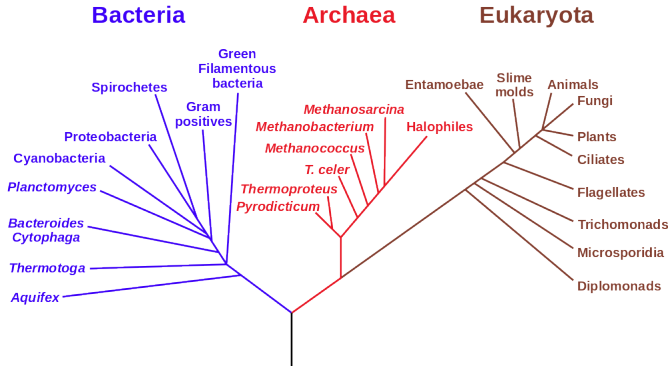


Operads and phylogenetic trees

John Baez and Nina Otter



Trees are used in biology to give a graphic representation of hierarchical relationships such as:

- Evolutionary relationships among species or genes
- Migration patterns of species
- Histories of populations
- Spread of viral infections
- Evolution of language

Trees that represent the evolution of species or genes are called **phylogenetic trees**.

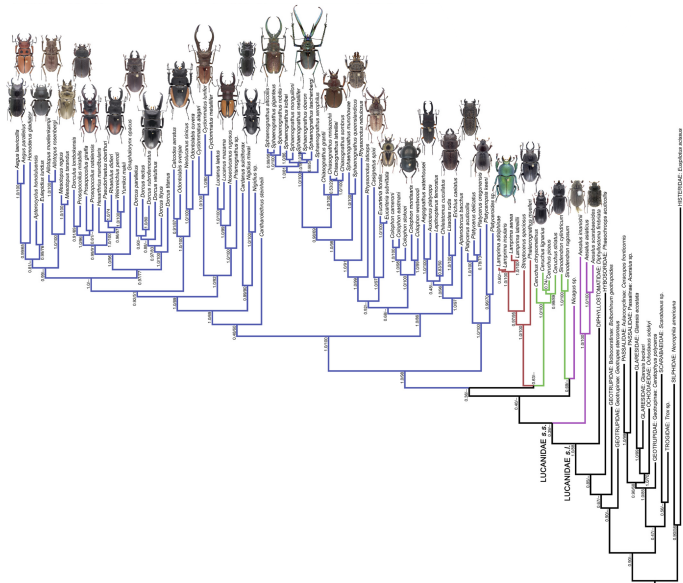
In the past, phylogenetic trees were constructed using morphological similarities among the species.

Trees that represent the evolution of species or genes are called **phylogenetic trees**.

In the past, phylogenetic trees were constructed using morphological similarities among the species.

Nowadays biologists use molecular sequences such as DNA or RNA.

Such trees are usually rooted and semi-labelled:



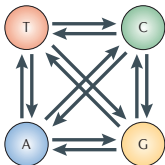
Phylogeny of world stag beetles (Coleoptera: Lucanidae) reveals a Gondwanan origin of Darwins stag beetle, S. I. Kim, B. D. Farrell, 2015

The construction of phylogenetic trees can be subdivided into roughly four steps:

- Alignment of sequences
- Selection of a mathematical model to describe sequence evolution
- Selection of a tree-building method
- Assessment of the quality of the resulting tree

Models of DNA evolution are usually Markov chain models of nucleotide substitution.

The simplest of these models is the Jukes-Cantor model: it assumes that at any time the nucleotides have the same frequency and that they have an equal substitution rate.



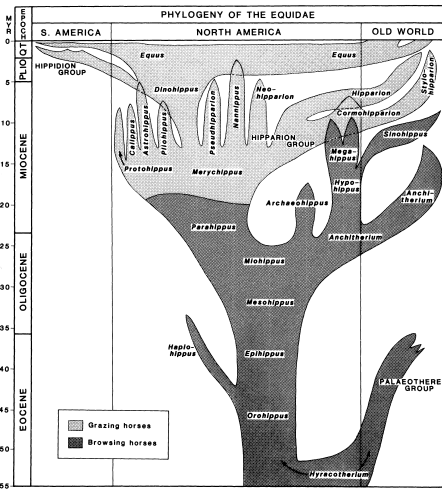
The thickness of the arrows represents the substitution rate of the nucleotides.

The size of the circles represents the frequency at equilibrium.

A simple model for tree reconstruction is the least-square model:

- First compute the evolutionary distances d_{ij} between any two sequences i and j
- Find the tree that minimizes an expression of the difference between the distances d_{ij} and the sum of branch lengths on the tree between the two sequences.

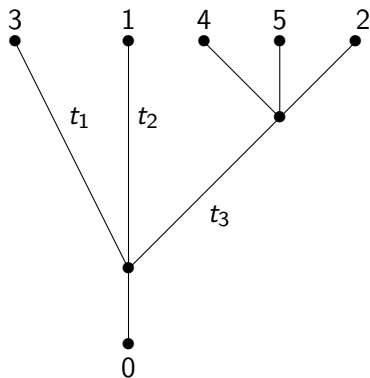
The resulting tree is often just a good approximation: uncertainty about branching order and relationship among leaves is sometimes represented by filling out the uncertain regions of the tree, as e.g.



Patterns of Phylogeny and Rates of Evolution in Fossil Horses, B. MacFadden, 1985

In 2001 Billera, Holmes and Vogtmann put forward a geometric model to parametrize the space of trees with fixed set of leaf labels.

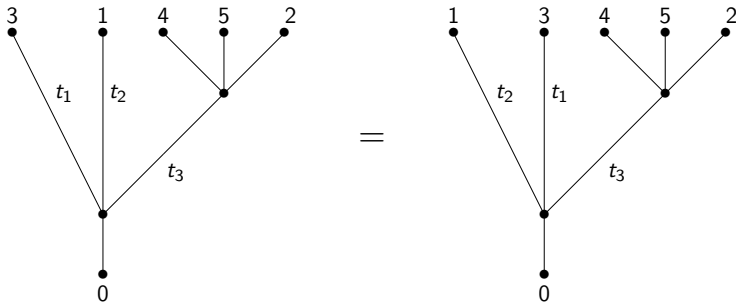
They define a **phylogenetic n -tree** to be a rooted tree with leaves labelled by natural numbers $1, \dots, n$ and internal edges labelled by numbers in $(0, \infty)$, e.g.:



Where $t_1, t_2, t_3 > 0$

The root is labelled 0, the leaves 1, 2, 3, 4, 5.

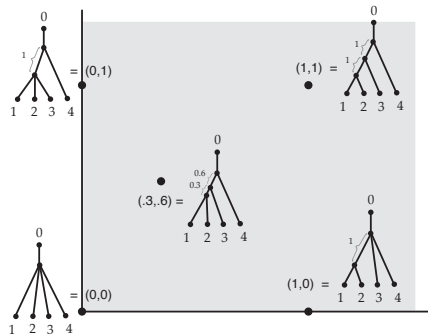
The embedding of the tree in the plane is irrelevant, so these are the same phylogenetic tree:



The space of trees

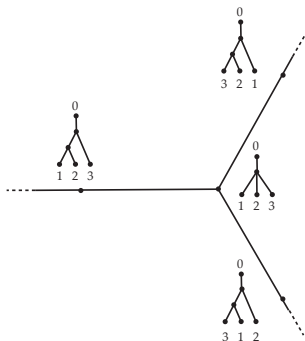
- A tree T with r internal edges is represented by a point in the r -dimensional orthant $(0, \infty)^r$.
- To each other point in the orthant associate the tree that is combinatorially the same as T but with different edge lengths.
- To points on the boundary of the orthant assign trees obtained from T by shrinking some interior edge of T to 0.

For example:



The space $\mathcal{T}_n$ of phylogenetic n -trees is given by taking an $(n - 2)$ -orthant for each binary phylogenetic n -tree and gluing them along their common faces.

For example this is $\mathcal{T}_3$:

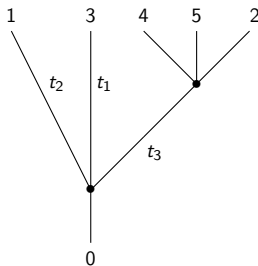


BHV put a metric on this space by taking distances to be shortest piecewise paths between orthants.

BHV put a metric on this space by taking distances to be shortest piecewise paths between orthants.

Note that:

- the trees in BHV have all edges with two adjacent vertices, but it doesn't make a difference to consider external edges with only one adjacent vertex, like this:



- They do not label the external edges; doing that would consist in taking $\mathcal{T}_n \times [0, \infty)^{n+1}$, so all the interesting topology lies in $\mathcal{T}_n$.

It was recently suggested that this space is *the* space for statistical analysis of phylogenetic trees.¹

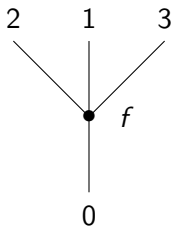
¹Point estimates in phylogenetic reconstructions, P. Benner, M. Bačák, P-Y. Bourguignon, 2014 

It was recently suggested that this space is *the* space for statistical analysis of phylogenetic trees.¹

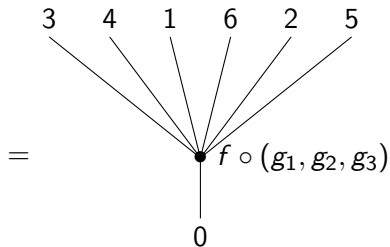
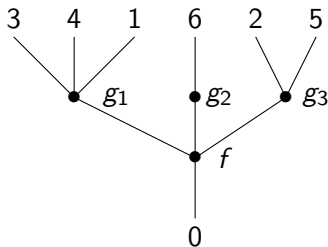
On the other hand, this space arises naturally in operad theory!

¹ *Point estimates in phylogenetic reconstructions*, P. Benner, M. Bačák, P-Y. Bourguignon, 2014

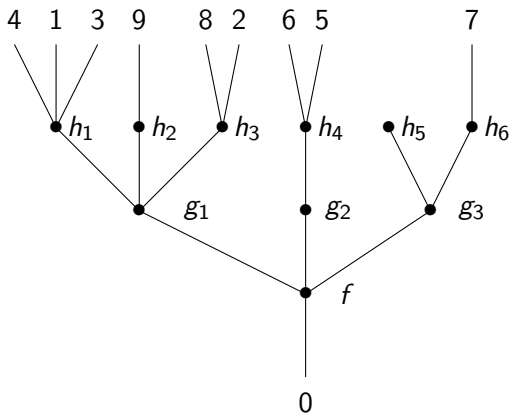
Roughly, an operad O is an algebraic structure where for each n we have a set O_n whose elements are called ' n -ary operations'...



...we can compose operations like this to get new operations...



...and an associative law holds, making this sort of composite unambiguous:



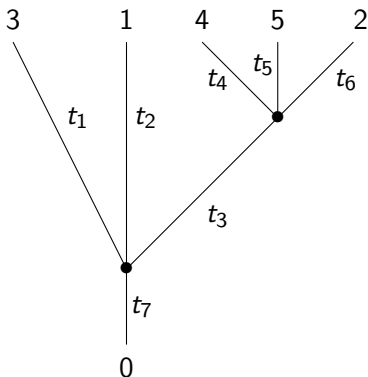
Operads here will always be ‘symmetric’, meaning we have an action of S_n on O_n , compatible with composition.

They will also be ‘topological’, meaning that O_n is a topological space, and composition and permutations act as continuous maps.

Define a **phylogenetic n -tree** to be a rooted tree with leaves labelled by numbers $1, 2, \dots, n$, root labelled by 0 and edges labelled by numbers in $[0, \infty)$. We require that:

- the length of every edge is positive, except perhaps for edges incident to a leaf or the root;
- no 2-ary vertices are allowed.

For example, here is a phylogenetic tree with 5 leaves:



where $t_1, t_2, t_3 > 0$ and $t_4, t_5, t_6, t_7 \geq 0$. The root is labelled 0, the leaves 1, 2, 3, 4, 5.

The embedding of a phylogenetic tree in the plane is irrelevant.

Denote by Phyl_n the space of phylogenetic n -trees.

Proposition.(JB, NO 2014) There is an operad Phyl , the **phylogenetic operad**, whose space of n -ary operations is Phyl_n .

Phyl is the coproduct of Com and $[0, \infty)$, where:

- Com is the operad with Com_n given by the one-point set for $n > 0$ and the empty set for $n = 0$,
- $[0, \infty)$ is the operad having only unary operations, one for each $t \in [0, \infty)$, with composition of operations given by addition.

Denote by Phyl_n the space of phylogenetic n -trees.

Proposition.(JB, NO 2014) There is an operad Phyl , the **phylogenetic operad**, whose space of n -ary operations is Phyl_n .

Phyl is the coproduct of Com and $[0, \infty)$, where:

- Com is the operad with Com_n given by the one-point set for $n > 0$ and the empty set for $n = 0$,
- $[0, \infty)$ is the operad having only unary operations, one for each $t \in [0, \infty)$, with composition of operations given by addition.

How is this related to the BHV space of trees $\mathcal{T}_n$?

The coproduct of two operads O and O' is an operad whose n -ary operations are equivalence classes of rooted trees with n -leaves and vertices labelled by operations in O and O' .

The coproduct of two operads O and O' is an operad whose n -ary operations are equivalence classes of rooted trees with n -leaves and vertices labelled by operations in O and O' .

If we endow the BHV space of trees with the topology induced by their metric and Phyl_n with the topology of the coproduct, we have:

Proposition.(JB NO 2015) For all $n > 1$ there is a homeomorphism

$$\text{Phyl}_n \cong \mathcal{T}_n \times [0, \infty)^{n+1}$$

The coproduct of two operads O and O' is an operad whose n -ary operations are equivalence classes of rooted trees with n -leaves and vertices labelled by operations in O and O' .

If we endow the BHV space of trees with the topology induced by their metric and Phyl_n with the topology of the coproduct, we have:

Proposition.(JB NO 2015) For all $n > 1$ there is a homeomorphism

$$\text{Phyl}_n \cong \mathcal{T}_n \times [0, \infty)^{n+1}$$

But there is more to this story:

Operads are interesting because they have algebras.

Operads are interesting because they have algebras.

An **algebra over an operad** O is a topological space A together with a continuous map

$$f: A^n \rightarrow A$$

for every $f \in O_n$. These maps satisfy some associativity and unitality conditions.

Operads are interesting because they have algebras.

An **algebra over an operad** O is a topological space A together with a continuous map

$$f: A^n \rightarrow A$$

for every $f \in O_n$. These maps satisfy some associativity and unitality conditions.

These conditions simply say that there is an operad morphism

$$\alpha: O \rightarrow \mathbf{End}(A)$$

where $\mathbf{End}(A)$ is the operad whose n -ary operations are maps $A^n \rightarrow A$.

Operads are interesting because they have algebras.

An **algebra over an operad** O is a topological space A together with a continuous map

$$f: A^n \rightarrow A$$

for every $f \in O_n$. These maps satisfy some associativity and unitality conditions.

These conditions simply say that there is an operad morphism

$$\alpha: O \rightarrow \mathbf{End}(A)$$

where $\mathbf{End}(A)$ is the operad whose n -ary operations are maps $A^n \rightarrow A$.

More generally we can consider an algebra of O in any symmetric monoidal category C .

Operads are interesting because they have algebras.

An **algebra over an operad** O is a topological space A together with a continuous map

$$f: A^n \rightarrow A$$

for every $f \in O_n$. These maps satisfy some associativity and unitality conditions.

These conditions simply say that there is an operad morphism

$$\alpha: O \rightarrow \mathbf{End}(A)$$

where $\mathbf{End}(A)$ is the operad whose n -ary operations are maps $A^n \rightarrow A$.

More generally we can consider an algebra of O in any symmetric monoidal category C .

A **coalgebra** of O in C is an algebra of O in C^{op} .

A **continuous time Markov process** on a finite set X is a collection of linear maps

$$\alpha(t): \mathbb{R}^X \rightarrow \mathbb{R}^X \quad t \in [0, \infty)$$

such that:

- ① $\alpha(t+s) = \alpha(s) \circ \alpha(t)$ for all $t, s \in [0, \infty)$
- ② $\alpha(0) = id_{\mathbb{R}^X}$
- ③ α depends continuously on t
- ④ if $\nu \in \mathbb{R}^X$ is a probability distribution on X , so is $\alpha(t)(\nu)$ for all t .

A **continuous time Markov process** on a finite set X is a collection of linear maps

$$\alpha(t): \mathbb{R}^X \rightarrow \mathbb{R}^X \quad t \in [0, \infty)$$

such that:

- ① $\alpha(t+s) = \alpha(s) \circ \alpha(t)$ for all $t, s \in [0, \infty)$
- ② $\alpha(0) = id_{\mathbb{R}^X}$
- ③ α depends continuously on t
- ④ if $\nu \in \mathbb{R}^X$ is a probability distribution on X , so is $\alpha(t)(\nu)$ for all t .

In other words, $\mathbb{R}^X$ is an algebra over $[0, \infty)$ in Top satisfying condition (4).

A **continuous time Markov process** on a finite set X is a collection of linear maps

$$\alpha(t): \mathbb{R}^X \rightarrow \mathbb{R}^X \quad t \in [0, \infty)$$

such that:

- 1 $\alpha(t+s) = \alpha(s) \circ \alpha(t)$ for all $t, s \in [0, \infty)$
- 2 $\alpha(0) = id_{\mathbb{R}^X}$
- 3 α depends continuously on t
- 4 if $\nu \in \mathbb{R}^X$ is a probability distribution on X , so is $\alpha(t)(\nu)$ for all t .

In other words, $\mathbb{R}^X$ is an algebra over $[0, \infty)$ in Top satisfying condition (4).

Since addition in $[0, \infty)$ is commutative, this is the same as a coalgebra over $[0, \infty)$ satisfying condition (4).

Algebras over $\mathbf{Com}$ are the topological commutative semigroups.

Algebras over $\mathbf{Com}$ are the topological commutative semigroups.

Coalgebras over $\mathbf{Com}$ are topological cocommutative coalgebras, not necessarily with counit.

Algebras over $\mathbf{Com}$ are the topological commutative semigroups.

Coalgebras over $\mathbf{Com}$ are topological cocommutative coalgebras, not necessarily with counit.

$\mathbb{R}^X$ is naturally a coalgebra over $\mathbf{Com}$: the unique n -ary operation μ_n acts by sending $v \in \mathbb{R}^X$ to $(v, \dots, v)$ in $\mathbb{R}^X \times \dots \times \mathbb{R}^X$.

Algebras over $\mathbf{Com}$ are the topological commutative semigroups.

Coalgebras over $\mathbf{Com}$ are topological cocommutative coalgebras, not necessarily with counit.

$\mathbb{R}^X$ is naturally a coalgebra over $\mathbf{Com}$: the unique n -ary operation μ_n acts by sending $v \in \mathbb{R}^X$ to $(v, \dots, v)$ in $\mathbb{R}^X \times \dots \times \mathbb{R}^X$.

By general abstract non-sense a coalgebra over $\mathbf{Com} + [0, \infty)$ is a topological space A that is both a coalgebra of $\mathbf{Com}$ and $[0, \infty)$.

Algebras over $\mathbf{Com}$ are the topological commutative semigroups.

Coalgebras over $\mathbf{Com}$ are topological cocommutative coalgebras, not necessarily with counit.

$\mathbb{R}^X$ is naturally a coalgebra over $\mathbf{Com}$: the unique n -ary operation μ_n acts by sending $v \in \mathbb{R}^X$ to $(v, \dots, v)$ in $\mathbb{R}^X \times \dots \times \mathbb{R}^X$.

By general abstract nonsense a coalgebra over $\mathbf{Com} + [0, \infty)$ is a topological space A that is both a coalgebra of $\mathbf{Com}$ and $[0, \infty)$.

We thus obtain:

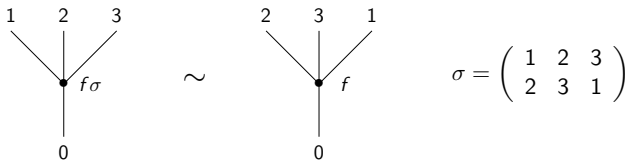
Proposition. (JB, NO 2015) Given a Markov process on a finite set X , the vector space $\mathbb{R}^X$ naturally becomes a coalgebra of $\mathbf{Phyl}$.

Furthermore, Phyl is closely related to the W -construction of Boardman and Vogt applied to the operad Com .

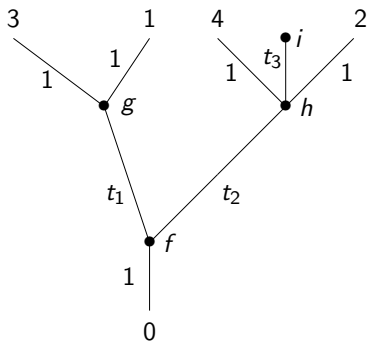
Furthermore, Phyl is closely related to the W -construction of Boardman and Vogt applied to the operad $\mathbf{Com}$.

For any operad O operations in $W(O)_n$ are equivalence classes of rooted planar trees with internal edges labelled by numbers in $(0, 1]$ and external edges labelled by 1, vertices labelled by operations in O .

The equivalence relation comes from permuting edges coming into a vertex, e.g.:



Here is an operation in $W(O)_4$:



where $t_1, t_2, t_3 > 0$.

Composition is given by grafting trees and assigning length 1 to the new edges arising.

Composition is given by grafting trees and assigning length 1 to the new edges arising.

$W(O)_n$ has an obvious topology, induced by the topology of the spaces in O and the topology of $[0, 1]$.

Berger and Moerdijk showed that if O is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of O in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

Berger and Moerdijk showed that if O is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of O in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(\text{Com})$ is not cofibrant, because the action of the symmetric group S_n on $W(\text{Com})_n$ is trivial.

Berger and Moerdijk showed that if O is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of O in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(\text{Com})$ is not cofibrant, because the action of the symmetric group S_n on $W(\text{Com})_n$ is trivial.

Algebras over $W(\text{Com})$ are commutative A_∞ algebras: associativity holds up to coherent homotopy, but commutativity is strict.

Berger and Moerdijk showed that if O is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of O in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(\text{Com})$ is not cofibrant, because the action of the symmetric group S_n on $W(\text{Com})_n$ is trivial.

Algebras over $W(\text{Com})$ are commutative A_∞ algebras: associativity holds up to coherent homotopy, but commutativity is strict.

However, $W(\text{Com})$ is still interesting:

Consider the closed interval $[0, \infty]$ and define addition on $[0, \infty]$ in the obvious way, where

$$\infty + t = t + \infty = \infty.$$

Then $[0, \infty]$ becomes a topological monoid, so there is an operad with only unary operations, one for each $t \in [0, \infty]$. Let's call this operad $[0, \infty]$.

Consider the closed interval $[0, \infty]$ and define addition on $[0, \infty]$ in the obvious way, where

$$\infty + t = t + \infty = \infty.$$

Then $[0, \infty]$ becomes a topological monoid, so there is an operad with only unary operations, one for each $t \in [0, \infty]$. Let's call this operad $[0, \infty]$.

Instead of labelling edges of trees by numbers in $[0, 1]$ we can adapt the Boardman Vogt resolution by labelling edges by numbers in $[0, \infty]$.

Consider the closed interval $[0, \infty]$ and define addition on $[0, \infty]$ in the obvious way, where

$$\infty + t = t + \infty = \infty.$$

Then $[0, \infty]$ becomes a topological monoid, so there is an operad with only unary operations, one for each $t \in [0, \infty]$. Let's call this operad $[0, \infty]$.

Instead of labelling edges of trees by numbers in $[0, 1]$ we can adapt the Boardman Vogt resolution by labelling edges by numbers in $[0, \infty]$.

We have:

Proposition. (JB, NO 2015) For any operad O there is a weak equivalence $W(O) \rightarrow O + [0, \infty]$ and the spaces underlying the operads are homotopy equivalent.

Consider the closed interval $[0, \infty]$ and define addition on $[0, \infty]$ in the obvious way, where

$$\infty + t = t + \infty = \infty.$$

Then $[0, \infty]$ becomes a topological monoid, so there is an operad with only unary operations, one for each $t \in [0, \infty]$. Let's call this operad $[0, \infty]$.

Instead of labelling edges of trees by numbers in $[0, 1]$ we can adapt the Boardman Vogt resolution by labelling edges by numbers in $[0, \infty]$.

We have:

Proposition. (JB, NO 2015) For any operad O there is a weak equivalence $W(O) \rightarrow O + [0, \infty]$ and the spaces underlying the operads are homotopy equivalent.

Moreover, for every $n > 1$ there is a homeomorphism $W(\text{Com})_n \cong \mathcal{T}_n$.

And the larger operad $\text{Com} + [0, \infty]$, a compactification of $\text{Phyl} = \text{Com} + [0, \infty)$, is also interesting.

And the larger operad $\text{Com} + [0, \infty]$, a compactification of $\text{Phyl} = \text{Com} + [0, \infty)$, is also interesting.

The reason is that any Markov process $\alpha: [0, \infty) \rightarrow \text{End}(\mathbb{R}^X)$ approaches a limit as $t \rightarrow \infty$. Indeed, it extends uniquely to a homomorphism of topological monoids $\alpha: [0, \infty] \rightarrow \text{End}(\mathbb{R}^X)$.

And the larger operad $\text{Com} + [0, \infty]$, a compactification of $\text{Phyl} = \text{Com} + [0, \infty)$, is also interesting.

The reason is that any Markov process $\alpha: [0, \infty) \rightarrow \text{End}(\mathbb{R}^X)$ approaches a limit as $t \rightarrow \infty$. Indeed, it extends uniquely to a homomorphism of topological monoids $\alpha: [0, \infty] \rightarrow \text{End}(\mathbb{R}^X)$.

We thus get:

Corollary. (JB NO 2015) Given a Markov process on a finite set X , the vector space $\mathbb{R}^X$ naturally becomes a coalgebra of $\text{Com} + [0, \infty]$.

Summary for topologists who don't care about applications:

For any operad O we have weak equivalences:

$$\begin{array}{ccc} O & \xrightarrow{\quad} & O + [0, \infty) \\ \uparrow & \searrow & \downarrow \\ W(O) & \xrightarrow{\quad} & O + [0, \infty] \end{array}$$

In particular, for $O = \text{Com}$ we have:

$$\begin{array}{ccccc}
 \text{Com} & \longrightarrow & \text{Com} + [0, \infty) & \longrightarrow & \text{Coend}(\mathbb{R}^X) \\
 \uparrow & & \downarrow & \nearrow & \\
 W(\text{Com}) & \longrightarrow & \text{Com} + [0, \infty] & &
 \end{array}$$

Finally: *the mystery of tropical trees!*



The **tropical semiring** is $\mathbb{R} \cup \{-\infty\}$ with addition given by taking the maximum and multiplication given by the classical addition. We can do algebraic geometry over this semiring and define ‘tropical curves’.

In 2007, Gathmann, Kerber and Markwig showed that a certain moduli space of genus 0 tropical curves with $n + 1$ marked points is the space of trees $\mathcal{T}_n$ studied by Billera, Holmes and Vogtmann.

Also in 2007, Mikhalkin showed this moduli space has a compactification that is a smooth compact tropical variety.

Proposition. (NO 2014) This compactification is $W(\text{Com})_n$. The operad structure on $W(\text{Com})$ corresponds to the ‘tropical clutching map’ described by Abramovich, Caporaso and Payne in 2012.

**The mystery: why are tropical curves related to phylogenetic trees?
Is this connection good for something?**

- *Geometry of the space of phylogenetic trees*, Louis J. Billera, Susan P. Holmes, Karen Vogtmann, 2001
- *Operads and phylogenetic trees*, John Baez and Nina Otter , 2015 (soon on the arXiv!)
- *The Boardman Vogt resolution and tropical moduli spaces*, Nina Otter, Master Thesis, 2014
- *The Boardman-Vogt resolution of operads in monoidal model categories*, Clemens Berger and Ieke Moerdijk, 2005
- *Tropical fans and the moduli spaces of tropical curves*, Andreas Gathmann, Michael Kerber and Hannah Markwig, 2007
- *Moduli spaces of rational tropical curves.*, Grigory Mikhalkin, 2007