The Number of Maximal Clades in Random Trees

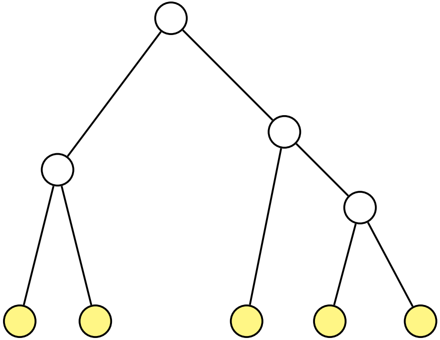
Michael Fuchs

Department of Applied Mathematics
National Chiao Tung University

September 2nd, 2016

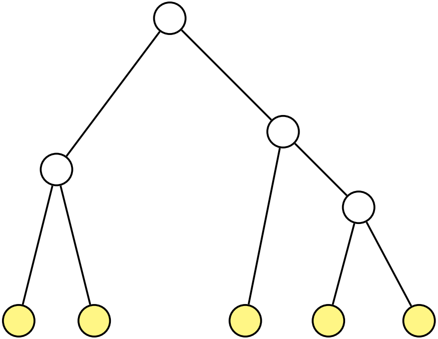
Clade and Maximal Clades

Phylogenetic Tree:



Clade and Maximal Clades

Phylogenetic Tree:

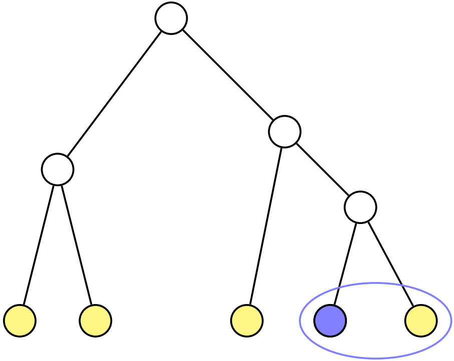


Clade of a leaf:

All leafs of the tree rooted at the parent.

Clade and Maximal Clades

Phylogenetic Tree:

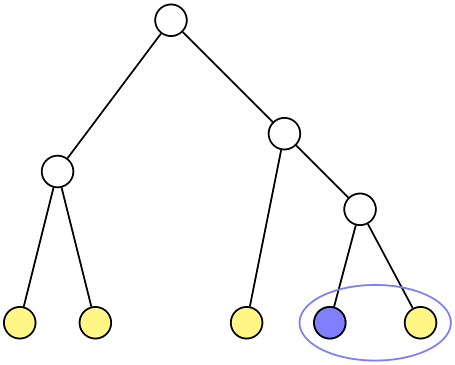


Clade of a leaf:

All leafs of the tree rooted at the parent.

Clade and Maximal Clades

Phylogenetic Tree:



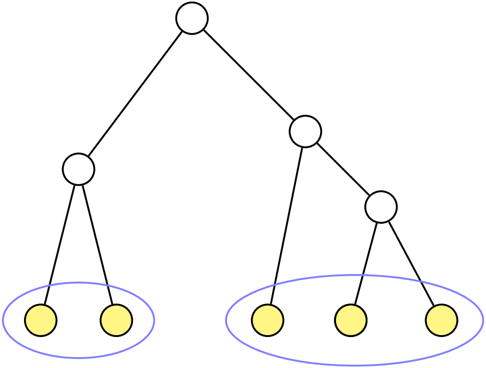
Clade of a leaf:

All leafs of the tree rooted at the parent.

# of Maximal Clades

Clade and Maximal Clades

Phylogenetic Tree:



Clade of a leaf:

All leafs of the tree rooted at the parent.

# of Maximal Clades $= 2$

of Maximal Clades

YHK Model (or BST Model):

Start with the root and two leaves (a cherry); choose one leaf uniformly at random and replace it by a cherry; etc.

of Maximal Clades

YHK Model (or BST Model):

Start with the root and two leaves (a cherry); choose one leaf uniformly at random and replace it by a cherry; etc.

$X_n = \#$ of maximal clades

of Maximal Clades

YHK Model (or BST Model):

Start with the root and two leaves (a cherry); choose one leaf uniformly at random and replace it by a cherry; etc.

$X_n = \#$ of maximal clades

Lemma

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 $I_n = \mathit{Uniform}\{1, \ldots, n-1\}$ *and* X^*_n *is an independent copy of* X_n.

Moments

Durand & François 2010: We have,

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

Moments

Durand & François 2010: We have,

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

Proposition (Y.-W. Lee 2012)

We have,

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

and for $k \geq 3$

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

Moments

Durand & François 2010: We have,

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

Proposition (Y.-W. Lee 2012)

We have,

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

and for $k \geq 3$

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

Remark: limit law cannot be found from the method of moments!

Limit Law

Theorem (Drmota, F., Lee; 2016)

We have,

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

Limit Law

Theorem (Drmota, F., Lee; 2016)

We have,

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

Sketch of Proof. Set

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

Then,

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

Limit Law

Theorem (Drmota, F., Lee; 2016)

We have,

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

Sketch of Proof. Set

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

Then,

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

This Riccati DE can be solved and from the solution asymptotic properties of $\mathbb{E}\left(e^{yX_n}\right)$ can be extracted. □

M-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

M-ary Search Trees

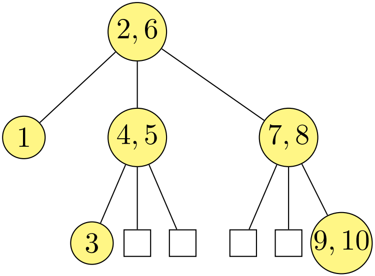
Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9

M-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9

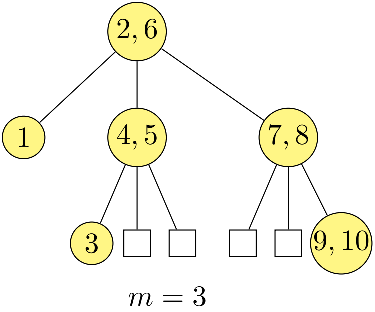


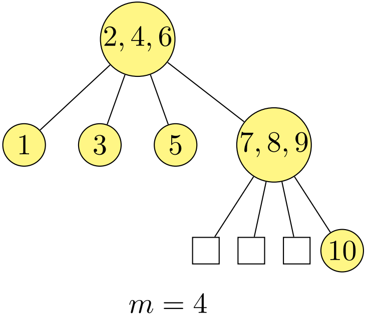
$m = 3$

M-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9

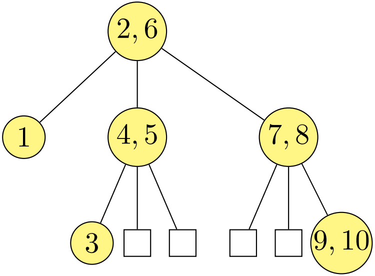




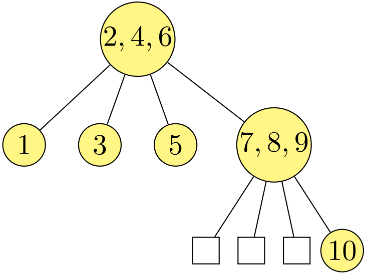
M-ary Search Trees

Proposed by Muntz and Uzgalis in 1971.

Input: 6, 2, 4, 8, 7, 1, 5, 3, 10, 9



$m = 3$



$m = 4$

If permutations are equally likely $\longrightarrow$ **random m-ary search trees**

of Maximal Clades

$X_n = \#$ of maximal clades

of Maximal Clades

$X_n = \#$ of maximal clades

Lemma

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{if } I_r = 0 \text{ for some } 1 \le r \le m; \\ X_{I_1}^{(1)} + \cdots X_{I_m}^{(m)}, & \text{otherwise,} \end{cases}$$

where $X_n^{(r)}, 1 \le r \le m$ *are independent copies of* X_n *and*

$$P\left(I_1^{(1)} = j_1, \ldots, I_m^{(m)} = j_m\right) = \frac{1}{\binom{n}{m-1}}.$$

of Maximal Clades

$X_n = \#$ of maximal clades

Lemma

We have,

$$X_n \stackrel{d}{=} \begin{cases} 1, & \text{if } I_r = 0 \text{ for some } 1 \le r \le m; \\ X_{I_1}^{(1)} + \cdots X_{I_m}^{(m)}, & \text{otherwise,} \end{cases}$$

where $X_n^{(r)}, 1 \le r \le m$ *are independent copies of* X_n *and*

$$P\left(I_1^{(1)} = j_1, \ldots, I_m^{(m)} = j_m\right) = \frac{1}{\binom{n}{m-1}}.$$

Remark: $X(y,z)$ satisfies a DE which in general is not solvable!

Moments

Proposition

We have,

$$\mathbb{E}(X_n) \sim cn,$$

for a (computable) c.

Moments

Proposition

We have,

$$\mathbb{E}(X_n) \sim cn,$$

for a (computable) c.

Moreover, for $2 \leq m \leq 26$,

$$\mathrm{Var}(X_n) \sim \frac{m(m-1)}{H_m - 1} c^2 n \log n$$

and for $k \geq 3$

$$\mathbb{E}(X_n - \mathbb{E}(X_n))^k \sim (-1)^k \frac{m(m-1)(m+k-2)!}{(m+k-2)! - m!(k-1)!} c^k n^{k-1}.$$

Limit Law

Conjecture

For $2 \leq m \leq 26$*,*

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

Limit Law

Conjecture

For $2 \leq m \leq 26$,

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

Question 1: asymptotic expansion of $\mathbb{E}(e^{yX_n})$ even though DE for $X(y,z)$ cannot be solved?

Limit Law

Conjecture

For $2 \leq m \leq 26$,

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

Question 1: asymptotic expansion of $\mathbb{E}(e^{yX_n})$ even though DE for $X(y,z)$ cannot be solved?

$\longrightarrow$ analytic approach to shape parameters of m-ary search trees?

Limit Law

Conjecture

For $2 \leq m \leq 26$,

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

Question 1: asymptotic expansion of $\mathbb{E}(e^{yX_n})$ even though DE for $X(y,z)$ cannot be solved?

$\longrightarrow$ analytic approach to shape parameters of m-ary search trees?

Question 2: what happens when $m > 26$?

Limit Law

Conjecture

For $2 \leq m \leq 26$,

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

Question 1: asymptotic expansion of $\mathbb{E}(e^{yX_n})$ even though DE for $X(y,z)$ cannot be solved?

$\longrightarrow$ analytic approach to shape parameters of m-ary search trees?

Question 2: what happens when $m > 26$?

Remark: for $m > 26$:

$$\text{order of } k\text{-th central moment} = \begin{cases} n^{k(\alpha-1)}, & \text{if } k-1 \leq k(\alpha-1); \\ n^{k-1}, & \text{if } k-1 > k(\alpha-1). \end{cases}$$

A Heuristic for $m = 3$ (i)

Denote by $\kappa_n^{(k)}$ the cumulants of X_n. Then,

$$\kappa_n^{(1)} = \mathbb{E}(X_n), \qquad \kappa_n^{(k)} \sim \mathbb{E}(X_n - \mathbb{E}(X_n))^k \qquad (k \geq 2).$$

A Heuristic for $m = 3$ (i)

Denote by $\kappa_n^{(k)}$ the cumulants of X_n. Then,

$$\kappa_n^{(1)} = \mathbb{E}(X_n), \qquad \kappa_n^{(k)} \sim \mathbb{E}(X_n - \mathbb{E}(X_n))^k \qquad (k \geq 2).$$

Thus,

$$\log \mathbb{E}\left(e^{X_n t}\right) = \sum_{k\geq 1} \kappa_n^{(k)} \frac{t^k}{k!} \approx cnt + \frac{18}{5} c^2 n (\log n) t^2 + \sum_{k \geq 3} \kappa_n^{(k)} \frac{t^k}{k!}.$$

A Heuristic for $m = 3$ (i)

Denote by $\kappa_n^{(k)}$ the cumulants of X_n. Then,

$$\kappa_n^{(1)} = \mathbb{E}(X_n), \qquad \kappa_n^{(k)} \sim \mathbb{E}(X_n - \mathbb{E}(X_n))^k \qquad (k \geq 2).$$

Thus,

$$\log \mathbb{E}\left(e^{X_n t}\right) = \sum_{k\geq 1} \kappa_n^{(k)} \frac{t^k}{k!} \approx cnt + \frac{18}{5} c^2 n(\log n)t^2 + \sum_{k\geq 3} \kappa_n^{(k)} \frac{t^k}{k!}.$$

Here,

$$\sum_{k\geq 3} \kappa_n^{(k)} \frac{t^k}{k!} \approx \frac{1}{n} \sum_{k\geq 3} \frac{6(k+1)}{(k-1)!(k^2+k-6)} (-cnt)^k.$$

A Heuristic for $m = 3$ (ii)

Note that

$$\sum_{k\geq 3}\frac{6(k+1)}{(k-1)!(k^2+k-6)}(-z)^k$$

$$= \frac{78z^5-75z^4-360}{25z^3}+\frac{18z^4+12z^3+36z^2+72z+72}{5z^3}e^{-z}$$

$$+\frac{18}{5}z^2\left(-\gamma-\ln(z)-\mathrm{Ei}(1,z)\right)$$

A Heuristic for $m = 3$ (ii)

Note that

$$\sum_{k\geq 3}\frac{6(k+1)}{(k-1)!(k^2+k-6)}(-z)^k$$
$$= \frac{78z^5-75z^4-360}{25z^3}+\frac{18z^4+12z^3+36z^2+72z+72}{5z^3}e^{-z}$$
$$+\frac{18}{5}z^2\left(-\gamma-\ln(z)-\mathrm{Ei}(1,z)\right)$$
$$\sim -\frac{18}{5}z^2\log z.$$

A Heuristic for $m = 3$ (ii)

Note that

$$\sum_{k\geq 3}\frac{6(k+1)}{(k-1)!(k^2+k-6)}(-z)^k$$
$$= \frac{78z^5-75z^4-360}{25z^3}+\frac{18z^4+12z^3+36z^2+72z+72}{5z^3}e^{-z}$$
$$+\frac{18}{5}z^2\left(-\gamma-\ln(z)-\mathrm{Ei}(1,z)\right)$$
$$\sim -\frac{18}{5}z^2\log z.$$

Thus,

$$\log\mathbb{E}\left(e^{X_nt/(3c\sqrt{2n\log n/5})}\right)\approx\frac{\sqrt{n}}{3\sqrt{2\log n/5}}t+t^2/2.$$

A Heuristic for $m = 3$ (ii)

Note that

$$\sum_{k\geq 3}\frac{6(k+1)}{(k-1)!(k^2+k-6)}(-z)^k$$
$$=\frac{78z^5-75z^4-360}{25z^3}+\frac{18z^4+12z^3+36z^2+72z+72}{5z^3}e^{-z}$$
$$+\frac{18}{5}z^2\left(-\gamma-\ln(z)-\mathrm{Ei}(1,z)\right)$$
$$\sim -\frac{18}{5}z^2\log z.$$

Thus,

$$\log\mathbb{E}\left(e^{X_n t/(3c\sqrt{2n\log n/5})}\right)\approx\frac{\sqrt{n}}{3\sqrt{2\log n/5}}t+t^2/2.$$

Question: can this made precise?