Ranked Tree Shapes and the Future Loss of Phylogenetic Diversity

Amaury Lambert

joint works with M. Steel, F. Gascuel and O. Maliet

June 26, 2018
Luminy

Outline

Loss of Phylogenetic Diversity

- Q : "What fraction of the underlying evolutionary history survives when k of n species in a taxon are lost ?" (Nee & May 1997)

Loss of Phylogenetic Diversity

- Q : "What fraction of the underlying evolutionary history survives when *k* of *n* species in a taxon are lost ?" (Nee & May 1997)

- Phylogenies as metric trees, carry a footprint of evolutionary history

Loss of Phylogenetic Diversity

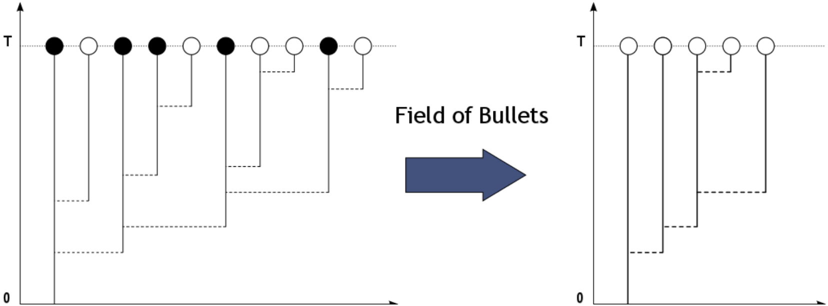
- Q : "What fraction of the underlying evolutionary history survives when *k* of *n* species in a taxon are lost ?" (Nee & May 1997)
- Phylogenies as metric trees, carry a footprint of evolutionary history
- *Phylogenetic Diversity PD* $=$ Total Length of Tree (Faith 1992)

Loss of Phylogenetic Diversity

- Q : "What fraction of the underlying evolutionary history survives when *k* of *n* species in a taxon are lost ?" (Nee & May 1997)
- Phylogenies as metric trees, carry a footprint of evolutionary history
- *Phylogenetic Diversity PD* $=$ Total Length of Tree (Faith 1992)
- Q becomes : "Can we predict how much Phylogenetic Diversity will remain in the face of present extinctions ?"

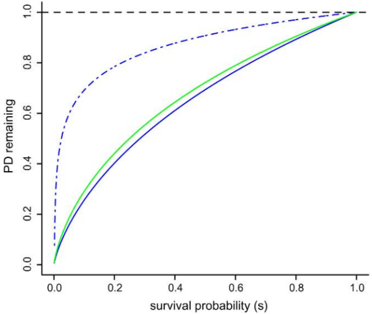
Field of Bullets Model (FoB)

- Take a phylogeny : tips = species
- Paint independently each tip in white w probability p, in black w probability $1 - p$
- White dot = extant/sampled
- Black dot = extinct/not sampled



The Loss of Phylogenetic Diversity

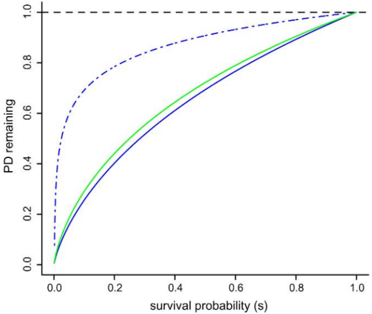
- Field of Bullets model : each species is removed independently, kept with probability *p*



(Mooers, Gascuel, Stadler, Li, Steel 2011)

The Loss of Phylogenetic Diversity

- Field of Bullets model : each species is removed independently, kept with probability p
- Remaining PD $S(p)$ = total length of tree spanned by extant/sampled species



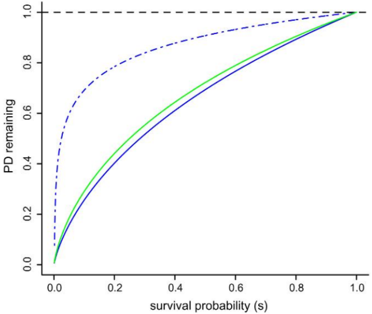
(Mooers, Gascuel, Stadler, Li, Steel 2011)

The Loss of Phylogenetic Diversity

- Field of Bullets model : each species is removed independently, kept with probability p
- Remaining PD $S(p)$ = total length of tree spanned by extant/sampled species
- For a given tree, $\mathbb{E}S(p)$ is increasing and concave (Faller, Pardi, Steel 2008)

$$\mathbb{E}S(p) = \sum_e \ell(e)\left(1-(1-p)^{n(e)}\right)$$

where :
$\ell(e)$ = length of edge e,
$n(e)$ = # tips descending from e



(Mooers, Gascuel, Stadler, Li, Steel 2011)

Nee & May *Science* 1997

Extinction and the Loss of Evolutionary History

Sean Nee* and Robert M. May

Extinction episodes, such as the anthropogenic one currently under way, result in a pruned tree of life. But what fraction of the underlying evolutionary history survives when k of n species in a taxon are lost? This is relevant both to how species loss has translated into a loss of evolutionary history and to assigning conservation priorities. Here it is shown that approximately 80 percent of the underlying tree of life can survive even when approximately 95 percent of species are lost, and that algorithms that maximize the amount of evolutionary history preserved are not much better than choosing the survivors at random. Given the political, economic, and social realities constraining conservation biology, these findings may be helpful.

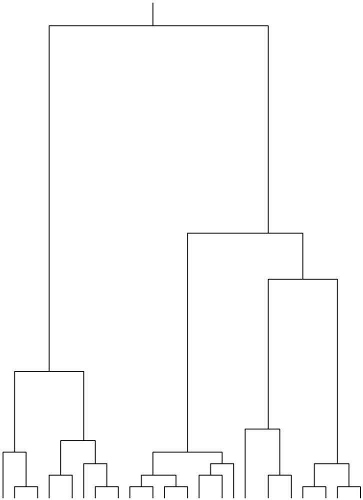
Nee & May *Science* 1997

"Approximately 80 percent of the underlying tree of life can survive even when approximately 95 percent of species are lost"

Rule of thumb :
$\mathbb{E}S_n(1) = \sum_{k=2}^{n} \frac{k}{\binom{k}{2}} \sim 2\log(n)$ so

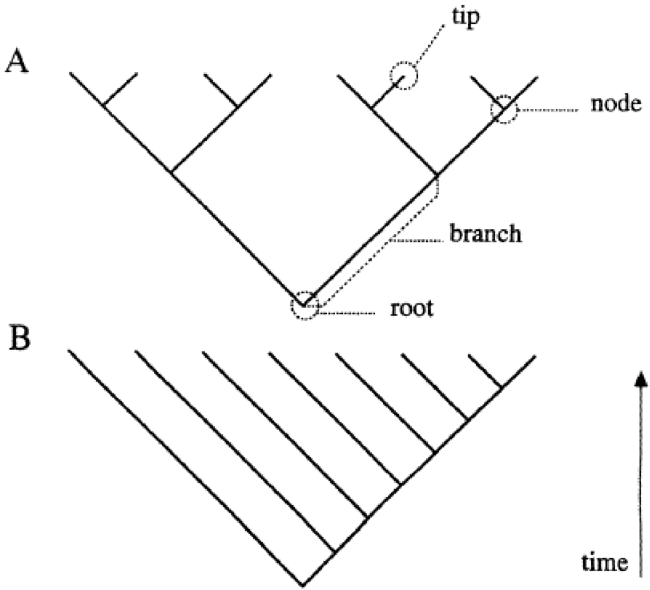
$$\frac{S_n(p)}{S_n(1)} \approx \frac{\log(pn)}{\log(n)} \approx 1$$

1. Field of Bullets
2. Very Small External Edges
3. Balanced



The Kingman Coalescent

Perfectly Balanced Tree (A) vs Caterpillar Tree (B)



Loss of PD in Random Trees

Remaining PD is...

- Low in imbalanced trees : more ‘distinctive’ sp

Loss of PD in Random Trees

Remaining PD is...

- Low in imbalanced trees : more ‘distinctive’ sp
- High for the Kingman coalescent (Nee & May *Science* 1997)

Loss of PD in Random Trees

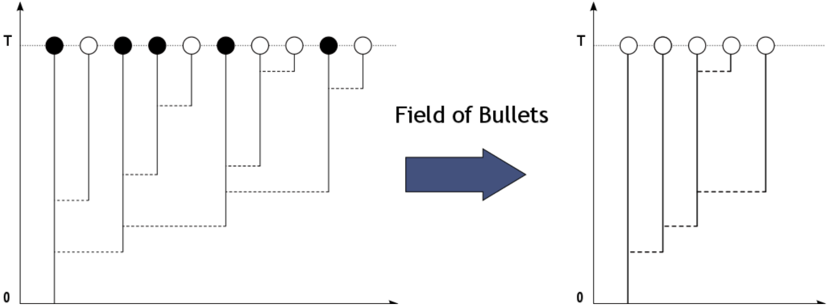
Remaining PD is...

- Low in imbalanced trees : more 'distinctive' sp
- High for the Kingman coalescent (Nee & May *Science* 1997)
- Lower for the Yule tree (Mooers, Gascuel, Stadler, Li, Steel *Syst Biol* 2011) :

$$\frac{\mathbb{E}S(p)}{\mathbb{E}S(1)} = \text{Ratio of expected remaining PD-to-Old PD} \approx -\frac{p \log p}{1-p}$$

Field of Bullets on a Birth-Death Tree

In a birth-death process stopped at time T, the **reduced tree** is a coalescent point process (CPP) : **node depths** are **i.i.d.**



Remaining PD for general Birth-Death Trees (1)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

- CPP (e.g., reconstructed birth-death tree) : **node depths** H_i are **i.i.d.**

Remaining PD for general Birth-Death Trees (1)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

- CPP (e.g., reconstructed birth-death tree) : **node depths** H_i are **i.i.d.**

- Conditional on n tips **before** FoB,

$$\lim_n \frac{S_n(p)}{S_n(1)} = p\,\frac{\mathbb{E}(B)}{\mathbb{E}(H)}$$

where

$$B := \max_{i=1,\dots,G} H_i,$$

and G is a geometric r.v. with success probability p.

Remaining PD for general Birth-Death Trees (1)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

- CPP (e.g., reconstructed birth-death tree) : **node depths** H_i are **i.i.d.**
- Conditional on n tips **before** FoB,

$$\lim_n \frac{S_n(p)}{S_n(1)} = p\,\frac{\mathbb{E}(B)}{\mathbb{E}(H)}$$

  where

$$B := \max_{i=1,\dots,G} H_i,$$

  and G is a geometric r.v. with success probability p.

- Simple argument :

Remaining PD for general Birth-Death Trees (1)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

- CPP (e.g., reconstructed birth-death tree) : **node depths** H_i are **i.i.d.**

- Conditional on n tips **before** FoB,

$$\lim_n \frac{S_n(p)}{S_n(1)} = p\,\frac{\mathbb{E}(B)}{\mathbb{E}(H)}$$

  where

$$B := \max_{i=1,\dots,G} H_i,$$

  and G is a geometric r.v. with success probability p.

- Simple argument :
  - After FoB, the phylogenetic tree is a CPP with node depth B and K_n tips

Remaining PD for general Birth-Death Trees (1)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

- CPP (e.g., reconstructed birth-death tree) : **node depths** H_i are **i.i.d.**
- Conditional on n tips **before** FoB,

$$\lim_n \frac{S_n(p)}{S_n(1)} = p\,\frac{\mathbb{E}(B)}{\mathbb{E}(H)}$$

  where

$$B := \max_{i=1,\dots,G} H_i,$$

  and G is a geometric r.v. with success probability p.

- Simple argument :
  - After FoB, the phylogenetic tree is a CPP with node depth B and K_n tips
  - By the SLLN, $S_n(1) \sim n\mathbb{E}(H)$ and $S_n(p) \sim K_n\mathbb{E}(B)$,

Remaining PD for general Birth-Death Trees (1)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

- CPP (e.g., reconstructed birth-death tree) : **node depths** H_i are **i.i.d.**
- Conditional on n tips **before** FoB,

$$\lim_n \frac{S_n(p)}{S_n(1)} = p\,\frac{\mathbb{E}(B)}{\mathbb{E}(H)}$$

  where

$$B := \max_{i=1,\dots,G} H_i,$$

  and G is a geometric r.v. with success probability p.

- Simple argument :
  - After FoB, the phylogenetic tree is a CPP with node depth B and K_n tips
  - By the SLLN, $S_n(1) \sim n\mathbb{E}(H)$ and $S_n(p) \sim K_n\mathbb{E}(B)$,
  - Conclude with $K_n/n \to p$.

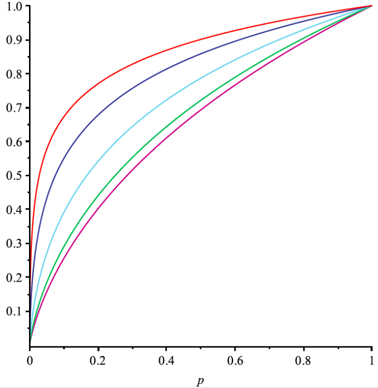
Remaining PD for general Birth-Death Trees (2)

Lambert & Steel "Predicting the Loss of Phylogenetic Diversity under Non-Stationary Diversification Models" *JTB* 2013

For a Birth-Death tree with sp rate b, ext rate d, div rate $r := b - d$

Remaining PD-to-old PD Ratio $= S(p)/S(1)$

$$= \begin{cases} \frac{dp}{bp-r}\frac{\ln(bp/r)}{\ln(b/r)} & \text{if } b > r \neq bp \\ -\frac{p\ln(p)}{1-p} & \text{if } b = r \\ -\frac{1-p}{\ln(p)} & \text{if } b > r = bp \end{cases}$$



Right : Slow progression towards the unit step function (from pure birth to critical) : $d/b = 0$ (the lowest curve) and then $d/b = 0.5, 0.9, 0.99, 0.999$.

Remaining Questions

1. What if trees are imbalanced ?
   $=\exists$ small clades $=$ unprotected nodes

Remaining Questions

1. What if trees are imbalanced?
   $=\exists$ small clades $=$ unprotected nodes

2. What if older clades are more species-poor?
   $=$ deep nodes are in small clades

Remaining Questions

1. What if trees are imbalanced?
   $=\exists$ small clades $=$ unprotected nodes

2. What if older clades are more species-poor?
   $=$ deep nodes are in small clades

3. What if older clades carry more extinct-prone species?
   $=$ deep nodes are in first hit clades

Remaining Questions

1. What if trees are imbalanced?
   $=\exists$ small clades $=$ unprotected nodes
2. What if older clades are more species-poor?
   $=$ deep nodes are in small clades
3. What if older clades carry more extinct-prone species?
   $=$ deep nodes are in first hit clades

Remaining Questions

1. What if trees are imbalanced ?
   $=\exists$ small clades $=$ unprotected nodes
2. What if older clades are more species-poor ?
   $=$ deep nodes are in small clades
3. What if older clades carry more extinct-prone species ?
   $=$ deep nodes are in first hit clades

$\Longrightarrow$ We need random tree models able to tune jointly :

Remaining Questions

1. What if trees are imbalanced ?
   $=\exists$ small clades $=$ unprotected nodes
2. What if older clades are more species-poor ?
   $=$ deep nodes are in small clades
3. What if older clades carry more extinct-prone species ?
   $=$ deep nodes are in first hit clades

$\Longrightarrow$ We need random tree models able to tune jointly :

1. Tree shape $\Longrightarrow \beta =$ tree imbalance

Remaining Questions

1. What if trees are imbalanced ?
   $=\exists$ small clades $=$ unprotected nodes

2. What if older clades are more species-poor ?
   $=$ deep nodes are in small clades

3. What if older clades carry more extinct-prone species ?
   $=$ deep nodes are in first hit clades

$\Longrightarrow$ We need random tree models able to tune jointly :

1. Tree shape $\Longrightarrow \beta =$ tree imbalance

2. Node depths' rankings $\Longrightarrow \alpha =$ age-richness index

Remaining Questions

1. What if trees are imbalanced ?
   $=\exists$ small clades $=$ unprotected nodes
2. What if older clades are more species-poor ?
   $=$ deep nodes are in small clades
3. What if older clades carry more extinct-prone species ?
   $=$ deep nodes are in first hit clades

$\Longrightarrow$ We need random tree models able to tune jointly :

1. Tree shape $\Longrightarrow \beta =$ tree imbalance
2. Node depths' rankings $\Longrightarrow \alpha =$ age-richness index
3. Abundances at tips $\Longrightarrow \eta =$ abundance-richness index

Outline

Aldous' program

Aldous "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today" *Statist. Sci. 2001*

1. What is a useful way to describe balance and imbalance in a general phylogenetic tree ?

Aldous' program

Aldous "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today" *Statist. Sci. 2001*

1. What is a useful way to describe balance and imbalance in a general phylogenetic tree ?

2. Is there some particular region of the balance-imbalance spectrum containing most actual phylogenetic trees ?

Aldous' program

Aldous "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today" *Statist. Sci. 2001*

1. What is a useful way to describe balance and imbalance in a general phylogenetic tree ?

2. Is there some particular region of the balance-imbalance spectrum containing most actual phylogenetic trees ?

3. If so, is there some mathematically simple and biologically plausible stochastic model for phylogenetic trees whose realizations mimic actual trees ?

Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

- We are given distributions q_n on $[n-1] = \{1, \ldots, n-1\}, \forall n \geq 2$

Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

- We are given distributions q_n on $[n-1] = \{1, \dots, n-1\}, \forall n \geq 2$
- Recursively split each k-subset of $[n]$ according to q_k independently :

Aldous' Markov branching model on binary tree shapes

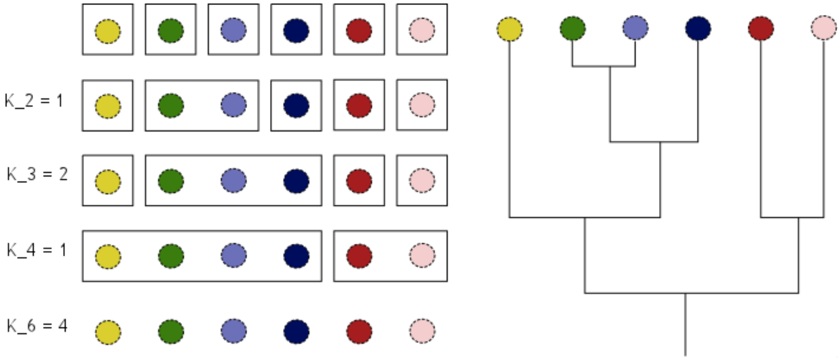
Aldous (1996, 2001)

- We are given distributions q_n on $[n-1] = \{1, \ldots, n-1\}, \forall n \geq 2$
- Recursively split each k-subset of $[n]$ according to q_k independently :

Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

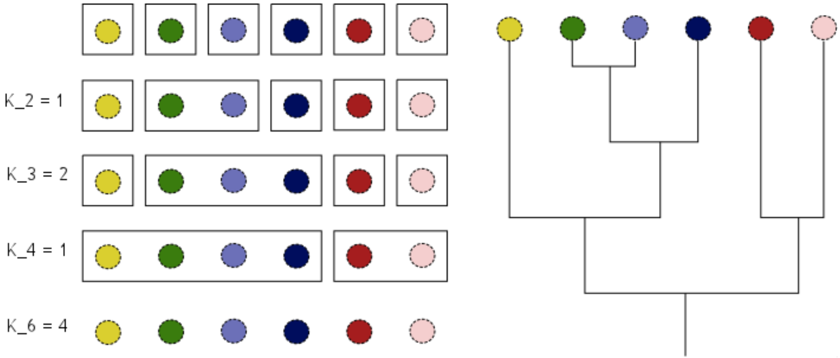
- We are given distributions q_n on $[n-1] = \{1, \ldots, n-1\}, \forall n \geq 2$
- Recursively split each k-subset of $[n]$ according to q_k independently :



Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

- We are given distributions q_n on $[n-1] = \{1, \ldots, n-1\}, \forall n \geq 2$
- Recursively split each k-subset of $[n]$ according to q_k independently :

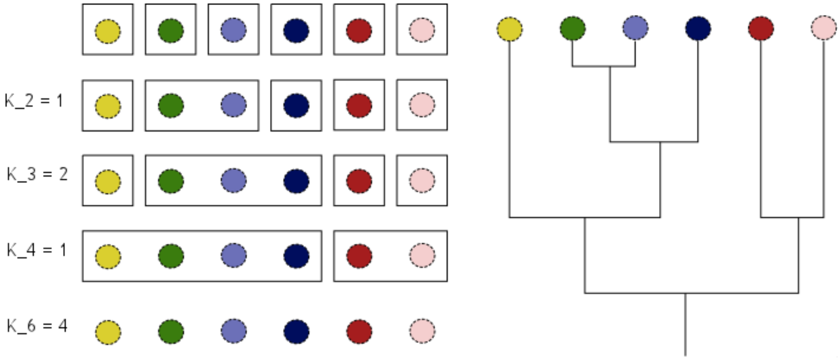


- Induces a law on **binary tree shapes with n labelled leaves**.

Aldous' Markov branching model on binary tree shapes

Aldous (1996, 2001)

- We are given distributions q_n on $[n-1] = \{1, \dots, n-1\}, \forall n \geq 2$
- Recursively split each k-subset of $[n]$ according to q_k independently :



- Induces a law on **binary tree shapes with n labelled leaves**.
- q_n uniform yields the same tree shape as the reduced tree of any birth–death process (e.g., Yule tree)

Sampling consistency

- A tree model is a family of probability distributions (P_n) on (exchangeably labelled) tree shapes with n tips

Sampling consistency

- A tree model is a family of probability distributions (P_n) on (exchangeably labelled) tree shapes with n tips
- Call T_n a random tree with law P_n

Sampling consistency

- A tree model is a family of probability distributions (P_n) on (exchangeably labelled) tree shapes with n tips
- Call T_n a random tree with law P_n
- Call T'_n the tree obtained by removing one tip from T_{n+1} (say the tip labelled $n + 1$)

Sampling consistency

- A tree model is a family of probability distributions (P_n) on (exchangeably labelled) tree shapes with n tips
- Call T_n a random tree with law P_n
- Call T'_n the tree obtained by removing one tip from T_{n+1} (say the tip labelled $n+1$)
- The model is said **sampling consistent** if T_n and T'_n have the same distribution.

Sampling consistency

- A tree model is a family of probability distributions (P_n) on (exchangeably labelled) tree shapes with n tips
- Call T_n a random tree with law P_n
- Call T'_n the tree obtained by removing one tip from T_{n+1} (say the tip labelled $n+1$)
- The model is said **sampling consistent** if T_n and T'_n have the same distribution.
- Example : Kingman coalescent.

Aldous' Markov branching model (Mbm)

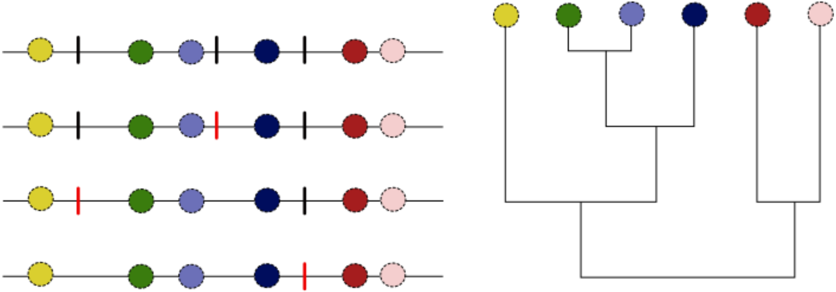
Theorem (Haas et al 2008, Lambert 2017)

A Mbm is sampling-consistent iff it there is a symmetric measure ν on $(0,1)$ s.t.

$$q_n(i) = a_n(\nu)^{-1} \binom{n}{i} \int_0^1 x^i (1-x)^{n-i} \nu(dx)$$

Construction

- Color dots are uniformly distributed in the interval
- Intervals are fragmented by r.v. R with law $\sim \nu$ (red dashes)



Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...

Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...
- ...and exchangeable = invariant by permutations ($\exists$ block frequencies) + distinct blocks have independent futures.

Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...
- ...and exchangeable = invariant by permutations ($\exists$ block frequencies) + distinct blocks have independent futures.
- Π is homogeneous if for any block B of $\Pi(0)$ and for any bijection $\varphi : B \to \mathbb{N}$,

$$(\varphi(\Pi_{|B}(t)); t \geq 0) \overset{\mathcal{L}}{=} (\Pi(t); t \geq 0).$$

Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...
- ...and exchangeable = invariant by permutations ($\exists$ block frequencies) + distinct blocks have independent futures.
- Π is homogeneous if for any block B of $\Pi(0)$ and for any bijection $\varphi : B \to \mathbb{N}$,

$$(\varphi(\Pi_{|B}(t)); t \geq 0) \overset{\mathcal{L}}{=} (\Pi(t); t \geq 0).$$

- Then $\Pi_{|[n]}$ is Markov $\forall n$

Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...
- ...and exchangeable = invariant by permutations ($\exists$ block frequencies) + distinct blocks have independent futures.
- Π is homogeneous if for any block B of $\Pi(0)$ and for any bijection $\varphi : B \to \mathbb{N}$,
$$(\varphi(\Pi_{|B}(t)); t \geq 0) \stackrel{\mathcal{L}}{=} (\Pi(t); t \geq 0).$$
- Then $\Pi_{|[n]}$ is Markov $\forall n$
- If Π is a binary, homogeneous fragmentation process, it is characterized by the fragmentation measure ν on $(0, 1)$, where...

Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...
- ...and exchangeable = invariant by permutations ($\exists$ block frequencies) + distinct blocks have independent futures.
- Π is homogeneous if for any block B of $\Pi(0)$ and for any bijection $\varphi : B \to \mathbb{N}$,
$$(\varphi(\Pi_{|B}(t)); t \geq 0) \stackrel{\mathcal{L}}{=} (\Pi(t); t \geq 0).$$
- Then $\Pi_{|[n]}$ is Markov $\forall n$
- If Π is a binary, homogeneous fragmentation process, it is characterized by the fragmentation measure ν on $(0, 1)$, where...
- ...each block fragments into two blocks with frequencies dx and $1 - dx$ at rate $\nu(dx)$.

Fragmentation processes

Bertoin "Homogeneous fragmentation processes" *PTRF* 2001

- A Markov process $(\Pi_t; t \geq 0)$ with values in the partitions of $\mathbb{N}$ is a fragmentation process if it is non-increasing...
- ...and exchangeable = invariant by permutations ($\exists$ block frequencies) + distinct blocks have independent futures.
- Π is homogeneous if for any block B of $\Pi(0)$ and for any bijection $\varphi : B \to \mathbb{N}$,
$$(\varphi(\Pi_{|B}(t)); t \geq 0) \overset{\mathcal{L}}{=} (\Pi(t); t \geq 0).$$
- Then $\Pi_{|[n]}$ is Markov $\forall n$
- If Π is a binary, homogeneous fragmentation process, it is characterized by the fragmentation measure ν on $(0, 1)$, where...
- ...each block fragments into two blocks with frequencies dx and $1 - dx$ at rate $\nu(dx)$.
- The previous theorem states the one-to-one correspondence :

$$\text{Sampling-consistent Mbm} \iff \Pi_{|[n]} \text{ without fragmentation times}$$

The β-splitting model

- The β-splitting model is for $\beta \in (-2, \infty)$: $\nu(dx) = cx^{\beta}(1-x)^{\beta}dx$

The β-splitting model

- The β-splitting model is for $\beta \in (-2, \infty)$: $\nu(dx) = cx^{\beta}(1-x)^{\beta}dx$
- Balance increases with β

The β-splitting model

- The β-splitting model is for $\beta \in (-2, \infty)$: $\nu(dx) = cx^{\beta}(1-x)^{\beta}dx$
- Balance increases with β

The β-splitting model

- The β-splitting model is for $\beta \in (-2, \infty)$: $\nu(dx) = cx^{\beta}(1-x)^{\beta}dx$
- Balance increases with β

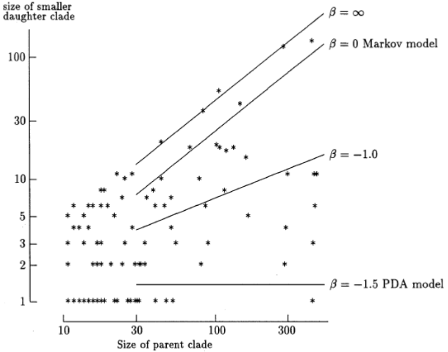
The β-splitting model

- The β-splitting model is for $\beta \in (-2, \infty)$: $\nu(dx) = cx^{\beta}(1-x)^{\beta}dx$
- Balance increases with β

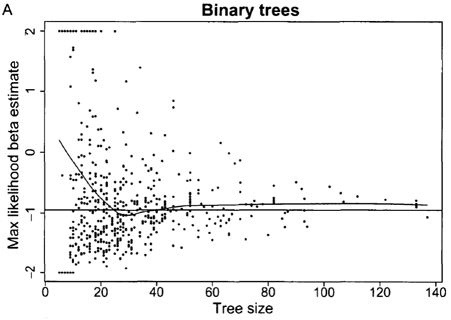
β	Description	Median split
-2	Completely unbalanced	1
-1.5	PDA model	1.5
-1	Unnamed	$\sqrt{m}$
0	Markov model	$m/4$
∞	An almost completely balanced model	$m/2$

Estimating β

S_{min} vs $S_{\text{min}} + S_{\text{max}}$ (Aldous 2001)



MLE of β (Blum & François 2006)



My favorite evolutionary conundrum : Why $\beta \approx -1$? (No answer here).

Outline

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$
- Iteratively **split nodes** from root to tips **w/ proba prop. to** X^{α}

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$
- Iteratively **split nodes** from root to tips **w/ proba prop. to** X^{α}
- Induces a sampling-consistent law on binary, **ranked** tree shapes with n labelled leaves.

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$
- Iteratively **split nodes** from root to tips **w/ proba prop. to** X^{α}
- Induces a sampling-consistent law on binary, **ranked** tree shapes with n labelled leaves.
- Yule tree/Kingman coalescent $\Leftrightarrow$ $\beta = 0$ and $\alpha = 1$.

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$
- Iteratively **split nodes** from root to tips **w/ proba prop. to** X^{α}
- Induces a sampling-consistent law on binary, **ranked** tree shapes with n labelled leaves.
- Yule tree/Kingman coalescent $\Leftrightarrow$ $\beta = 0$ and $\alpha = 1$.
- $\alpha > 0$: deeper nodes in rich clades ('phylogenetic redundancy')

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

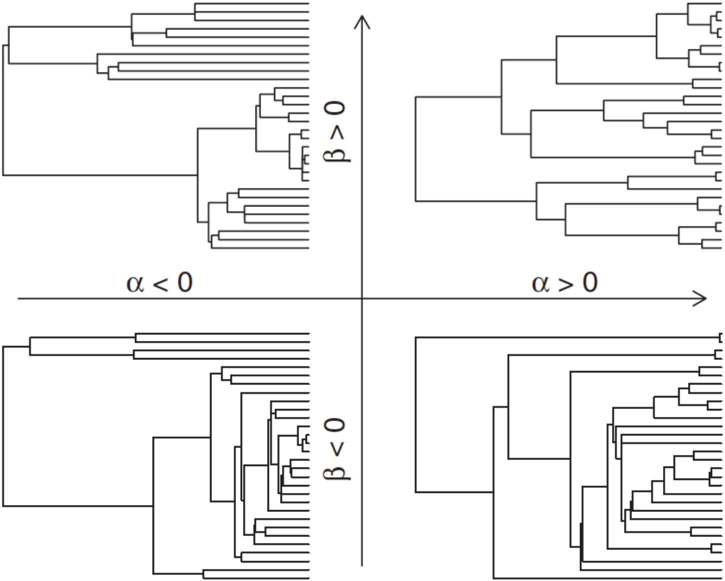
Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$
- Iteratively **split nodes** from root to tips **w/ proba prop. to** X^{α}
- Induces a sampling-consistent law on binary, **ranked** tree shapes with n labelled leaves.
- Yule tree/Kingman coalescent $\Leftrightarrow$ $\beta = 0$ and $\alpha = 1$.
- $\alpha > 0$: deeper nodes in rich clades ('phylogenetic redundancy')
- $\alpha < 0$: deeper nodes in poor clades = PD at risk

ALPHA : age/richness index, $\alpha \in \mathbb{R}$

Richness = Number of species in clade

- Recall each subclade is associated w/ an interval of width X
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$
- Iteratively **split nodes** from root to tips **w/ proba prop. to** X^{α}
- Induces a sampling-consistent law on binary, **ranked** tree shapes with n labelled leaves.
- Yule tree/Kingman coalescent $\Leftrightarrow$ $\beta = 0$ and $\alpha = 1$.
- $\alpha > 0$: deeper nodes in rich clades ('phylogenetic redundancy')
- $\alpha < 0$: deeper nodes in poor clades = PD at risk
- See also : Sainudiin & Véber "A Beta-splitting model for evolutionary trees" *Royal Society Open Science* 2016



β > 0
α < 0
α > 0
β < 0

Self-similar fragmentation processes

Bertoin "Self-similar fragmentation processes" *Annales de l'IHP* 2002

- Recall the fragmentation process $(\Pi(t); t \geq 0)$ with values in the partitions of $\mathbb{N}$.

Self-similar fragmentation processes

Bertoin "Self-similar fragmentation processes" *Annales de l'IHP* 2002

- Recall the fragmentation process $(\Pi(t); t \geq 0)$ with values in the partitions of $\mathbb{N}$.
- Π^α is self-similar with index α if for any block B of $\Pi^\alpha(0)$ with frequency x and for any bijection $\varphi : B \to \mathbb{N}$,

$$(\varphi(\Pi^\alpha_{|B}(t)); t \geq 0) \stackrel{\mathcal{L}}{=} (\Pi^\alpha(x^\alpha t); t \geq 0).$$

Self-similar fragmentation processes

Bertoin "Self-similar fragmentation processes" *Annales de l'IHP* 2002

- Recall the fragmentation process $(\Pi(t); t \geq 0)$ with values in the partitions of $\mathbb{N}$.
- Π^α is self-similar with index α if for any block B of $\Pi^\alpha(0)$ with frequency x and for any bijection $\varphi : B \to \mathbb{N}$,

$$(\varphi(\Pi^\alpha_{|B}(t)); t \geq 0) \overset{\mathcal{L}}{=} (\Pi^\alpha(x^\alpha t); t \geq 0).$$

- Homogeneous fragmentation : $\alpha = 0$.

Self-similar fragmentation processes

Bertoin "Self-similar fragmentation processes" *Annales de l'IHP* 2002

- Recall the fragmentation process $(\Pi(t); t \geq 0)$ with values in the partitions of $\mathbb{N}$.
- Π^α is self-similar with index α if for any block B of $\Pi^\alpha(0)$ with frequency x and for any bijection $\varphi : B \to \mathbb{N}$,

$$(\varphi(\Pi^\alpha_{|B}(t)); t \geq 0) \stackrel{\mathcal{L}}{=} (\Pi^\alpha(x^\alpha t); t \geq 0).$$

- Homogeneous fragmentation : $\alpha = 0$.
- Now $(\Pi^\alpha_{|[n]})$ is not Markov

Self-similar fragmentation processes

Bertoin "Self-similar fragmentation processes" *Annales de l'IHP* 2002

- Recall the fragmentation process $(\Pi(t); t \geq 0)$ with values in the partitions of $\mathbb{N}$.
- Π^α is self-similar with index α if for any block B of $\Pi^\alpha(0)$ with frequency x and for any bijection $\varphi : B \to \mathbb{N}$,

$$(\varphi(\Pi^\alpha_{|B}(t)); t \geq 0) \stackrel{\mathcal{L}}{=} (\Pi^\alpha(x^\alpha t); t \geq 0).$$

- Homogeneous fragmentation : $\alpha = 0$.
- Now $(\Pi^\alpha_{|[n]})$ is not Markov
- But $(\Pi_{|[n]}, R_n)$ is Markov, where R_n is the vector of n-tagged fragments = frequencies of blocks containing elements of $[n]$

SC Mbm w/ age-richness index $\alpha \iff \Pi^\alpha_{|[n]}$ w/ relative fragmentation times

ETA : abundance/richness index, $\eta \in \mathbb{R}$

Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**

ETA : abundance/richness index, $\eta \in \mathbb{R}$

Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1-R)X$ and

$$A_{left} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{R^\eta}{R^\eta + (1-R)^\eta} A$$

$$A_{right} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A$$

ETA : abundance/richness index, $\eta \in \mathbb{R}$

Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1-R)X$ and

$$A_{left} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{R^\eta}{R^\eta + (1-R)^\eta} A$$
$$A_{right} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A$$

- Induces a sampling-consistent law on binary tree shapes with (n labelled leaves and) a **probability measure on its leaf set**.

ETA : abundance/richness index, $\eta \in \mathbb{R}$

Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$ and

$$A_{left} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{R^\eta}{R^\eta + (1-R)^\eta} A$$

$$A_{right} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A$$

- Induces a sampling-consistent law on binary tree shapes with (n labelled leaves and) a **probability measure on its leaf set**.
- $\eta < 1$: higher abundances in small clades

ETA : abundance/richness index, $\eta \in \mathbb{R}$

Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$ and

$$A_{left} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{R^\eta}{R^\eta + (1-R)^\eta} A$$
$$A_{right} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A$$

- Induces a sampling-consistent law on binary tree shapes with (n labelled leaves and) a **probability measure on its leaf set**.
- $\eta < 1$: higher abundances in small clades
- $\eta = 1$: total clade abundance prop. to richness

ETA : abundance/richness index, $\eta \in \mathbb{R}$

Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1 - R)X$ and

$$A_{left} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{R^\eta}{R^\eta + (1-R)^\eta} A$$
$$A_{right} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A$$

- Induces a sampling-consistent law on binary tree shapes with (n labelled leaves and) a **probability measure on its leaf set**.
- $\eta < 1$: higher abundances in small clades
- $\eta = 1$: total clade abundance prop. to richness
- $\eta > 1$: lower abundances in small clades = PD at risk

ETA : abundance/richness index, $\eta \in \mathbb{R}$

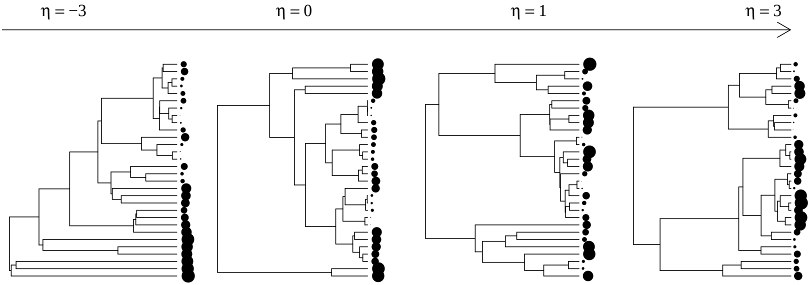
Abundance = Relative abundance in number of individuals of clade

- Each subclade assoc. w/ interval of width X, now also **abundance A**
- Daughter subclades have widths $X_{left} = RX$ and $X_{right} = (1-R)X$ and

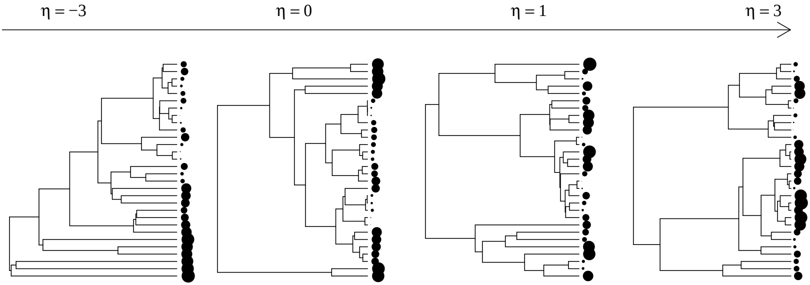
$$A_{left} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{R^\eta}{R^\eta + (1-R)^\eta} A$$
$$A_{right} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A$$

- Induces a sampling-consistent law on binary tree shapes with (n labelled leaves and) a **probability measure on its leaf set**.
- $\eta < 1$: higher abundances in small clades
- $\eta = 1$: total clade abundance prop. to richness
- $\eta > 1$: lower abundances in small clades = PD at risk
- **Sampling (keeping extant) in prop to abundance**

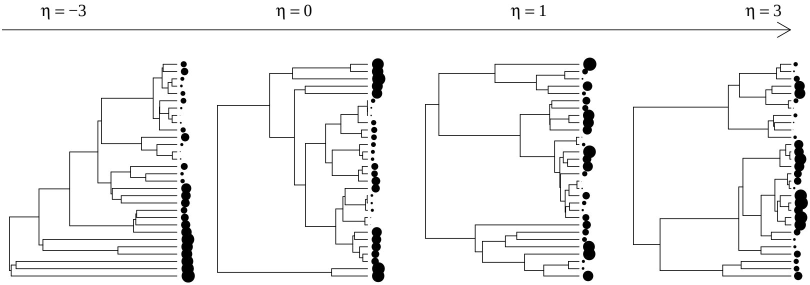
  $\approx$ Field of Bullets iff $\eta = 1$



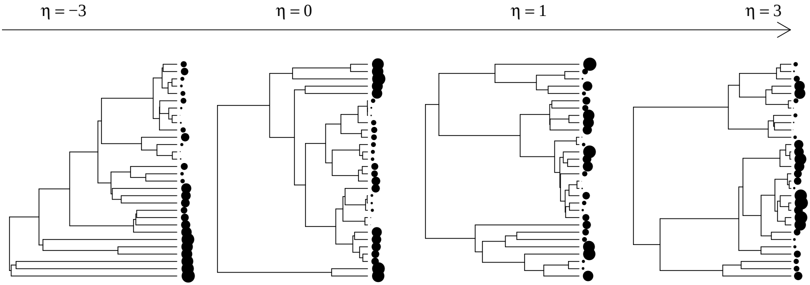
- $\eta = -3$, rare species are all in large clades



- $\eta = -3$, rare species are all in large clades
- $\eta = 0$, 'equal-sharing' measure (inv. prop. to number of splits)



- $\eta = -3$, rare species are all in large clades
- $\eta = 0$, 'equal-sharing' measure (inv. prop. to number of splits)
- $\eta = 1$, species have equal abundances on average



- $\eta = -3$, rare species are all in large clades
- $\eta = 0$, ‘equal-sharing’ measure (inv. prop. to number of splits)
- $\eta = 1$, species have equal abundances on average
- $\eta = 3$, the few species in poor clades are rare

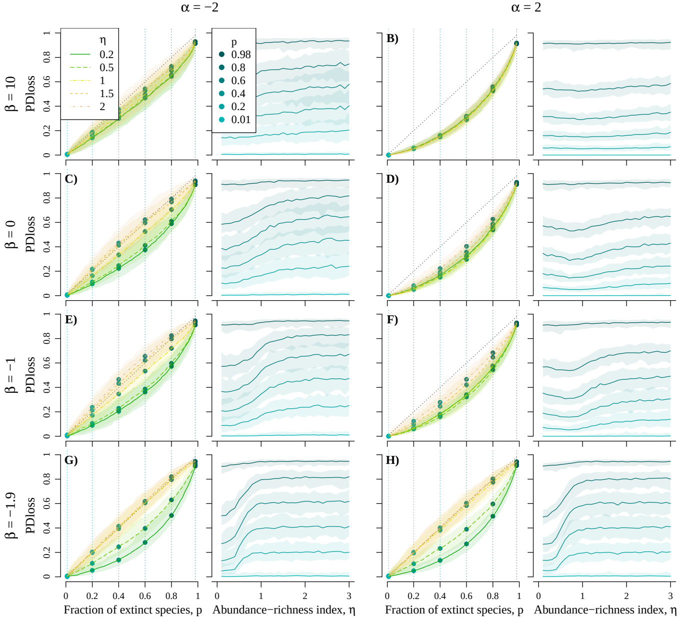
Warning!

The next slide represents $1 - S =$ PD loss...

...as a function of $1 - p =$ fraction of extinct species.

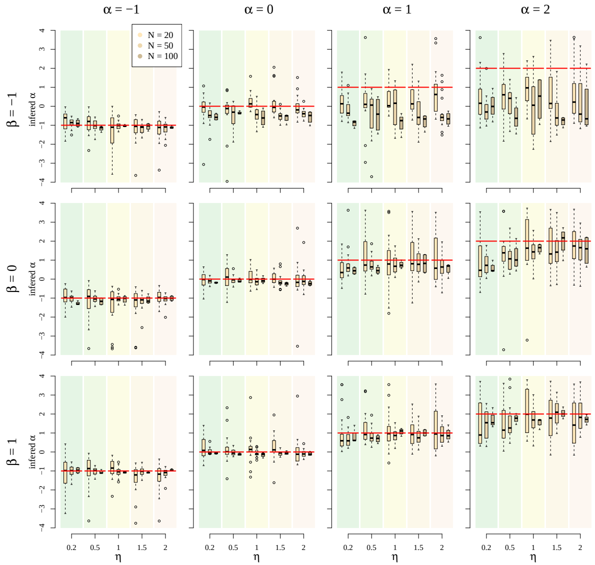
This function is expected to be increasing and convex in the FoB model.

'Danger zone' : $\alpha < 0, \eta > 1$

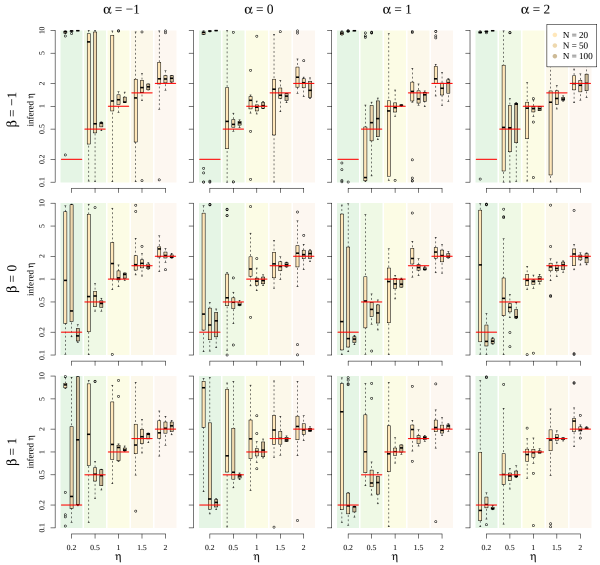


Outline

Inference of α

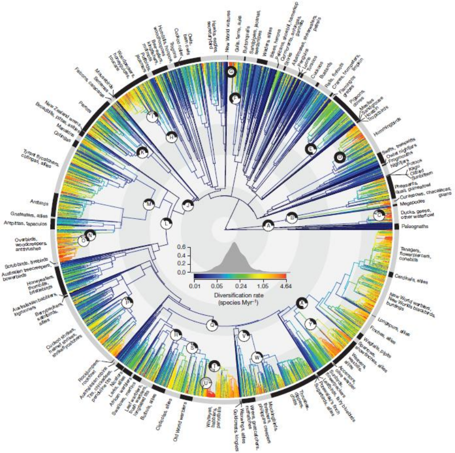


Inference of η



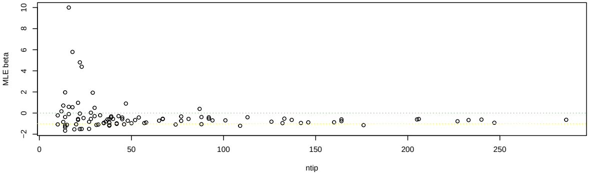
Data

Jetz et al "The global diversity of birds in space and time" *Nature* 2012

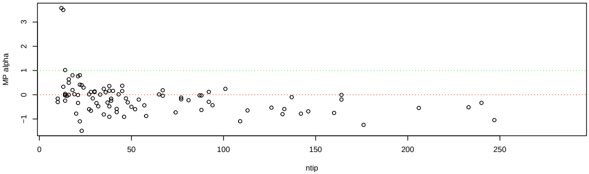


- Phylogenies of ≈ 100 clades in the class *Aves* (a.k.a. birds)
- Information used for inference :
  - Shapes of trees
  - Relative positions of nodes (but not the exact datation estimates)
  - Species range relative to sum of all species ranges.

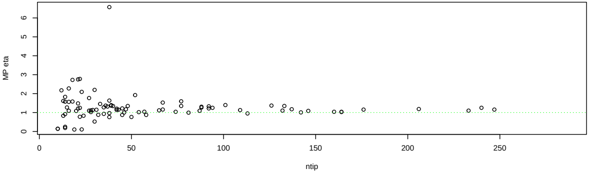
Inference from $\approx$ 100 bird clades



$\beta < 0$ and $\beta \approx -1$
(real trees are imbalanced)

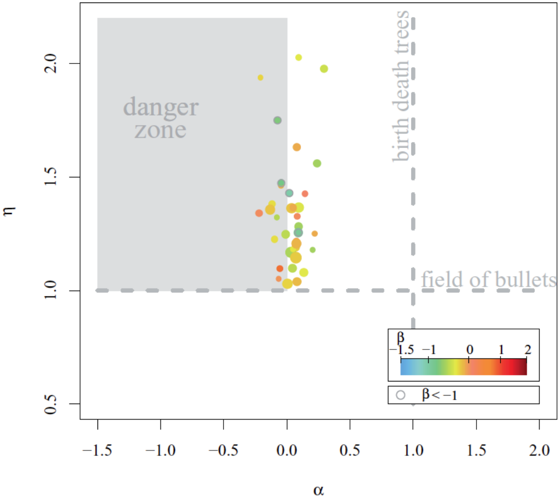


$\alpha < 0$
(deep nodes $\in$ small clades)



$\eta \geq 1$ and $\eta \approx 1$
(rare species $\in$ small clades)

Inference from ≈ 100 bird clades



Conclusion

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- New stochastic tree model with 3 parameters, tuning :
  - tree shape : $\beta =$ tree balance
  - ranked node depths : $\alpha =$ age-richness index
  - abundances at tips : $\eta =$ abundance-richness index
- Danger zone : $\beta < 0, \alpha < 0, \eta > 1$ = PD at risk
- Implemented in R-package *apTreeshape* (maintained by M.J. Blum)
- Large bird clades have $\beta \approx -1, \alpha < 0, \eta \approx 1$
- Bird clades in danger zone !

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^\alpha$

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^{\alpha}$
  - $M(b)$ is the mass measure of block b, instead of $\sim |b|^{\eta}$

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^\alpha$
  - $M(b)$ is the mass measure of block b, instead of $\sim |b|^\eta$
- To explain the triple conundrum :

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^\alpha$
  - $M(b)$ is the mass measure of block b, instead of $\sim |b|^\eta$
- To explain the triple conundrum :
  - $\beta \approx -1$: real trees are imbalanced

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^\alpha$
  - $M(b)$ is the mass measure of block b, instead of $\sim |b|^\eta$
- To explain the triple conundrum :
  - $\beta \approx -1$: real trees are imbalanced
  - $\alpha < 0$: real trees have deeper nodes in smaller clades

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^\alpha$
  - $M(b)$ is the mass measure of block b, instead of $\sim |b|^\eta$
- To explain the triple conundrum :
  - $\beta \approx -1$: real trees are imbalanced
  - $\alpha < 0$: real trees have deeper nodes in smaller clades
  - $\eta \geq 1$: real trees have rarer species in smaller clades, but $\eta \approx 1$: not much phylogenetic signal for species' abundances.

Perspectives

Maliet, O., Gascuel, F., Lambert, A. (2018) "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity" *Systematic Biology* (in press)

- To extend the fragmentation process Π to a random fragmentation triple (Π, W, M) (Jean-Jil Duchamps, work in progress), where :
  - $W(b)$ is the fragmentation rate of block b, instead of $|b|^\alpha$
  - $M(b)$ is the mass measure of block b, instead of $\sim |b|^\eta$
- To explain the triple conundrum :
  - $\beta \approx -1$: real trees are imbalanced
  - $\alpha < 0$: real trees have deeper nodes in smaller clades
  - $\eta \geq 1$: real trees have rarer species in smaller clades, but $\eta \approx 1$: not much phylogenetic signal for species' abundances.
- To confirm the last two findings with more data.

Collaborators

Fanny Gascuel (Institut Curie) . . . .

Odile Maliet (ENS) . . . .

Mike Steel (U Canterbury) . . . .

SMILE : an interdisciplinary group in Paris

SCIENCES SORBONNE UNIVERSITÉ

COLLÈGE DE FRANCE 1530

S.M.I.L.E

Stochastic Models for the Inference of Life Evolution

SMILE = Stochastic Models for the Inference of Life Evolution

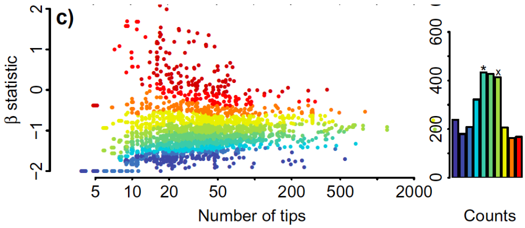
A positive answer to $\beta \approx -1$?

Hagen, Hartmann, Steel, Stadler "Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies" *Systematic Biology (2015)*

- Birth-death process with age-dependent birth rate $b = b(a)$ parameterized by

$$b(a) = ca^{\phi-1}$$

- Estimates of ϕ lie in $(0, 1)$: speciation rate decreases with age



For $\phi = 0.6$, the reconstructed tree has $\beta \approx -1$.

Q : "Why $\beta \approx -1$?"

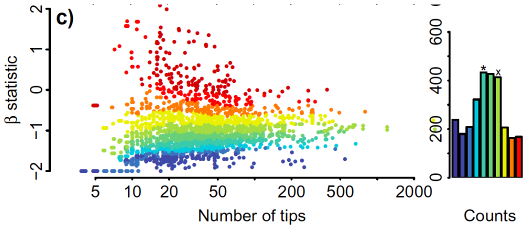
A positive answer to $\beta \approx -1$?

Hagen, Hartmann, Steel, Stadler "Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies" *Systematic Biology (2015)*

- Birth-death process with age-dependent birth rate $b = b(a)$ parameterized by

$$b(a) = ca^{\phi-1}$$

- Estimates of ϕ lie in $(0, 1)$: speciation rate decreases with age



For $\phi = 0.6$, the reconstructed tree has $\beta \approx -1$.

Q : "Why $\beta \approx -1$?"

— "Because $\phi \approx 0.6$" ;-)

Balance of incomplete trees

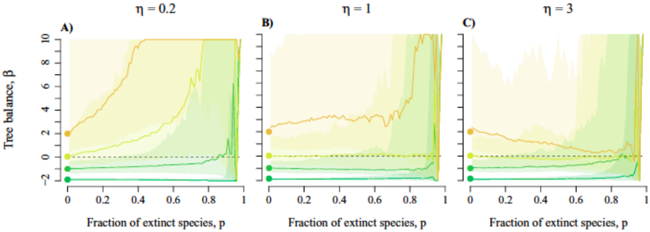


FIGURE 8. Effect of abundance-richness index η on the balance of phylogenetic trees after extinctions (Maximum Likelihood Estimate $\hat{\beta}$ of β). Initial tree balance β ranges from 10 (brown dots and lines, "bush trees") to -1.9 (green dots and lines, "caterpilar trees"). Extinction fraction p increases from 0.01 to 0.98 (from left to right). Results are based on 100 simulation replicates: plain lines give median values and light areas give 95% confidence intervals. Other parameter values: number of species $N = 100$, approximation parameter $\varepsilon = 0.001$, $\alpha = 0$.