Steven N. Evans
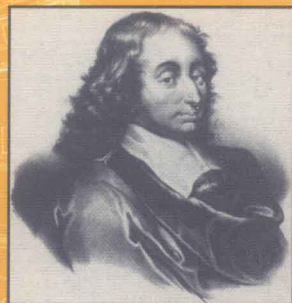
Probability and Real Trees

**Ecole d'Eté de Probabilités
de Saint-Flour XXXV – 2005**

Steven N. Evans

Probability and Real Trees

École d'Été de Probabilités
de Saint-Flour XXXV - 2005

Author

Steven Neil Evans
Department of Statistics # 3860
367 Evans Hall
University of California at Berkeley
Berkeley, CA 94720-3860
USA
e-mail: evans@stat.Berkeley.EDU
URL: http://www.stat.berkeley.edu/users/evans

Cover: Blaise Pascal (1623-1662)

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

For Ailan Hywel, Ciaran Leuel and Huw Rhys

Preface

These are notes from a series of ten lectures given at the Saint–Flour Probability Summer School, July 6 – July 23, 2005.

Berkeley, California, U.S.A.

Steven N. Evans
October 2006

Contents

1

Introduction

The *Oxford English Dictionary* provides the following two related definitions of the word *phylogeny:*

1. The pattern of historical relationships between species or other groups resulting from divergence during evolution.
2. A diagram or theoretical model of the sequence of evolutionary divergence of species or other groups of organisms from their common ancestors.

In short, a phylogeny is the "family tree" of a collection of units designated generically as *taxa*. Figure 1.1 is a simple example of a phylogeny for four primate species. Strictly speaking, phylogenies need not be trees. For instance, biological phenomena such as hybridization and horizontal gene transfer can lead to non-tree-like *reticulate* phylogenies for organisms. However, we will only be concerned with trees in these notes.

Phylogenetics (that is, the construction of phylogenies) is now a huge enterprise in biology, with several sophisticated computer packages employed extensively by researchers using massive amounts of DNA sequence data to study all manner of organisms. An introduction to the subject that is accessible to mathematicians is [67], while many of the more mathematical aspects are surveyed in [125].

It is often remarked that a tree is the only illustration Charles Darwin included in *The Origin of Species*. What is less commonly noted is that Darwin acknowledged the prior use of trees as representations of evolutionary relationships in historical linguistics – see Figure 1.2. A recent collection of papers on the application of computational phylogenetic methods to historical linguistics is [69].

The diversity of life is enormous. As J.B.S Haldane often remarked[1] in various forms:

---

[1] See Stephen Jay Gould's essay "A special fondness for beetles" in his book [77] for a discussion of the occasions on which Haldane may or may not have made this remark.
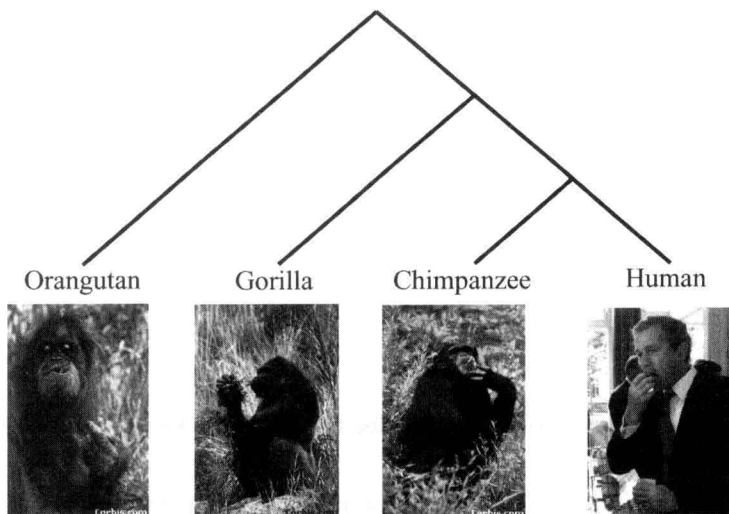
Fig. 1.1. The phylogeny of four primate species. Illustrations are from the *Tree of Life Web Project* at the University of Arizona

> I don't know if there is a God, but if He exists He must be inordinately fond of beetles.

Thus, phylogenetics leads naturally to the consideration of very large trees – see Figure 1.3 for a representation of what the phylogeny of all organisms might look like and browse the Tree of Life Web Project web-site at `http://www.tolweb.org/tree/` to get a feeling for just how large the phylogenies of even quite specific groups (for example, beetles) can be.

Not only can phylogenetic trees be very large, but the number of possible phylogenetic trees for even a moderate number of taxa is enormous. Phylogenetic trees are typically thought of as rooted bifurcating trees with only the leaves labeled, and the number of such trees for n leaves is $(2n-3) \times (2n-5) \times \cdots \times 7 \times 5 \times 3 \times 1$ – see, for example, Chapter 3 of [67]. Consequently, if we try to use statistical methods to find the "best" tree that fits a given set of data, then it is impossible to exhaustively search all possible trees and we must use techniques such as Bayesian Markov Chain Monte Carlo and simulated annealing that randomly explore tree space in some way. Hence phylogenetics leads naturally to the study of large random trees and stochastic processes that move around spaces of large trees.
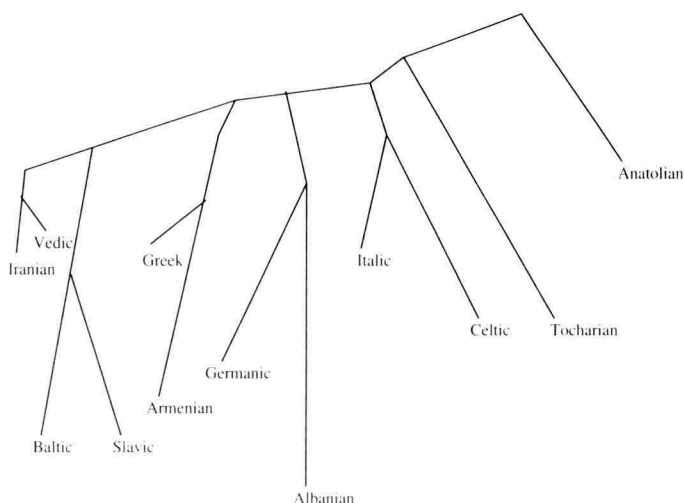
Fig. 1.2. One possible phylogenetic tree for the Indo-European family of languages from [118]

Although the investigation of random trees has a long history stretching back to the eponymous work of Galton and Watson on branching processes, a watershed in the area was the sequence of papers by Aldous [12, 13, 10]. Previous authors had considered the asymptotic behavior of numerical features of an ensemble of random trees such as their height, total number of vertices, average branching degree, etc. Aldous made sense of the idea of a sequence of trees converging to a limiting "tree-like object", so that many such limit results could be read off immediately in a manner similar to the way that limit theorems for sums of independent random variables are straightforward consequences of Donsker's invariance principle and known properties of Brownian motion. Moreover, Aldous showed that, akin to Donsker's invariance principle, many different sequences of random trees have the same limit, the *Brownian continuum random tree*, and that this limit is essentially the standard Brownian excursion "in disguise".

We briefly survey Aldous's work in Chapter 2, where we also present some of the historical development that appears to have led up to it, namely the probabilistic proof of the Markov chain tree theorem from [21] and the algorithm of [17, 35] for generating uniform random trees that was inspired by that proof. Moreover, the asymptotic behavior of the tree-generating algorithm
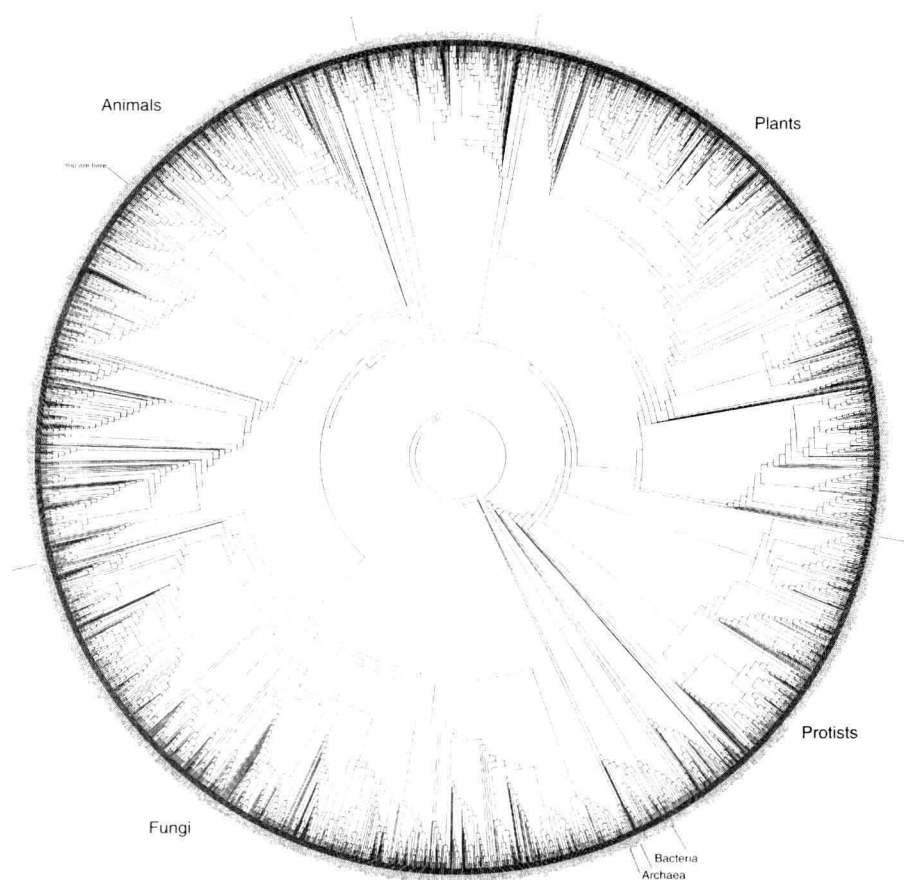
Fig. 1.3. A somewhat impressionistic depiction of the phylogenetic tree of all life produced by David M. Hillis, Derrick Zwickl, and Robin Gutell, University of Texas

when the number of vertices is large is the subject of Chapter 5, which is based on [63].

Perhaps the key conceptual difficulty that Aldous had to overcome was how to embed the collection of finite trees into a larger universe of "tree-like objects" that can arise as re-scaling limits when the number of vertices goes to infinity. Aldous proposed two devices for doing this. Firstly, he began with a classical bijection, due to Dyck, between rooted planar trees and suitable lattice paths (more precisely, the sort of paths that can appear as the "positive excursions" of a simple random walk). He showed how such an encoding of trees as continuous functions enables us to make sense of weak convergence of random trees as just weak convergence of random functions (in the sense of weak convergence with respect to the usual supremum norm). Secondly, he

noted that a finite tree with edge lengths is naturally isomorphic to a compact subset of ℓ^1, the space of absolutely summable sequences. This enabled him to treat weak convergence of random trees as just weak convergence of random compact sets (where compact subsets of ℓ^1 are equipped with the Hausdorff distance arising from the usual norm on ℓ^1).

Although Aldous's approaches are extremely powerful, the identification of trees as continuous functions or compact subsets of ℓ^1 requires, respectively, that they are embedded in the plane or leaf-labeled. This embedding or labeling can be something of an artifact when the trees we are dealing with don't naturally come with such a structure. It can be particularly cumbersome when we are considering tree-valued stochastic processes, where we have to keep updating an artificial embedding or labeling as the process evolves. Aldous's perspective is analogous to the use of coordinates in differential geometry: explicit coordinates are extremely useful for many calculations but they may not always offer the smoothest approach. Moreover, it is not clear *a priori* that every object we might legitimately think of as tree-like necessarily has a representation as an excursion path or a subset of ℓ^1. Also, the topologies inherited from the supremum norm or the Hausdorff metric may be too strong for some purposes.

We must, therefore, seek more intrinsic ways of characterizing what is meant by a "tree-like object". Finite combinatorial trees are just graphs that are connected and acyclic. If we regard the edges of such a tree as intervals, so that a tree is a cell complex (and, hence, a particular type of topological space), then these two defining properties correspond respectively to connectedness in the usual topological sense and the absence of subspaces that are homeomorphic to the circle. Alternatively, a finite combinatorial tree thought of as a cell complex has a natural metric on it: the distance between two points is just the length of the unique "path" through the tree connecting them (where each edge is given unit length). There is a well-known characterization of the metrics that are associated with trees that is often called *(Buneman's) four point condition* – see Chapter 3. Its significance seems to have been recognized independently in [149, 130, 36] – see [125] for a discussion of the history.

These observations suggest that the appropriate definition of a "tree-like object" should be a general topological or metric space with analogous properties. Such spaces are called $\mathbb{R}$-*trees* and they have been studied extensively – see [46, 45, 137, 39]. We review some of the relevant theory and the connection with 0-*hyperbolicity* (which is closely related to the four point condition) in Chapter 3.

We note in passing that $\mathbb{R}$-trees, albeit ones with high degrees of symmetry, play an important role in geometric group theory – see, for example, [126, 110, 127, 30, 39]. Also, 0-hyperbolic metric spaces are the simplest example of the δ-hyperbolic metric spaces that were introduced in [79] as a class of spaces with global features similar to those of complete, simply connected manifolds of negative curvature. For more on the motivation and subsequent history of this notion, we refer the reader to [33, 39, 80]. Groups with a natural

δ-hyperbolic metric have turned out to be particularly important in a number of areas of mathematics, see [79, 20, 40, 76].

In order to have a nice theory of random $\mathbb{R}$-trees and $\mathbb{R}$-tree-valued stochastic processes, it is necessary to metrize a collection of $\mathbb{R}$-trees, and, since $\mathbb{R}$-trees are just metric spaces with certain special properties, this means that we need a way of assigning a distance between two metric spaces (or, more correctly, between two isometry classes of metric spaces). The *Gromov-Hausdorff distance* – see [80, 37, 34] – does exactly this and turns out to be very pleasant to work with. The particular properties of the Gromov-Hausdorff distance for collections of $\mathbb{R}$-trees have been investigated in [63, 65, 78] and we describe some of the resulting theory in Chapter 4.

Since we introduced the idea of using the formalism of $\mathbb{R}$-trees equipped with the Gromov-Hausdorff metric to study the asymptotics of large random trees and tree-valued processes in [63, 65], there have been several papers that have adopted a similar point of view – see, for example, [49, 101, 102, 103, 50, 81, 78].

As we noted above, stochastic processes that move through a space of finite trees are an important ingredient for several algorithms in phylogenetic analysis. Usually, such chains are based on a set of simple rearrangements that transform a tree into a "neighboring" tree. One standard set of moves that is implemented in several phylogenetic software packages is the set of *subtree prune and re-graft* (SPR) moves that were first described in [134] and are further discussed in [67, 19, 125]. Moreover, as remarked in [19],

> The SPR operation is of particular interest as it can be used to model biological processes such as horizontal gene transfer and recombination.

Section 2.7 of [125] provides more background on this point as well as a comment on the role of SPR moves in the two phenomena of lineage sorting and gene duplication and loss. Following [65], we investigate in Chapter 9 the behavior when the number of vertices goes to infinity of the simplest Markov chain based on SPR moves.

Tree-valued Markov processes appear in contexts other than phylogenetics. For example, a number of such processes appear in combinatorics associated with the random graph process, stochastic coalescence, and spanning trees – see [115]. One such process is the *wild chain*, a Markov process that appears as a limiting case of tree–valued Markov chains arising from pruning operations on Galton–Watson and conditioned Galton–Watson trees in [16, 14].

The state space of the wild chain is the set $\mathbf{T}^*$ consisting of rooted $\mathbb{R}$-trees such that each edge has length 1, each vertex has finite degree, and if the tree is infinite there is a single path of infinite length from the root. The wild chain is reversible (that is, symmetric). Its equilibrium measure is the distribution of the critical Poisson Galton–Watson branching process (we denote this probability measure on rooted trees by PGW(1)). When started in a state that is a finite tree, the wild chain holds for an exponentially distributed

amount of time and then jumps to a state that is an infinite tree. Then, as must be the case given that the PGW(1) distribution assigns all of its mass to finite trees, the process instantaneously re-enters the set of finite trees. In other words, the sample–paths of the wild chain bounce backwards and forwards between the finite and infinite trees.

As we show in Chapter 6 following [15], the wild chain is a particular instance of a general class of symmetric Markov processes that spend Lebesgue almost all of their time in a countable, discrete part of their state-space but continually bounce back and forth between this region and a continuous "boundary". Other processes in this general class are closely related to the Markov processes on totally disconnected Abelian groups considered in [59]. A special case of these latter group-valued processes, where the group is the additive group of a local field such as the p-adic numbers, is investigated in [4, 5, 7, 6, 2, 8, 9, 87, 131, 68].

Besides branching models such as Galton–Watson processes, another familiar source of random trees is the general class of coalescing models – see [18] for a recent survey and bibliography.

Kingman's coalescent was introduced in [90, 89] as a model for genealogies in the context of population genetics and has since been the subject of a large amount of applied and theoretical work – see [136, 144, 83] for an indication of some of the applications of Kingman's coalescent in genetics.

Families of coalescing Markov processes appear as duals to interacting particle systems such as the voter model and stepping stone models . Motivated by this connection, [22] investigated systems of coalescing Brownian motions and the closely related coalescing Brownian flow . Coalescing Brownian motion has recently become a topic of renewed interest, primarily in the study of filtrations and "noises" – see, for example, [140, 132, 138, 55].

In Chapter 8 we show, following [60, 44], how Kingman's coalescent and systems of coalescing Brownian motions on the circle are each naturally associated with random compact metric spaces and we investigate the fractal properties of those spaces. A similar study was performed in [28] for trees arising from the beta-coalescents of [116]. There has been quite a bit of work on fractal properties of random trees constructed in various ways from Galton–Watson branching processes; for example, [82] computed the Hausdorff dimension of the boundary of a Galton–Watson tree equipped with a natural metric – see also [104, 96].

We observe that Markov processes with *continuous* sample paths that take values in a space of continuous excursion paths and are reversible with respect to the distribution of standard Brownian excursion have been investigated in [148, 147, 146]. These processes can be thought of as $\mathbb{R}$-tree valued diffusion processes that are reversible with respect to the distribution of the Brownian continuum random tree.

Moving in a slightly different but related direction, there is a large literature on random walks with state-space a given infinite tree: [145, 105] are excellent bibliographical references. In particular, there is a substantial

amount of research on the Martin boundary of such walks beginning with [52, 38, 122].

The literature on diffusions on tree–like or graph–like structures is more modest. A general construction of diffusions on graphs using Dirichlet form methods is given in [141]. Diffusions on tree–like objects are studied in [42, 93] using excursion theory ideas, local times of diffusions on graphs are investigated in [53, 54], and an averaging principle for such processes is considered in [71]. One particular process that has received a substantial amount of attention is the so-called *Walsh's spider*. The spider is a diffusion on the tree consisting of a finite number of semi–infinite rays emanating from a single vertex – see [142, 26, 139, 25].

A higher dimensional diffusion with a structure somewhat akin to that of the spider, in which regions of higher dimensional spaces are "glued" together along lower dimensional boundaries, appears in the work of Sowers [133] on Hamiltonian systems perturbed by noise – see also [111]. A general construction encompassing such processes is given in [64]. This construction was used in [24] to build diffusions on the interesting fractals introduced in [95] to answer a question posed in [84].

In Chapter 7 we describe a particular Markov process with state–space an $\mathbb{R}$-tree that does not have any leaves (in the sense that any path in the tree can be continued indefinitely in both directions). The initial study of this process in [61] was motivated by Le Gall's *Brownian snake* process – see, for example, [97, 98, 99, 100]. One agreeable feature of this process is that it serves as a new and convenient "test bed" on which we can study many of the objects of general Markov process theory such as Doob h-transforms, the classification of entrance laws, the identification of the Martin boundary and representation of excessive functions, and the existence of non-constant harmonic functions and the triviality of tail σ-fields.

We use Dirichlet form methods in several chapters, so we have provided a brief summary of some of the more salient parts of the theory in Appendix A. Similarly, we summarize some results on Hausdorff dimension, packing dimension and capacity that we use in various places in Appendix B.

Around the Continuum Random Tree

2.1 Random Trees from Random Walks

2.1.1 Markov Chain Tree Theorem

Suppose that we have a discrete time Markov chain $X = \{X_n\}_{n \in \mathbb{N}_0}$ with state space V and irreducible transition matrix P. Let π be the corresponding stationary distribution. The *Markov chain tree theorem* gives an explicit formula for π, as opposed to the usual implicit description of π as the unique probability vector that solves the equation $\pi P = \pi$. In order to describe this result, we need to introduce some more notation.

Let $G = (V, E)$ be the directed graph with vertex set V and directed edges consisting of pairs of vertices (i, j) such that $p_{ij} > 0$. We call p_{ij} the *weight* of the edge (i, j).

A *rooted spanning tree of G* is a directed subgraph of G that is a spanning tree as an undirected graph (that is, it is a connected subgraph without any cycles that has V as its vertex set) and is such that each vertex has out-degree 1, except for a distinguished vertex, the *root*, that has out-degree 0. Write $\mathcal{A}$ for the set of all rooted spanning trees of G and $\mathcal{A}_i$ for the set of rooted spanning trees that have i as their root.

The *weight* of a rooted spanning tree T is the product of its edge weights, which we write as weight(T).

Theorem 2.1. *The stationary distribution π is given by*

$$\pi_i = \frac{\sum_{T \in \mathcal{A}_i} \text{weight}(T)}{\sum_{T \in \mathcal{A}} \text{weight}(T)}.$$

Proof. Let $\bar{X} = \{\bar{X}_n\}_{n \in \mathbb{Z}}$ be a two-sided stationary Markov chain with the transition matrix P (so that $\bar{X}_n$ has distribution π for all $n \in \mathbb{Z}$).

Define a map $f : V^{\mathbb{Z}} \to \mathcal{A}$ as follows – see Figure 2.2.

- The root of $f(x)$ is x_0.