The Critical Beta-Splitting Random Tree Model: Results and Open Problems

David Aldous

20 November 2023

Many models of size-n random trees have been studied, with exact results by combinatorialists and $n \to \infty$ asymptotics by probabilists. This talk concerns a "new" model with quite different asymptotics.

- Some (weak) real-world motivation.
- Being a "random tree" model, there are many aspects to study, and **many different techniques** one can attempt to use.
- Can compare and contrast with the known continuum random tree limits of other models.
- A classical analysis technique (analysis of recursions) gets surprisingly sharp answers to certain basic questions.
- We have many results and many open problems **suitable for some smart young person**.

2 long preprints on arXiv
*The Critical Beta-splitting Random Tree, I and II*

David Aldous   The Critical Beta-Splitting Random Tree Model: Results and Open Problems
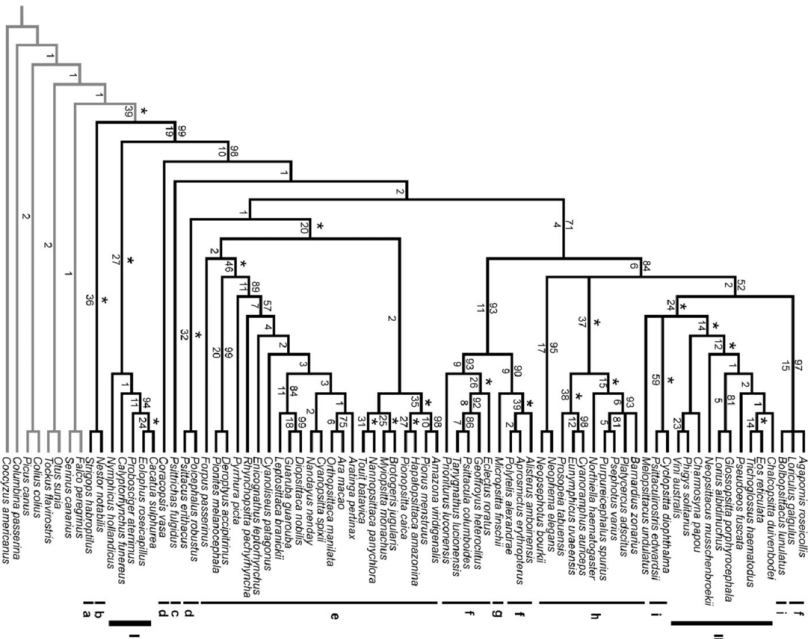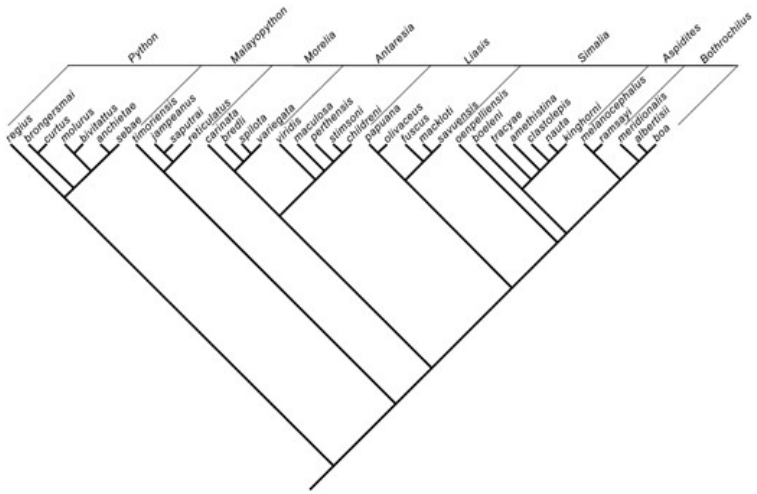
This "uneven split" property holds for most phylogenies – just type "xxx phylogeny" into Google Images.

To demonstrate, having already said "dead parrot" let's continue the Monty Python theme by showing phylogenetic trees for

- Brontosaurus
- Swallows
- Pythons
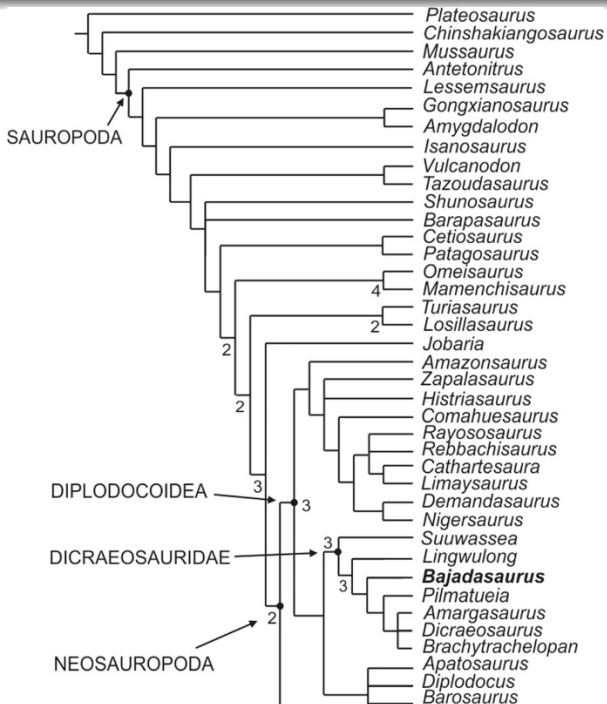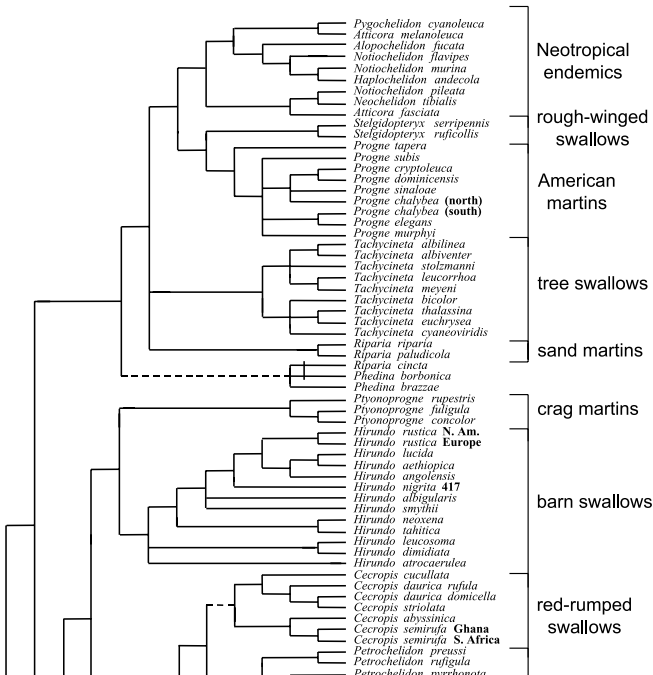
- Is there a simple probability model that replicates this "uneven splits" aspect of real cladograms?

At each split within a cladogram, a clade (sub-tree) of size m species is split into clades of sizes i and $m - i$. Data often shows (Aldous, Stat. Sci, 2001) that the median size of the smaller subtree scales as roughly $m^{1/2}$. Simple probability models used before 2000 would predict median size $O(1)$ or $O(\log m)$ or $\Omega(m)$.

One could invent models with several real parameters, and then see if any parameter values gave order $m^{1/2}$.

- Is there a simple model that predicts $m^{1/2}$?

A class of probability models for n-leaf rooted binary trees.

Recursive, but in opposite direction from e.g. Galton-Watson.

For each $m \geq 2$, specify a probability distribution $(q(m, i), 1 \leq i \leq m-1)$ with the symmetry condition $q(m, i) \equiv q(m, m-i)$.

Given n, construct the random tree by specifying that there is a left edge and a right edge at the root, leading to a left subtree which will have L_n leaves and a right subtree which will have $R_n = n - L_n$ leaves, where L_n (and also R_n, by symmetry) has distribution $q(n, \cdot)$.

Continue recursively; a subtree which will have $m \geq 2$ leaves is split into two subtrees of random size from the distribution $q(m, \cdot)$; continue until reaching subtrees of size 1, which are leaves.
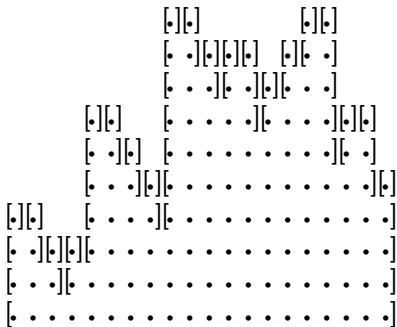
Figure: Representation as discrete interval-splitting: $n = 20$.

Specialize to a 1-parameter family, which we call *beta*-splitting: roughly it is

$$q(n, i) \propto i^{\beta}(n-i)^{\beta}, \quad 1 \le i \le n-1$$

which is interesting for $-2 \le \beta \le \infty$.

In this model the height of a typical leaf (number of edges to the root) grows as

$(\beta > -1)$: order $\log n$

$(\beta < -1)$: order $n^{-\beta-1}$.

We will study the *critical* case $\beta = -1$.

Two motivations:

(i) Will fit the order $m^{1/2}$ data.

(ii) A stochastic model, with a "phase transition" separating qualitatively different behaviors, often has mathematically interesting special properties at the critical value of the parameter.

This project was proposed in (Aldous, Probability Distributions on Cladograms, 1995) but not followed up.

Simulation of model, drawn as cladogram.
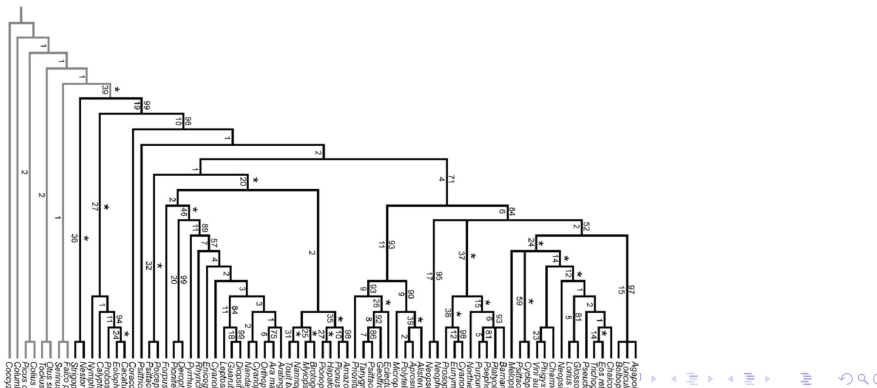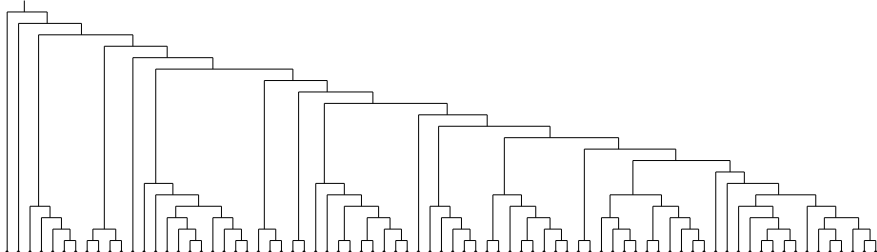
OK, forget the biology, now onto the mathematics.

Our model: splitting probability $q(n, i) \propto \frac{1}{i(n-i)}$.

Note $\frac{1}{i(n-i)} = \frac{1}{n}(\frac{1}{i} + \frac{1}{n-i})$, so we get the normalization constant

$$q(n, i) = \frac{n}{2h_{n-1}} \frac{1}{i(n - i)}, \;\; 1 \le i \le n - 1$$

where $h_n = \sum_{i=1}^{n} \frac{1}{i} \sim \log n$. So the median size of the smaller split is essentially $n^{1/2}$ because when we sum over $1 \le i \le n^{1/2}$

$$2 \times \frac{1}{2h_{n-1}} \times \sum_{i=1}^{n^{1/2}} \frac{1}{i} \approx \frac{1}{\log n} \times \log n^{1/2} \approx \frac{1}{2}.$$

So now, what does the random tree look like drawn from the root? First we have to think how we will draw a tree.
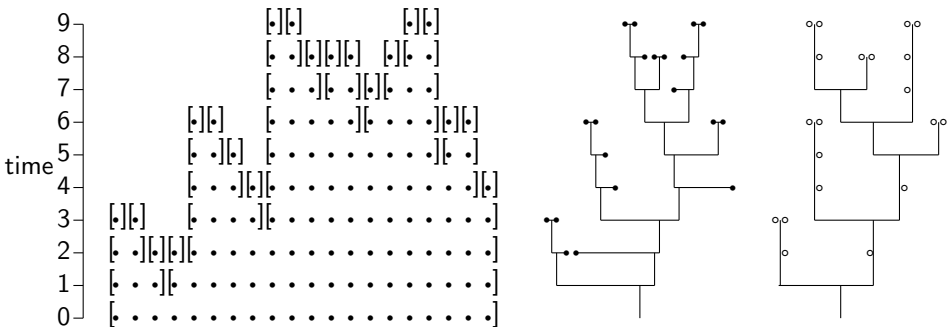
Figure: Equivalent representations of a realization of DTCS(20).

The tree on the right has some specific structure: leaves occur as pairs at the end of a stem, or as a singleton on one side of a branch. This "pruned" form turns out to be mathematically convenient when we switch to continuous time.

Overview of results

- There is a canonical way to embed the discrete-time model into a continuous-time model (which we call CTCS(n)) by specifying that a clade of size $m \geq 2$ is split at rate h_{m-1}.

- For the height (time reached) D_n of a uniform random leaf in the CTCS(n) model, $\mathbb{E}[D_n] \sim \frac{6}{\pi^2} \log n$ and also there is a Gaussian limit distribution. **Paper #1 shows many related results of surprising sharpness, obtained via analysis of recursions**, for both DTCS(n) and CTCS(n).

- We can describe the limit fringe distribution of CTCS(n), that is the local weak limit relative to a random leaf.

- There is a non-obvious consistency property of $(CTCS(n), n \geq 2)$ in its "pruned" form: given CTCS(n+1), delete a random leaf and prune; this gives CTCS(n). In reverse this gives an explicit algorithm for growing CTCS(n+1) from CTCS(n).

- There is a scaling limit of $(CTCS(n), n \geq 2)$, as a process of splitting the continuous interval $(0, 1)$, with a corresponding continuum tree. The pruned spanning tree on n random points is CTCS(n).

Our discrete time construction was:

At each unit time, split a size m clade into $(i, m - i)$ clades with probability

$$q(m, i) = \frac{m}{2h_{m-1}} \frac{1}{i(m - i)}, \quad 1 \leq i \leq m - 1$$

Instead we will work with a continuous time model CTCS(n) where we split size m clades at rate h_{m-1} instead. That is:

Split rate is $= \frac{m}{2} \frac{1}{i(m-i)}, \quad 1 \leq i \leq m - 1$

This turns out to be mathematically more tractable.

We will mostly be doing $n \to \infty$ asymptotics, so what does a tree on 400 leaves look like?
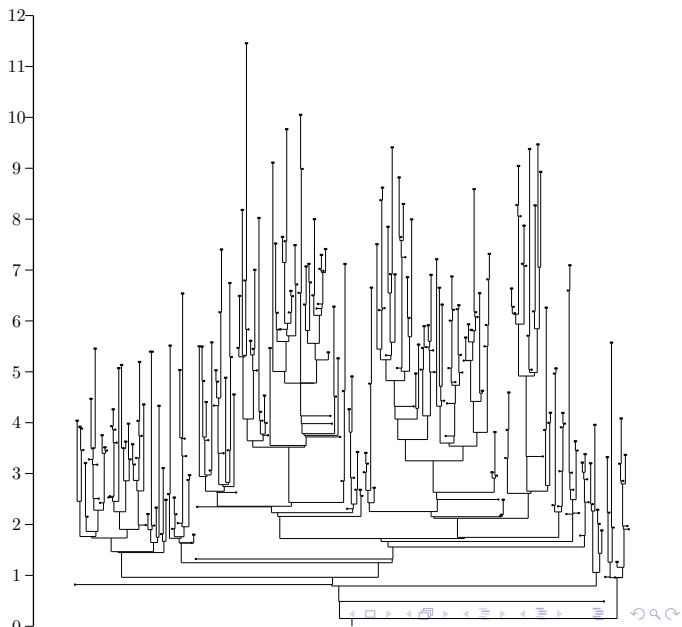
Note that in our discrete-time model, there's no direct relation between the trees for n and $n+1$, we have to start over with the construction. Somewhat magically, there is a simple connection for the continuous-time model:

Given CTCS(n+1), delete a random leaf and prune; this gives CTCS(n).

Here's a discussion.

To formulate a consistency property, first consider spanning sub-trees on a given set of leaves within a large tree.
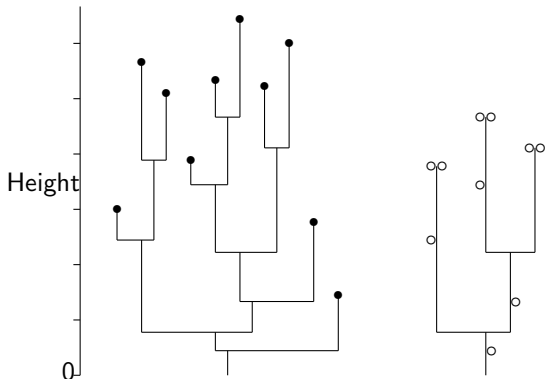

Figure: A spanning tree on $k = 10$ leaves within CTCS(n) for some $n \gg k$ (left) and the corresponding pruned tree PRU(n,k) (right).

The key consequence of the continuous-time embedding is (via a simple calculation)

(*) Within CTCS(n), the time S_n at which the paths to 2 different random leaves diverge satisfies

$$(*) \quad S_n \text{ has exactly Exponential(1) distribution.}$$

This makes it intuitively clear that, for the pruned tree PRU(n,k) on k random leaves, there must be some limit

$$\text{PRU(n,k)} \to \mathbb{T}(k) \text{ as } n \to \infty$$

because (*) says we already did the right order of scaling. By construction, the family $(\mathbb{T}(k), k \geq 2)$ must be consistent under "delete random leaf and prune".

Is this $\mathbb{T}(k)$ the same as CTCS(k)?

Is this $\mathbb{T}(k)$ the same as CTCS(k)?

Yes: there is an abstract-but-strangely-unconvincing proof. This implies the family is consistent under "delete random leaf and re-prune". But more informative to check by explicit formulas for the distribution of shape/density-of-edge-lengths, which leads to the following inductive construction of (CTCS(n), $n \geq 2$)

Algorithm: given CTCS(k)

- Pick uniform random leaf; move up path from root toward that leaf. A "stop" event occurs at rate = 1/(size of subclade from current position).
- If "stop" before reaching target leaf, make a side-leaf.
- Otherwise, extend target leaf into a twig of Exponential(1) length to make a leaf-pair.
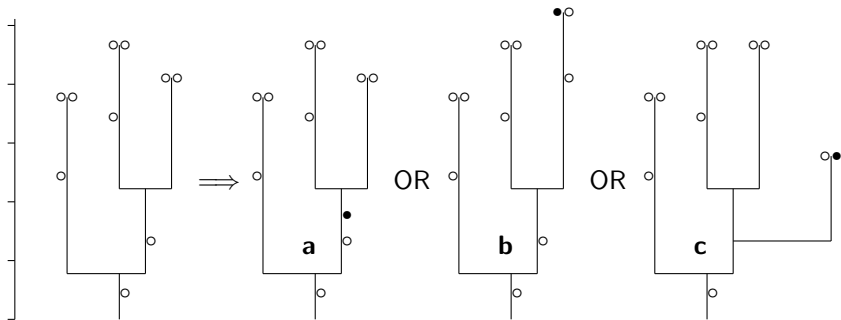
Figure: The possible transitions from CTCS(10) to CTCS(11): the added leaf is ●.

Challenge #1. Is this construction useful for doing calculations? Is there some relevant martingale and a.s. limits?

Most of our (ongoing collaboration with Boris Pittel, who did the actual math) actual results start from detailed study of

$$D_n := \text{height of random leaf } \ell \text{ in CTCS(n).}$$

Along the path from the root to ℓ, at each time t we are in a clade of some size X_t. By size-biasing of the split probabilities $q(m, i)$ we find that the process X_t is the decreasing continuous-time Markov chain on $\{n, n-1, n-2, \ldots, 1\}$ started at n, absorbing at 1, with transition rates

$$\lambda(j, i) = \frac{1}{j - i}, \quad 1 \le i < j \le n.$$

Let's call X_t the explorer chain.

[Needs a better name. Has it been studied before in some other context ???]

A simple "stochastic calculus" argument shows

$$\mathbb{E}[X_t] = 1 + (n-1)e^{-t}$$

implying

$$(*) \qquad \mathbb{E}[D_n] \leq 1 + \log(n-1).$$

In fact this is not the right way to study $\mathbb{E}[D_n]$, but allows me to introduce an alternative proof of (*) via recursions.

Because of the recursive structure of the model, $\mathbb{E}[D_n]$ is determined by a certain recurrence:

$$\mathbb{E}[D_1] = 0$$

$$\mathbb{E}[D_n] = \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1} \frac{\mathbb{E}[D_k]}{n-k}\right), \quad n \geq 2. \tag{1}$$

We will prove

$$\mathbb{E}[D_n] \leq f(n) := 1 + \log(n-1). \tag{2}$$

It is enough to show that $f(n)$ satisfies

$$f(n) \geq \frac{1}{h_{n-1}}\left(1 + \sum_{i=1}^{n-1} \frac{f(i)}{n-i}\right), \quad n \geq 2. \tag{3}$$

Since $f(x)$ is *concave* for $x > 1$, we have

$$\frac{1}{h_{n-1}}\left(1 + \sum_{i=1}^{n-1} \frac{f(i)}{n-i}\right) \leq \frac{1}{h_{n-1}} + f\left(\sum_{i=1}^{n-1} \frac{i}{n-i}\right)$$
$$= \frac{1}{h_{n-1}} + f\left(n - \frac{n-1}{h_{n-1}}\right) \leq \frac{1}{h_{n-1}} + f(n) - f'(n)\left(\frac{n-1}{h_{n-1}}\right),$$

which is exactly $f(n)$, since $f'(x) = \frac{1}{x-1}$ for $x > 1$.

To most of this audience, that "recurrence" argument is less informative/natural than the "probability" argument, which established an exact result on the way. But the simplicity of the probability argument in this case is purely lucky.

In the context of probability-on-trees, (and many analysis-of-algorithms settings), one can often set up such recurrences. And anything defined by a recurrence can in principle be bounded by inductively verifying a bound.

This talk focusses on probabilistic proofs (in preprint #2), whereas preprint #1 proves a variety of refinements based on the recurrence method above.

A key insight is that the explorer chain X_t is decreasing in some "multiplicative" way. Recall the elementary textbook example:

What is the behavior of $M_n := \prod_{i=1}^{n} U_i$ for i.i.d. U[0,1] RVs U_i ?

First answer: $\mathbb{E}[M_n] = 2^{-n}$.

Better answer: $M_n \approx e^{-n}$ because $\log M_n = \sum_{i=1}^{n} \log U_i$ and so $n^{-1} \log M_n \to \mathbb{E}[\log U] = e^{-1}$.

So let's go back to our explorer chain and take *logs*.

$X(t)$ is the continuous-time Markov chain on $\{1, 2, 3, \ldots, n\}$ started at n, absorbing at 1, with rates

$$\lambda(j, i) = \frac{1}{j - i}, \quad 1 \le i < j \le n.$$

Study $Z(t) := \log X(t)$. A transition $z \to z - a$ is a transition

$$x = e^z \to e^{z-a} = xe^{-a} = x - x(1 - e^{-a}).$$

So rate of transitions $z \to [0, z - a]$ is

$$\sum_{i=x(1-e^{-a})}^{x} 1/i \sim -\log(1 - e^{-a}).$$

which does not depend on z.

This says that the process $\log X(t)$ is essentially just a (continuous time) random walk. More specifically:

There is a σ-finite measure ψ on $(0, \infty)$ with $\psi[a, \infty) = -\log(1 - e^{-a})$. Write $Y(t)$ for the subordinator with Levy measure ψ. Then, for $X^{(n)}(t)$ the chain started at n,

$$\log X^{(n)}(t) \approx \log n - Y(t) \text{ until this is } O(1). \qquad (4)$$

We are studying

$$D_n := \inf\{t : X^{(n)}(t) = 1\}.$$

But we have a SLLN and CLT for the subordinator. Assuming the approximation (4) is good enough:

$$t^{-1}Y(t) \to \rho := \int_0^\infty \psi[a, \infty) \, da$$

and so $D_n \sim \rho^{-1} \log n$. By a classical identity $\rho = \zeta(2) = \pi^2/6$ so our simple bound $\mathbb{E}D_n \leq 1 + \log(n-1)$ is upgraded to $\mathbb{E}D_n \sim 6\pi^{-2} \log n$. And finally the CLT:

Theorem

$$\frac{D_n - \mu \log n}{\sqrt{\log n}} \to_d \text{Normal}(0, \mu^3 \sigma^2)$$

where

$$\mu := 1/\zeta(2) = 6/\pi^2 = 0.6079...; \quad \sigma^2 := 2\zeta(3) = 2.4040.....$$

So in outline this is just the textbook CLT for renewal processes, but the technical work is in justifying the approximation (4). Our proof (preprint # 2) seems a Horrible Hack: there must be some better way . . . .
Challenge #2.

In parallel, preprint #1 gives an analytic proof based on the recurrence for the Laplace transform. But also technically intricate.

Sharp results by analysis of recurrences

$\mathbb{E}[D_n]$ is determined by the recurrence: $\mathbb{E}[D_1] = 0$ and

$$\mathbb{E}[D_n] = \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1} \frac{\mathbb{E}[D_k]}{n-k}\right), \quad n \geq 2. \tag{5}$$

Theorem

$$\mathbb{E}[D_n] = \frac{6}{\pi^2} \log n + O(1) \text{ as } n \to \infty.$$

Proposition

Assuming the h-ansatz, there exists a constant c_0 such that

$$\mathbb{E}[D_n] = \frac{6}{\pi^2} \log n + c_0 - \frac{3}{\pi^2} n^{-1} + O(n^{-2}). \tag{6}$$

One can calculate $\mathbb{E}[D_n]$ numerically via the basic recurrence, and doing so up to $n = 400,000$ gives a good fit to (6) with $c_0 = 0.7951556604.....$ Yes, really 10 significant digits ......

Proofs are long and technical; they depend on sharp estimates like

Lemma

$$\sum_{k=1}^{n-1} \frac{\log(k/n)}{n-k} = -\frac{\pi^2}{6} + \frac{\log(2\pi e)}{2n} + \frac{\log n}{12n^2} + O(n^{-2}).$$

which are proved, in the spirit of Knuth's *Concrete Mathematics*, via ingredients such as Euler's summation formula: if $f(x)$ is a smooth differentiable function for $x \in [a, b]$ such that the even derivatives $f^{(2)}, f^{(4)}, \ldots$ are all of the same sign, then for every $m \geq 1$

$$\begin{aligned}
\sum_{a \leq k < b} f(k) &= \int_a^b f(x)\, dx - \tfrac{1}{2} f(x)\Big|_a^b \\
&+ \sum_{\ell=1}^m \frac{B_{2\ell}}{(2\ell)!} f^{(2\ell-1)}(x)\Big|_a^b + \theta_m \frac{B_{2m+2}}{(2m+2)!} f^{(2m+1)}(x)\Big|_a^b. \quad (7)
\end{aligned}$$

Here θ_m is some real in $(0, 1)$ and the $\{B_{2\ell}\}$ are the even Bernoulli numbers, defined by $\frac{z}{e^z-1} = \sum_{\mu \geq 0} B_\mu \frac{z^\mu}{\mu!}$.

In the tree model, D_n arises from two levels of randomness, as the distance $d(U_n, \mathbb{T}_n)$ within a random tree $\mathbb{T}_n$ from the root to a uniform random leaf U_n of that tree. Write $a(\mathbb{T}_n)$ for the average height of the n leaves of $\mathbb{T}_n$. The law of total variance says

$$\text{var}[D_n] = \mathbb{E}[\text{var}(d(U_n, \mathbb{T}_n)|\mathbb{T}_n)] + \text{var}[a(\mathbb{T}_n)]. \tag{8}$$

As statisticians say, the first term of the right indicates the "within tree" variability of leaf height, and the second term indicates the "between trees" variability. As a standard technique, one can calculate the proportion of "between trees" variance

$$r_n := \frac{\text{var}[a(\mathbb{T}_n)]}{\text{var}[D_n]}$$

because it is essentially the correlation between the D's of two random leaves from the same realization of $\mathbb{T}_n$.

Theorem

Assuming the h-ansatz: for Euler's constant γ

$$\lim_{n \to \infty} r_n = \frac{\gamma \zeta(2)}{2\zeta(3)} = 0.3949404179\ldots,$$

Theorem

Assuming the h-ansatz: for Euler's constant γ

$$\lim_{n \to \infty} r_n = \frac{\gamma \zeta(2)}{2\zeta(3)} = 0.3949404179\ldots,$$

If we could find a nicer "probability" proof of the CLT for D_n, that could presumably be extended to a bivariate Gaussian limit in this setting.

[repeat earlier slide]

Overview of results/problems

- ✓ There is a canonical way to embed the discrete-time model into a continuous-time model (which we call CTCS(n)) by specifying that a clade of size $m \geq 2$ is split at rate h_{m-1}.

- ✓ For the height (time reached) D_n of a uniform random leaf in the CTCS(n) model, $\mathbb{E}[D_n] \sim \frac{6}{\pi^2} \log n$ and also there is a Gaussian limit distribution. Many related results of surprising sharpness can be obtained via analysis of recursions.

- We can describe the limit fringe distribution of CTCS(n), that is the local weak limit relative to a random leaf.

- ✓ There is a non-obvious consistency property of $(CTCS(n), n \geq 2)$ in its "pruned" form: given CTCS(n+1), delete a random leaf and prune; this gives CTCS(n). In reverse this gives an explicit algorithm for growing CTCS(n+1) from CTCS(n).

- There is a scaling limit of $(CTCS(n), n \geq 2)$, as a process of splitting the continuous interval $(0, 1)$, with a corresponding continuum tree. The pruned spanning tree on n random points is CTCS(n).

The fringe process

Consider the quantity

$a(n, i) := \mathbb{P}(\textit{explorer chain started at state n is ever in state i})$

[same for the discrete or continuous models.] By a coupling argument

Proposition

The limit $a_i := \lim_{n \to \infty} a(n, i)$ exists, $i = 1, 2, \ldots$.

Ongoing work by Svante Janson shows

$$a_i = \frac{6h_{i-1}}{\pi^2(i-1)}.$$

The motivation for this Proposition involves the distribution of the asymptotic *fringe process* for the tree model, that is the description of the tree relative to a typical leaf, which in the $n \to \infty$ limit can be described in terms of the a_i as follows.

a_i is the probability that the typical leaf in in some clade of size i. So Bayes rule tells us that the sequence of clade sizes as one moves away from the leaf is the discrete time Markov chain started at state 1, whose "upward" transition probabilities are

$$q^{\uparrow}(i,j) = \frac{i a_j}{j a_i}(q(j,i) + q(j,j-i)), \quad j > i. \tag{9}$$

At each such step $i \to j$ there is the sibling clade of size $j - i$, and this clade is simply distributed as DTCS(j-i).

In particular, the first step (to the clade that the leaf first joins) is given by

$$q^{\uparrow}(1,j) = \frac{a_j}{j}(q(j,1) + q(j,j-1)), \quad j > 1$$

for which

$$q^{\uparrow}(1,j) \sim \frac{12}{\pi^2}\frac{1}{j^2}. \tag{10}$$

If we could code the fringe tree as some kind of random walk and then take a scaling limit, would it be something like a 1-sided Cauchy process?

David Aldous   The Critical Beta-Splitting Random Tree Model: Results and Open Problems
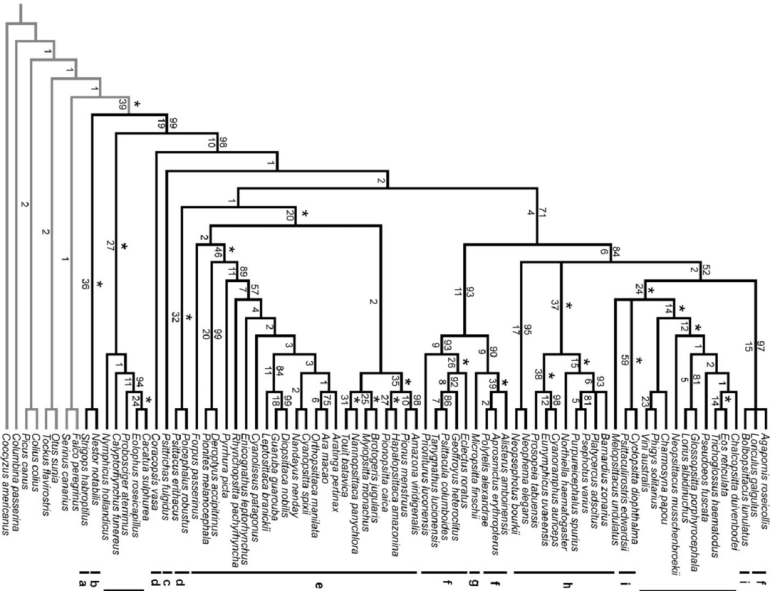
The scaling limit

- Heuristically, there is a scaling limit of $(CTCS(n), n \geq 2)$, as a process of splitting the continuous interval $(0, 1)$, with a corresponding continuum tree. The pruned spanning tree on n random points is CTCS(n).

Challenge #4.. Think rigorously about this, and connections with below.

- There is a classical "applied probability" literature on interval-splitting, focussed on the distribution of fragment lengths, which in our model would be clade sizes at a given time.
- A more recent approach is via exchangeable partitions, see e..g. Haas - Miermont - Pitman - Winkel 2008.
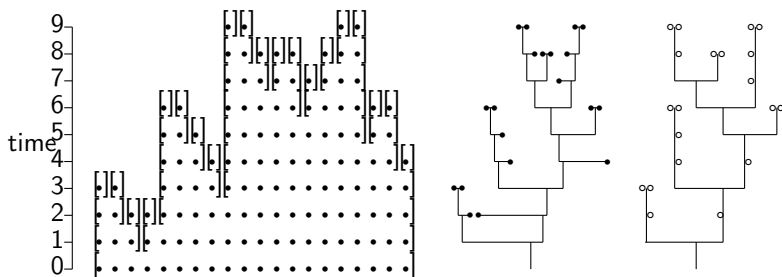
Consider a process of splitting the continuous unit interval $[0, 1]$.

An interval of length x is split into sub-intervals of lengths $(y, x - y)$ at σ-finite rate $\frac{x}{2y(x-y)} \, dy$. Hard to draw a good picture, but the induced spanning tree on k points has the previous type of structure (right side).

There are many analogs/differences between this setting and the theory around the Brownian CRT. Perhaps a "Cauchy" analog?