathrm{unif}}(S_n,\Phi_n) = \frac{\frac{52-15\pi}{60}}{\sqrt{\frac{10-3\pi}{3}\,\frac{56-15\pi}{240}}} \sim 0.965.$$

The Quadratic Colless index

- Introduced in [Bartoszek et al. 2020].
- Only sound for bifurcating trees.
- Let $u \in \mathring{V}(T)$, and call u_1, u_2 its two children. Let $\kappa(u_i)$ be the number of leaves of T under u_i.
- Then,

$$C^{(2)}(T) = \sum_{u \in \mathring{V}(T)} (\kappa(u_1) - \kappa(u_2))^2.$$

The Quadratic Colless index

- Introduced in [Bartoszek et al. 2020].
- Only sound for bifurcating trees.
- Let $u \in \mathring{V}(T)$, and call u_1, u_2 its two children. Let $\kappa(u_i)$ be the number of leaves of T under u_i.
- Then,

$$C^{(2)}(T) = \sum_{u \in \mathring{V}(T)} (\kappa(u_1) - \kappa(u_2))^2.$$

In other words, it has the same intuitive justification as the Colless index, but the square instead of the absolute value makes it much more easy to manipulate.

The Quadratic Colless index

The Quadratic Colless index has the undeniable quality of being intuitive, as it sums up all the "local imbalances" of a tree.

- Its maximum value for a tree with n leaves is $\frac{n(n-1)(2n-1)}{6}$ and it is attained exactly by the caterpillars [Bartoszek et al. 2020].
- Its minimum value is the same of the Colless index. It is attained exactly by the maximally balanced trees [Bartoszek et al. 2020].

The Quadratic Colless index

The Quadratic Colless index has the undeniable quality of being intuitive, as it sums up all the "local imbalances" of a tree.

- Its maximum value for a tree with n leaves is $\frac{n(n-1)(2n-1)}{6}$ and it is attained exactly by the caterpillars [Bartoszek et al. 2020].
- Its minimum value is the same of the Colless index. It is attained exactly by the maximally balanced trees [Bartoszek et al. 2020]. In contrast with the difficult characterization of the trees attaining the minimum Colless index.

The Quadratic Colless index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_{α}	σ^2_{α}	E_{β}	σ^2_{β}
Colless	✓	✓	O	×	×	×	×	×
Sackin	✓	✓	✓	✓	×	×	×	×
Cophenetic	✓	✓	✓	✓	×	×	×	×
Q. Colless	✓ [1]	✓ [1]	✓ [1]	✓ [4]	×	×	×	×

[1] Bartoszek et al. 2020

The Quadratic Colless index: limit behaviour under the Yule model

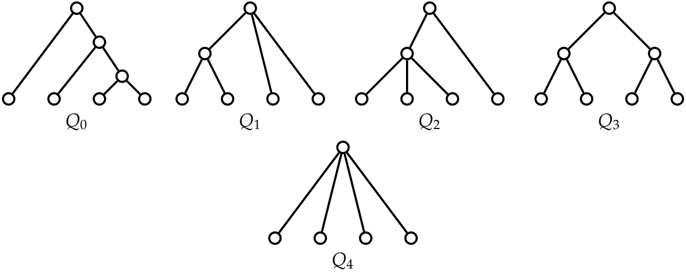
Set $Y := \frac{C^{(2)} - E_{\text{Yule}}(C_n^{(2)})}{n^2}$.
As $n \to \infty$, the distribution under the Yule model of Y is such that

$$Y \to \tau^2 Y' + (1-\tau)^2 Y'' + (1 + 6\tau^2 - 6\tau),$$

in distribution, where $\tau \sim \text{Unif}[0,1]$ and Y', Y'' are independent and distributed according to the same law as the limit of Y [Bartoszek et al. 2020].

The rooted Quartet index

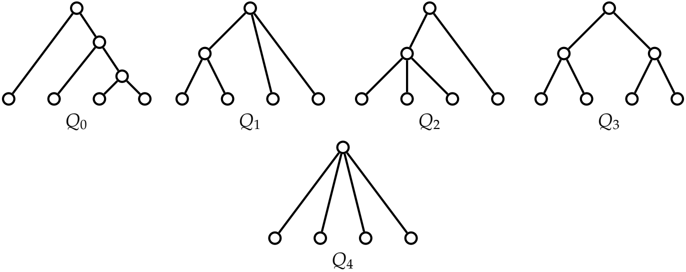
There are five different trees with five leaves.



Figur: The five tree shapes in $\mathcal{T}_4$.

The rooted Quartet index

There are five different trees with five leaves.



Figur: The five tree shapes in $\mathcal{T}_4$.

They are ordered according to their number of automorphisms, and assigned a number q_i increasing on it.

The rooted Quartet index

- Introduced in [Coronado, Mir, Rosselló, and Valiente 2019].
- Can be defined (and makes sense) for all trees.
- Defined as

$$\mathrm{QI}(T) = \sum_{i=0}^{4} |\{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\}| \cdot q_i.$$

The rooted Quartet index

- Introduced in [Coronado, Mir, Rosselló, and Valiente 2019].
- Can be defined (and makes sense) for all trees.
- Defined as

$$\mathrm{QI}(T) = \sum_{i=0}^{4} |\{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\}| \cdot q_i.$$

Notice that, in this case, the value “increases with balance”, where in the other cases more “balanced” trees had smaller index values.

The rooted Quartet index

- Its maximum value for a multifurcating tree with n leaves is $\binom{n}{4}q_4$ and it is attained exactly by the stars [Coronado, Mir, Rosselló, and Valiente 2019].
- Its maximum value for a bifurcating tree with n leaves can be found in Sloane's Encyclopedia of Integer Sequences [Sloane 1964], seq. A300445. It is attained exactly by the maximally balanced trees [Coronado, Mir, Rosselló, and Valiente 2019].
- Its minimum value is 0, and it is attained exactly by the caterpillars.

The rooted Quartet index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_α	σ^2_α	E_β	σ^2_β	$E_{\alpha,\gamma}$	$\sigma^2_{\alpha,\gamma}$
Colless	✓	✓	O	×	×	×	×	×		
Sackin	✓	✓	✓	✓	×	×	×	×		
Cophenetic	✓	✓	✓	✓	×	×	×	×		
Q. Colless	✓	✓	✓	✓	×	×	×	×		
r. Quartet	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]

[1] Coronado, Mir, Rosselló, and Valiente 2019

The rooted Quartet index

- The rooted Quartet index is the only balance index presented whose first and second moments are known under all the probabilistic models presented so far.

The rooted Quartet index

- The rooted Quartet index is the only balance index presented whose first and second moments are known under all the probabilistic models presented so far.
- The backbone of the above proofs is that the α-γ-model is **sampling consistent**.

The rooted Quartet index

- The rooted Quartet index is the only balance index presented whose first and second moments are known under all the probabilistic models presented so far.
- The backbone of the above proofs is that the α-γ-model is **sampling consistent**.
- Indeed: for any sampling consistent probabilistic model (P_n^*) of trees [Coronado, Mir, Rosselló, and Valiente 2019],

$$E_P(\mathrm{QI}_n) = \binom{n}{4} \sum_{i=1}^{4} P_4^*(Q_i) q_i.$$

$$\sigma_P^2(\mathrm{QI}_n) = \binom{n}{4} \sum_{i=1}^{4} q_i^2 P_4^*(Q_i) - \binom{n}{4}^2 \left(\sum_{i=1}^{4} q_i P_4^*(Q_i) \right)^2$$

$$\sum_{i=1}^{4} \sum_{j=1}^{4} q_i q_j \left(\sum_{k=5}^{8} \binom{n}{k} \sum_{T \in \mathcal{T}_k} \Theta_{ij}(T) P_k^*(T) \right),$$

The rooted Quartet index

Under the α-γ-model

- The expected value of the rooted Quartet index under the α-γ-model is known [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_{\alpha,\gamma}(\mathrm{QI}_n) = \Big(\frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_4 + \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}q_3$$

$$\frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_2 + \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_1\Big)\binom{n}{4}.$$

The rooted Quartet index

Under the α-γ-model

- The expected value of the rooted Quartet index under the α-γ-model is known [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_{\alpha,\gamma}(\mathrm{QI}_n) = \Big(\frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_4 + \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}q_3 \\ \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_2 + \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_1\Big)\binom{n}{4}.$$

- When $\alpha=\gamma$ (α-model), we get $E_\alpha(\mathrm{QI}_n) = \frac{(1-\alpha)(2-\alpha)}{(3-\alpha)(2-\alpha)}\binom{n}{4}q_3$.

The rooted Quartet index

Under the α-γ-model

- The expected value of the rooted Quartet index under the α-γ-model is known [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_{\alpha,\gamma}(\mathrm{QI}_n) = \Big(\frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_4 + \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}q_3 \\ \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_2 + \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}q_1\Big)\binom{n}{4}.$$

- When $\alpha=\gamma$ (α-model), we get $E_\alpha(\mathrm{QI}_n) = \frac{(1-\alpha)(2-\alpha)}{(3-\alpha)(2-\alpha)}\binom{n}{4}q_3$.
- Yule model: $\alpha=0$. Uniform model: $\alpha=1/2$.

The rooted Quartet index

- The variance under the α-γ model is also known, but the formula is too long! [Coronado, Mir, Rosselló, and Valiente 2019]

The rooted Quartet index

Under the β-model

- The β-model is also sampling consistent.

The rooted Quartet index

Under the β-model

- The β-model is also sampling consistent.
- That gives us the expected value of QI_n under the β-model, too [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_\beta(\mathrm{QI}_n) = \frac{3\beta+6}{7\beta+18}.$$

The rooted Quartet index

Under the β-model

- The β-model is also sampling consistent.
- That gives us the expected value of QI_n under the β-model, too [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_\beta(\mathrm{QI}_n) = \frac{3\beta+6}{7\beta+18}.$$

and its variance (again, too long!) [Coronado, Mir, Rosselló, and Valiente 2019].

The rooted Quartet index

Under the β-model

- The β-model is also sampling consistent.
- That gives us the expected value of QI_n under the β-model, too [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_\beta(\mathrm{QI}_n) = \frac{3\beta+6}{7\beta+18}.$$

  and its variance (again, too long!) [Coronado, Mir, Rosselló, and Valiente 2019].
- Yule model: $\beta = 0$. Uniform model $\beta = -3/2$.

The rooted Quartet index: limit behaviour under the β-model

An interesting result about the limit distribution of the Quartet index under the β-model, $\beta \geq 0$, can be found in [Bartoszek 2018b].
It shows that it converges weakly to a distribution that can be characterized as the fixed point of a contraction operator on a class of distributions.

1 What is a phylogenetic tree?
   - Balance

2 Probabilistic models for phylogenetic trees
   - The Yule model
   - The Uniform model
   - The α and α-γ models
   - The β-model

3 Balance indices
   - The Colless index
   - The Sackin index
   - The Cophenetic index
   - The Quadratic Colless index
   - The rooted Quartet index

4 Conclusions

5 References

LINKÖPING
UNIVERSITY

Conclusions

- We know some things

Conclusions

- We know some things
- but we ignore some other things.

The rooted Quartet index: what do we know?

index	E_{Yule}	σ^2_{Yule}	E_{unif}	σ^2_{unif}	E_α	σ^2_α	E_β	σ^2_β	$E_{\alpha,\gamma}$	$\sigma^2_{\alpha,\gamma}$
Colless	✓ [1]	✓ [2]	O [3]	×	×	×	×	×		
Sackin	✓ [4]	✓ [2]	✓ [5]	✓ [6]	×	×	×	×		
Cophenetic	✓ [5]	✓ [2]	✓ [5]	✓ [6]	×	×	×	×		
Q. Colless	✓ [7]	✓ [7]	✓ [7]	✓ [7]	×	×	×	×		
r. Quartet	✓ [8]	✓ [8]	✓ [8]	✓ [8]	✓ [8]	✓ [8]	✓ [8]	✓ [8]	✓ [8]	✓ [8]

[1] Heard 1992

[2] Cardona, Mir, and Rosselló 2013

[3] Blum, François, and Janson 1996

[4] Kirkpatrick and Slatkin 1993

[5] Mir, Rotger, and Rosselló 2013

[6] Coronado, Mir, Rosselló, and Rotger 2020

[7] Bartoszek et al. 2020

[8] Coronado, Mir, Rosselló, and Valiente 2019

1 What is a phylogenetic tree?
    - Balance

2 Probabilistic models for phylogenetic trees
    - The Yule model
    - The Uniform model
    - The α and α-γ models
    - The β-model

3 Balance indices
    - The Colless index
    - The Sackin index
    - The Cophenetic index
    - The Quadratic Colless index
    - The rooted Quartet index

4 Conclusions

5 References

- Schröder, E. (1870). “Vier Combinatorische Probleme”. In: *Z.Math. Phys.* 15, pp. 361–376.
- Sloane, N. (1964). *Online Encyclopedia of Integer Sequences*. `https://oeis.org/`.
- Colless, D. H. (1982). “Review of Phylogenetics: the theory and practice of phylogenetic systematics”. In: *Systematic Zoology* 31, pp. 100–104.
- Sokal, R. R. (1983). “A phylogenetic analysis of the Caminalcules I: The data base”. In: *Systematic Biology* 32, pp. 159–184.
- Heard, S. B. (1992). “Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees”. In: *Evolution* 46, pp. 1818–1826.
- Kirkpatrick, M. and M. Slatkin (1993). “Searching for evolutionary patterns in the shape of a phylogenetic tree”. In: *Evolution* 47, pp. 1171–1181.
- Aldous, D. J. (1996). “Probability distributions on cladograms”. In: *Random discrete structures*, pp. 1–18.

Blum, M. B., O. François, and S. Janson (1996). "The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance". In: *The Annals of Applied Probability* 16.4, pp. 2195–2214.

Flajolet, P. and G. Louchard (2001). "Analytic variations on the Airy distribution". In: *Algorithmica* 31, pp. 361–377.

Semple, C. and M. Steel (2003). *Phylogenetics*. Oxford University Press.

Ford, D. J. (2005). *Probabilities on cladograms: introduction to the alpha model*. https://arxiv.org/abs/math/0511246v1.

Drummond, A. J. et al. (2006). "On Sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index". In: *BMC Bioinformatics* 4.88.

Chen, B., D. J. Ford, and M. Winkel (2009). "A new family of Markov branching trees: the alpha-gamma model". In: *Electron. J. Probab.* 14, pp. 400–430.

Cardona, G., A. Mir, and F. Rosselló (2013). "Exact formulas for the variance of several balance indices under the Yule model". In: *Journal of Mathematical Biology* 67, pp. 1833–1846.

- Mir, A., L. Rotger, and F. Rosselló (2013). “A new balance index for phylogenetic trees”. In: *Mathematical Biosciences* 241.1, pp. 125–136.
- Bartoszek, K. (2018a). “Exact and approximate limit behaviour of the Yule tree’s Cophenetic index”. In: *Mathematical Biosciences* 303, pp. 26–45.
- – (2018b). “Limit distribution of the quartet index for Aldous’s $\beta \geq 0$-model”. In: *biorxiv*.
- Fischer, M. (2018). *Extremal values of the Sackin balance index for rooted binary trees*. https://arxiv.org/abs/1801.10418.
- Coronado, T. M., A. Mir, F. Rosselló, and G. Valiente (2019). “A balance index for phylogenetic trees based on rooted quartets”. In: *Journal of Mathematical Biology* 79, pp. 1105–1148.
- Bartoszek, K. et al. (2020). “Squaring within the Colless index yields a better balance index”. In: *arXiv*.
- Coronado, T. M., M. Fischer, et al. (2020). “On the minimum value of the Colless index and the bifurcating trees that achieve it”. In: *Journal of Mathematical Biology* 80, pp. 1993–2054.

Coronado, T. M., A. Mir, F. Rosselló, and L. Rotger (2020). “On Sackin’s original proposal: the variance of the leaves’ depths as a phylogenetic balance index”. In: *BMC Bioinformatics* 21.154.

Tack för idag!

www.liu.se

LINKÖPING
UNIVERSITY