Talk 1. Strolling in the Stochastic Arboretum.

Talk 2. Continuous Limit Trees
(critical branching processes, Brownian excursion, superprocesses, diffusions on fractals)

Talk 3. The View from the Fringe
(stable type structure, discrete infinite limit trees, stationary processes).

Preprint "Asymptotic Fringe Distributions . . ." covers **3** and its examples overlap **1**. For these conference proceedings I'll write up more details of **2**.

(Joke) Unlike the theory of random graphs, which has 500 papers on 1 model, the subject of random trees has 1 paper on each of 500 models.

Purpose of this talk is to mention 11 examples, chosen to be mathematically clean. Assume familiar with BP (branching process) models of single-sex populations, which are really just probability models of “family trees”. My examples are not overtly BPs, but some are BPs in disguise.

Default convention: what I say is not new or mathematically deep.

Example. Wright-Fisher ancestor tree (discrete coalescent)

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ common ancestor

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ previous gens

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ present gen.

Probability model: children pick parents at random (uniformly).

The "downwards" description of the tree is roughly like a density-dependent BP.

Example. Random recursive tree
(Yule tree)

Start with vertex 1 .
Add vertex n by an edge to a uniformly-chosen existing vertex.

[Lots of variations]

Example. (name?)

n vertices
Initially no edges.
At each step, add an edge (i, j) whose endpoints are chosen uniformly from pairs in different components.
After $n - 1$ steps we get a random tree.

Equivalent description. Start with complete graph. Put i.i.d. weights on edges. Form minimum-weight spanning tree.

Example. Euclidean MST.

Given n points in R^d there is a minimum-length spanning tree connecting them.

Probability model: points of a Poisson process (rate 1) in $[-L, L]^d$.

Asymptotics of functionals (e.g. total length) classical, up to non-explicit constants. Asymptotics of **tree** hard – continuum percolation.

Examples. Models for Spanning Trees.

G finite connected graph.
A spanning tree has same vertices, fewer edges.
Here are two ways to get a spanning tree.

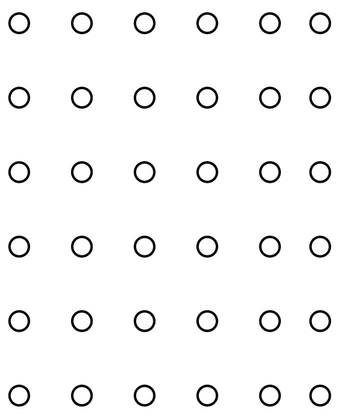
1. Start with G; repeat operation
"delete some edge whose deletion does not disconnect graph".

2. Start with no edges; repeat operation
"add some edge of G which does not create a cycle".

Choosing uniformly at each step gives models for random spanning tree. These models are different from each other and from

3. Uniform random spanning tree.
(Not obvious how to construct, but can be done via random walk on G)

Loosely, there are many “growth processes” definable on a general graph G which for $G =$ complete graph give combinatorial models like the preceeding, and for $G =$ integer lattice Z^d related to standard “hard models” not usually regarded as trees.



Pemantle studied asymptotics of the “uniform spanning tree” model on finite lattices, and shows limit on infinite lattice is tree [forest] for $d \leq 4$ [$d \geq 5$].

People store files in alphabetical order.
This is less convenient for computers.
Knuth's The Art of Computer Programming discusses many sorting, storing and searching methods, some tree-based.

Want algorithms which associate storage locations [think of mailboxes] with names. Seek to minimize "costs" such as
(i) time to find location, given a name
(ii) time to allocate location to new name
(iii) time to delete name (which typically requires subsequent rearrangements)
(iv) preallocated total storage space.

Convert name to “key” using arbitrary hash function.

name	key	(rank)
MERCURY	.01110	6
VENUS	.01011	4
EARTH	.11010	10
MARS	.01100	5
ASTEROIDS	.00110	2
JUPITER	.10111	8
SATURN	.10000	7
URANUS	.00010	1
NEPTUNE	.11001	9
PLUTO	.01010	3

(ranks are just for our visual purposes – the algorithm uses the keys).

Imagine storage locations arranged as binary tree.

Example. Binary Search Tree.
"Find" by comparing key with number at root: if less, move left – if greater, move right.

Probability model. n insertions, no deletions. All orders equally likely.

Equivalent description of random tree. The n'th insertion is uniform random over the n possible places it could go.

[order statistics fact – exploited by Devroye].

Example. 2 – 3 Trees.
Here each vertex represents 2 storage locations.

Here all leaves are at a fixed level – preallocated space requirement less.
Insertion rule: insert at bottom level – if this makes 3 items, push up middle item.

Probability model as before.

Example. Digital Search Tree.
Use binary digits to search left/right.

Probability model: the bits of each key are independent uniform.

Under this stronger assumption, more efficient.

Example. Trie.

Vertices prelabelled as finite binary strings. Insert key at shortest prefix which distinguishes it from the others (may be necessary to move other key).

Sample trie doesn't depend on insertion order.
Deletions easy.
Probability model as before.

Other Areas Using Random Tree Models.

1. Polymerization of molecules
2. River networks
3. Statistical data analysis (discrimination, classification)
4. Phylogenic (evolutionary) trees.

$\mathcal{T}_n$ random tree of size n, arising from some concrete process (maybe with extra structure). Some "cost" functional c. Interested in $Ec(\mathcal{T}_n)$. For CS trees, cost functionals are
$c(t) =$ ave number of comparisons to "find"
$c(t) =$ ave number of moves to "insert"
$c(t) =$ height of tree.

Often random trees have recursive structure, and can seek exact solutions via generating function/ recurrence equations (c.f. Galton-Watson BP: extinction probs, total population size). Large literature in combinatorics/CS takes this line.

My personal angle is (i) $n \to \infty$ asymptotics without exact formulas.
(ii) convergence of trees themselves, not just individual functionals of the trees.

Combinatorial Random Trees. Assume all trees (of some specified size and type) equally likely. This requires precise notion of which trees are to be considered the same.

1. Ordered Trees (= "planar" = "birth-order (left-to-right) counts").

Trees above are the same, considered as

2. Unordered, unlabelled trees.

Another possibility is

3. Unordered, labelled trees.

Galton-Watson Branching Process.

[Peeve: Authors who say “the BP is $(X_i; i \geq 0)$”, where $X_i =$ size of generation i.]

Write $\xi =$ number of offspring
$\mathcal{T} =$ family tree of entire BP.
Interested in $\mathcal{T}$ conditioned on total population size $|\mathcal{T}| = n$: call this conditioned BP $\mathcal{T}_n$.

Note $\mathcal{T}_n$ unaffected by changing ξ to $\widehat{\xi}$:

$$P(\widehat{\xi} = i) = c\theta^i P(\xi = i)$$

Since $P(|\mathcal{T}| = n) \to 0$ exponentially fast in sub- and super-critical case, natural to consider w.l.o.g. the <u>critical</u> case $E\xi = 1$.

Connection. Many combinatorial models of uniform random n-trees are the same as the conditioned critical GWBP for suitable choice of offspring dist ξ.

ordered trees: $P(\xi = i) = (1/2)^i, i \geq 0$.
unordered labelled trees: ξ is Poisson(1)

Variants – allow with at most K offspring:
ordered: $P(\xi = i) = c\theta^i, i \leq K$
unordered labelled: $P(\xi = i) = c\theta^i/i!, i \leq K$

But no similar story known for uniform unordered unlabelled trees.

Global asymptotics given by an invariance principle – here is informal statement.

$\mathcal{T}_n$ conditioned critical GWBP with

$$0 < \sigma^2 \equiv \mathrm{var}(\xi) < \infty.$$

We can draw tree with edge-lengths $1/\sqrt{n}$ so that trees converge in dist to a limit $\mathcal{S}$ (scaled by $1/\sigma$).

Here $\mathcal{S}$ is a particular "continuous random tree".

[Analogy: random walks and Brownian motion]

The 2 fundamental combinatorial random trees can each be studied via explicit constructions.

Deterministic walk round rooted ordered n-tree, choosing left-most edges first, can be written as a string of length $2n$

$$ab \downarrow cd \downarrow e \downarrow\downarrow fg \downarrow\downarrow\downarrow$$

f means "walk away from root to vertex f
$\downarrow$ means "walk toward root"
Now view each letter as a $+1$ step, and $\downarrow$ as a -1 step, and we get a simple deterministic walk on the integers, with first return to 0 at time $2n$.

Can reconstruct the tree from the walk. In other words, there is a $1-1$ correspondence between such walks and ordered n-trees. So we can construct the "uniform random ordered n-tree" from simple symmetric RW conditioned on first return to 0 at time $2n$.

Now intuitively obvious we can take $n \to \infty$ asymptotics.

- conditioned RW → Brownian excursion of duration 1

- so $\mathcal{T}_n \to$ some continuous tree $\mathcal{S}$ built out of Brownian excursion.

But not clear exactly how to think of $\mathcal{S}$. We return to this later.

Unordered labelled n-trees. Cayley's formula: there are n^{n-1} such (rooted) trees. There are several explicit bijections between such trees and sequences

$$(a_1, a_2, \ldots, a_{n-1}) : 1 \leq a_i \leq n$$

which can be used for simulation. Here is an alternative **algorithm.**

Connect vertex 2 to root vertex 1.
stage 1. For $3 \leq i \leq n$ connect vertex i to vertex $V_i = \min(U_i, i-1)$, where $U_2, \ldots, U_n$ are independent and uniform on $1, \ldots, n$.
stage 2. Relabel vertices using uniform random permutation of $(1, \ldots, n)$.

Most questions involve only the shape of the tree, so can omit stage 2 and drop labels.

Pictures drawn using the rule:
start new branch if $U_i \leq i - 1$.

Explicit construction of limit tree.

Half line $[0, \infty)$. Make cuts according to a Poisson process, rate $r(t) = t$.

Each new segment is attached (orthogonally in l_1) to a uniform random point in the existing tree.

The closure of this entire process is *compact* in l_1.

Finite tree has *graph distance*

$$d(v,w) = \text{ \# vertices on path } v \text{ to } w$$

Consider sequence space l_1, distance

$$||\mathbf{x}-\mathbf{y}|| = \sum_i |x_i - y_i|.$$

Possible to represent the vertices of a finite tree as points in l_1 such that

$$d(v,w) \equiv ||v-w||.$$

Call this a *set-representation*. Related is the uniform probability distribution on the vertices: the *measure-representation*. The point is

- l_1 permits linear rescaling
- we have natural notions of convergence for sets and measures in l_1

Without explicit algorithm, how to show a family $(\mathcal{T}_n)$ of random n-trees satisfies

$$\text{rescaled } \mathcal{T}_n \to \text{ some } \mathcal{S} \text{ ?}$$

Natural to rescale so that

$$E \text{ (distance root to random vertex)} = 1.$$

Fix k. Take k random vertices of $\mathcal{T}_n$. These and root define a subtree (with edge-lengths) $\mathcal{S}_{n,k}$ say.

Suppose (analog of "convergence of f.d.d.'s")

$$(a) \quad \mathcal{S}_{n,k} \overset{d}{\to} \mathcal{S}_k \text{ say, as } n \to \infty.$$

Then $(\mathcal{S}_k)$ satisfies a consistency condition in k, so corresponds loosely to some $\mathcal{S}_\infty$.

Theorem 1 *$\mathcal{S}_\infty$ can be represented as a random p.m. on l_1 iff*
(b) vertex 1 not isolated in $(\mathcal{S}_k, k \geq 1)$.
Moreover, given (a) and (b) we have rescaled $\mathcal{T}_n \overset{d}{\to} \mathcal{S}_\infty$ in sense of measure-representations.

Use this to prove invariance principle: general conditioned GWBPs can be rescaled to converge to $\mathcal{S}$.

Theorem looks nice but . . . . . . most examples of random trees fail (b), e.g. supercritical BPs at first time population $= n$.

Curious fact. Subtree $\mathcal{S}_k$ of first k line-segments in algorithm has same dist as subtree from k random vertices of $\mathcal{S}$.

This roundabout argument leads to construction of $\mathcal{S}$ in terms of Brownian excursion $(B_t : 0 \leq t \leq 1)$ and independent $U(0,1)$ r.v.'s (U_i).

Neveu, Pitman, Le Gall have given direct constructions of trees in Brownian excursion.

In terms of the BE building the whole tree, the height profile diffusion (Y_x) is local time at level x.

(c.f. Le Gall & Yor)

We now have many ways to study explicit distributions associated with $\mathcal{S}$. Consider for instance the random distances

$D = d(\text{root}, V)$ for uniform random $V \in \mathcal{S}$

$H = \max_v \ d(\text{root}, v)$ "height"

$\Delta = \max_{v_1,v_2} \ d(v_1, v_2)$ "diameter"

Distance D_n in uniform random unordered tree has simple exact distribution

$$P(D_n = i) = \frac{(i+1)(n-1)!}{n^i(n-i-1)!}$$

and taking limits gives

$$D \text{ has density } f_D(x) = xe^{-x^2/2}.$$

Other interpretations are

$$f_D(x) = EY_x \qquad Y \text{ height profile diffusion}$$

$$f_D(x)dx = E\int_0^1 P(B_t \in dx)dt$$

Combinatorialists long known asymptotic distribution of H_n, but easiest to use BE construction (Durrett Kesten & Waymire)

$$H = \max_t B_t.$$

The fact $EH = 2ED$ is clear by symmetry (under time-reversal) of height profile process (Y_x).

In terms of Brownian excursion

$$\Delta = \max_{t_1<t_2<t_3} (B(t_1) - B(t_2)) + (B(t_3) - B(t_2)).$$

Combinatorialists know a complicated exact formula for density of Δ

$$3/\sqrt{2\pi}\ f_\Delta(x) = \sum_{m=1}^{\infty} \exp(-b_{m,x})$$

$$\{\frac{64}{x^4}(4b_{m,x}^4 - 36b_{m,x}^3 + 75b_{m,x}^2 - 30b_{m,x})$$

$$+\frac{8}{x^2}(4b_{m,x}^3 - 10b_{m,x}^2)\}$$

$$\text{where } b_{m,x} = (\pi m/x)^2.$$

and thereby calculated $E\Delta = \frac{4}{3}EH$. The latter can be derived from a symmetry argument on $\mathcal{S}$.

<u>Open Problem</u> Derive density of Δ from BE description.

Superprocesses. In studying superprocess associated with a continuous-time Σ-valued Markov process, helpful to separate the “family tree” structure from the “spatial movement” structure. W.l.o.g. condition on total population size before extinction – then the family tree is just $\mathcal{S}$.

There is obvious construction of a realization of Σ-valued $(X_t : t \in \mathcal{S})$ along with the construction of $\mathcal{S}$.

In the same way, we can construct the superprocess conditioned on non-extinction by using its family tree $\mathcal{S}_\infty$, which is the limit of combinatorial random trees under a different rescaling.

At $\tau = 0$, semi-infinite line.
During $[\tau, \tau + d\tau]$, each segment dx of existing tree has chance $dx d\tau$ to grow a branch, with length exponential(τ).
Closure of $\tau = \infty$ limit is a self-similar random tree $\mathcal{S}_\infty$.

Interpolating between work of Kesten and of Barlow-Perkins, we can consider $\mathcal{S}$ or $\mathcal{S}_\infty$-valued diffusions. These are examples of "diffusions on fractals" where (a) existence is easy (b) can do explicit calculations.

(a) Build $\mathcal{S}$ from finite line segments; put BM on each. Adding new segment does not change local time on existing tree, so (without work) there exists "local time for the limit process".

Finite rooted tree. Each vertex v defines a subtree rooted at v. Picking v at random (uniformly) gives the random fringe subtree.

$T = \{$ finite rooted trees $\}$, countable set.
$\mathcal{T}_n$: random tree, size n.
$\mathcal{F}_n$: random fringe subtree of $\mathcal{T}_n$.

Empirical Observation. In most examples, easy bare-hands arguments show

$$\mathcal{F}_n \ \stackrel{d}{\rightarrow}\ \mathcal{F} \text{ (say), as } n \to \infty$$

where limit is a.s. finite random tree.
(Maybe asymptotic cycling instead.)

Call $\mathcal{F}$ the asymptotic fringe distribution. Informally, $\mathcal{F}$ describes limits of all "local" properties of $\mathcal{T}_n$. For instance

$P(\mathcal{F} = \bullet) =$ asymptotic prop. of leaves in $\mathcal{T}_n$.

degree of root of $\mathcal{F} =$ asymptotic
distribution of out-degrees in $\mathcal{T}_n$.

The Pessimist's View of Life

1. You're born; you have a random number of children at random times; you die.

2. Your children behave in the same way, independently of you.

The mathematical model is alternatively called the Crump-Mode-Jagers general continuous-time supercritical branching process.

Model. BP where each individual has C children ($EC > 1$) at times $(\xi_1, \xi_2, \ldots, \xi_C)$ (arbitrary distribution) after own birth.

Standard facts. "Stable type structure".

1. (Number born before t) $\sim Ze^{\theta t}$ for a certain Malthusian constant θ.

2. Pick individual at random from those born before T; look at descendants born before T. As $T \to \infty$ this "random family tree" has the following limit.

Start the BP with 1 individual and watch for an exponential (θ) time.

My point. Limits often exist for models of random trees with no BP structure.

Aside. Some of the initial examples can be regarded as the "jump processes" of continuous-time BPs.

1. Linear pure birth process (Yule process) gives the random recursive tree (Yule tree).

2. The continuous-time BP where each individual has exactly 2 offspring, at (concurrent) exponential times after own birth, gives the random binary search tree.

Typically, continuous-time supercritical BPs, and Markovian models of tree growth in discrete time, the relation between height and size of tree is

$$\text{height}(\mathcal{T}) \approx \log |\mathcal{T}|$$

This differs from conditioned critical GWBPs, where

$$\text{height}(\mathcal{T}) \approx |\mathcal{T}|^{1/2}$$

By analogy with $(S_n - a_n)/b_n \to Y$, ask:

"What distributions π on T occur as limits of random fringe subtrees $\mathcal{F}_n$? "

Answer: π occurs iff $\pi Q = \pi$, where $Q = Q(t, t^*)$ is the matrix counting first-generation subtrees.
$Q(t, \quad) = 2.$
$Q(t, \quad) = 1.$
$Q(t, \quad) = 1.$

Call such π's <u>fringe distributions</u>. Given such a π, define a Markov transition matrix by

$$P_\pi(t, t^*) = \pi(t^*)Q(t^*, t)/\pi(t).$$

Run this chain with initial dist. π ...

This defines a discrete infinite rooted tree with unique infinite path from the root. I call such a tree a **sin**-tree, for **s**ingle **in**finite path. Alternative "tree with one end".

Fringe distributions on finite trees correspond precisely to invariant distributions on sin-trees: invariant under map: move root up one step.

1. Local limits in $(\mathcal{T}_n)$ around random vertex can be described via limit sin-tree – this looks toward the root, as well as away from the root.

2. For conditioned critical GWBPs it's natural to consider discrete limit trees around the root. For general families this is not interesting.

Work towards stating a theorem.

Abstract BP with arbitrary types.

Type space S. Individual, type s, has random number/type of children.

$$R(s, \cdot) = E(\# \text{ offspring with type } \in \cdot).$$

Given a p.m. ν on S, define random tree $\mathcal{F}$ to be family tree of descendants of individual, type $\sim \nu$. Suppose finite.

Fact. For $\mathcal{F}$ as above, dist($\mathcal{F}$) has a fringe dist. iff $\nu R = \nu$.

Empirical observation. (N.B. selection bias!) In all examples where asymptotic fringe dist. is known, it has description as abstract BP with simple type-space, even though underlying trees are not BP's.

Random trees $\mathcal{T}_n$, random fringe subtrees $\mathcal{F}_n$. Suppose asymptotic fringe $\mathcal{F}$ exists. This says

$$E\phi(\mathcal{F}_n) \to E\phi(\mathcal{F}) \text{ (all bounded } \phi\text{).}$$

Want to improve "expected" to "empirical".

Theorem. In order that

$$\phi(\mathcal{F}_n) \to_p \; E\phi(\mathcal{F}) \text{ (all bounded } \phi\text{),}$$

it suffices that the "ancestor chain"
$s \to$ type of ancestor of s

$$R^*(s, ds^*) = \nu(ds^*) R(s^*, ds)/\nu(ds)$$

has trivial invariant σ-field.

Remarks. 1. For CMJBP, ancestor chain is renewal process, and Choquet-Deny theorem applies.
2. Otherwise, seek to verify by coupling.
3. Proof is non-trivial: show dist($\mathcal{F}$) is extreme amongst all fringe distributions.

This gives an abstract way to prove WLLNs, for example for empirical proportion of leaves in $\mathcal{T}_n$. (can be hard to estimate variances in order to use Chebyshev's inequality). Thinking about CLTs is harder.

Given $(\mathcal{T}_n)$ with asymptotic fringe distribution $\mathcal{F}$, we sometimes have a special property

dist. of $\mathcal{F}$ given $|\mathcal{F}| = n$ is same as dist. of $\mathcal{T}_n$.

This coherence holds for
(a) conditioned GWBPs
(b) random recursive trees
(c) random binary search trees
(d) random tries.

Recall we can identify ordered n-trees with $\{-1, 1\}$-valued sequences $(x_1, \ldots, x_{2n})$ for which first return to 0 of partial sums is at time $2n$. We can identify sin-trees with certain infinite $\{-1, 1\}$-valued sequences, and limiting invariant sin-trees with stationary randon sequences (X_i). In the concrete examples, these stationary processes are somewhat unusual. For instance, with "supercritical BP" type models we often find that the partial sum process $\sum_{i=1}^{n} X_i$ is stochastically bounded as $n \to \infty$.

Point. "Coherence" can be defined in terms of the stationary (X_i) as: the tree represented by

$(X_1, \ldots, X_{2n})$ given min$\{m > 0 : S_m \leq 0\} = 2n$

is exactly $\mathcal{T}_n$.

Then CLTs for empirical local quantities associated with random trees (e.g. proportion of leaves) become CLTs for sums

$$\sum_{i=1}^{m} f(\theta^{-i} \circ \mathbf{X}) \qquad\qquad (f \text{ finite width })$$

conditioned as above. (Not yet clear if this approach is useful.)