Two Combinatorial Problems Arising from Phylogenetics

(joint with M. Drmota, E. Y. Jin, C.-H. Lee, Y.-W. Lee and A. R. Paningbatan)

Michael Fuchs

Department of Applied Mathematics
Chiao Tung University AND Chengchi University

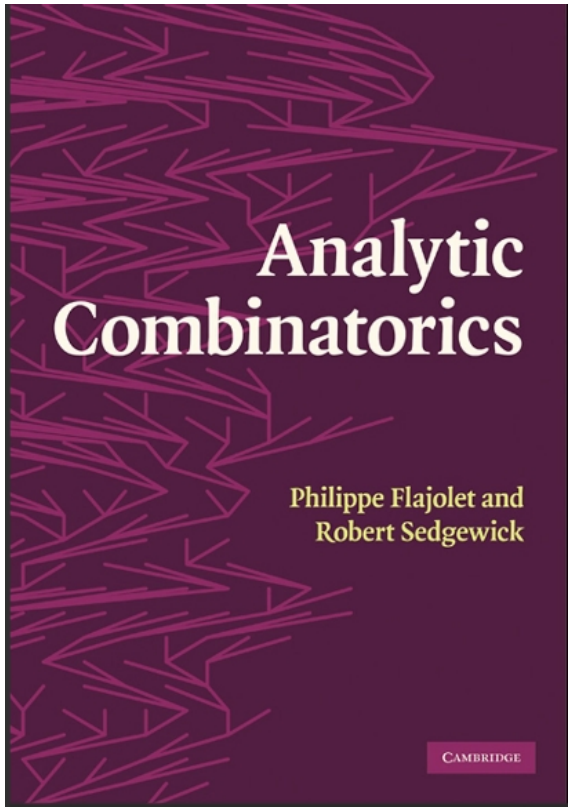
June 14th, 2019

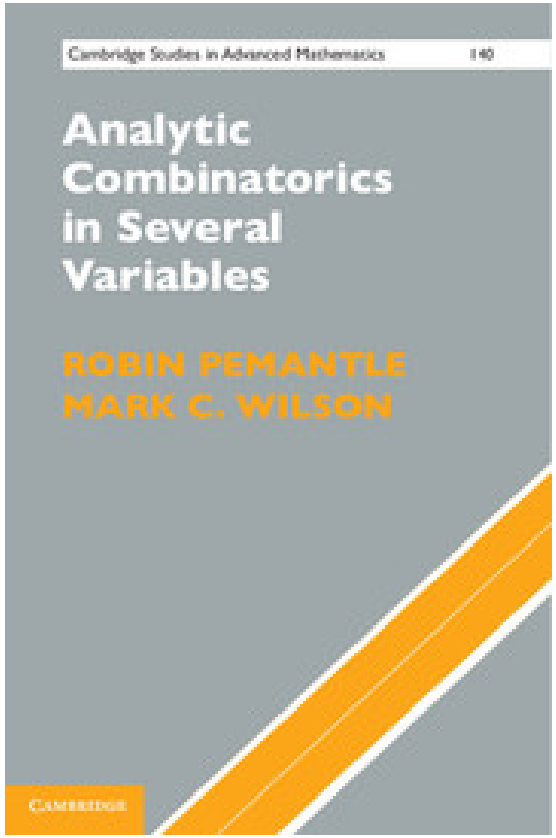
Analytic Combinatorics (i)

Combinatorialists use recurrence, generating functions, and such transformations as the Vandermonde convolution; others, to my horror, use contour integrals, differential equations, and other resources of mathematical analysis.

- John Riordan (1968).

Analytic Combinatorics (ii)





Outline of the Talk

Outline of the Talk

1. Introduction
2. Biodiversity Indices

F. and Jin (2015). Equality of Shapley value and fair proportion index in phylogenetic trees.

F. and Paningbatan (2019+). Correlation between Shapley values of rooted phylogenetic trees under the beta-splitting model.

Outline of the Talk

1. Introduction
2. Biodiversity Indices

   F. and Jin (2015). Equality of Shapley value and fair proportion index in phylogenetic trees.

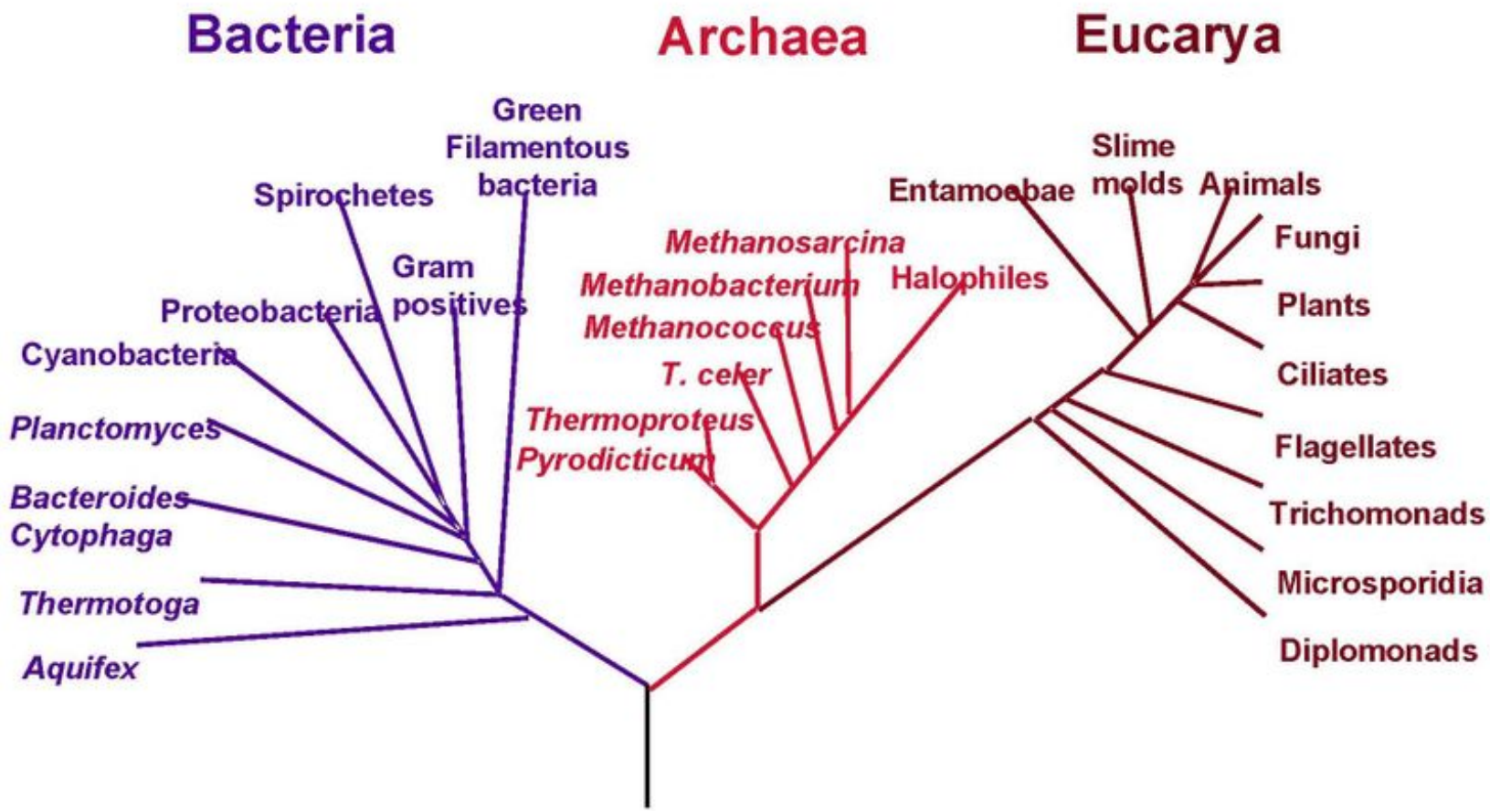
   F. and Paningbatan (2019+). Correlation between Shapley values of rooted phylogenetic trees under the beta-splitting model.

3. Group Pattern Formation of Social Animals

   Drmota, F., Y.-W. Lee (2016). Stochastic analysis of the extra clustering model for animal grouping.

   F., C.-H. Lee, Paningbatan (2019+). Distributional analysis of the extra clustering model with uniformly generated phylogenetic trees.

Evolutionary or Phylogenetic tree (=PT)



Phylogenetic tree of size n: rooted, plane, unlabelled binary tree with n external nodes (and consequently $n-1$ internal nodes).

Aldous β-splitting Model (i)

- Let f be probability density on $[0,1]$ which is symmetric (i.e. $f(x) = f(1-x)$)

Aldous β-splitting Model (i)

- Let f be probability density on $[0,1]$ which is symmetric (i.e. $f(x) = f(1-x)$)
- Throw n balls uniformly at random into $[0,1]$.

Aldous β-splitting Model (i)

- Let f be probability density on $[0,1]$ which is symmetric (i.e. $f(x) = f(1-x)$)
- Throw n balls uniformly at random into $[0,1]$.
- Split $[0,1]$ into two subintervals according to f; if one subinterval contains no ball repeat.

Aldous β-splitting Model (i)

- Let f be probability density on $[0,1]$ which is symmetric (i.e. $f(x) = f(1-x)$)
- Throw n balls uniformly at random into $[0,1]$.
- Split $[0,1]$ into two subintervals according to f; if one subinterval contains no ball repeat.
- Recursively continue with the subintervals, where a subinterval $[a,b]$ is split at $a + X(b-a)$ with X having distribution f.

Aldous β-splitting Model (i)

- Let f be probability density on $[0,1]$ which is symmetric (i.e. $f(x) = f(1-x)$)
- Throw n balls uniformly at random into $[0,1]$.
- Split $[0,1]$ into two subintervals according to f; if one subinterval contains no ball repeat.
- Recursively continue with the subintervals, where a subinterval $[a,b]$ is split at $a + X(b-a)$ with X having distribution f.
- Stop when a subinterval contains only one ball.

Aldous β-splitting Model (i)

- Let f be probability density on $[0,1]$ which is symmetric (i.e. $f(x) = f(1-x)$)
- Throw n balls uniformly at random into $[0,1]$.
- Split $[0,1]$ into two subintervals according to f; if one subinterval contains no ball repeat.
- Recursively continue with the subintervals, where a subinterval $[a,b]$ is split at $a + X(b-a)$ with X having distribution f.
- Stop when a subinterval contains only one ball.

$\longrightarrow$ This gives a probability distribution on PTs of size n.

Aldous β-splitting Model (ii)

T ... random PT.

Choose a β-distribution ($\beta > -1$):

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \qquad x \in [0,1].$$

Aldous β-splitting Model (ii)

T ... random PT.

Choose a β-distribution ($\beta > -1$):

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \qquad x \in [0,1].$$

Let $\pi_{n,i}$ be the probability that left subtree has size i. Then,

$$\pi_{n,i} = \frac{1}{\pi_n(\beta)} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{i!(n-i)!}, \qquad 1 \le i \le n-1,$$

where $\pi_n(\beta)$ is a suitable constant.

Aldous β-splitting Model (ii)

T ... random PT.

Choose a β-distribution ($\beta > -1$):

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \qquad x \in [0,1].$$

Let $\pi_{n,i}$ be the probability that left subtree has size i. Then,

$$\pi_{n,i} = \frac{1}{\pi_n(\beta)} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{i!(n-i)!}, \qquad 1 \le i \le n-1,$$

where $\pi_n(\beta)$ is a suitable constant.

Note that the above expression makes also sense for $-2 < \beta \le -1$.

Special Cases

- $\beta = 0$: Yule-Harding model:

$$\pi_{n,i} = \frac{1}{n-1}, \qquad 1 \le i \le n-1.$$

Is also generated by a continuous-time pure birth process..

Special Cases

- $\beta = 0$: Yule-Harding model:

$$\pi_{n,i} = \frac{1}{n-1}, \qquad 1 \le i \le n-1.$$

Is also generated by a continuous-time pure birth process..

- $\beta = -3/2$: Uniform or PDA model:

$$\pi_{n,i} = \frac{C_{i-1}C_{n-i-1}}{C_{n-1}}, \qquad 1 \le i \le n-1,$$

where $C_n = \binom{2n}{n}/(n+1)$ are the Catalan numbers.

Special Cases

- $\beta = 0$: Yule-Harding model:

$$\pi_{n,i} = \frac{1}{n-1}, \qquad 1 \le i \le n-1.$$

Is also generated by a continuous-time pure birth process..

- $\beta = -3/2$: Uniform or PDA model:

$$\pi_{n,i} = \frac{C_{i-1}C_{n-i-1}}{C_{n-1}}, \qquad 1 \le i \le n-1,$$

where $C_n = \binom{2n}{n}/(n+1)$ are the Catalan numbers.

- $\beta = -1$: with H_n the harmonic numbers:

$$\pi_{n,i} = \frac{n}{2H_{n-1}} \cdot \frac{1}{i(n-i)}, \qquad 1 \le i \le n-1.$$

This model seems to have the best match with real trees.

Lloyd Shapley

Lloyd Shapley
(1923-2016)

Lloyd Shapley

Lloyd Shapley
(1923-2016)

Shapley value:

Measure of importance of each player in a cooperative game

Lloyd Shapley

Lloyd Shapley
(1923-2016)

Shapley value:

Measure of importance of each player in a cooperative game

$\longrightarrow$ recently used as prioritization tool of taxa in biodiversity

Shapley Value and Fair Proportion Index

Let T be a PT and a a taxon (=leaf) of T.

Shapley Value and Fair Proportion Index

Let T be a PT and a a taxon (=leaf) of T.

Rooted Shapley Value $\mathrm{SV}_T^{[r]}(a)$:

$$\mathrm{SV}_T^{[r]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S| - 1)!(n - |S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})),$$

where $\mathrm{PD}(S)$ is the size of the ancestor tree of S.

Shapley Value and Fair Proportion Index

Let T be a PT and a a taxon (=leaf) of T.

Rooted Shapley Value $\mathrm{SV}_T^{[r]}(a)$:

$$\mathrm{SV}_T^{[r]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S| - 1)!(n - |S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})),$$

where $\mathrm{PD}(S)$ is the size of the ancestor tree of S.

Fair Proportion Index $\mathrm{FP}_T(a)$:

$$\mathrm{FP}_T(a) = \sum_e D_e^{-1},$$

where D_e the number of taxa below e.

Shapley Value and Fair Proportion Index

Let T be a PT and a a taxon (=leaf) of T.

Rooted Shapley Value $\mathrm{SV}_T^{[r]}(a)$:

$$\mathrm{SV}_T^{[r]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S| - 1)!(n - |S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})),$$

where $\mathrm{PD}(S)$ is the size of the ancestor tree of S.

Fair Proportion Index $\mathrm{FP}_T(a)$:

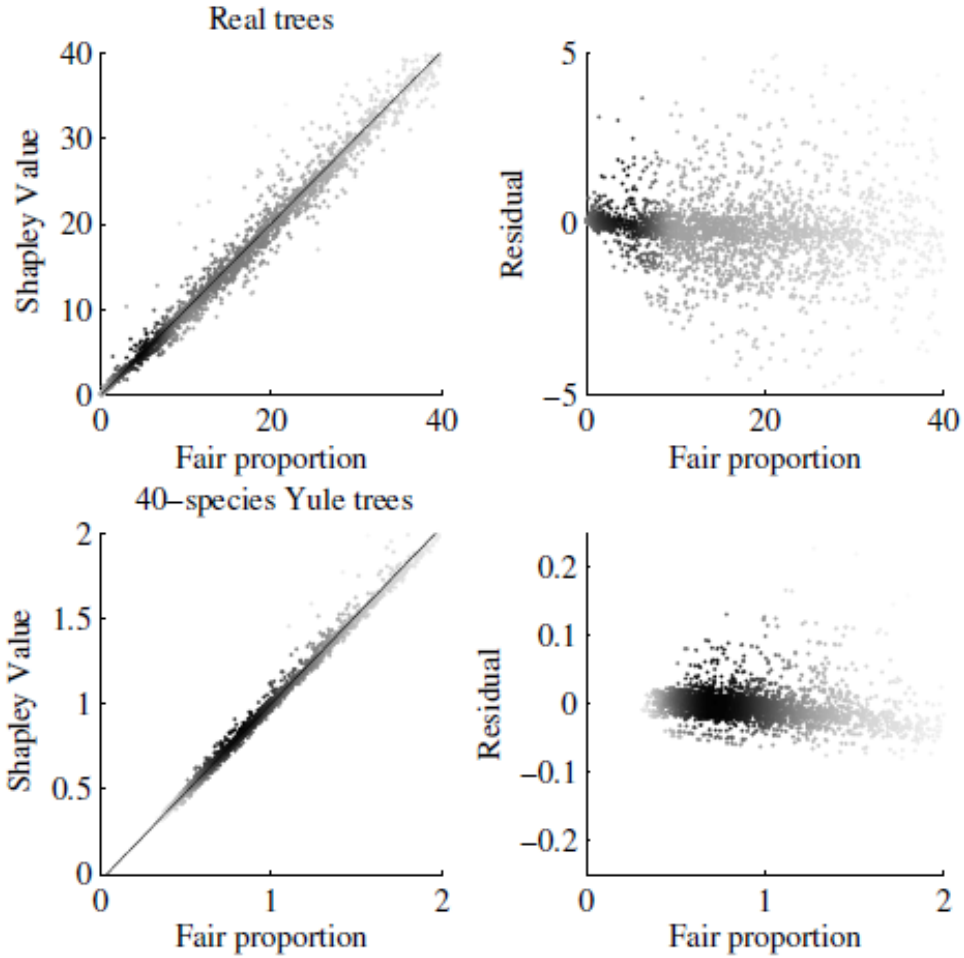
$$\mathrm{FP}_T(a) = \sum_e D_e^{-1},$$

where D_e the number of taxa below e.

Used (somehow arbitrarily) for EDGE of Existence conservation program!

Correlation between $\mathrm{SV}^{[r]}$ and FP

Hartmann (2013):



Assume a is in left subtree T_ℓ and $|T_\ell| = i$.

Assume a is in left subtree T_ℓ and $|T_\ell| = i$.

Lemma

We have,

$$\mathrm{FP}_T(a) = \frac{1}{i} + \mathrm{FP}_{T_\ell}(a).$$

Assume a is in left subtree T_ℓ and $|T_\ell| = i$.

Lemma

We have,

$$\mathrm{FP}_T(a) = \frac{1}{i} + \mathrm{FP}_{T_\ell}(a).$$

and

$$\mathrm{SV}_T^{[r]}(a) = \frac{1}{i} + \mathrm{SV}_{T_\ell}^{[r]}(a)$$

$\mathrm{SV}^{[r]} = \mathrm{FP}$

Assume a is in left subtree T_ℓ and $|T_\ell| = i$.

Lemma

We have,

$$\mathrm{FP}_T(a) = \frac{1}{i} + \mathrm{FP}_{T_\ell}(a).$$

and

$$\mathrm{SV}_T^{[r]}(a) = \frac{1}{i} + \mathrm{SV}_{T_\ell}^{[r]}(a)$$

Theorem (F. and Jin; 2015)

We have,

$$\mathrm{SV}_T^{[r]}(a) = \mathrm{FP}_T(a).$$

Modified Shapley Value

Which Shapley value did Hartmann use?

Modified Shapley Value

Which Shapley value did Hartmann use?

Modified Shapley Value $\widetilde{\mathrm{SV}}_T(a)$:

$$\widetilde{\mathrm{SV}}_T(a) = \frac{1}{n!} \sum_{|S|\geq 2, a\in S} (|S|-1)!(n-|S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S\setminus\{a\})),$$

where $\mathrm{PD}(S)$ is as before.

Modified Shapley Value

Which Shapley value did Hartmann use?

Modified Shapley Value $\widetilde{\mathrm{SV}}_T(a)$:

$$\widetilde{\mathrm{SV}}_T(a) = \frac{1}{n!} \sum_{|S|\geq 2, a\in S} (|S|-1)!(n-|S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S\setminus\{a\})),$$

where $\mathrm{PD}(S)$ is as before.

Theorem (F. and Jin; 2015)

Under the PDA model and the Yule-Harding model,

$$\lim_{n\to\infty} \rho(\widetilde{\mathrm{SV}}_n, \mathrm{FP}_n) = 1,$$

where ρ denotes the correlation coefficient.

Unrooted Shapley Value

It turned out that Hartmann used yet another Shapley value!

Unrooted Shapley Value

It turned out that Hartmann used yet another Shapley value!

Unrooted Shapley Value $\mathrm{SV}_T^{[u]}(a)$:

$$\mathrm{SV}_T^{[u]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S|-1)!(n-|S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})),$$

where $\mathrm{PD}(S)$ is the size of the Steiner tree of S.

Unrooted Shapley Value

It turned out that Hartmann used yet another Shapley value!

Unrooted Shapley Value $\mathrm{SV}_T^{[u]}(a)$:

$$\mathrm{SV}_T^{[u]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S| - 1)!(n - |S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})),$$

where $\mathrm{PD}(S)$ is the size of the Steiner tree of S.

Theorem (F. and Paningbatan; 2019+)

Under the β-splitting model with $\beta > -1$,

$$\lim_{n \to \infty} \rho(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) = 1,$$

where ρ denotes the correlation coefficient.

Outline of the Proof

- Bounds for moments of additive shape parameters under the β-splitting model.

Outline of the Proof

- Bounds for moments of additive shape parameters under the β-splitting model.

  For this, we study

$$a_n = 2\sum_{i=1}^{n-1} \pi_{n,i} a_i + b_n$$

  for varying toll-sequence b_n.

Outline of the Proof

- Bounds for moments of additive shape parameters under the β-splitting model.

  For this, we study

$$a_n = 2\sum_{i=1}^{n-1} \pi_{n,i} a_i + b_n$$

  for varying toll-sequence b_n.

- An expression for the difference:

$$\mathrm{SV}_T^{[u]}(a) - \mathrm{FP}_T(a) = \mathrm{SV}_T^{[u]}(a) - \mathrm{SV}_T^{[r]}(a).$$

Outline of the Proof

- Bounds for moments of additive shape parameters under the β-splitting model.

  For this, we study

$$a_n = 2\sum_{i=1}^{n-1} \pi_{n,i} a_i + b_n$$

  for varying toll-sequence b_n.

- An expression for the difference:

$$\mathrm{SV}_T^{[u]}(a) - \mathrm{FP}_T(a) = \mathrm{SV}_T^{[u]}(a) - \mathrm{SV}_T^{[r]}(a).$$

- Using the above two steps to bound the correlation coefficient with a bound which tends to 0.

Master Theorems

Let a_n satisfy the recurrence from the last slide with $\beta > -1$.

Master Theorems

Let a_n satisfy the recurrence from the last slide with $\beta > -1$.

Proposition

Assume that

$$b_n = \mathcal{O}(n^\gamma \log^\delta n)$$

for integers $\gamma, \delta \geq 0$. *Then,*

(i) *if* $\gamma = 1$, *then* $a_n = \mathcal{O}(n \log^{\delta+1} n)$;

(ii) *if* $\gamma > 1$, *then* $a_n = \mathcal{O}(n^\gamma \log^\delta n)$.

Master Theorems

Let a_n satisfy the recurrence from the last slide with $\beta > -1$.

Proposition

Assume that

$$b_n = \mathcal{O}(n^\gamma \log^\delta n)$$

for integers $\gamma, \delta \geq 0$. *Then,*

(i) *if* $\gamma = 1$, *then* $a_n = \mathcal{O}(n \log^{\delta+1} n)$;

(ii) *if* $\gamma > 1$, *then* $a_n = \mathcal{O}(n^\gamma \log^\delta n)$.

Proposition

If b_n *is non-negative and* $b_{n_0} > 0$ *for at least one* n_0, *then*

$$b_n = \Omega(n).$$

Bounds for Moments

Consider the additive shape parameters:

Bounds for Moments

Consider the additive shape parameters:

- Sackin Index S_n: sum over all taxon-root distances;
- Depth D_n: distance to root of a random taxon.

Bounds for Moments

Consider the additive shape parameters:

- Sackin Index S_n: sum over all taxon-root distances;
- Depth D_n: distance to root of a random taxon.

Corollary

For $\beta > -1$, we have

$$\mathbb{E}(S_n) = \mathcal{O}(n \log n), \qquad \mathbb{E}(S_n^2) = \mathcal{O}(n^2 \log^2 n);$$
$$\mathbb{E}(D_n) = \mathcal{O}(\log n), \qquad \mathbb{E}(D_n^2) = \mathcal{O}(\log^2 n).$$

Bounds for Moments

Consider the additive shape parameters:

- Sackin Index S_n: sum over all taxon-root distances;
- Depth D_n: distance to root of a random taxon.

Corollary

For $\beta > -1$, we have

$$\mathbb{E}(S_n) = \mathcal{O}(n \log n), \qquad \mathbb{E}(S_n^2) = \mathcal{O}(n^2 \log^2 n);$$
$$\mathbb{E}(D_n) = \mathcal{O}(\log n), \qquad \mathbb{E}(D_n^2) = \mathcal{O}(\log^2 n).$$

Corollary

For $\beta > -1$, we have

$$\mathrm{Var}(\mathrm{FP}_n) = \Omega(1).$$

More Additive Shape Parameters

Two more additive shape parameters:

More Additive Shape Parameters

Two more additive shape parameters:

- $X_T^{[i]}$: sum of all distances between the root and common ancestor of sets of size i;

More Additive Shape Parameters

Two more additive shape parameters:

- $X_T^{[i]}$: sum of all distances between the root and common ancestor of sets of size i;
- $Y_T^{[i]}(a)$: sum of all distances between the common ancestors of a set of size i and the set together with a.

More Additive Shape Parameters

Two more additive shape parameters:

- $X_T^{[i]}$: sum of all distances between the root and common ancestor of sets of size i;
- $Y_T^{[i]}(a)$: sum of all distances between the common ancestors of a set of size i and the set together with a.

We have,

$$X_T^{[i]} = X_{T_\ell}^{[i]} + X_{T_r}^{[i]} + \binom{|T_\ell|}{i} + \binom{|T_r|}{i}$$

and

$$Y_T^{[i]}(a) = \begin{cases} Y_{T_\ell}^{[i]}(a) + X_{T_r}^{[i]} + \binom{|T_r|}{i}, & \text{if } a \in T_\ell; \\ Y_{T_r}^{[i]}(a) + X_{T_\ell}^{[i]} + \binom{|T_\ell|}{i}, & \text{if } a \in T_r. \end{cases}$$

Difference between $\mathrm{SV}^{[u]}$ and FP

Proposition

For $a \in T_\ell$, we have

$$\begin{aligned}\mathrm{SV}_T^{[u]}(a) - \mathrm{SV}_T^{[r]}(a) = &-\frac{1}{n} D_T(a) \\ &+ \frac{1}{n!} \sum_{i=1}^{|T_r|} i!(n-i-1)! \left(X_{T_r}^{[i]} + \binom{|T_r|}{i} \right) \\ &+ \frac{1}{n!} \sum_{i=1}^{|T_\ell|-1} i!(n-i-1)! Y_{T_\ell}^{[i]}(a).\end{aligned}$$

Difference between $\mathrm{SV}^{[u]}$ and FP

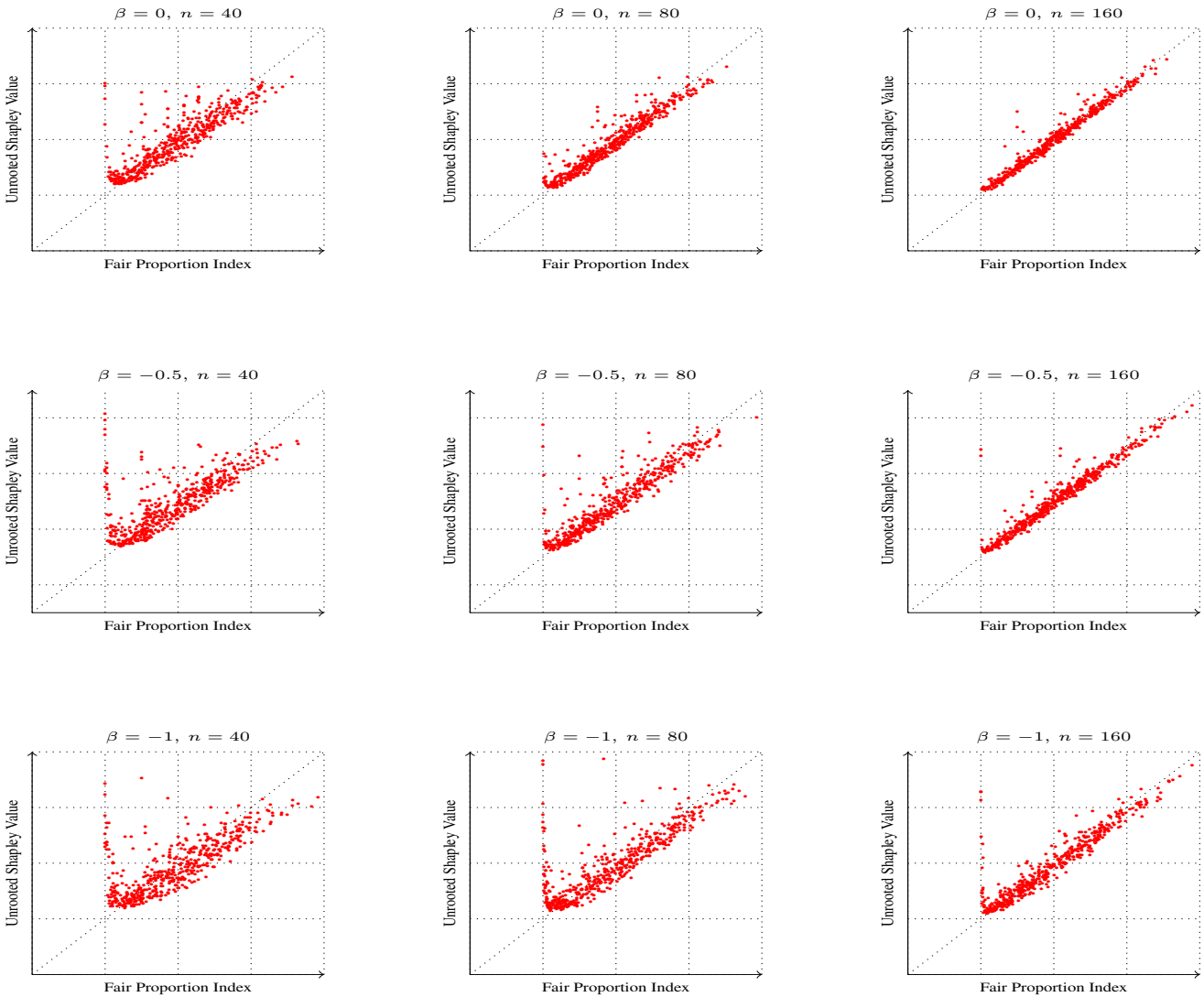
Proposition

For $a \in T_\ell$, we have

$$\begin{aligned}\mathrm{SV}_T^{[u]}(a) - \mathrm{SV}_T^{[r]}(a) = &-\frac{1}{n} D_T(a) \\ &+ \frac{1}{n!} \sum_{i=1}^{|T_r|} i!(n-i-1)! \left(X_{T_r}^{[i]} + \binom{|T_r|}{i} \right) \\ &+ \frac{1}{n!} \sum_{i=1}^{|T_\ell|-1} i!(n-i-1)! Y_{T_\ell}^{[i]}(a).\end{aligned}$$

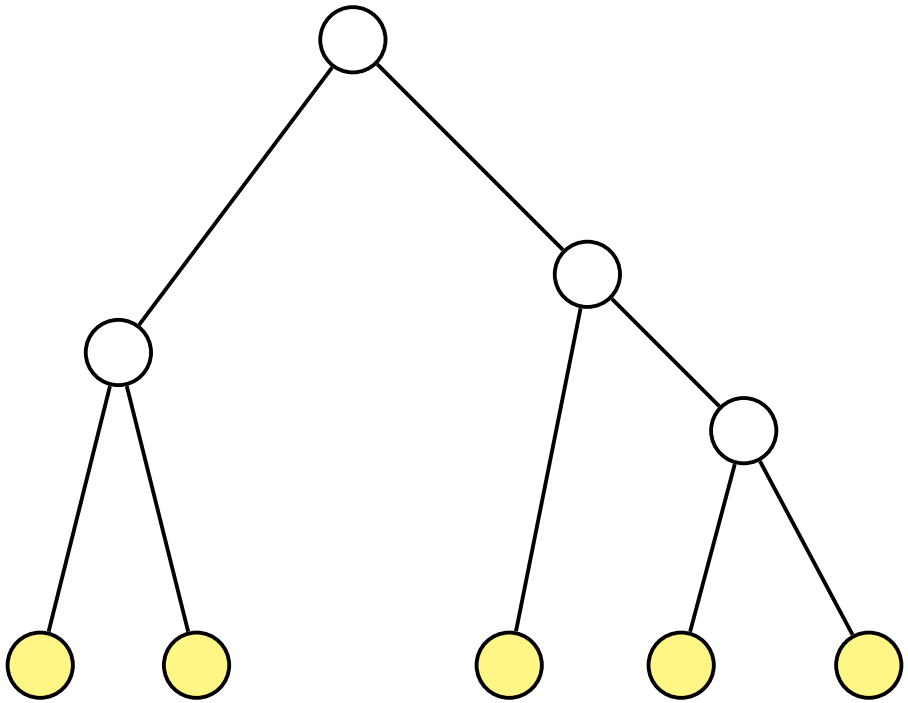
Since $\mathrm{FP}_T, D_T(a), X_T^{[i]}, Y_T^{[i]}(a)$ can all be computed recursively, $\mathrm{SV}_T^{[u]}$ can be computed efficiently.

Does our Theorem extend to $\beta = -1$ (and beyond)?



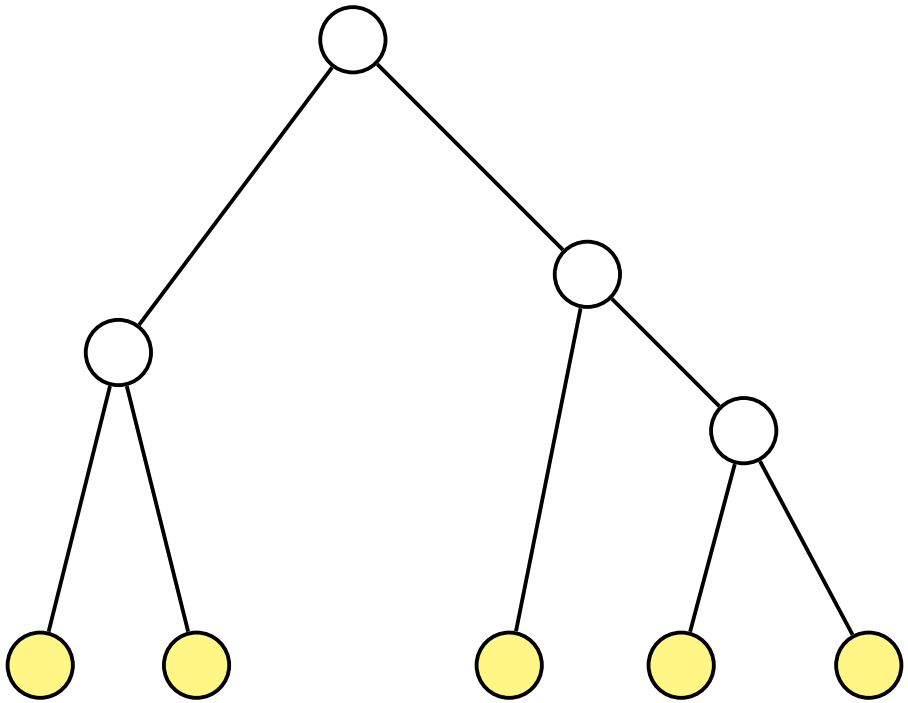
Phylogenetic Trees and Animal Grouping

Let the leaves represent social animals.



Phylogenetic Trees and Animal Grouping

Let the leaves represent social animals.



Describes the genetic relatedness of the animals.

Animal Groups

Durand, Blum and François (2007):

Groups contain more likely animals which are genetically related.

Animal Groups

Durand, Blum and François (2007):

Groups contain more likely animals which are genetically related.

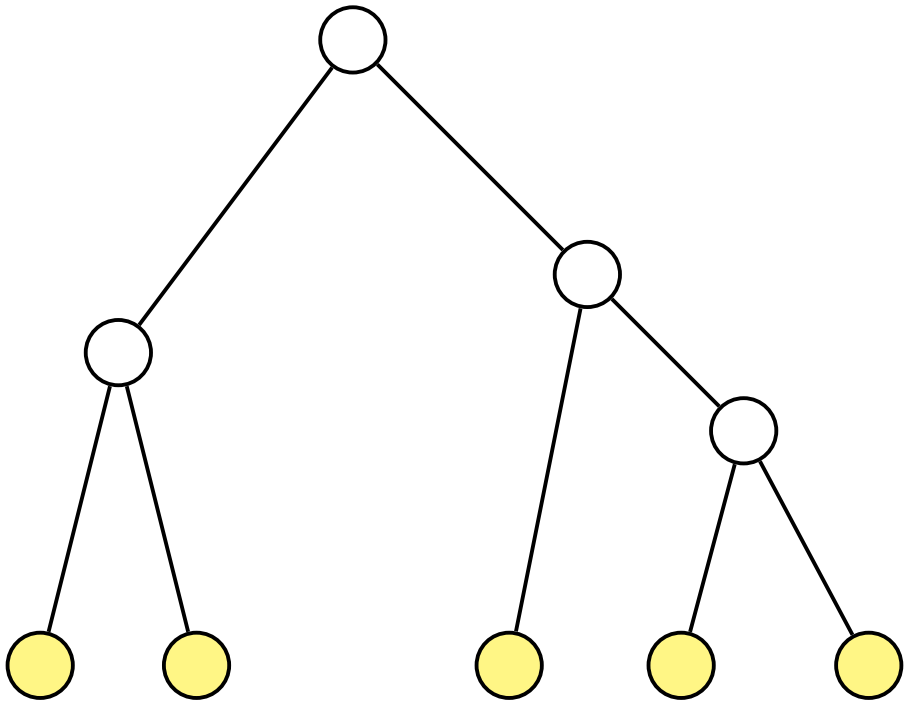
$$\longrightarrow \quad \text{neutral model.}$$

Animal Groups

Durand, Blum and François (2007):

Groups contain more likely animals which are genetically related.

$\longrightarrow$ neutral model.



Clade of a leaf:

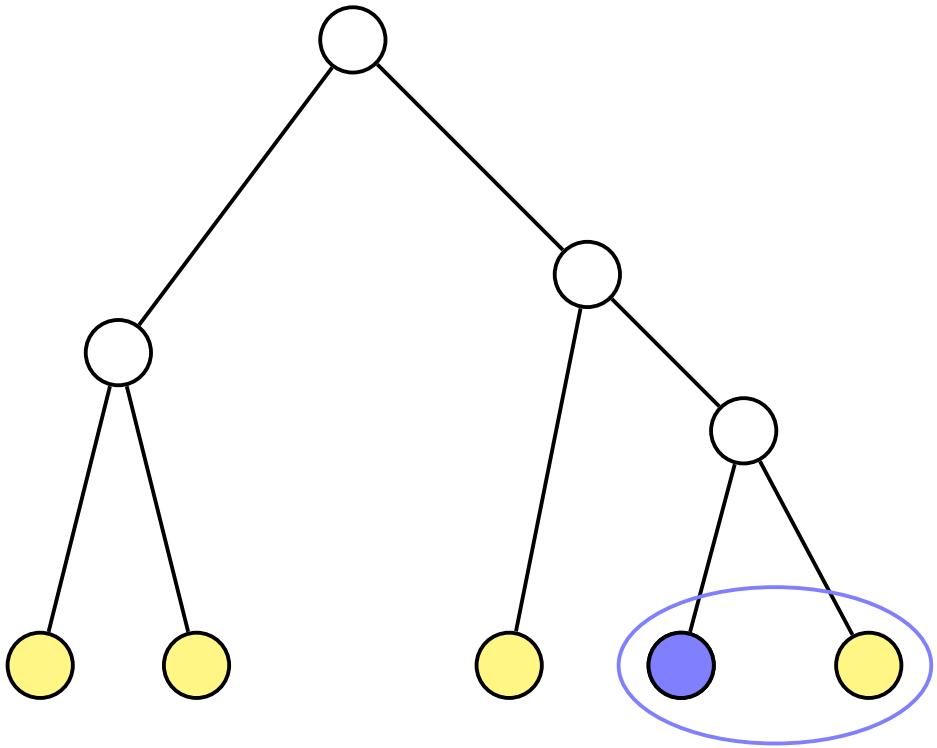
All leaves of the tree rooted at the parent.

Animal Groups

Durand, Blum and François (2007):

Groups contain more likely animals which are genetically related.

$\longrightarrow$ neutral model.



Clade of a leaf:

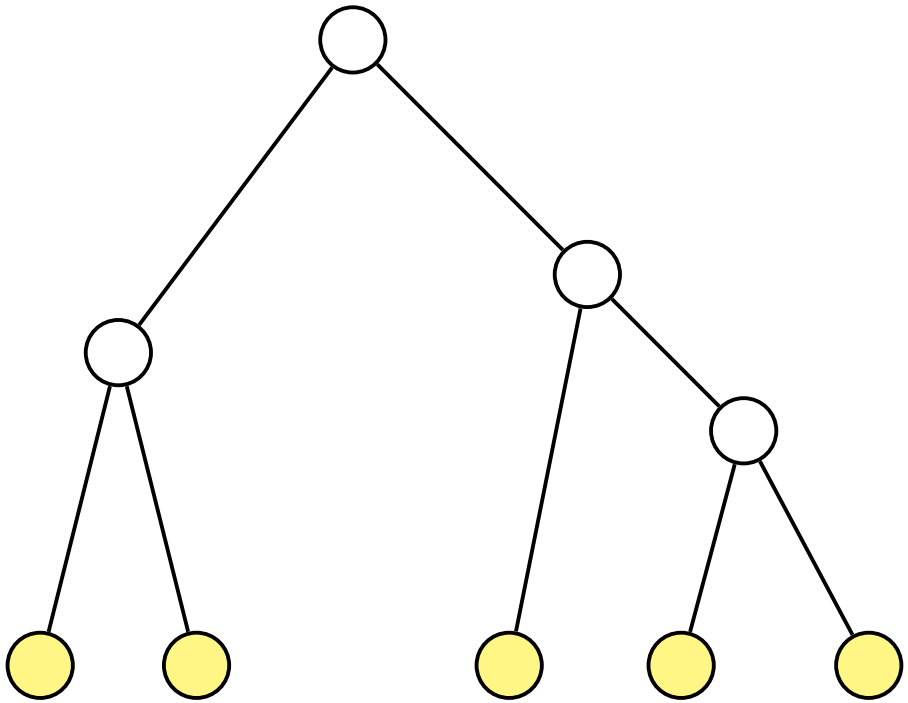
All leaves of the tree rooted at the parent.

Animal Groups

Durand, Blum and François (2007):

Groups contain more likely animals which are genetically related.

$\longrightarrow$ neutral model.



Clade of a leaf:

All leaves of the tree rooted at the parent.

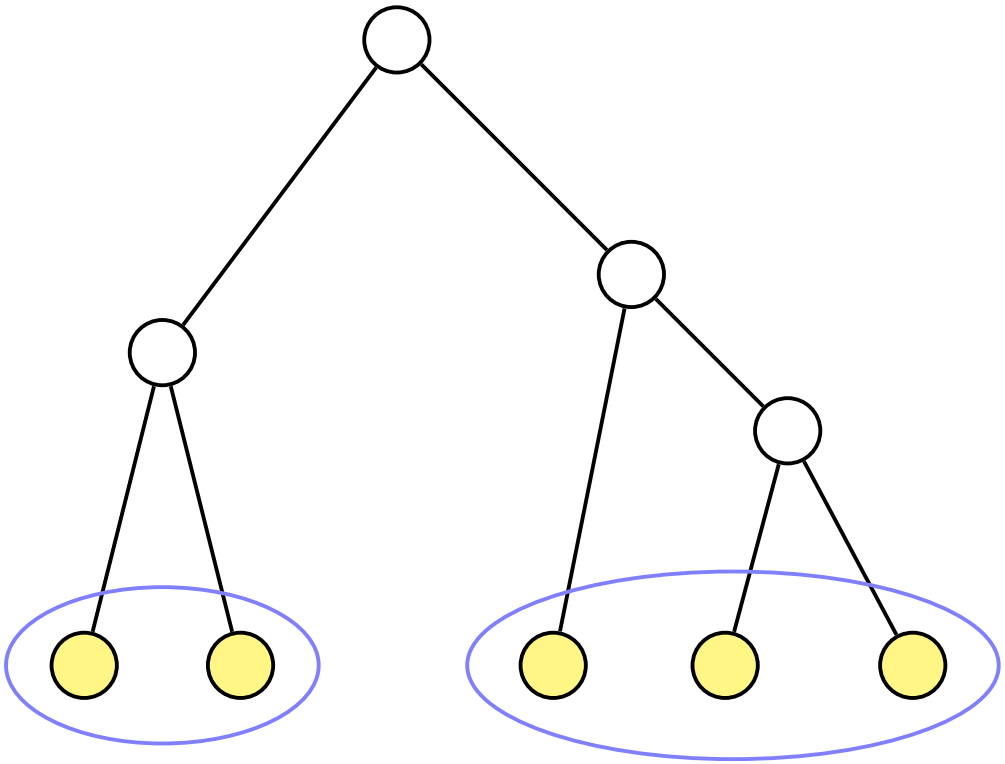
Maximal Clades = Groups

Animal Groups

Durand, Blum and François (2007):

Groups contain more likely animals which are genetically related.

$\longrightarrow$ neutral model.



Clade of a leaf:

All leaves of the tree rooted at the parent.

Maximal Clades = Groups

of Groups

$X_n = \#$ of groups

of Groups

$X_n = \#$ of groups

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{if } I_n = 1 \text{ or } I_n = n-1, \\ X_{I_n} + X^*_{n-I_n}, & \text{otherwise,} \end{cases}$$

where X^*_n is an independent copy of X_n.

of Groups

$X_n = \#$ of groups

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{if } I_n = 1 \text{ or } I_n = n-1, \\ X_{I_n} + X^*_{n-I_n}, & \text{otherwise,} \end{cases}$$

where X_n^* is an independent copy of X_n.

Extra Clustering Model: $0 \leq p < 1$

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{with probability } p \\ \text{same as neutral model}, & \text{otherwise.} \end{cases}$$

of Groups

$X_n = \#$ of groups

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{if } I_n = 1 \text{ or } I_n = n-1, \\ X_{I_n} + X^*_{n-I_n}, & \text{otherwise,} \end{cases}$$

where X_n^* is an independent copy of X_n.

Extra Clustering Model: $0 \leq p < 1$

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{with probability } p \\ \text{same as neutral model}, & \text{otherwise.} \end{cases}$$

For $p = 0$ this is the neutral model.

Expected Number of Groups – YH Model

Theorem (Durand and François; 2010)

We have,

$$\mathbb{E}(X_n) \sim \begin{cases} \dfrac{c(p)}{\Gamma(2(1-p))} n^{1-2p}, & \text{if } p < 1/2; \\ \dfrac{\log n}{2}, & \text{if } p = 1/2; \\ \dfrac{p}{2p-1}, & \text{if } p > 1/2, \end{cases}$$

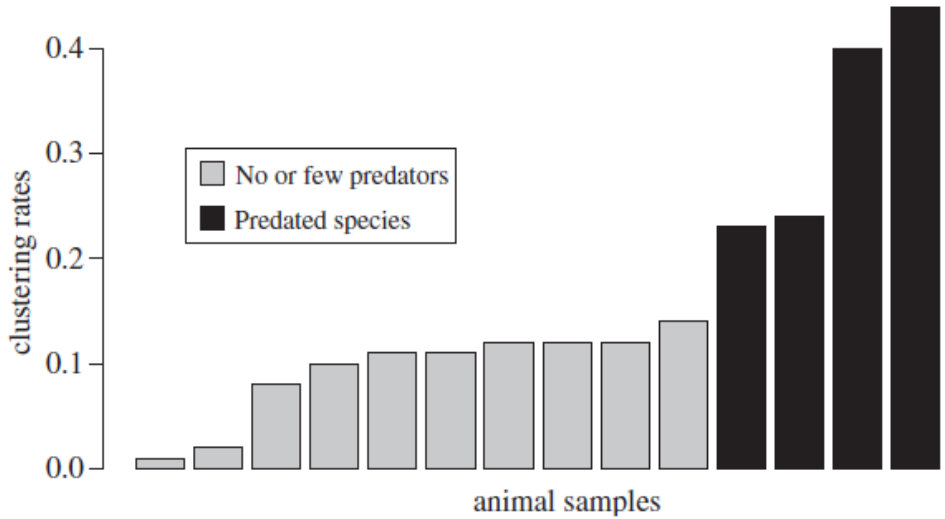
where

$$c(p) := \frac{1}{e^{2(1-p)}} \int_0^1 (1-t)^{-2p} e^{2(1-p)t} \left(1 - (1-p)t^2\right) \mathrm{d}t.$$

Testing for the Neutral Model

Durand, Blum and François (2007):

	Sample size	Number of herds	Rate $\hat{p}$
(A)			
Springboks (browsers)	149	6	0.40
Springboks (graze)	1064	40	0.24
Fallow deers	349	22	0.23
Grant's gazelles	221	6	0.44
Wild camels	227	27	0.14
Kangaroos	348	41	0.12
African savannah elephants	304	45	0.08
	Sample size	Number of packs/prides	Rate $\hat{p}$
(B)			
Yellowstone Wolves 2002	90	14	0.11
Yellowstone Wolves 2004	112	16	0.12
Alaska Wolves	151	30	0.02
Scandinavian wolf	76	12	0.11
Zambia Kafue lions	95	14	0.12
Selous Game lions	51	13	0.00
Serengeti lions	100	16	0.10



$p = 0$ – YH Model

Theorem (Lee; 2012)

We have,

$$\mathrm{Var}(X_n) \sim \frac{(1-e^{-2})^2}{4} n \log n$$

and for $k \geq 3$,

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k \sim (-1)^k \frac{2k}{k-2} \left(\frac{1-e^{-2}}{4}\right)^k n^{k-1}.$$

$p = 0$ – YH Model

Theorem (Lee; 2012)

We have,

$$\mathrm{Var}(X_n) \sim \frac{(1-e^{-2})^2}{4} n \log n$$

and for $k \geq 3$,

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k \sim (-1)^k \frac{2k}{k-2} \left(\frac{1-e^{-2}}{4}\right)^k n^{k-1}.$$

Theorem (Drmota, F., Lee; 2016)

We have,

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)/2}} \xrightarrow{d} N(0,1).$$

$0 < p < 1/2$ – YH model

Theorem (Drmota, F., Lee; 2016)

We have,

$$\frac{X_n}{n^{1-2p}} \xrightarrow{d} X,$$

with convergence of all moments.

$0 < p < 1/2$ – YH model

Theorem (Drmota, F., Lee; 2016)

We have,

$$\frac{X_n}{n^{1-2p}} \xrightarrow{d} X,$$

with convergence of all moments.

Here, the law of X is the sum of a discrete law with mass $p/(1-p)$ at 0 and a continuous law on $[0,\infty)$ with density

$$f(x) = -\delta(p)\frac{1-2p}{1-p}\sum_{k\geq 0}\frac{\delta(p)^k}{k!\Gamma(2(k+1)p-k)}x^k,$$

where

$$\delta(p) = \frac{(1-2p)^2 W_{p,(1-2p)/p}\left(-2(1-p)\right)}{e^{2\pi i p}4^{p-1}(1-p)^{2p}M_{p,(1-2p)/p}\left(-2(1-p)\right)}.$$

$1/2 \leq p < 1$ – YH Model

Theorem (Drmota, F., Lee; 2016)

We have,

$$X_n \xrightarrow{d} X,$$

with convergence of all moments for $1/2 < p < 1$.

$1/2 \leq p < 1$ – YH Model

Theorem (Drmota, F., Lee; 2016)

We have,

$$X_n \xrightarrow{d} X,$$

with convergence of all moments for $1/2 < p < 1$. *Here,* X *is the discrete law with*

$$\mathbb{E}\left(u^X\right) = \frac{1 - \sqrt{1 - 4p(1-p)u}}{2(1-p)}.$$

$1/2 \leq p < 1$ – YH Model

Theorem (Drmota, F., Lee; 2016)

We have,

$$X_n \xrightarrow{d} X,$$

with convergence of all moments for $1/2 < p < 1$. *Here,* X *is the discrete law with*

$$\mathbb{E}\left(u^X\right) = \frac{1 - \sqrt{1 - 4p(1-p)u}}{2(1-p)}.$$

Theorem (Drmota, F., Lee; 2016)

For $p = 1/2$, *we have*

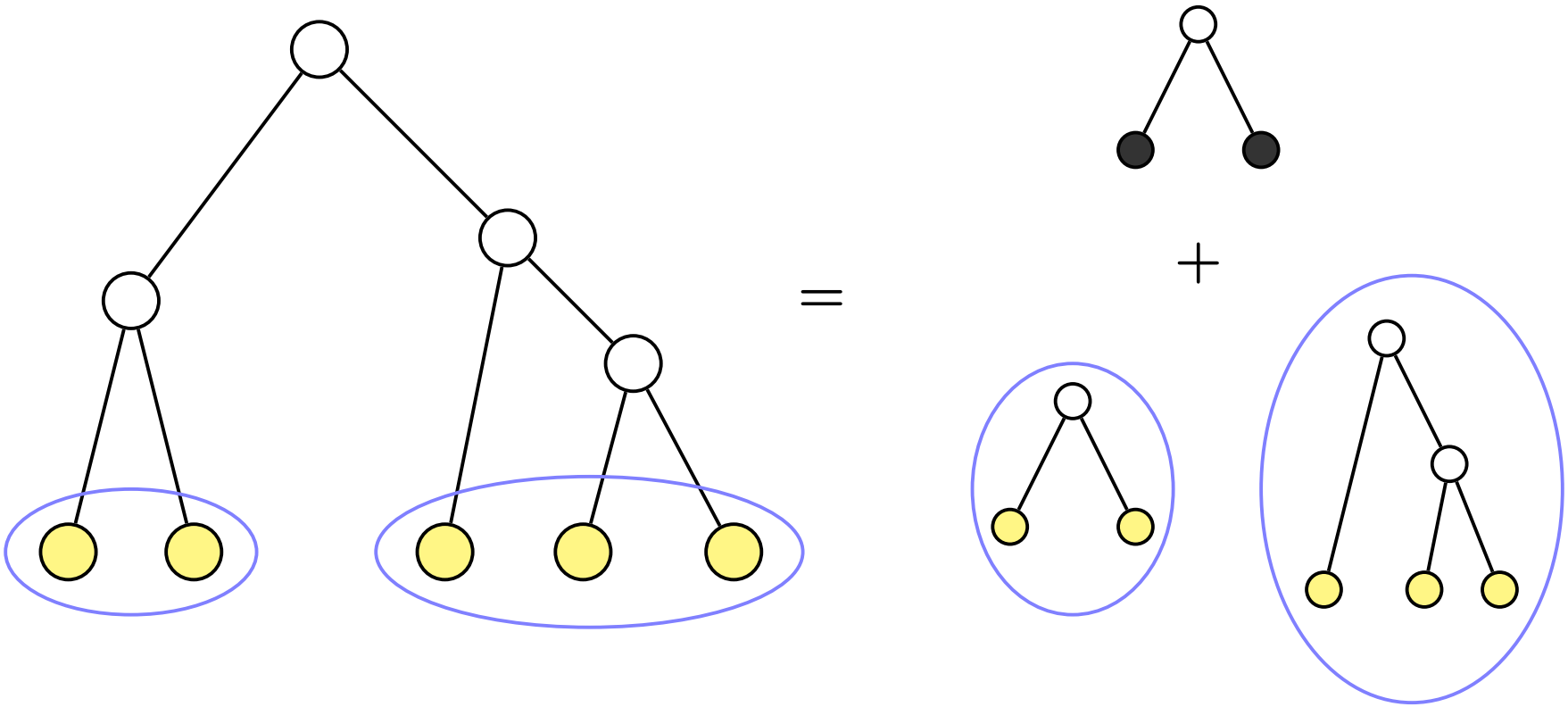
$$\mathbb{E}(X_n^k) \sim \frac{k! J_{2k-1}}{(2k-1)! 2^{2k-1}} \log^{2k-1} n, \qquad J_{2k-1} = (2k-1)! [z^{2k-1}] \tan(z).$$

A Decomposition

Every PT can be decomposed as:

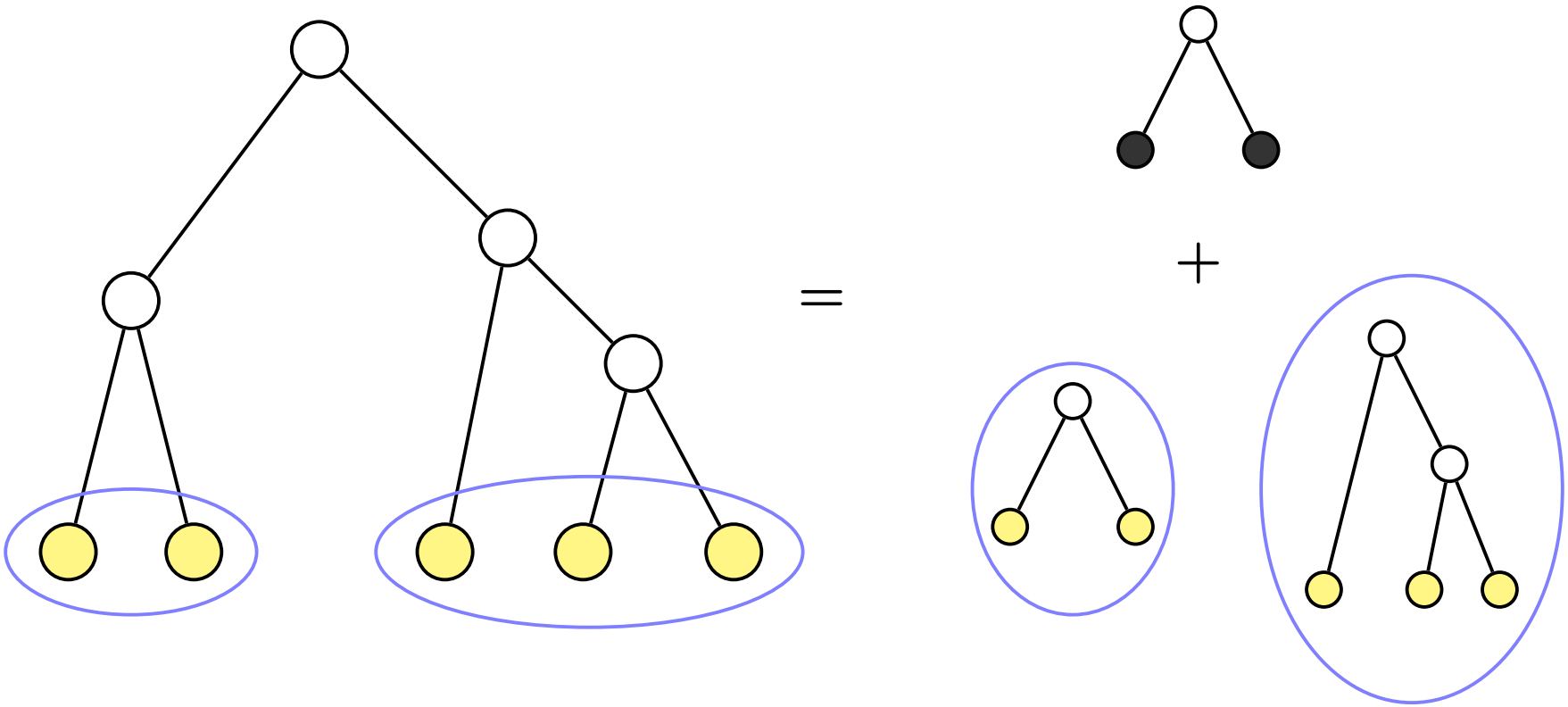
A Decomposition

Every PT can be decomposed as:



A Decomposition

Every PT can be decomposed as:



For the extra clustering model, one has to introduce weights!

Two Generating Functions

Weighted binary trees: internal nodes are weighted by $q := 1 - p$,

$$G(z) = \sum_{n \geq 1} q^{n-1} C_{n-1} z^n = zC(qz),$$

where $C(z) = (1 - \sqrt{1 - 4z})/(2z)$.

Two Generating Functions

Weighted binary trees: internal nodes are weighted by $q := 1 - p$,

$$G(z) = \sum_{n \geq 1} q^{n-1} C_{n-1} z^n = zC(qz),$$

where $C(z) = (1 - \sqrt{1-4z})/(2z)$.

Maximal clades:

$$\begin{aligned} H(z) &= z^2 + \sum_{n \geq 3} (pC_{n-1} + 2qC_{n-2}) z^n \\ &= z^2 + pz(C(z) - 1 - z) + 2qz^2(C(z) - 1). \end{aligned}$$

Two Generating Functions

Weighted binary trees: internal nodes are weighted by $q := 1 - p$,

$$G(z) = \sum_{n \geq 1} q^{n-1} C_{n-1} z^n = zC(qz),$$

where $C(z) = (1 - \sqrt{1-4z})/(2z)$.

Maximal clades:

$$\begin{aligned} H(z) &= z^2 + \sum_{n \geq 3} (pC_{n-1} + 2qC_{n-2}) z^n \\ &= z^2 + pz(C(z) - 1 - z) + 2qz^2(C(z) - 1). \end{aligned}$$

Lemma

We have, $G(H(z)) = z(C(z) - 1)$

of Groups – PDA Model

We have,

$$\mathbb{P}(X_n = k) = \frac{[u^k z^n] G(uH(z))}{C_{n-1}}.$$

of Groups – PDA Model

We have,

$$\mathbb{P}(X_n = k) = \frac{[u^k z^n] G(uH(z))}{C_{n-1}}.$$

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$X_n \xrightarrow{d} X := \mathrm{NB}\left(\frac{1}{2}, \frac{3-2p-p^2}{4}\right) + 1$$

with convergence of all moments.

of Groups – PDA Model

We have,

$$\mathbb{P}(X_n = k) = \frac{[u^k z^n] G(uH(z))}{C_{n-1}}.$$

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$X_n \xrightarrow{d} X := \mathrm{NB}\left(\frac{1}{2}, \frac{3-2p-p^2}{4}\right) + 1$$

with convergence of all moments.

Corollary

We have,

$$\mathbb{E}(X_n) \sim \frac{5+2p+p^2}{2+4p+2p^2}.$$

of Groups of Size m – PDA Model

With $a_m = pC_{m-1} + (2 - \delta_{2,m})qC_{m-2}$,

$$\mathbb{E}(X_n^{[m]} = k) = \frac{[u^k z^n] G(a_m(u-1)z^m + H(z))}{C_{n-1}}.$$

of Groups of Size m – PDA Model

With $a_m = pC_{m-1} + (2 - \delta_{2,m})qC_{m-2}$,

$$\mathbb{E}(X_n^{[m]} = k) = \frac{[u^k z^n] G(a_m(u-1)z^m + H(z))}{C_{n-1}}.$$

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$X_n^{[m]} \xrightarrow{d} X^{[m]} := \mathrm{NB}\left(\frac{1}{2}, \frac{4^{2-m}qa_m}{(1+p)^2 + 4^{2-m}qa_m}\right)$$

with convergence of all moments.

of Groups of Size m – PDA Model

With $a_m = pC_{m-1} + (2 - \delta_{2,m})qC_{m-2}$,

$$\mathbb{E}(X_n^{[m]} = k) = \frac{[u^k z^n] G(a_m(u-1)z^m + H(z))}{C_{n-1}}.$$

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$X_n^{[m]} \xrightarrow{d} X^{[m]} := \mathrm{NB}\left(\frac{1}{2}, \frac{4^{2-m}qa_m}{(1+p)^2 + 4^{2-m}qa_m}\right)$$

with convergence of all moments.

Corollary

We have $\sum_{m \geq 2} \mathbb{E}(X^{[m]}) = \mathbb{E}(X) + 1$.

Largest Groups Size – PDA Model (i)

For the largest group size M_n, we have

$$\mathbb{P}(M_n = n-k) = \frac{[z^k]G'(H(z))[z^{n-k}]H(z)}{C_{n-1}},$$

where $0 \le k < n/2$.

Largest Groups Size – PDA Model (i)

For the largest group size M_n, we have

$$\mathbb{P}(M_n = n-k) = \frac{[z^k]G'(H(z))[z^{n-k}]H(z)}{C_{n-1}},$$

where $0 \le k < n/2$.

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$n - M_n \xrightarrow{d} M,$$

where M has probability generating function $(1+p)/(2F(u/4))$ with

$$F(u) = \sqrt{r(u) - 2p(p-2qu)\sqrt{1-4u}},$$

where $r(u) = 1 - 2p + 2p^2 - 4(1-2p)qu + 4q^2u^2$.

Largest Group Size – PDA Model (ii)

In the limit theorem for M_n, moments do not converge.

Largest Group Size – PDA Model (ii)

In the limit theorem for M_n, moments do not converge.

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$\mathbb{E}(M_n) = n - \frac{2q}{(1+p)\sqrt{\pi}} n^{1/2} + o(n^{1/2})$$

and for $\ell \geq 2$,

$$\mathbb{E}((M_n - \mathbb{E}(M_n))^\ell) \sim (-1)^\ell d_\ell n^{\ell - 1/2},$$

where

$$d_\ell = \frac{q}{(1+p)\sqrt{\pi}} \int_0^{1/2} x^{\ell - 3/2} (1-x)^{-3/2} \mathrm{d}x.$$

Largest Group Size – PDA Model (ii)

In the limit theorem for M_n, moments do not converge.

Theorem (F., Lee, Paningbatan; 2019+)

We have,

$$\mathbb{E}(M_n) = n - \frac{2q}{(1+p)\sqrt{\pi}} n^{1/2} + o(n^{1/2})$$

and for $\ell \geq 2$,

$$\mathbb{E}((M_n - \mathbb{E}(M_n))^\ell) \sim (-1)^\ell d_\ell n^{\ell - 1/2},$$

where

$$d_\ell = \frac{q}{(1+p)\sqrt{\pi}} \int_0^{1/2} x^{\ell - 3/2}(1-x)^{-3/2} \mathrm{d}x.$$

Proof uses singularity analysis and Euler-Maclaurin summation formula.

Summary and Open Problems

Summary and Open Problems

- We studied the correlation coefficient of biodiversity indices.

Summary and Open Problems

- We studied the correlation coefficient of biodiversity indices.
- We studied the extra clustering model when trees are generated by the PDA model.

Summary and Open Problems

- We studied the correlation coefficient of biodiversity indices.
- We studied the extra clustering model when trees are generated by the PDA model.

  Our result shows that on average there is only a finite number of groups all of which are small except one group which contains almost all animals.

Summary and Open Problems

- We studied the correlation coefficient of biodiversity indices.
- We studied the extra clustering model when trees are generated by the PDA model.

  Our result shows that on average there is only a finite number of groups all of which are small except one group which contains almost all animals.
- How about the number of groups of fixed size and largest group size under the YH model?

Summary and Open Problems

- We studied the correlation coefficient of biodiversity indices.
- We studied the extra clustering model when trees are generated by the PDA model.

  Our result shows that on average there is only a finite number of groups all of which are small except one group which contains almost all animals.
- How about the number of groups of fixed size and largest group size under the YH model?

  Mean for number of groups of fixed size was studied by Durand and François (2010).

Summary and Open Problems

- We studied the correlation coefficient of biodiversity indices.
- We studied the extra clustering model when trees are generated by the PDA model.

  Our result shows that on average there is only a finite number of groups all of which are small except one group which contains almost all animals.
- How about the number of groups of fixed size and largest group size under the YH model?

  Mean for number of groups of fixed size was studied by Durand and François (2010). Refined results will appear in:

  A. Paningbatan (2020). Three Combinatorial Topics Arising from Phylogenetics, PhD thesis, in preparation.

謝謝聆聽!

谢谢聆听!