APPLICATIONS OF ANALYTIC COMBINATORICS IN MATHEMATICAL BIOLOGY

(joint with H. Chang, M. Drmota, E. Y. Jin, and Y.-W. Lee)

Michael Fuchs

Department of Applied Mathematics
National Chiao Tung University

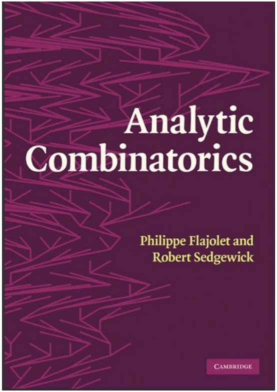
June 30th, 2015

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).

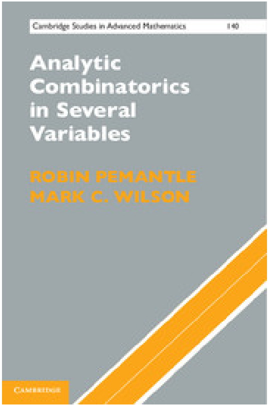
Analytic Combinatorics (ii)



Analytic Combinatorics

Philippe Flajolet and Robert Sedgewick

CAMBRIDGE



Cambridge Studies in Advanced Mathematics 140

Analytic Combinatorics in Several Variables

ROBIN PEMANTLE
MARK C. WILSON

CAMBRIDGE

Outline of the Talk

1. Introduction

Outline of the Talk

1. Introduction

2. Patterns in Phylogenetic Trees

   H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

Outline of the Talk

1. Introduction
2. Patterns in Phylogenetic Trees

   H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

3. Shapley Value and Fair Proportion Index

   M. Fuchs and E. Y. Jin (2015). Equality of Shapley value and fair proportion index in phylogenetic trees, *J. Math. Biol.*, in press.

Outline of the Talk

1. Introduction
2. Patterns in Phylogenetic Trees

   H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

3. Shapley Value and Fair Proportion Index

   M. Fuchs and E. Y. Jin (2015). Equality of Shapley value and fair proportion index in phylogenetic trees, *J. Math. Biol.*, in press.

4. Number of Groups formed by Social Animals

   M. Drmota, M. Fuchs, Y.-W. Lee (2016+). Stochastic analysis of the extra clustering model for animal grouping, in revision.

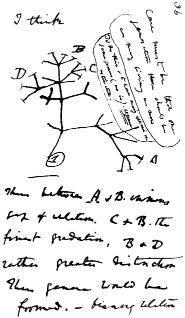
Evolutionary Biology

Charles Darwin
(1809-1882)

Evolutionary Biology

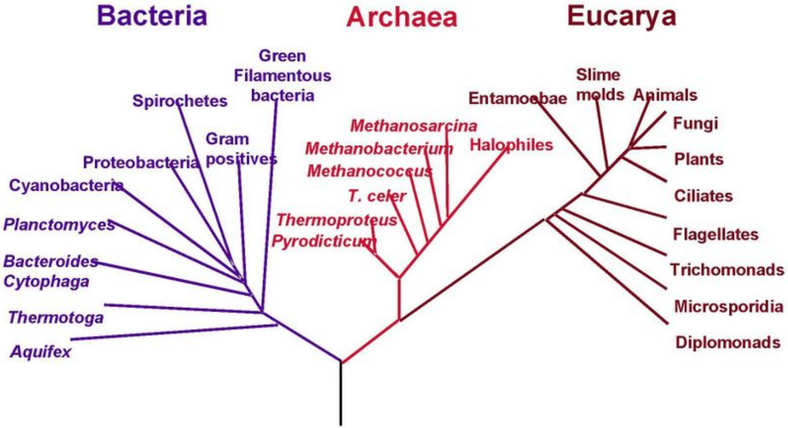
Charles Darwin
(1809-1882)

First notebook on Transmutation of Species (1837)

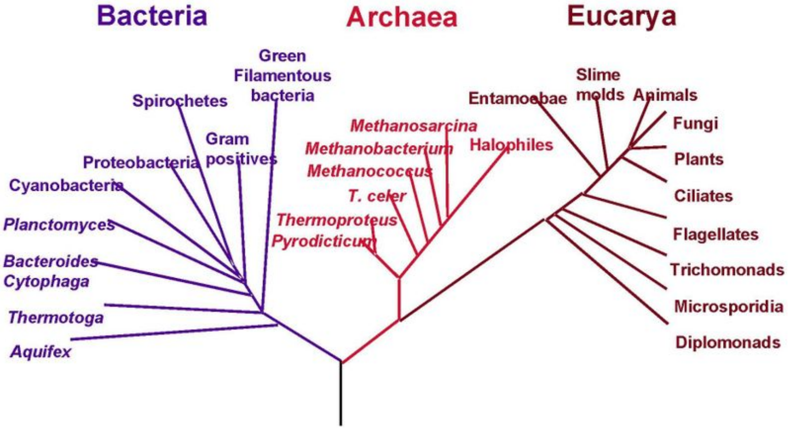


Then between A & B. immense
gap of relation. C & B. the
finest gradation, B & D
rather greater distinction
Thus genera would be
formed. – bearing relation

Phylogenetic trees (=PTs)



Phylogenetic trees (=PTs)



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

Applications of PTs

- Understanding genetic relatedness of species

Applications of PTs

- Understanding genetic relatedness of species
- Understanding the underlying evolutionary process

Applications of PTs

- Understanding genetic relatedness of species
- Understanding the underlying evolutionary process
- Predicting possible future outcomes

Applications of PTs

- Understanding genetic relatedness of species
- Understanding the underlying evolutionary process
- Predicting possible future outcomes
- Testing appropriateness of random models

Applications of PTs

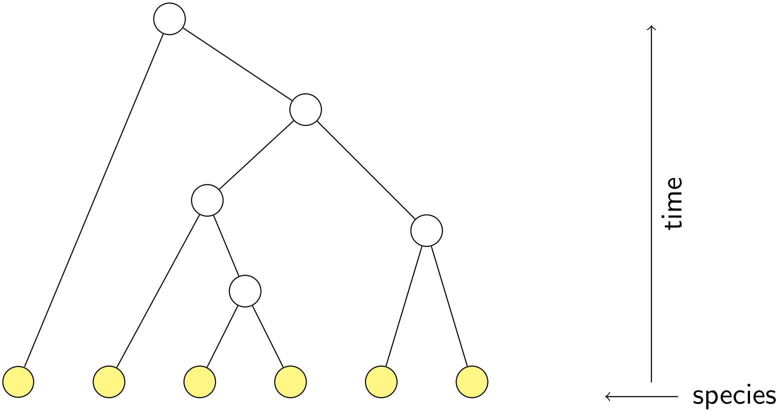
- Understanding genetic relatedness of species
- Understanding the underlying evolutionary process
- Predicting possible future outcomes
- Testing appropriateness of random models
- Making conservation decisions in genetics

Applications of PTs

- Understanding genetic relatedness of species
- Understanding the underlying evolutionary process
- Predicting possible future outcomes
- Testing appropriateness of random models
- Making conservation decisions in genetics
- Modeling the group formation process of social animals
- Etc.

Yule-Harding model or Kingman Coalescent

Example:



Yule-Harding model or Kingman Coalescent

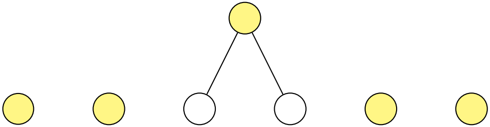
Example:

Random Model 1:

At every time point, two yellow nodes uniformly coalescent.

Yule-Harding model or Kingman Coalescent

Example:

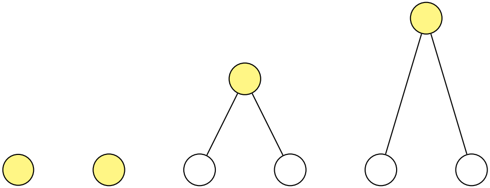


Random Model 1:

At every time point, two yellow nodes uniformly coalescent.

Yule-Harding model or Kingman Coalescent

Example:

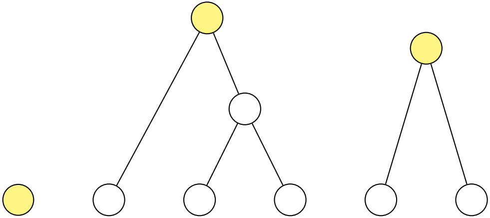


Random Model 1:

At every time point, two yellow nodes uniformly coalescent.

Yule-Harding model or Kingman Coalescent

Example:

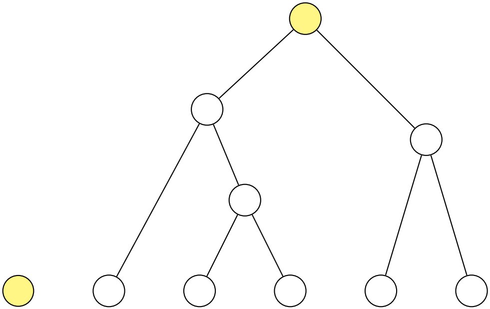


Random Model 1:

At every time point, two yellow nodes uniformly coalescent.

Yule-Harding model or Kingman Coalescent

Example:

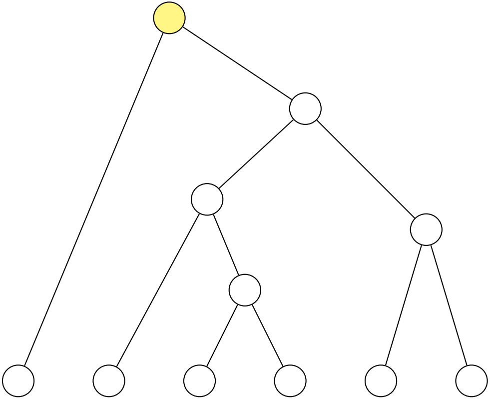


Random Model 1:

At every time point, two yellow nodes uniformly coalescent.

Yule-Harding model or Kingman Coalescent

Example:



Random Model 1:

At every time point, two yellow nodes uniformly coalescent.

Yule-Harding model or Increasing Binary Trees

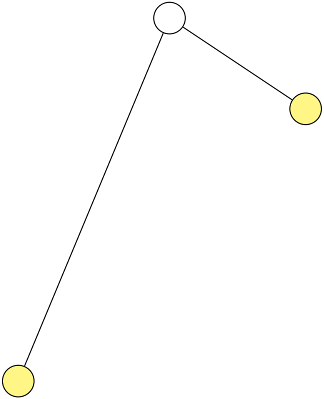
Example:

Random Model 2:

At every time point,
a yellow node is
replaced by a cherry.

Yule-Harding model or Increasing Binary Trees

Example:

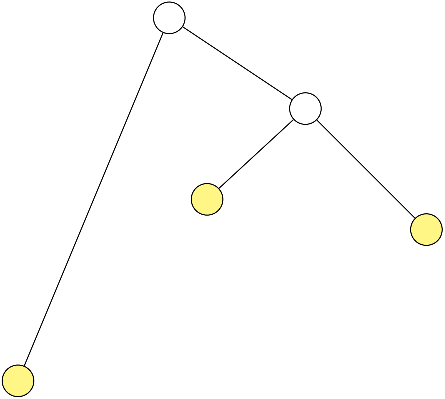


Random Model 2:

At every time point,
a yellow node is
replaced by a cherry.

Yule-Harding model or Increasing Binary Trees

Example:

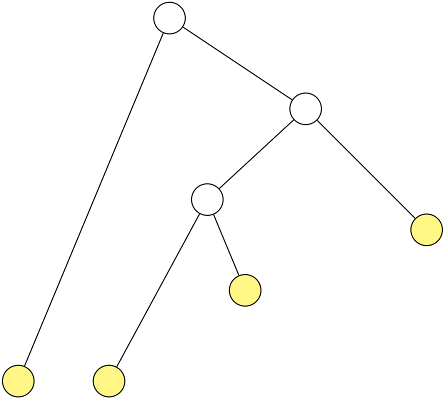


Random Model 2:

At every time point, a yellow node is replaced by a cherry.

Yule-Harding model or Increasing Binary Trees

Example:

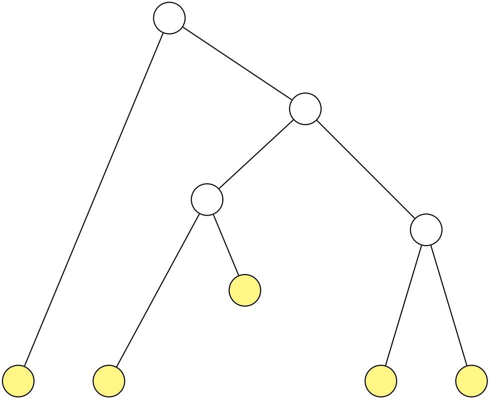


Random Model 2:

At every time point, a yellow node is replaced by a cherry.

Yule-Harding model or Increasing Binary Trees

Example:

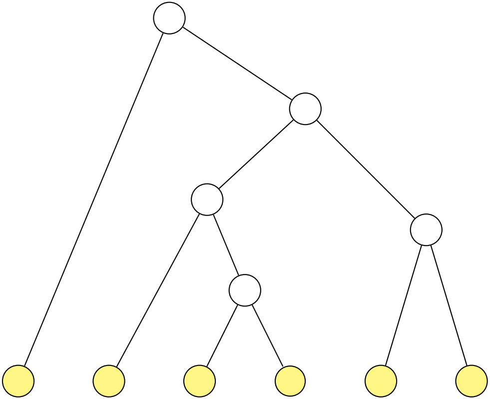


Random Model 2:

At every time point, a yellow node is replaced by a cherry.

Yule-Harding model or Increasing Binary Trees

Example:

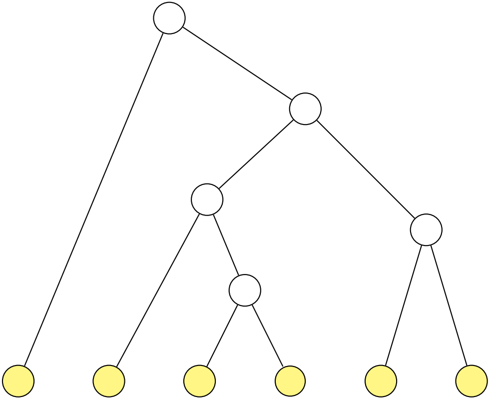


Random Model 2:

At every time point, a yellow node is replaced by a cherry.

Yule-Harding model or Increasing Binary Trees

Example:



Random Model 2:

At every time point, a yellow node is replaced by a cherry.

Random model 1 and random model 2 are the same.

Binary Search Trees

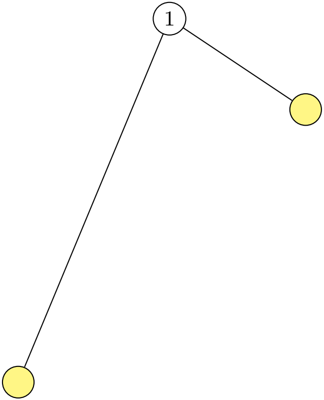
Example: Input $1, 4, 2, 5, 3$

Random Model 3:

All permutations of the input are equally equally.

Binary Search Trees

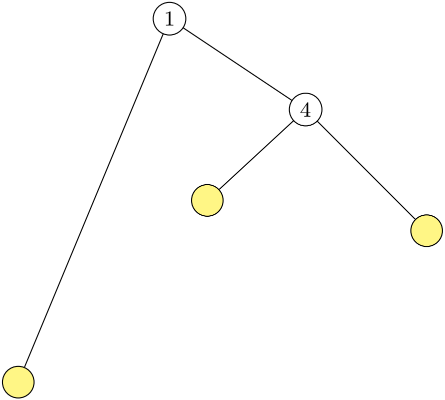
Example: Input $1, 4, 2, 5, 3$



Random Model 3:

All permutations of the input are equally equally.

Example: Input $1, 4, 2, 5, 3$

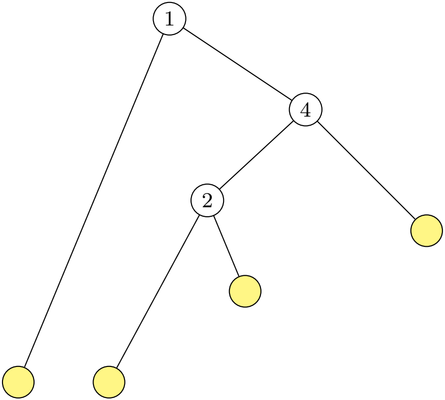


Random Model 3:

All permutations of the input are equally equally.

Binary Search Trees

Example: Input $1, 4, 2, 5, 3$

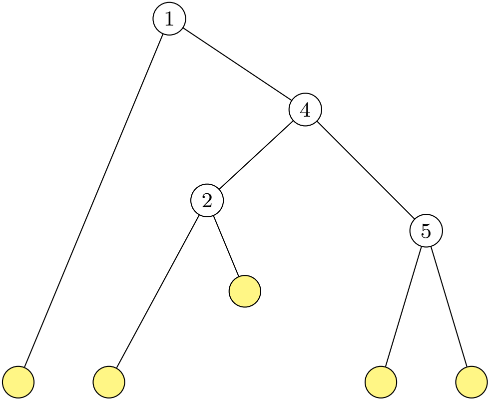


Random Model 3:

All permutations of the input are equally equally.

Binary Search Trees

Example: Input $1, 4, 2, 5, 3$

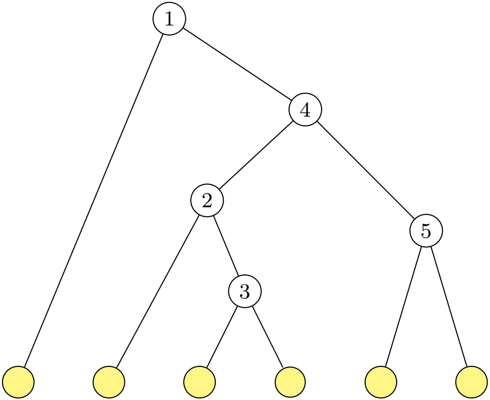


Random Model 3:

All permutations of the input are equally equally.

Binary Search Trees

Example: Input $1, 4, 2, 5, 3$

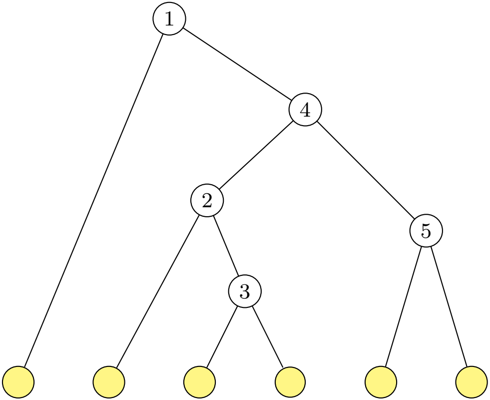


Random Model 3:

All permutations of the input are equally equally.

Binary Search Trees

Example: Input $1, 4, 2, 5, 3$



Random Model 3:

All permutations of the input are equally equally.

Again the same as random model 1 and random model 2.

Patterns in PTs

- k-pronged node (McKenzie and Steel 2000; Rosenberg 2006):

  Node with an induced subtree of size k.

Patterns in PTs

- k-pronged node (McKenzie and Steel 2000; Rosenberg 2006):

  Node with an induced subtree of size k.

- k-caterpillar (Rosenberg 2006):

  Induced subtree of size k with an internal node which is descendent of all other internal nodes.

Patterns in PTs

- k-pronged node (McKenzie and Steel 2000; Rosenberg 2006):

  Node with an induced subtree of size k.

- k-caterpillar (Rosenberg 2006):

  Induced subtree of size k with an internal node which is descendent of all other internal nodes.

- Node with minimal clade size $k \geq 3$ (Blum and François (2005)):

  Node with induced subtree of size k and either right or left subtree is an external node.

Mean and variance of k-pronged nodes

$X_{n,k} = \#$ of k-pronged nodes in random PT of size n.

Mean and variance of k-pronged nodes

$X_{n,k} = \#$ of k-pronged nodes in random PT of size n.

Rosenberg 2006: we have

$$\mu_{n,k} := \mathbb{E}(X_{n,k}) = \frac{2n}{k(k+1)}, \qquad (n > k)$$

and

$$\sigma_{n,k}^2 := \mathrm{Var}(X_{n,k}) = \begin{cases} \dfrac{2(4k^2-3k-4)(k-1)n}{k(k+1)^2(2k-1)(2k+1)}, & \text{if } n > 2k; \\ \dfrac{2(5k-7)(k-1)}{(k+1)^2(2k-1)}, & \text{if } n = 2k; \\ \dfrac{2(k^2+k-2n)n}{k^2(k+1)^2}, & \text{if } 2k > n > k. \end{cases}$$

Mean and variance of k-pronged nodes

$X_{n,k} = \#$ of k-pronged nodes in random PT of size n.

Rosenberg 2006: we have

$$\mu_{n,k} := \mathbb{E}(X_{n,k}) = \frac{2n}{k(k+1)}, \qquad (n > k)$$

and

$$\sigma^2_{n,k} := \mathrm{Var}(X_{n,k}) = \begin{cases} \dfrac{2(4k^2-3k-4)(k-1)n}{k(k+1)^2(2k-1)(2k+1)}, & \text{if } n > 2k; \\ \dfrac{2(5k-7)(k-1)}{(k+1)^2(2k-1)}, & \text{if } n = 2k; \\ \dfrac{2(k^2+k-2n)n}{k^2(k+1)^2}, & \text{if } 2k > n > k. \end{cases}$$

This result + central limit theorem also obtained by Devroye in 1991!

“The phase change”

Question: What happens for $k \to \infty$ as $n \to \infty$?

“The phase change”

Question: What happens for $k \to \infty$ as $n \to \infty$?

Theorem (Feng, Mahmoud, Panholzer; 2008)

(i) *(Normal range) Let* $k = o(\sqrt{n})$. *Then,*

$$\frac{X_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} \mathcal{N}(0,1).$$

(ii) *(Poisson range) Let* $k \sim c\sqrt{n}$. *Then,*

$$X_{n,k} \xrightarrow{d} \mathrm{Po}(2c^{-2}).$$

(iii) *(Degenerate range) Let* $k < n$ *and* $\sqrt{n} = o(k)$. *Then,*

$$X_{n,k} \xrightarrow{L_1} 0.$$

A General Framework

A *pattern* of size k is a set of induced subtrees of size k.

A General Framework

A *pattern* of size k is a set of induced subtrees of size k.

$X_{n,k} = \#$ of occurrence of a pattern of size k in random PT of size n.

A General Framework

A *pattern* of size k is a set of induced subtrees of size k.

$X_{n,k} = \#$ of occurrence of a pattern of size k in random PT of size n.

We have,

$$X_{n,k} \overset{d}{=} X_{I_n,k} + X^*_{n-I_n,k},$$

where $X_{k,k} = \text{Bernoulli}(p_k)$, $X_{I_n,k}$ and $X^*_{n-I_n,k}$ are conditionally independent given I_n, and $I_n = \text{Unif}\{1, \ldots, n-1\}$

A General Framework

A *pattern* of size k is a set of induced subtrees of size k.

$X_{n,k} = \#$ of occurrence of a pattern of size k in random PT of size n.

We have,

$$X_{n,k} \stackrel{d}{=} X_{I_n,k} + X^*_{n-I_n,k},$$

where $X_{k,k} = \text{Bernoulli}(p_k)$, $X_{I_n,k}$ and $X^*_{n-I_n,k}$ are conditionally independent given I_n, and $I_n = \text{Unif}\{1, \ldots, n-1\}$

Here,

p_k	shape parameter
1	$\#$ of k-pronged nodes
$2/(k-1)$	$\#$ of nodes with minimal clade size k
$2^{k-2}/(k-1)!$	$\#$ of k caterpillars

Mean Value and Variance

We have

$$\mu_{n,k} = \frac{2p_k n}{k(k+1)}, \qquad (n > k),$$

and

$$\sigma_{n,k}^2 = \frac{2(4k^3 + 4k^2 - k - 1 - (11k^2 - 5)p_k)p_k n}{k(k+1)^2(2k-1)(2k+1)}$$

for $n > 2k$.

Mean Value and Variance

We have

$$\mu_{n,k} = \frac{2p_k n}{k(k+1)}, \qquad (n > k),$$

and

$$\sigma_{n,k}^2 = \frac{2(4k^3 + 4k^2 - k - 1 - (11k^2 - 5)p_k)p_k n}{k(k+1)^2(2k-1)(2k+1)}$$

for $n > 2k$.

Note that

$$\mu_{n,k} \sim \sigma_{n,k}^2 \sim \frac{2p_k n}{k^2}$$

for $n > 2k$ and $k \to \infty$.

Poisson Approximation

Theorem (Chang and F.; 2010)

Let $k < n$ and $k \to \infty$. Then,

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \begin{cases} \mathcal{O}\left(p_k/k\right), & \text{if } \mu_{n,k} \geq 1; \\ \mathcal{O}\left(p_k/k \cdot \mu_{n,k}\right), & \text{if } \mu_{n,k} < 1. \end{cases}$$

Poisson Approximation

Theorem (Chang and F.; 2010)

Let $k < n$ and $k \to \infty$. Then,

$$d_{TV}(X_{n,k}, \mathrm{Po}(\mu_{n,k})) = \begin{cases} \mathcal{O}\left(p_k/k\right), & \textit{if } \mu_{n,k} \geq 1; \\ \mathcal{O}\left(p_k/k \cdot \mu_{n,k}\right), & \textit{if } \mu_{n,k} < 1. \end{cases}$$

Recently re-proved by Holmgren and Janson with Stein's method:

C. Holmgren and S. Janson (2015). Limit laws for functions of fringe trees for binary search trees and recursive trees, *Electronic J. Probability*, **20:4**, 1–51.

Berry-Esseen bound and LLT

Theorem (Chang and F.; 2010)

For $\mu_{n,k} \to \infty$,

$$\sup_{x\in\mathbb{R}} \left| P\left(\frac{X_{n,k}-\mu_{n,k}}{\sigma_{n,k}} < x\right) - \Phi(x) \right| = \mathcal{O}\left(\frac{k}{\sqrt{p_k n}}\right).$$

Berry-Esseen bound and LLT

Theorem (Chang and F.; 2010)

For $\mu_{n,k} \to \infty$,

$$\sup_{x\in\mathbb{R}} \left| P\left(\frac{X_{n,k}-\mu_{n,k}}{\sigma_{n,k}} < x\right) - \Phi(x) \right| = \mathcal{O}\left(\frac{k}{\sqrt{p_k n}}\right).$$

Theorem (Chang and F.; 2010)

For $\mu_{n,k} \to \infty$,

$$P(X_{n,k} = \lfloor \mu_{n,k} + x\sigma_{n,k}^2 \rfloor) = \frac{e^{-x^2/2}}{\sqrt{2\pi}\sigma_{n,k}}\left(1 + \mathcal{O}\left((1+|x|^3)\frac{k}{\sqrt{p_k n}}\right)\right),$$

uniformly in $x = o((p_k n)^{1/6}/k^{1/3})$.

Short Summary

We have proved:

- Explicit expressions for mean and variance.
- Local limit theorem + Berry-Esseen bound for the largest possible range of k.
- Poisson approximation + rate whenever $k \to \infty$.

Short Summary

We have proved:

- Explicit expressions for mean and variance.
- Local limit theorem + Berry-Esseen bound for the largest possible range of k.
- Poisson approximation + rate whenever $k \to \infty$.

So, for $k \ll n$ the number of occurrences of the pattern can be approximated by the normal distribution, whereas for the remaining range a Poisson random variable should be used.

Short Summary

We have proved:

- Explicit expressions for mean and variance.
- Local limit theorem + Berry-Esseen bound for the largest possible range of k.
- Poisson approximation + rate whenever $k \to \infty$.

So, for $k \ll n$ the number of occurrences of the pattern can be approximated by the normal distribution, whereas for the remaining range a Poisson random variable should be used.

In particular, the phase change occurs much earlier than predicted by Feng, Mahmoud, and Panholzer.

Lloyd Shapley

Lloyd Shapley
(1923-)

Lloyd Shapley

Lloyd Shapley
(1923-)

Shapley value:

Measure of importance of each player in a cooperative game

Lloyd Shapley

Lloyd Shapley
(1923-)

Shapley value:

Measure of importance of each player in a cooperative game

⟶ recently used as prioritization tool of taxa in phylogenetics

Shapley Value and Modified Shapley Value

T . . . PT;

a . . . taxon (=leaf) of T.

Shapley Value and Modified Shapley Value

T . . . PT;

a . . . taxon (=leaf) of T.

Shapley value $\mathrm{SV}_T(a)$:

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

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

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

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

Fair Proportion Index

Fair proportion index $\mathrm{FP}_T(a)$:

$$\mathrm{FP}_T(a) = \sum_e \frac{1}{D_e}$$

with D_e the number of taxa below e.

Fair Proportion Index

Fair proportion index $\mathrm{FP}_T(a)$:

$$\mathrm{FP}_T(a) = \sum_e \frac{1}{D_e}$$

with D_e the number of taxa below e.

Selected (somehow arbitrarily) by Zoological Society of London for EDGE of Existence conservation program!

Fair Proportion Index

Fair proportion index $\mathrm{FP}_T(a)$:

$$\mathrm{FP}_T(a) = \sum_e \frac{1}{D_e}$$

with D_e the number of taxa below e.

Selected (somehow arbitrarily) by Zoological Society of London for EDGE of Existence conservation program!

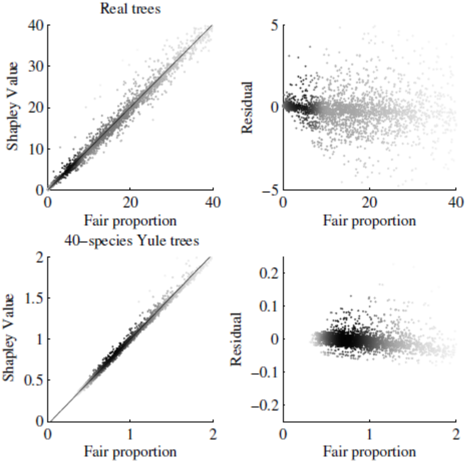
FP_n = fair proportion index of random taxon in random PT of size n:

$$\mathrm{FP}_n|(I_n = j) = \begin{cases} \frac{1}{j} + \mathrm{FP}_j, & \text{with probability } j/n; \\ \frac{1}{n-j} + \mathrm{FP}_{n-j}, & \text{with probability } (n-j)/n, \end{cases}$$

where $I_n = \mathrm{Unif}\{1, \ldots, n-1\}$.

Strong Correlation between $\widetilde{\mathrm{SV}}$ and FP

Hartmann (2013):



$\mathrm{SV} = \mathrm{FP}$

Assume a is in left subtree T_l and $|T_l| = j$.

$\mathrm{SV} = \mathrm{FP}$

Assume a is in left subtree T_l and $|T_l| = j$.

Lemma

We have,

$$\mathrm{SV}_T(a) = \frac{1}{j} + \mathrm{SV}_{T_l}(a)$$

and

$$\mathrm{FP}_T(a) = \frac{1}{j} + \mathrm{FP}_{T_l}(a).$$

$\mathrm{SV} = \mathrm{FP}$

Assume a is in left subtree T_l and $|T_l| = j$.

Lemma

We have,

$$\mathrm{SV}_T(a) = \frac{1}{j} + \mathrm{SV}_{T_l}(a)$$

and

$$\mathrm{FP}_T(a) = \frac{1}{j} + \mathrm{FP}_{T_l}(a).$$

Theorem (F. and Jin; 2015)

We have,

$$\mathrm{SV}_T(a) = \mathrm{FP}_T(a).$$

$\widetilde{\mathrm{SV}}$ and FP (i)

$D_T(a)$ = depth of a in T:

$$\mathrm{FP}_T(a) = \mathrm{SV}_T(a) = \widetilde{\mathrm{SV}}_T(a) + \frac{D_T(a)}{n}.$$

$\widetilde{\mathrm{SV}}$ and FP (i)

$D_T(a)$ = depth of a in T:

$$\mathrm{FP}_T(a) = \mathrm{SV}_T(a) = \widetilde{\mathrm{SV}}_T(a) + \frac{D_T(a)}{n}.$$

D_n = depth of random taxon in random PT of size n.

Lemma

We have,

$$\mathrm{Var}(\mathrm{FP}_n) = 10 - 6H_{n-1}^{(2)} - \frac{6}{n} - \frac{4}{n^2} \sim 10 - \pi^2$$
$$\mathrm{Var}(D_n) = 2H_n - 4H_n^{(2)} + 2 \sim 2\log n,$$

where $H_n = \sum_{1\le j\le n} 1/j$ *and* $H_n^{(2)} = \sum_{1\le j\le n} 1/j^2$.

$\widetilde{\mathrm{SV}}$ and FP (ii)

Lemma

We have,

$$\mathrm{Cov}(\mathrm{FP}_n, D_n) = 4H_{n-1}^{(2)} - 6 + \frac{2}{n} + \frac{4}{n^2} \sim \frac{2\pi^2}{6} - 6.$$

$\widetilde{\mathrm{SV}}$ and FP (ii)

Lemma

We have,

$$\mathrm{Cov}(\mathrm{FP}_n, D_n) = 4H_{n-1}^{(2)} - 6 + \frac{2}{n} + \frac{4}{n^2} \sim \frac{2\pi^2}{6} - 6.$$

Theorem (F. and Jin; 2015)

The correlation coefficient $\rho(\widetilde{\mathrm{SV}}_n, \mathrm{FP}_n)$ *of modified Shapley value and fair proportion index tends to* 1, *i.e.,*

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

Short Summary

- Shapley value was introduced for the purpose of making conservation decisions in genetics.

Short Summary

- Shapley value was introduced for the purpose of making conservation decisions in genetics.
- Fair proportion index was used by Zoological Society of London but its biodiversity value was unclear.

Short Summary

- Shapley value was introduced for the purpose of making conservation decisions in genetics.
- Fair proportion index was used by Zoological Society of London but its biodiversity value was unclear.
- Strong correlation between modified Shapley value and fair proportion index was observed by Hartmann and others.

Short Summary

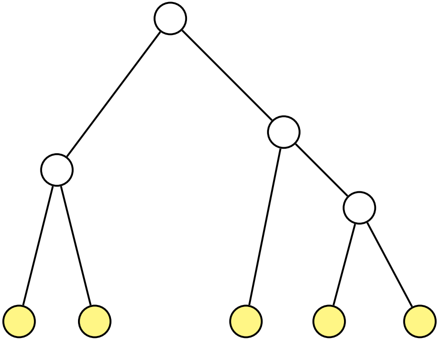
- Shapley value was introduced for the purpose of making conservation decisions in genetics.
- Fair proportion index was used by Zoological Society of London but its biodiversity value was unclear.
- Strong correlation between modified Shapley value and fair proportion index was observed by Hartmann and others.
- We proved that Shapley value equals fair proportion index.

Short Summary

- Shapley value was introduced for the purpose of making conservation decisions in genetics.
- Fair proportion index was used by Zoological Society of London but its biodiversity value was unclear.
- Strong correlation between modified Shapley value and fair proportion index was observed by Hartmann and others.
- We proved that Shapley value equals fair proportion index.
- We showed that correlation coefficient of modified Shapley value and fair proportion index tends to 1 in Yule-Harding model and other random models.

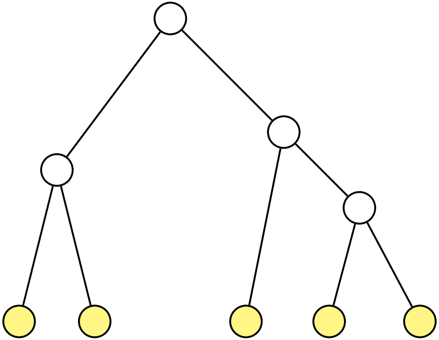
Social Animals

Consider n animals of a class of social animals. Construct a random PT with leaves representing the animals.



Social Animals

Consider n animals of a class of social animals. Construct a random PT with leaves representing the animals.



Describes the genetic relatedness of the animals.

Animal Groups

Durand, Blum and François (2007):

Groups are formed more likely by animals which are genetically related.

Animal Groups

Durand, Blum and François (2007):

Groups are formed more likely by animals which are genetically related.

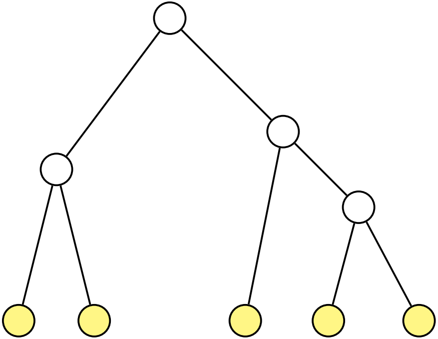
$\longrightarrow$ neutral model.

Animal Groups

Durand, Blum and François (2007):

Groups are formed more likely by animals which are genetically related.

⟶ neutral model.



Clade of a leaf:

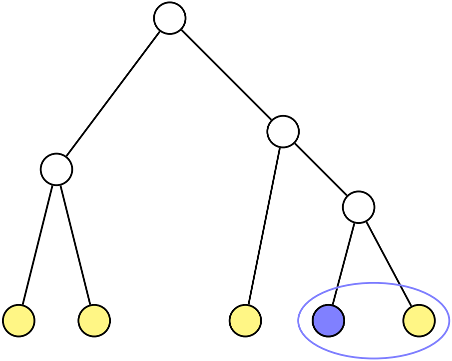
All leafs of the tree rooted at the parent.

Animal Groups

Durand, Blum and François (2007):

Groups are formed more likely by animals which are genetically related.

⟶ neutral model.



Clade of a leaf:

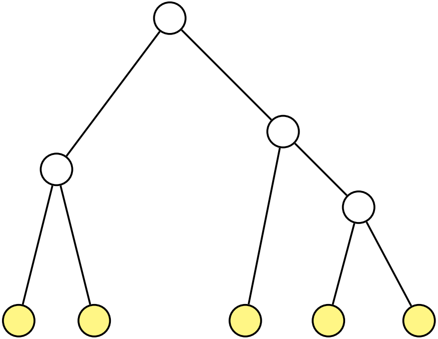
All leafs of the tree rooted at the parent.

Animal Groups

Durand, Blum and François (2007):

Groups are formed more likely by animals which are genetically related.

$\longrightarrow$ neutral model.



# of groups

$\Updownarrow$

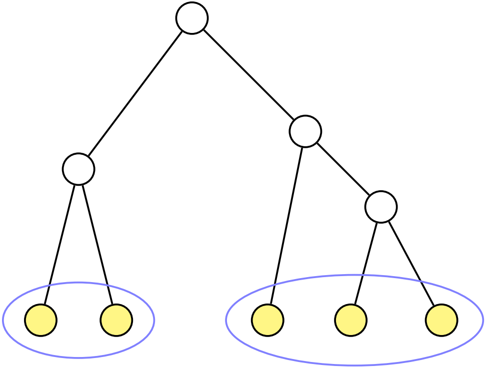
# of maximal clades

Animal Groups

Durand, Blum and François (2007):

Groups are formed more likely by animals which are genetically related.

$\longrightarrow$ neutral model.



# of groups

$\Updownarrow$

# of maximal clades

$\Updownarrow$

2

of Groups

$X_n = \#$ of groups under the Yule Harding model

of Groups

$X_n = \#$ of groups under the Yule Harding model

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 $I_n = \text{Uniform}\{1, \dots, n-1\}$ is the # of animals in the left subtree and X^*_n is an independent copy of X_n.

of Groups

$X_n = \#$ of groups under the Yule Harding model

We have,

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

where $I_n = \text{Uniform}\{1, \dots, n-1\}$ is the # of animals in the left subtree and X_n^* is an independent copy of X_n.

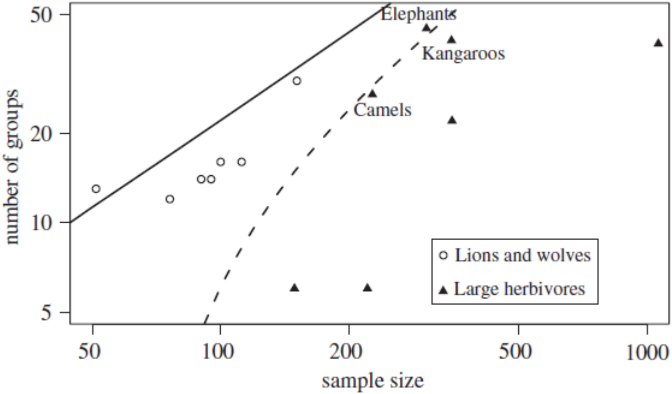
Theorem (Durand and François; 2010)

We have,

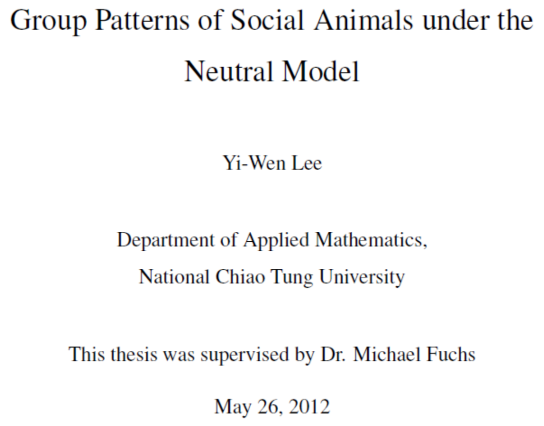
$$\mathbb{E}(X_n) \sim an \qquad \left(a := \frac{1-e^{-2}}{4}\right).$$

Comparison with Real-life Data

Durand, Blum and François (2007) presented the following data:



Yi-Wen’s Thesis (2012)



Group Patterns of Social Animals under the Neutral Model

Yi-Wen Lee

Department of Applied Mathematics,
National Chiao Tung University

This thesis was supervised by Dr. Michael Fuchs

May 26, 2012

Variance and SLLN

Theorem (Lee; 2012)

We have,

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

Variance and SLLN

Theorem (Lee; 2012)

We have,

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

Theorem (Lee; 2012)

We have,

$$P\left(\lim_{n\to\infty}\left|\frac{X_n}{\mathbb{E}(X_n)} - 1\right| = 0\right) = 1.$$

For SLLN, X_n is constructed on the same probability space via the tree evolution process underlying the Yule-Harding model.

Method of Moments

Theorem

Assume that $\mathbb{E}(X_n^k) \longrightarrow \mathbb{E}(X^k)$ *for all* $k \geq 1$ *and that* X *is uniquely characterised by its sequence of moments. Then,*

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

Method of Moments

Theorem

Assume that $\mathbb{E}(X_n^k) \longrightarrow \mathbb{E}(X^k)$ *for all* $k \geq 1$ *and that* X *is uniquely characterised by its sequence of moments. Then,*

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

Many sufficient conditions for X being uniquely characterized by its moments are known, e.g.,

$$\sum_{k\geq 1} \mathbb{E}(X^k)\frac{z^k}{k!}$$

has a positive radius of convergence.

Higher Moments

Theorem (Lee; 2012)

For all $k \geq 3$,

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k \sim (-1)^k \frac{2k}{k-2} a^k n^{k-1}.$$

Higher Moments

Theorem (Lee; 2012)

For all $k \geq 3$,

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k \sim (-1)^k \frac{2k}{k-2} a^k n^{k-1}.$$

This implies that all moments larger than two of

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)}}$$

tend to infinity!

Higher Moments

Theorem (Lee; 2012)

For all $k \geq 3$,

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k \sim (-1)^k \frac{2k}{k-2} a^k n^{k-1}.$$

This implies that all moments larger than two of

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)}}$$

tend to infinity!

Question: Is there a limit distribution?

Random Recursive Trees

Unordered, rooted trees.

Random Recursive Trees

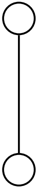
Unordered, rooted trees.

Uniformly choose one of the nodes and attach a child.

Random Recursive Trees

Unordered, rooted trees.

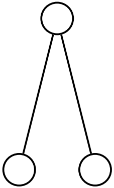
Uniformly choose one of the nodes and attach a child.



Random Recursive Trees

Unordered, rooted trees.

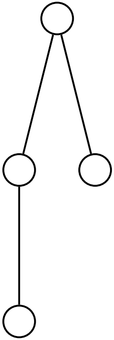
Uniformly choose one of the nodes and attach a child.



Random Recursive Trees

Unordered, rooted trees.

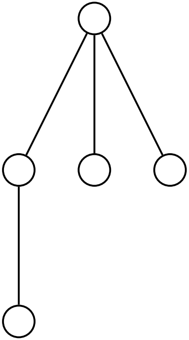
Uniformly choose one of the nodes and attach a child.



Random Recursive Trees

Unordered, rooted trees.

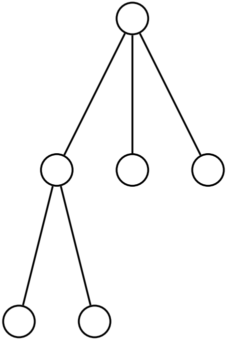
Uniformly choose one of the nodes and attach a child.



Random Recursive Trees

Unordered, rooted trees.

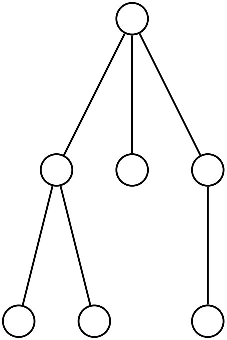
Uniformly choose one of the nodes and attach a child.



Random Recursive Trees

Unordered, rooted trees.

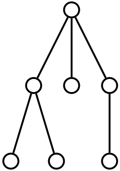
Uniformly choose one of the nodes and attach a child.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

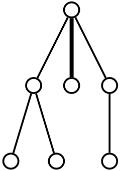
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

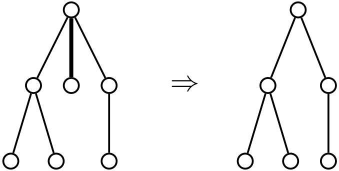
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

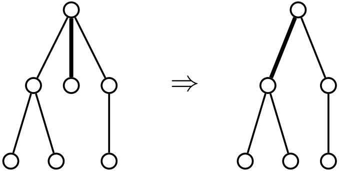
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

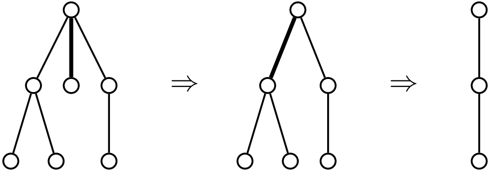
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

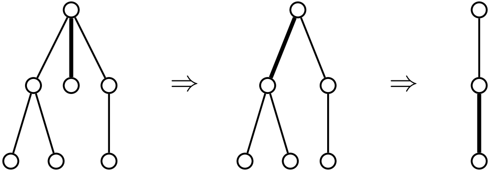
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

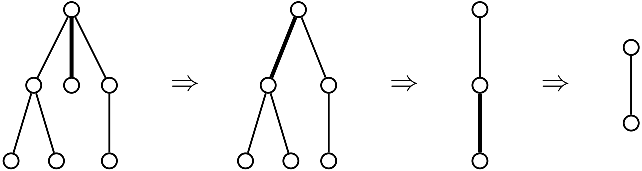
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

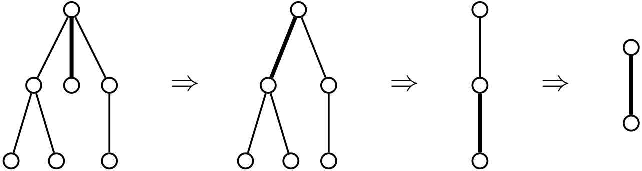
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

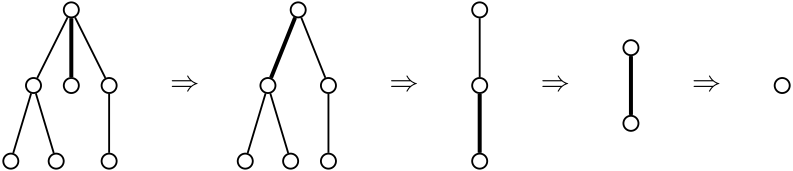
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

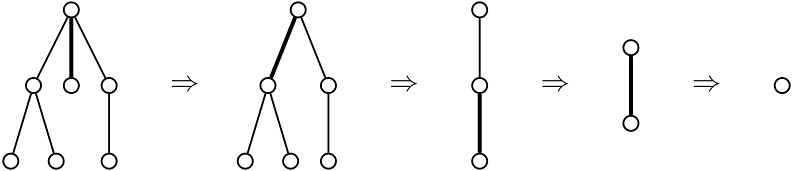
Randomly pick an edge and remove it; retain the tree containing the root.



Cutting Down Random Recursive Trees

Meir and Moon (1974):

Randomly pick an edge and remove it; retain the tree containing the root.



$$\begin{aligned} Y_n &= \text{ number of steps until tree is destroyed} \\ &= \text{ number of edges cut} = 4. \end{aligned}$$

Mean, Variance and Higher Moments

Theorem (Panholzer; 2004)

We have,

$$\mathbb{E}(Y_n) \sim \frac{n}{\log n}$$

and for $k \geq 2$

$$\mathbb{E}(Y_n - \mathbb{E}(Y_n))^k \sim \frac{(-1)^k}{k(k-1)} \cdot \frac{n^k}{\log^{k+1} n}.$$

Mean, Variance and Higher Moments

Theorem (Panholzer; 2004)

We have,

$$\mathbb{E}(Y_n) \sim \frac{n}{\log n}$$

and for $k \geq 2$

$$\mathbb{E}(Y_n - \mathbb{E}(Y_n))^k \sim \frac{(-1)^k}{k(k-1)} \cdot \frac{n^k}{\log^{k+1} n}.$$

Thus, again the limit law of

$$\frac{Y_n - \mathbb{E}(Y_n)}{\sqrt{\mathrm{Var}(Y_n)}}$$

cannot obtained from the method of moments!

Limit Law

Theorem (Drmota, Iksanov, Moehle, Roessler; 2009)

We have,

$$\frac{\log^2 n}{n} Y_n - \log n - \log\log n \xrightarrow{d} Y$$

with

$$\mathbb{E}(e^{i\lambda Y}) = e^{i\lambda \log|\lambda| - \pi|\lambda|/2}.$$

The law of Y is spectrally negative stable with index of stability 1.

Limit Law

Theorem (Drmota, Iksanov, Moehle, Roessler; 2009)

We have,

$$\frac{\log^2 n}{n} Y_n - \log n - \log\log n \xrightarrow{d} Y$$

with

$$\mathbb{E}(e^{i\lambda Y}) = e^{i\lambda \log|\lambda| - \pi|\lambda|/2}.$$

The law of Y is spectrally negative stable with index of stability 1.

Different proofs of this result exist.

Limit Law of X_n

Theorem (Drmota, F., Lee; 2014)

We have,

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

Limit Law of X_n

Theorem (Drmota, F., Lee; 2014)

We have,

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

For the proof, we use singularity perturbation analysis.

Limit Law of X_n

Theorem (Drmota, F., Lee; 2014)

We have,

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

For the proof, we use singularity perturbation analysis.

A probabilistic proof explaining the curious normalization was given recently.

S. Janson (2015). Maximal clades in random binary search trees, *Electronic J. Combinatorics*, **22**:1, paper 31.

Some Ideas of the Proof (i)

Set

$$X(y,z)=\sum_{n\geq 2}\mathbb{E}\left(e^{yX_n}\right)z^n.$$

Then,

$$z\frac{\partial}{\partial z}X(y,z)=X(y,z)+X^2(y,z)+e^yz^2\frac{2e^yz^3}{1-z}.$$

This is a Riccati DE.

Some Ideas of the Proof (i)

Set

$$X(y,z)=\sum_{n\ge 2}\mathbb{E}\left(e^{yX_n}\right)z^n.$$

Then,

$$z\frac{\partial}{\partial z}X(y,z)=X(y,z)+X^2(y,z)+e^yz^2\frac{2e^yz^3}{1-z}.$$

This is a Riccati DE.

Set

$$\tilde{X}(y,z)=\frac{X(y,z)}{z}.$$

Then,

$$\frac{\partial}{\partial z}\tilde{X}(y,z)=\tilde{X}^2(y,z)+e^y\frac{1+z}{1-z}.$$

Some Ideas of the Proof (ii)

Set

$$\tilde{X}(y,z) = -\frac{V'(y,z)}{V(y,z)}.$$

Then,

$$V''(y,z) + e^y \frac{1+z}{1-z} V(y,z) = 0.$$

This is Whittaker's DE.

Some Ideas of the Proof (ii)

Set

$$\tilde{X}(y,z) = -\frac{V'(y,z)}{V(y,z)}.$$

Then,

$$V''(y,z) + e^{y}\frac{1+z}{1-z}V(y,z) = 0.$$

This is Whittaker's DE.

Solution is given by

$$V(y,z) = M_{-e^{y/2},1/2}\left(2e^{y/2}(z-1)\right) + c(y)W_{-e^{y/2},1/2}\left(2e^{y/2}(z-1)\right),$$

where

$$c(y) = -\frac{\left(e^{y/2}-1\right)M_{-e^{y/2}+1,1/2}\left(-2e^{y/2}\right)}{W_{-e^{y/2}+1,1/2}\left(-2e^{y/2}\right)}.$$

Some Ideas of the Proof (iii)

Lemma

$V(y,z)$ is analytic in

$$\Delta = \{z \in \mathbb{C} \; : \; |z| < 1+\delta\} \backslash \{\textit{branch cut from } 1 \textit{ to } \infty\}$$

for all $|y| < \eta$.

Moreover, $V(y,z)$ has only one (simple) zero with

$$z_0(y) = 1 - ay + 2a^2y^2 \log y + \mathcal{O}(y^2).$$

Some Ideas of the Proof (iii)

Lemma

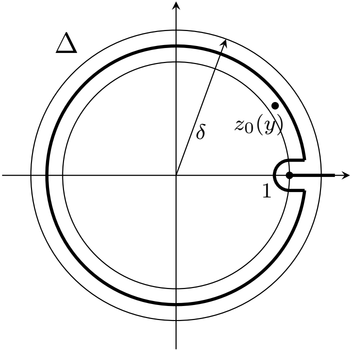
$V(y,z)$ is analytic in

$$\Delta = \{z \in \mathbb{C} \ : \ |z| < 1+\delta\} \setminus \{\textit{branch cut from } 1 \textit{ to } \infty\}$$

for all $|y| < \eta$.

Moreover, $V(y,z)$ has only one (simple) zero with

$$z_0(y) = 1 - ay + 2a^2y^2 \log y + \mathcal{O}(y^2).$$



Some Ideas of the Proof (iv)

Let $y = it/(2a\sqrt{n\log n})$. Then,

$$\mathbb{E}\left(e^{yX_n}\right) = \frac{1}{2\pi i}\int_{\mathcal{C}} \frac{X(y,z)}{z^{n+1}}\mathrm{d}z.$$

Some Ideas of the Proof (iv)

Let $y = it/(2a\sqrt{n \log n})$. Then,

$$\mathbb{E}\left(e^{yX_n}\right) = \frac{1}{2\pi i}\int_{\mathcal{C}} \frac{X(y,z)}{z^{n+1}}\mathrm{d}z.$$

Lemma

We have,

$$\mathbb{E}\left(e^{yX_n}\right) = z_0(y)^{-n} + \mathcal{O}\left(\frac{\log^3 n}{n}\right).$$

Some Ideas of the Proof (iv)

Let $y = it/(2a\sqrt{n \log n})$. Then,

$$\mathbb{E}\left(e^{yX_n}\right) = \frac{1}{2\pi i}\int_{\mathcal{C}} \frac{X(y,z)}{z^{n+1}}\mathrm{d}z.$$

Lemma

We have,

$$\mathbb{E}\left(e^{yX_n}\right) = z_0(y)^{-n} + \mathcal{O}\left(\frac{\log^3 n}{n}\right).$$

This together with the expansion of $z_0(y)$ yields

$$\mathbb{E}\left(e^{yX_n}\right) = \exp\left(\frac{it\sqrt{n}}{2\sqrt{\log n}} - \frac{t^2}{4}\right)\left(1 + \mathcal{O}\left(\frac{\log\log n}{\log n}\right)\right).$$

Summary

Summary

- We gave a comprehensive study of pattern occurrences in random phylogenetic trees.

Summary

- We gave a comprehensive study of pattern occurrences in random phylogenetic trees.
- We explained the strong correlation between Shapley value and fair proportion index in phylogenetic trees. This is of great interest for people working in biodiversity.

Summary

- We gave a comprehensive study of pattern occurrences in random phylogenetic trees.
- We explained the strong correlation between Shapley value and fair proportion index in phylogenetic trees. This is of great interest for people working in biodiversity.
- We found a curious central limit theorem for the number of groups formed by social animals under the neutral model.

Summary

- We gave a comprehensive study of pattern occurrences in random phylogenetic trees.
- We explained the strong correlation between Shapley value and fair proportion index in phylogenetic trees. This is of great interest for people working in biodiversity.
- We found a curious central limit theorem for the number of groups formed by social animals under the neutral model.
- Analytic combinatorics is useful in studying mathematical problems for random phylogenetic trees. We expect many more applications.

Summary

- We gave a comprehensive study of pattern occurrences in random phylogenetic trees.
- We explained the strong correlation between Shapley value and fair proportion index in phylogenetic trees. This is of great interest for people working in biodiversity.
- We found a curious central limit theorem for the number of groups formed by social animals under the neutral model.
- Analytic combinatorics is useful in studying mathematical problems for random phylogenetic trees. We expect many more applications.
- More input from biologists is needed to make our results more relevant from the point of view of applications.