Statistical Tests for Large Tree-Structured Data

Karthik Bharath\textsuperscript{a}, Prabhanjan Kambadur\textsuperscript{b}, Dipak K. Dey\textsuperscript{c}, Arvind Rao\textsuperscript{d}, and Veerabhadran Baladandayuthapani\textsuperscript{e}

\textsuperscript{a}School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom; \textsuperscript{b}Bloomberg LP, New York, NY; \textsuperscript{c}Department of Statistics, University of Connecticut, Storrs, CT; \textsuperscript{d}Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX; \textsuperscript{e}Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX

\textbf{ABSTRACT}

We develop a general statistical framework for the analysis and inference of large tree-structured data, with a focus on developing asymptotic goodness-of-fit tests. We first propose a consistent statistical model for binary trees, from which we develop a class of invariant tests. Using the model for binary trees, we then construct tests for general trees by using the distributional properties of the continuum random tree, which arises as the invariant limit for a broad class of models for tree-structured data based on conditioned Galton–Watson processes. The test statistics for the goodness-of-fit tests are simple to compute and are asymptotically distributed as $\chi^2$ and $F$ random variables. We illustrate our methods on an important application of detecting tumor heterogeneity in brain cancer. We use a novel approach with tree-based representations of magnetic resonance images and employ the developed tests to ascertain tumor heterogeneity between two groups of patients. Supplementary materials for this article are available online.

1. Introduction

Rapid advancements in technology have led to the emergence of datasets in which the underlying quantities of interest are non-Euclidean objects. Increasingly encountered across several disciplines is data that generate tree-like structures or hierarchical representations. Following a suitable representation of trees, the assumption then is to treat the tree-structured object as an observable quantity that represents the statistical atom. Some central challenges have stymied the systematic development of tools for statistical inference in such settings: the non-Euclidean nature of the set of trees induces an unreasonable dependence of models on the choice of representation and geometry; comparing and sampling trees with different topological and branch length information is not straightforward; the labeling scheme of the vertices and branches influence inference.

Attempts to address these issues, at least from a modeling perspective, have hitherto been characterized by nonparametric or heuristic/algorithmic approaches specifically motivated by applications. For example, in the context of trees used to model blood vessels in a human brain, Shen et al. (2014) used nonparametric functional data analysis methods; Wang et al. (2012) developed a nonparametric regression model with a tree-structured response variable; and Aylward and Bullitt (2002) and Aydin et al. (2009) used principal component analysis (PCA) methods using a suitable metric on trees. Yang, Kalnis, and Tung (2005) and Tatikonda and Parthasarathy (2010) employed the Edit distance with algorithmic tools for detection of duplicate XML records represented as trees. See also Le, Nussinov, and Maizel (1989), Schmitt and Waterman (1994), and Kolbe and Eddy (2009) for use of algorithmic methods on secondary structures of RNA represented as trees. Despite the flexibility associated with such nonparametric and algorithmic methods for tree-structured data, depending on the choice of representation and metric, developing robust inferential methods, and determining distributions of the statistics of interest are difficult. Critically, without a probability model supporting the simulation of trees through a generative model, it becomes hard to assess the generality of the methods across tree-structured datasets.

In contrast, parametric probability models for trees were considered by Steele (1987) and Aldous (1996) wherein the parameters were designed to capture topological features and branch length information. Such probability models form the basis of modeling hierarchical data using Bayesian methods with nonparametric hierarchical priors such as the Dirichlet diffusion tree (Neal 2001), beta-coalescent trees (Hu et al. 2013) and stick-breaking priors (Adams, Ghahramani, and Jordan 2010), and tasks involving clustering (Heller and Ghahramani 2005; Teh, Daume, and Roy 2008). Although sampling classes of trees are convenient with such methods, posterior inference is nonstandard, and assessing the quality of posterior samples is difficult. Furthermore, extensions to general classes of trees require substantial revamping of the computational setup; for example, modifying the Dirichlet diffusion tree prior for binary trees to handle trees with an arbitrary number of children requires significant changes in the probability model and computational techniques.

Such considerations support an amalgamation of the simplicity of parametric models, in terms of interpretation and...
simulation, and the flexibility and generality offered by non-parametric or algorithmic methods to develop coherent and principled inferential methods for tree-structured data. This forms the leitmotif of this article. The key challenges in constructing valid probability models for inference are intimately related to the structural components of a tree-structured datum: internal vertices, leaves or terminal vertices, and labels. Candidate probability distributions are necessarily multivariate and capture hierarchical relationships between the vertices and the edge length information; in contrast to nonhierarchical relationships, marginalization with respect to any subset of the vertex set will not be meaningful. Moreover, the distributions should be unaffected by the choice of labeling scheme on the vertices. This issue is best understood on binary trees (each node, except the root, has either 0 or 2 children only) with n leaves, 2n − 1 edges, and 2n vertices (including the root and the leaves), when the root has only one child. For simplicity, in this example let us ignore edge length information. Since the removal of an internal vertex destroys the tree structure, it is natural to define a distribution using n leaves. As a consequence, if \( f_n \) is an n-dimensional distribution on the tree with n leaves, then the removal of one leaf should ensure that the resulting distribution on the binary tree with \( n - 1 \) leaves is \( f_{n-1} \). Additionally, \( f_n \) should be exchangeable with respect to the leaves, thereby ensuring that the leaf labels are irrelevant. Figure 1 further elucidates these issues using binary trees. Considering the complexity arising from different sources of variation the following requirements are crucial to any statistical method on tree-structured data:

(i) Uniform sampling: modeling and inference should be compatible with notions of random sampling from a population of trees;

(ii) Projective property: when increasing or decreasing the dimensionality of a tree by adding or deleting leaves, the probability distribution should be compatible with the notion of obtaining marginal distributions from higher-dimensional ones;

(iii) Interpretability of parameters: parameters should retain interpretability when moving from a subtree to the full tree, or vice versa, when adding or deleting leaves;

(iv) Lack of dependence on labeling: inference should be impervious to labeling of leaves and edges, since the labeling scheme adopted on tree-structured data is usually arbitrary.

In view of the requirements, we first propose a generative model for binary trees that satisfies (i)–(iv), resulting in a density that is exchangeable with respect to the leaf labels. Such a property fits into the general framework for exchangeable random structures developed by Orbanz and Roy (2015) in the context of Bayesian models. Using the density, we construct exact parametric one- and two-sample tests that are invariant to permutation of the leaves. This is meaningful because the number of leaves in a binary tree determines its topological structure. The density of binary trees is vital in the generalization to arbitrary trees. We model general trees using the genealogical trees of Galton–Watson branching processes with an offspring distribution, conditioned on the total progeny, and referred to as conditioned Galton–Watson trees (see Section 2 for definition). These models are quite versatile and can be used for various types of trees. However, the distribution of a conditioned Galton–Watson tree incorporates only topological information and does not satisfy all requirements (i)–(iv). To remedy this, we move to the asymptotic limit (as the number of vertices grow without bound) of such trees by seeking recourse to the abstract notion of a limit tree, referred to as the continuum random tree (CRT) from Aldous (1991a) and Aldous (1993). By suitably modifying and extending the theory of the CRT to accommodate realistic statistical models, we show how the model for binary trees can be extended to a valid probability model for a rich class of general trees. Conditioned Galton–Watson trees, regardless of the offspring distribution, converge weakly to the same limit tree, the CRT; using the two different distributional characterizations of the CRT, we propose goodness-of-fit tests for conditioned Galton–Watson tree models. We design a simulation technique based on an expected linear run-time algorithm by Devroye (2012), which enables the efficient generation of large samples of large conditioned Galton–Watson trees. We then illustrate our methodology through a case study in cancer imaging, where it is well established that a systematic classification of intra- and intertumor heterogeneity is crucial for drug development and accurate assessment of response to treatment (Just 2014). Obtaining tree-based representations of magnetic resonance (MR) images of brain cancer, we use the proposed tests to detect tumor heterogeneity between cancer patients with long versus short survival times.

The article is structured as follows. In Section 2, we establish the notations and review technical details used in the article. In
Section 3, we illustrate the motivation for using the CRT through a simple model for binary trees that leads to the definition of a consistent family of densities for the tree-structured data, and the development of invariant tests for tree distributions. In Section 4, we review and suitably modify the key ingredients of the CRT, restricting our attention to relevant aspects of their relationships. In Section 5, we construct goodness-of-fit tests for conditioned Galton–Watson tree models using a class of binary subtrees. In view of requirement (iii), we then consider the problem of estimating and interpreting the variance parameter in the offspring distribution of conditioned Galton–Watson tree models. In Section 6, we present results from numerical images. In Section 7, we comment on possible extensions and generalizations, and point out some shortcomings of the current framework.

2. Notation and Technical Preliminaries

To make the article self-contained, in this section we detail concepts that figure repeatedly in our methodological development. This section can be skimmed at first reading and referred to later as required.

Tree representation: A tree is a connected, acyclic graph with a distinguished vertex or node called the root. Nodes are connected through edges that have nonzero lengths. Trees are allowed to have unequal numbers of vertices or nodes. We confine our attention to finite, rooted, ordered trees: trees with a root, containing a finite number of vertices, and having an ordering among the offspring at a parent node. This leads to the notion of a left and right child in the case of binary trees. Ordered or planar representations of unordered trees can be obtained using the scheme by Aldous and Pitman (2000).

Following the notation used by Aldous (1991a), we represent a finite rooted tree \( \tau_n \) with \( n \) vertices, including the root, as a point \( \tau_n \) in the product space \( \mathcal{T}_n \times \mathbb{R}^{n-1} \), where \( \mathcal{T}_n \) is the set of all finite trees on \( n \) vertices. This implies that the connectivity information between the vertices is contained in \( \mathcal{T}_n \). For simplicity, we choose the notation \( \tau_n = (\mathcal{V}(\tau_n), E(\tau_n)) \), where \( \mathcal{V}(\tau_n) = (\text{root}, v_1, \ldots, v_{n-1}) \) is the set of vertices and \( E(\tau_n) = (e_1, \ldots, e_{n-1}) \) represents the set of edge lengths. It is tacitly assumed that the notation contains the information that relates each edge to a unique pair of vertices, and we always view a tree \( \tau_n \) as a point in the product space containing topological and edge length information, and use the notation of the vertex and edge length set for convenience. We denote by \( \tau_n \) the tree with \( n \) vertices including the root, and by \( \tau(n) \) the tree with \( n \) terminal vertices or leaves.

Conditioned Galton–Watson trees: A Galton–Watson process \( \{X_n\}_{n \geq 0} \) with offspring distribution \( \{\pi_k, k = 0, 1, \ldots\} \) is a \( \mathbb{Z}_+ \)-valued discrete-time Markov chain with transition function

\[
P(X_{n+1} = k | X_n = m) = \pi_k^m,
\]

where \( * \) denotes convolution operator. If \( \mu = \sum_k k\pi_k \), the process \( X_n \) is critical if \( \mu = 1 \), sub-critical if \( \mu < 1 \), and super-critical if \( \mu > 1 \).

A Galton–Watson tree \( \tau \) corresponding to \( X_n \) is constructed by recursively starting with the root and giving each node \( v_i \), independent of other nodes, a number \( o(v, \tau) \) of children with probability \( \pi_{o(v, \tau)} \), where \( o(v, \tau) \in \{0, 1, \ldots\} \) is the out-degree or the number of children of vertex \( v \) in the vertex set \( \mathcal{V}(\tau) \) of \( \tau \). \( \{\pi_k\} \) is referred to as the offspring distribution. When conditioned to have \( n \)-vertices, the resulting tree \( \tau_n \) is known as a conditioned Galton–Watson tree with distribution

\[
P(\tau_n = t) \propto \prod_{v \in \mathcal{V}(t)} \pi_{o(v, \tau)} \quad \text{on} \{ t : \text{cardinality of } \mathcal{V}(t) = n \}.
\]

Least common ancestor tree (L-tree): For a tree \( \tau_n = (\mathcal{V}(\tau_n), E(\tau_n)) \), we define its least common ancestor tree (L-tree) in the following manner: choose a subset \( B \) of \( \mathcal{V}(\tau_n) \) for vertices \( v_1 \) and \( v_2 \) in \( B \) find their last common ancestor, or the branch point after which the paths to \( v_1 \) and \( v_2 \) from the root diverge or branch out. Then, the L-tree corresponding to the subset \( B \) of the vertices of \( \tau_n \) is the tree, denoted as \( L(\tau_n, B) \), that contains the root, the vertices of \( B \), and all the branch points with distances from the root to the vertices of \( B \) preserved. Figure 2 illustrates this with \( B = \{v_1, v_2, v_3, v_4\} \); the branch points are \( b_1 \) and \( b_2 \), and to preserve the distances from the root to the vertices of \( B \), the new branch from the branch points to the elements of \( B \) has a length that is equal to the sum of the edges along the path from the root to the elements of \( B \) in the original tree.

3. A Generative Model and Test for Binary Trees

We begin with binary trees for two reasons. We intend to define probability distributions on trees using leaf-level information,
which for binary trees determines the topology completely. Second, the distributional properties of the CRT, which we intend to use to define asymptotic models on general trees, are completely characterized by random binary subtrees. Consequently, a systematic development of methods with desirable properties for binary trees provides similarly attractive ones for general trees. The results of this section are pivotal in developing inferential models for general trees. We first describe a generative model and its properties and then construct one- and two-sample tests that are invariant to leaf-labeling.

Note that a binary tree with \( k \) leaves \( \tau(k) \) is isomorphic to the binary tree \( \tau_{2k} \) with 2 \( k \) vertices. Suppose \( t_1, t_2, \ldots \) are the arrival times of a nonhomogenous Poisson process with rate \( \lambda(t) = t \). Let \( \tau(1) \) be a tree with one leaf formed by attaching an edge of length \( t_1 \) to the root. This tree will have a density that corresponds to the time until the first event of the nonhomogenous Poisson process, with mean function \( m(t) = \int_0^t sds = t^2/2 \). A tree with two leaves \( \tau(2) \) is then constructed by adding a second edge of length \( t_2 - t_1 \), the interarrival time, to a point on \( \tau(1) \) chosen according to a uniform distribution on \([0, t_1]\). Recursively, the tree \( \tau(k + 1) \) is obtained from \( \tau(k) \) by attaching an edge of length \( t_{k + 1} - t_k \) to a point chosen randomly on the tree (i.e., according to a uniform distribution on the sum of the branch lengths of \( \tau(k) \)) and labeling the new leaf \( k + 1 \). Since we are interested in only ordered trees, an edge can be attached to the left or to the right of the randomly chosen point with equal probability. Observe that \( \tau(k) \) has 2 \( k \) - 1 branches with lengths, say \((x_1, \ldots, x_{2k-1})\), and \( \tau(k + 1) \) has 2 \( k + 1 \) branches of lengths \((y_1, \ldots, y_{2k+1})\), with \( \tau(k + 1) \) formed by the above construction by splitting a branch of length \( x_j \) into two branches \( y_{j1} \) and \( y_{j2} \) with \( x_j = y_{j1} + y_{j2} \). Therefore, through induction, the “density” of \( \tau(k + 1) \) is

\[
\frac{\gamma_k(\tau(k))}{\gamma_{k+1}(\tau(k+1))} = \text{(density of } k \text{th interarrival time)} \times \text{(probability of attaching the new edge to } \tau(k) \text{)} \times \text{(probability of a left or right attachment)}
\]

Note that \( g(\cdot) \) at this state is a density (nonnormalized) with respect to the product Lebesgue measure on the edge lengths only, with the topological information of the tree being captured only in the decision to add an edge to the left or right with equal probability. Denoting the sum of the branch lengths of \( \tau(k) \) by \( s_k \), and noting from the construction that the third term is \( 1/s_k \) and the fourth is \( 1/2 \), we obtain, \( g_{k+1}(\tau(k+1)) = g_k(\tau(k)) \times (s_k)^{-1} \times \text{(probability of attaching the new edge to } \tau(k) \times \text{(probability of a left or right attachment)}) \)

\[
g_k(\tau(k)) = \frac{1}{2^{k-1}}s_k^{-1/2}, \quad s = \sum_{i=1}^{2k-1} x_i.
\]

The function \( g(\cdot) \) is nonnegative but does not integrate to one. By virtue of the construction mechanism, since at each step \( i \) there are \( 2i - 1 \) possible tree topologies, the structural information of the tree is captured in the normalizing constant:

\[
\int \cdots \int \frac{1}{s_k^{1/2}} dx_1 \cdots dx_{2k-1} = \prod_{i=1}^{k-1} \frac{1}{2i-1}.
\]

Therefore, the probability density of the ordered and rooted binary tree with \( k \) leaves is

\[
f_k(\tau(k)) = \prod_{i=1}^{k-1} \frac{1}{2i-1} g_k(\tau(k)) = \frac{1}{2^{k-1}}s_k^{-1/2}, \quad s = \sum_{i=1}^{2k-1} x_i.
\]

### 3.1. Interpretation and Properties of the Density

The density \( f_k \) can be interpreted as assigning mass to a tree with branch lengths \((x_1, \ldots, x_{2k-1})\) with a canonical binary tree topology on \( k \) leaves or \( 2k \) vertices. That is, the mass is the same for rooted binary trees with \( k \) leaves of all possible topologies with branch lengths \((x_1, \ldots, x_{2k-1})\). We hence observe that, conditional on \( k \), \( f_k \) is a density with respect to the product measure \( u_k \otimes dx \) on \( \mathbb{T}_{2k} \times \mathbb{R}_{2k-1}^+ \), where \( u_k \) is a uniform measure on all rooted binary trees on \( k \) leaves, and \( dx \) is the Lebesgue measure on \( \mathbb{R}_{2k-1}^+ \). The function \( f_k \) is interpreted as a density conditional on \( K = k \), the number of interevent times of the Poisson process considered. Indeed, \( K \) is Poisson distributed with mean \( \frac{2}{2} \).

The model is particularly useful in applications where branch lengths are perhaps more important than topology (Fu and Li 1993; Lanfear 2010). We introduce a parametric form of the density obtained by defining \( \lambda(t, \theta) = \theta t \) that leads to a parametric family \( \{f_k; \theta \in \Theta\} \), wherein trees of varying branch lengths can be generated by varying \( \theta \) in \( \Theta \). The reason for the simple
multiplicative form lies in our insistence that the density satisfy properties (i)–(iv) in the Introduction. The projective property of the induced probability measures (via the construction) is easily violated when complicated parametric forms are considered for the rate function of the Poisson process. Based on the requirements in the Introduction, we briefly summarize four attractive properties of the density \( f_{k, \theta} \) as a statistical model. Formal statements of the properties and their proofs are available in Section 2.1 (Theorem 2.1 and Proposition 2.1) of supplementary material.

(i) The uniform measure \( u_k \) on the topology ensures that the particular binary tree has been randomly sampled from a population of binary trees on \( k \) leaves.

(ii) For every \( \theta \in \Theta \), \( f_{k, \theta} \) is compatible with marginalization over the number of leaves \( k \).

(iii) For \( \tau(k) \), the density \( f_{k, \theta} \) depends on the branch lengths \( x_1, \ldots, x_{2k-1} \) only through their sum \( \sum x_i \), which is Gamma distributed with shape \( k \) and scale 2.

(iv) From (1), it is evident that \( f_{k, \theta} \) is exchangeable with respect to the branch lengths.

### 3.2. Invariant One- and Two-Sample Tests

Hypothesis tests for trees entail testing for a generative probability model in which the probability distribution depends on the number of vertices and branches. When we state that an independent sample is generated from a density, we are essentially viewing the trees as independent trees generated from the same probability model, regardless of the number of vertices. Since the density \( f_{k, \theta} \) is uniform on all topologies of \( k \)-leaved binary trees, for a chosen tree topology, exchangeability of the leaves implies that the data space \( X_k := \mathbb{R}^{2k-1} \) is invariant to the symmetric group \( G \) of all possible permutations of the branch lengths. Thus every \( g \in G \) acting on \( X_k \) induces an action \( g' \) on \( \Theta \), that is, if \( f_{k, \theta} \) is the resulting density following the transformation \( gX_k \) (with density \( f_{k, \theta} \)), then \( g'\theta = \theta' \). It is easy to observe from the Poisson process construction that \( g' \) is the identity function (i.e., \( g' \theta = \theta' = \theta \) for every \( \theta \in \Theta \)), and therefore \( \Theta \) is also invariant under any \( g \in G \). This leads to a testing procedure for binary trees that is invariant to the action of \( G \) on the leaves. Technical details ensuring this are in Proposition 2.2 of supplementary material.

**Theorem 1.** Suppose \( \tau(n) = (\tau(n_1), \ldots, \tau(n_p)) \) and \( \eta(m) = (\eta(m_1), \ldots, \eta(m_q)) \) are independent samples of binary trees from \( f_{0, \theta} \) and \( f_{0, \theta} \), respectively.

1. Consider the critical function

\[
\phi(\theta', n, \alpha) = \begin{cases} 
1 & \text{if } \theta' \sum_{i=1}^{p} s_i < \chi_{a-2, \sum_{i=1}^{p} n_i} \text{ or } \theta' \sum_{i=1}^{p} s_i > \chi_{1-a-2, \sum_{i=1}^{p} n_i}, \\
0 & \text{otherwise},
\end{cases}
\]

where \( s_i \) is the sum of the branch lengths of \( \tau(n_i) \) and \( \chi_{a, b} \) denotes the \( \alpha \)th percentile of a \( \chi^2 \) distribution with \( b \) degrees of freedom. For the hypotheses \( H_0 : \theta_0 = \theta' \) against \( H_1 : \theta_0 \neq \theta' \), \( \phi(\cdot) \) is invariant to the permutation of leaves, and is such that \( E_{H_0} \phi(\theta', n, \alpha) = \alpha \).

2. Let \( r_j \) denote the sum of the branch lengths of \( \eta(m_j) \). Then, the critical function

\[
\psi(n, m, \alpha) = \begin{cases} 
1 & \text{if } \frac{\sum_{j=1}^{q} r_j}{\sum_{j=1}^{q} m_j} > \frac{1}{a-2 \sum_{j=1}^{q} m_j} F_{F_{a, 2 \sum_{j=1}^{q} m_j}}, \\
0 & \text{otherwise},
\end{cases}
\]

where \( F_{a, b} \) is the \( \alpha \)th percentile of an \( F \) distribution with \( a \) and \( b \) degrees of freedom, for testing \( H_0 : \theta_0 = \theta_1 \) against \( H_1 : \theta_0 \neq \theta_1 \), and is such that \( E_{H_0} \psi(n, m, \alpha) = \alpha \).

In Theorem 1 for the particular case in which \( \theta' = 1 \) without explicitly specifying an alternative hypothesis, the test is a goodness-of-fit test for the nonhomogenous Poisson process model generating trees with intensity \( \lambda(t) = t \). It can easily be checked that the invariance to the action of the symmetry group on the leaves is satisfied.

### 4. Conditioned Galton–Watson Tree Models and the CRT

The tests in Theorem 1 can neither be used to detect topological variation within or across samples, nor to test nonbinary trees. In this section, we describe the ingredients that allow us to extend the use of the density in (1) to nonbinary trees.

#### 4.1. Conditioned Galton–Watson Tree Models

Conditioned Galton–Watson trees are genealogical trees of Galton–Watson processes conditioned on the total progeny. We refer to the definition in Section 2 of a conditioned Galton–Watson tree \( \tau_n \) with \( n \) vertices. The key property of interest is that for a fixed offspring distribution \( \pi_n \), the corresponding conditioned Galton–Watson tree can be viewed as being picked according to a uniform distribution on certain types of trees with \( n \) vertices. For example, if we wish to choose a binary tree with \( n \) vertices according to a uniform distribution on the space of \( n \)-vertex binary trees, we can equivalently construct a conditioned Galton–Watson tree with an offspring distribution of \( 1 - p \) and \( p \) each for 0 and 2 children, respectively, where \( p \) is the probability of having 2 children. Other examples include geometric offspring distribution with probability 1/2 for ordered trees with unrestricted degree; Binomial distribution on 2 trials with probability 1/2 for trees with 0,1 or 2 vertex degrees; uniform distribution on \( \{0, 1, 2\} \) for unordered and unlabeled trees with 0,1, or 2 vertex degrees . For a detailed look at the properties of conditioned Galton–Watson trees, and their relationship with other tree models, we refer the interested reader to Janson (2012).

Evidently, conditioned Galton–Watson trees can be constructed from critical, sub-critical, and super-critical Galton–Watson process. However, asymptotic behavior of such trees (as number of vertices tends to infinity) blurs the decision between trees from critical and noncritical Galton–Watson processes: in the sub-critical or super-critical case with offspring distribution \( \pi_n \), as long as there exists a \( \lambda > 0 \) such that \( \sum_{i, \theta} \pi_i \lambda^i < \infty \), the asymptotic behavior resembles that of a critical Galton–Watson process with finite offspring variance. For modeling
purposes such a property is useful since a broad class of distributions can be used as offspring distributions for the Galton–Watson process, giving rise to a rich class of trees. The following Proposition clarifies this for a few offspring distributions.

**Proposition 1.** Conditioned Galton–Watson trees with the following offspring distributions can be modeled as critical conditioned Galton–Watson trees for every $0 < p < 1$ in (i)–(iii) and $0 < p_0, p_1 < 1$ in (iv).

(i) $\pi_i = (1 - p)^{i-1}p$ for $i = 1, 2, \ldots$ and $0 < p < 1$;

(ii) $\pi_i = \frac{2i}{(2i-1)!!} p^{2i-1}$ for $i = 0, 1, 2$ and $0 < p < 1$;

(iii) $\pi_0 = 1 - p$, $\pi_2 = p$ and $0 < p < 1$;

(iv) $\pi_0 = p_0$, $\pi_1 = p_1$, $\pi_2 = 1 - p_0 - p_1$ and $0 < p_0, p_1 < 1$.

At this point we may choose to view inference on conditioned Galton–Watson trees as inference on Galton–Watson processes; a good source for available methods is Guttormsg (1991). However, we move to the asymptotic setting for at least two reasons: to incorporate branch length information within the conditioned Galton–Watson tree models; and to obtain knowledge of the distributions of local structural aspects like height and variations in branching structure, through weak convergence techniques based on the CRT.

### 4.2. Continuum Random Tree and L-tree Models

A terse and heuristic definition of the CRT is given in Section 2; see Aldous (1993) for a formal definition. The CRT arises as a “continuous” limit of conditioned Galton–Watson trees as the number of vertices tends to infinity, regardless of the offspring distribution (up to a scaling factor). In this setting, the variance parameter $\sigma^2$ of the offspring distribution of conditioned Galton–Watson trees appears in the limit. The distribution of the CRT can be characterized in two equivalent ways: as limit of randomly chosen binary subtrees of conditioned Galton–Watson trees; as a weak limit of a continuous function constructed by a walk on conditioned Galton–Watson trees. Both characterizations will be profitably used to construct goodness-of-fit tests.

The class of binary subtrees that characterize the distribution of the CRT arises as the limit of randomly chosen binary subtrees of conditioned Galton–Watson trees known as L-trees used in various applications (Aho et al. 1981; Gronau and Moran 2007); for a constructive definition of an L-tree see Section 2. For a conditioned Galton–Watson tree $\tau_n$, fix $k < n$ and choose $k$ leaves according to a uniform distribution on $V(\tau_n)$. Therefore, the L-tree of a conditioned Galton–Watson tree that is a binary tree with $k$ leaves and $2k - 1$ edges. The following characterization of the CRT based on the limit of the L-trees, modified suitably for the case of ordered conditioned Galton–Watson trees, is useful for our purpose. The limit binary trees can be viewed as “marginals” of the CRT or its finite-dimensional projections.

**Lemma 1** (Aldous 1993). From a conditioned Galton–Watson tree $\tau_n = (V(\tau_n), E(\tau_n))$ generated from an offspring distribution with variance $\sigma^2 < \infty$, for a fixed $k < n$, let $L(\tau_n, \{l_1, \ldots, l_k\})$ be an L-tree obtained by leaves $l_i \in V(\tau_n)$ with $s = e_1 + \cdots + e_{2k-1}$. Then, as $n \to \infty$, for a fixed $k$, there exists a consistent family $(C(k), k \geq 1)$ of binary trees with $k$ leaves that define the CRT that has the density

$$f_{k, \sigma^2}(c(k)) = \left[\prod_{i=1}^{k-1} \frac{1}{2i - 1}\right]^{-1} \frac{1}{2k-1} (\sigma^2)^{k} e^{-\frac{c(k)x_k^2}{2}} \quad \sigma^2 \in \mathcal{S}.$$  

The L-trees provide the link through which the density for binary trees obtained using the Poisson process construction is related to conditioned Galton–Watson trees and the CRT. The density in (2) coincides with density obtained for binary trees in (1) from the nonhomogenous Poisson process construction, and can be used to approximate the distribution of L-trees of conditioned Galton–Watson trees from any offspring distribution with a finite variance $\sigma^2$. Bearing in mind that the binary tree model satisfies requirements (i)–(iv), the idea then is to model L-trees of observed trees with the probability density in (2), conduct inference, and then extend the resulting conclusions to the whole tree.

The construction of an L-tree incorporates both topological and branch length information. Lemma 1 implies that the asymptotic distribution of the sequence of L-trees of any conditioned Galton–Watson tree model can be approximated by the sequence in (2). The freedom associated with the choice of the number of leaves used to construct an L-tree of a given tree helps in reducing the dimensionality of the inferential problem involving large trees. Two tree populations can be distinguished by constructing low-dimensional summary statistics by choosing a small proportion of the leaves while constructing an L-tree, contingent on consistent estimation of $\sigma^2$. An important issue is the interpretability of $\sigma^2$ in the context of using $f_{k, \sigma^2}$ as a parametric statistical model on the L-trees. Formal statement of this property and its proof can be found in Proposition 2.3. in the supplementary material.

### 5. Goodness-of-Fit Tests for Conditioned Galton–Watson Trees

The employment of conditioned Galton–Watson trees models, through their connection to the CRT, permits us to extend the use of the consistent parametric family of densities on binary trees to more general settings. Importantly, the CRT, characterized by the limiting L-trees, is the invariant limit for conditioned Galton–Watson trees from any offspring distribution with finite variance $\sigma^2$. Consequently, the density in (2) can be used to approximate the density of a binary L-tree of any conditioned Galton–Watson tree. In contrast to the situation with binary trees, bearing in mind the invariant limit (CRT), the appropriate test would be a goodness-of-fit test for conditioned Galton–Watson trees, where $\sigma^2$ is viewed as a nuisance parameter that needs to be estimated consistently. We therefore develop tests to check whether the data have been generated from a conditioned Galton–Watson tree model. The null hypothesis would be that the samples of trees are independent copies of a conditioned Galton–Watson tree with finite-variance offspring distribution. This generalizes the test in Theorem 1 which is applicable only to binary trees.

Starting with a tree $\tau_n$ with $n_l$ number of leaves, the requisite one-sample test invariant to the permutation of leaves, is constructed as follows: (1) construct a consistent estimator
\[ \hat{\sigma}^2 \] of \( \sigma^2 \); (2) for each tree, construct an L-tree by randomly choosing a subset of the leaves; (3) using the test statistic defined as the product of \( \hat{\sigma}^2 \) and the sum of the branch lengths of the L-trees, and Slutsky’s theorem, construct an asymptotic rejection region based on the \( \chi^2 \) distribution. The invariance property is interpreted conditional on the number of leaves chosen randomly. The extension to the two-sample case is straightforward.

Theorem 2. Suppose \( \tau_n = (\tau_{n_1}, \ldots, \tau_{n_k}) \) and \( \eta_m = (\eta_{m_1}, \ldots, \eta_{m_l}) \) are independent samples of conditioned Galton–Watson trees from \( \pi_\tau \) and \( \pi_\eta \), respectively, with \( \hat{\sigma}_r^2 \) and \( \hat{\sigma}_n^2 \) as offspring variances, with respective consistent estimators \( \hat{\sigma}_r \) and \( \hat{\sigma}_n \).

Let \( k = (k_1, \ldots, k_p) \) be subsets of the leaves of \( \tau_{n_k} \), where \( k_i \) is chosen according to a uniform distribution on the leaves of \( \tau_{n_k} \) and let \( k = |k| \) denote the cardinality of set \( k \).

1. For a fixed \( k \), define the critical function

\[
\phi(k, \alpha) = \begin{cases} 
1 & \text{if } \frac{1}{|k|} \sum_{i=1}^{|k|} s_i > \chi_{1-\alpha, 2\sum_{i=1}^{|k|}|k|}; \\
0 & \text{otherwise},
\end{cases}
\]

where \( s_i \) are the total path lengths of \( L(\tau_{n_k}, k_i) \).

Then, given \( k \), for the pair of hypotheses \( H_0 : \pi_\tau = \pi_\eta \) versus \( H_1 : \pi_\tau \neq \pi_\eta \), where \( \pi \) is the density of a conditioned Galton–Watson tree, the test given by \( \phi(k, \alpha) \) is such that

\[
\lim_{|k| \to \infty} \mathbb{E}_{\pi_\eta} \phi(k, \alpha) = \alpha.
\]

2. Choose \( g = (g_1, \ldots, g_q) \) as the subset of leaves from \( \eta_m \) in a similar manner with cardinality \( |g_j| \). Then, the critical function

\[
\psi(k, g, \alpha) = \begin{cases} 
1 & \text{if } \frac{\hat{\sigma}_r^2}{\hat{\sigma}_n^2} \sum_{j=1}^q s_j > \left( \frac{\sum_{i=1}^{|k|} |k_i|}{\sum_{i=1}^{|k|} |g_j|} \right) F_{1-\alpha, 2\sum_{i=1}^{|k|} |k_i|, 2\sum_{j=1}^q |g_j|}; \\
0 & \text{otherwise},
\end{cases}
\]

where \( r_j \) are the total path lengths of \( L(\eta_{m_1}, g_j) \), defined, assuming no loss of generality that the numerator exceeds the denominator, for testing \( H_0 : \pi_\tau = \pi_\eta \) against \( H_1 : \pi_\tau \neq \pi_\eta \). The critical function \( \psi \) is such that \( \min(n_i, m_j) \to \infty \) for every \( i, j \), \( \mathbb{E}_{\pi_\eta} \psi(k, g, \alpha) \to \alpha \).

5.1. Consistent Estimation of Offspring Variance

For binary trees, under the Poisson process model, the parameter \( \theta \) is estimated quite easily using the principle of maximum likelihood (using the density in (1)), since the size of the tree remains fixed. Under the conditioned Galton–Watson tree model, for a tree \( \tau_n \), the density (2) is the asymptotic density of an L-tree with \( k \) leaves, as \( n \to \infty \). It is difficult to examine the behavior of an estimator of \( \hat{\sigma}^2 \), which is intricately dependent on \( k \). It would be desirable to construct an estimator that does not depend on the number of leaves chosen to construct the L-tree. This can be achieved through the characterization of the CRT using a mapping of trees to a function space, known as Dyck paths.

Any rooted ordered tree of \( n \) vertices can be uniquely coded by a traversal of the tree. When the traversal is a depth-first walk, one can construct a function, referred to as a Dyck path, which is bijective to the tree in the following manner. Imagine the motion of a particle that starts at time \( t = 0 \) from the root of the tree and then explores the tree from the left to the right, moving continuously along the edges at unit speed until all the edges have been explored and the particle has come back to the root. Each edge will be crossed twice in this evolution, hence the total time needed to explore the tree is \( 2n \), where \( n \) is the total path length, or sum of the branch lengths. The walk can be represented as the value \( H_n(s) \) of a continuous function \( H_n : [0, 2n] \to \mathbb{R}_{>0} \) at time \( s \in [0, 2n] \) such that \( H_n(s) = d(\text{root}, v) \) where \( v \) is the vertex obtained during the walk such that the sum of the edges traversed until \( v \) is \( s \), and \( d(\text{u}, v) \) is the length of the unique path from vertex \( u \) to vertex \( v \). Figure 4, taken from Pitman (2006), offers a more intuitive description. The map from \( \tau_n \) to its Dyck path is a bijection. The Dyck path approach to analysis of tree-structured data was adopted by Shen et al. (2014). Our interest in the Dyck path representation is captured in the following result for conditioned Galton–Watson trees on \( n \) vertices; thus branch lengths are unit length with corresponding Dyck path \( H_n : [0, 2n] \to \mathbb{R}_{>0} \).

Theorem 3. (Aldous 1993) Let \( \tau_n \) be a conditioned Galton–Watson tree conditioned with an offspring distribution with mean 1 and variance \( \sigma^2 < \infty \). Let \( H_n(k), 0 \leq k \leq 2n \) be the Dyck path associated with \( \tau_n \). Then, as \( n \to \infty \),

\[
\frac{1}{\sqrt{n}} H_n(|2nt|), 0 \leq t \leq 1 \Rightarrow \left\{ \frac{B^{ce}}{\sigma} : 0 \leq t \leq 1 \right\},
\]

where \( B^{ce} \) is the standard Brownian excursion, and \( \Rightarrow \) denotes weak convergence in \( C[0, 1] \).

Figure 4. A tree with root at the bottom and its corresponding Dyck path. The x axis ranges from 0 to twice the sum of lengths of the edges; the Dyck path is constructed by traversing the tree in a depth-first manner at unit speed.
Theorem 3 implies an invariance principle with the Brownian excursion as the weak limit regardless of the offspring distribution as long as it is critical with finite variance. Proposition 1 extends the result to certain sub-and super-critical cases. From this, asymptotic properties of the functionals of conditioned Galton–Watson tree models can be examined. The following Theorem presents a consistent estimator of $\sigma^2$. Recall that $d(u, v)$ represented the length of the unique path connecting vertices $u$ and $v$ in a tree.

**Theorem 4.** Let $\tau_n = (\tau_{n_1}, \ldots, \tau_{n_p})$ be a random sample of conditioned Galton–Watson trees.

1. On each $\tau_n$, suppose $v_i$ is a vertex chosen according to a uniform distribution on $V(\tau_n)$. Then, the random variable $n_i^{-1/2} d(\text{root}, v_i) \xrightarrow{d} W$, as $n_i \to \infty$, where $W$ is a Rayleigh distributed random variable with scale $1/\sigma$.

2. Let $W_i$ be random variables that denote the normalized distance of a randomly chosen vertex from the root. Then, as $n_i \to \infty$ for each $i$, then $\hat{\sigma}_i^2 = 2p(\sum_{i=1}^p W_i^2)^{-1}$ is a consistent estimator of $\sigma^2$.

### 6. Numerical Illustrations

#### 6.1. Simulations

We use an efficient method to simulate conditioned Galton–Watson trees by employing the algorithm provided by Devroye (2012) with a linear expected time. This enables us to simulate a large number of conditioned Galton–Watson trees, each of which contains a large number of vertices, and each tree is generated in expected linear time. Details of the simulation procedures can be found in Section 3 of supplementary material. Table 1 reports the performances of tests in Theorem 2 and a competing permutation test with the same test statistic. Rejection rates were computed by averaging over multiple permutations of the chosen leaves. Geo(0.5) denotes a Geometric distribution with probability 0.5; Bin(2,0.35) denotes subcritical conditioned Galton–Watson tree from a Binomial distribution with 2 trials and success probability 0.35; GW-Bin(2,0.5) denotes unconditioned Galton–Watson trees with Bin(2,0.5) distribution; Phylo.bd and Phylo.coal correspond to phylogenetic trees based on birth–death processes on fixed taxa with specification rate 2 and Kingman’s coalescent process (see Section 1 of supplementary material for definition), respectively. The poor power against trees generated from a coalescent process is due to their connection with the CRT (Haulk 2011), and this will be exploited in the data application. In Section 4 of supplementary material, we provide detailed results from examinations of the asymptotic behavior of test statistics, estimator of the offspring variance, and performances of the tests on binary trees from Poisson model. The results largely corroborate the theoretical findings.

#### 6.2. Data Application: Detection of Tumor Heterogeneity Using Magnetic Resonance Images

We illustrate the utility of the proposed tests through a novel approach to detecting tumor heterogeneity in brain cancer by constructing binary trees obtained from Magnetic Resonance (MR) images.

Data structure, preprocessing, and key scientific question. In this study, we used presurgical, T1-weighted post-contrast and T2-weighted/FLAIR images of 82 patients (26 women and 56 men) with histologically confirmed glioblastoma multiforme (GBM)—an aggressive form of brain cancer—from The Cancer Genome Atlas (TCGA) database. The images were downloaded from The Cancer Imaging Archive at https://www.cancerimagingarchive.net/ and are publicly available. We preprocessed the MR images and obtained three-dimensional (3D) tumor volumes. Specifically, the images were registered spatially, followed by intensity bias correction using Medical Image Processing Analysis and Visualization software (MIPAV v6.0.0). The tumor region was segmented semiautomatically in 3D using the Medical Image Interaction Toolkit (MITK.org). Tumor regions were defined as a combination of the T1-contrast enhancing region as well as the FLAIR hyperintense region, or specifically, the regions common to the T1-enhancing signal and FLAIR hyperintensity; this captures a combination of the tumor’s enhancing component as well as the infiltrative edema component. The T1-weighted post-contrast and FLAIR tumor regions were delineated separately by a qualified neurosurgeon using the Medical Image Interaction Toolkit. The in-plane resolution of the image was $1\text{mm} \times 1\text{mm}$. Our analysis was based on T2-weighted intensities from only the segmented regions; see image in Figure 5 where the segmented region is outlined in black.

Tumor heterogeneity expressed though pixel intensities indicates a latent ordering of groups of pixels with similar intensities that represent similar etiologies. The problem of interest is to appropriately characterize tumor heterogeneity in the brain.

| Distribution | One-sample | Two-sample |
|--------------|------------|------------|
|              | $N = 10$   | $N = 100$  | $N = 1000$ | $N = 10$   | $N = 100$  | $N = 1000$ |
|              | $\chi^2$ perm | $\chi^2$ perm | $\chi^2$ perm | $F$ perm | $F$ perm | $F$ perm |
| Geo(0.5)     | 0.09       | 0.15       | 0.05       | 0.08       | 0.03       | 0.09       | 0.11       | 0.10       | 0.13       | 0.08       | 0.04       | 0.02       |
| Bin(2,0.5)   | 0.13       | 0.08       | 0.04       | 0.03       | 0.01       | 0.01       | 0.08       | 0.14       | 0.08       | 0.09       | 0.03       | 0.06       |
| Bin(2,0.35)  | 0.10       | 0.16       | 0.12       | 0.07       | 0.06       | 0.08       | 0.21       | 0.14       | 0.13       | 0.10       | 0.04       | 0.01       |
| GW-Bin(2,0.5)| 0.78       | 0.91       | 0.91       | 0.97       | 0.99       | 1.00       | 0.82       | 0.87       | 0.88       | 0.93       | 0.95       | 1.00       |
| Phylo.bd     | 0.81       | 0.83       | 0.89       | 0.91       | 0.98       | 0.94       | 0.92       | 0.83       | 0.96       | 0.97       | 0.99       | 1.00       |
| Phylo.coal   | 0.26       | 0.37       | 0.14       | 0.21       | 0.11       | 0.08       | 0.37       | 0.28       | 0.23       | 0.18       | 0.19       | 0.17       |
specifically, we consider detection of tumor heterogeneity between patients with GBM who have long survival times versus those who have short survival times.

**Methods.** Current approaches to this problem are based on simple summaries of the entire image such as skewness or kurtosis of the probability density of the intensities, which fail to take into account the structural complexities of the pixel-level intensities, while also neglecting spatial information; see Just (2014) for a detailed review. In contrast, our approach is based on exploring the clustering properties inherent in the density, extracted through the number and size of modes. Our approach, illustrated in Figure 5, is as follows.

1. For each patient, from a single axial slice, obtain the pixels from the segmented tumor.
2. Construct a binary tree/dendrogram by implementing an agglomerative hierarchical clustering algorithm on the pixels.
3. Randomly choose a subset of the leaves, construct the corresponding $L$-tree, and compute its total path length.

On the tree, the individual intensities appear as leaves, clusters of similarly valued intensities are the internal vertices, and branch lengths represent distances between the clusters. Such a representation expresses heterogeneity in the tumor image through the clustering of intensity values. A rationale for using vectorized pixel intensities, disregarding the spatial information, is that when the chief objective is to characterize or classify tumors, the location of the heterogenous regions on the image is of little relevance: the classifying procedure needs to only detect the heterogeneity in the entire image in relation to another image. Our view of heterogeneity is closely linked to the density-cluster tree (Hartigan 1975; Azzalini and Torelli 2007) from density-based clustering, wherein the focus is on the multiple modes within the probability density of the intensity values: each branch represents the high-density clusters within a single mode of the density, that is, groupings of pixels that are not separated easily. Branch points of the tree represent values at which a new mode of the density, or a new cluster, emerges. Hence, on the scale of the data at multiple resolutions, the branch lengths are indicative of how long a particular mode within the density lasts before being broken up; the sum of the branch lengths is the corresponding cumulative measure on the scale of the data, capturing heterogeneity through the number and size of modes in the density of the pixel intensities.

Subsequently, the test in *Theorem 2*, with the test statistic that relates the pixel-cluster distances to the branch lengths of an $L$-tree, can be used to detect group differences between patients with GBM who experience long survival times ($> 12$ months) and those who experience short survival times ($\leq 12$ months). As an illustration, consider the images of the patients who, respectively, correspond to the short (0.723 months) and long (57.8 months) survival times, as shown in the top panel of Figure 6. Their respective trees have total path length 3.69 and 2.13, respectively. Intuitively, the MR image for the patient with the shorter survival time, with the tumor appearing to be in an advanced stage in the image on the right in Figure 6, should have pixel intensities with richer and varied clustering tendencies, which are evident in the form of more branches with smaller branch lengths. From the analysis of the rejection rates of the tests in Table 1, we note that the two proposed test techniques do not have good power against ultrametric trees such as coalescent trees. These trees are characterized as possessing leaves that are equidistant from the root. Dendrograms arising from agglomerative hierarchical clustering are ultrametric, and a surprising connection exists, in the asymptotic regime, between the dendrograms and the $L$-trees that characterize the CRT; see Haulk (2011) for details.

### 6.3. Results

We use agglomerative hierarchical clustering in MATLAB with Single linkage function to obtain dendrograms of the image intensities. Carlsson and Memoli (2010) noted that the single linkage function is the most stable choice for hierarchical clustering in the sense that the distance is small between the original tree and the tree obtained after a small perturbation of the input data. Similar constancy results under Single linkage function are available for the density-cluster (Hartigan 1975). However, the large sizes of the images considered in our setting led to trees that did not appear significantly different for different choices of the linkage functions; this was corroborated by the results of the $F$ and permutation tests, which were unchanged in relation to different choices of linkage function. To use the test prescribed in *Theorem 2*, we estimate the variance $\sigma^2$ with $\hat{\sigma}^2$ from *Theorem 5*. The small sample sizes—31 and 42 patients in groups who experienced short and long survival times, respectively—do not guarantee a consistent estimate of $\sigma^2$; however, this does not appear to affect the efficacy of the tests.

Using the total path length, scaled by the estimate of the variance, of an $L$-tree constructed by choosing various percentages of leaves at random as the test statistic, the $F$ test from *Theorem 2* rejected the null hypothesis of equality of distribution, at 1% significance level, between the groups with long versus short survival times, and failed to reject $H_0$ when the labels were randomly permuted. Within each group, when two subgroups were randomly chosen and compared for group differences, the test failed reject the null hypothesis of equality of distribution at 1% significance level. Details are reported in Table 2 for the comparison of groups with long versus short survival times, and within groups with survival times. Note that the choice of the
proportion of leaves used to construct the $L$-trees does not seem to affect the conclusions of the test. This is particularly encouraging since computation of $L$-trees from observed trees can be computationally expensive. The permutation test provided identical results for the same test statistic at 2000, 5000, and 10,000 permutations. Importantly, the conclusions of the tests were unaltered when the linkage function was changed. Scientifically, this points to the existence of significant GBM tumor heterogeneity among patients with different durations of survival.

7. Discussion

Aldous’ articles on the CRT and variants (see, in addition, Aldous 1991b, 1994) provide useful distributional results and connections to common stochastic processes, which in principle can be harnessed in modeling tree functionals (e.g., total height, Wiener index) by corresponding functionals of the limit stochastic processes. However, the absence of information on the rates of convergence of conditioned Galton–Watson trees to the CRT prevents us from obtaining a clear idea as to the size of the trees that assure the validity of our results.

Apart from the need to retain the projective property, another reason for using the multiplicative model $\lambda(t) = \theta t$ for the intensity function is to model the dependence of the CRT on the variance $\sigma^2$ of the offspring distribution of the conditioned Galton–Watson trees. The parameter $\sigma^2$, as can be seen from the construction and Lemma 1, appears naturally in the limiting density of the $L$-trees as a scale parameter; this is very useful for its consistent estimation using a maximum likelihood estimate, described in Theorem 5.

An approach for testing not explored in this article is based on using the normalized Dyck paths to define likelihoods. Recall that $\frac{2}{\sigma} B^{\text{ex}}$ is the limit of normalized Dyck paths that uniquely code conditioned Galton–Watson trees. The CRT is connected to $\frac{2}{\sigma} B^{\text{ex}}$ in the following manner: Consider uniform order statistics $U_{1:k} < \cdots < U_{k:k}$ and set $V_i = \min U_{1:k+2-i} \leq U_{k+2-i}$ and $\frac{2}{\sigma} B^{\text{ex}}(t)$. Then the $(2n + 2)$-dimensional vector taking values in $\mathbb{R}^{2n+2}$ as $X_n = (\frac{2}{\sigma} B^{\text{ex}}(U_{1:n}), \frac{2}{\sigma} B^{\text{ex}}(V_{1:n}))$ has the

Table 2. At different values for the percentage of leaves chosen to construct the $L$-trees of the dendrograms from the images, the results of the $F$ test from Theorem 2 to test the null hypothesis of equality of distribution of the two groups.

| % of leaves | Null dist. | $F$-statistic | Decision at 1% |
|-------------|------------|---------------|---------------|
| 10          | $F_{0.016, 0.075}$ | 0.780         | Reject        |
| 20          | $F_{0.0330, 0.0255}$ | 0.775         | Reject        |
| 30          | $F_{0.0713, 0.0328}$ | 0.774         | Reject        |
| 40          | $F_{0.1646, 0.0470}$ | 0.772         | Reject        |

| % of leaves | Null dist. | $F$-statistic | Decision at 1% |
|-------------|------------|---------------|---------------|
| 10          | $F_{0.0494, 0.1814}$ | 0.990         | Do not reject |
| 20          | $F_{0.0721, 0.2015}$ | 0.985         | Do not reject |
| 30          | $F_{0.0959, 0.2785}$ | 0.997         | Do not reject |
| 40          | $F_{0.1832, 0.3020}$ | 1.001         | Do not reject |
same distribution as $\frac{\sigma (\Omega _{V} > V_i)}{\Gamma (n + 1, \frac{1}{\gamma_i})}$ \(i = 1, \ldots , n\), where \(\gamma_i \geq 1\), is a Gamma random variable with shape \(n + 1\) and scale 1 (Pitman 1999). Although in principle, it would be reasonable to define a parametric class, the distribution of \(X_i\) is not easy to compute.

The data application considered in this article represents, to our knowledge, the first attempt at characterizing tumor heterogeneity from images that use tree representations. However, important extensions such as establishing correspondence between images, incorporating spatial information of the pixels, using covariate information for each patient, and developing methods for images obtained from a longitudinal study of the patients, are part of our current work. Much remains to be done in this direction.

**Supplementary Material**

Supplementary material available online includes technical details on the properties of the tree models, notes on generating conditioned Galton–Watson trees, detailed results from simulations, and proofs of all results.

**Acknowledgments**

The authors thank Joonsang Lee, Juan Martinez, Shivali Narang, and Ganesh Rao for their assistance with MR images in the data application, and the referees for helpful comments.

**References**

Adams, R., Graharamani, Z., and Jordan, M.I. (2010), “Tree-Structured Stick Breaking for Hierarchical Data,” in NIPS, pp. 19–27. [1733]

Aho, A., Sagiv, Y., Szymanski, T. G., and Ullman, J. D. (1981), “Inferring a Tree from Lowest Common Ancestors with an Application to the Optimization of Relational Expressions,” SIAM Journal on Computing, 10, 405–421. [1738]

Aldous, D. (1991a), "The Continuum Random Tree I," Annals of Probability, 19, 1–28. [1734,1735]

——— (1991b), "Asymptotic Fringe Distributions for General Families of Random Trees," Annals of Applied Probability, 1, 228–266. [1742]

——— (1993), "The Continuum Random Tree III," Annals of Probability, 21, 248–289. [1734,1738,1739]

——— (1994), "Recursive Self-similarity for Random Trees, Random Trangulations and Brownian Excursion," Annals of Probability, 22, 527–545. [1742]

——— (1996), "Probability Distributions on Cladograms," in Random Discrete Structures (The IMA Volumes in Mathematics and its Applications, Vol. 76), eds. D. Aldous and R. Pemantle, New York: Springer, pp. 1–18. [1733]

Aldous, D., and Pitman, J. (2000), "Inhomogeneous Continuum Random Trees and the Entrance Boundary of the Additive Coalescent," Probability Theory and Related Fields, 118, 455–482. [1735]

Aydin, B., Pataki, G., Wang, H., Bullitt, E., and Marron, J. S. (2009), "A Principal Component Analysis for Trees," Annals of Applied Statistics, 3, 1597–1615. [1733]

Ayldward, S., and Bullitt, E. (2002), "Initialization, Noise, Singularities and Scale in Height Ridge Traversal for Tubular Object Centerline Extraction," IEEE Transactions of Medical Imaging, 21, 61–75. [1733]

Azzalini, A., and Torelli, N. (2007), "Clustering via Nonparametric Density Estimation," Statistics and Computing, 17, 71–80. [1741]

Carlsson, G., and Mémoli, F. (2010), "Characterization, Stability and Convergence of Hierarchical Clustering Methods," Journal of Machine Learning Research, 11, 1425–1470. [1741]

Devroye, L. (2012), "Simulating Size-Constrained Galton–Watson Trees," SIAM Journal of Computing, 41, 1–11. [1734,1740]

Fu, Y. X., and Li, W. H. (1993), "Statistical Tests of Neutrality of Mutations," Genetics, 133, 693–709. [1736]

Gronau, L., and Moran, S. (2007), "Neighbor Joining Algorithms for Inferring Phylogenies via LCA Distances," Journal of Computational Biology, 14, 1–15. [1738]

Gurtrop, P. (1991), Statistical Inference for Branching Processes (Wiley Series in Probability and Statistics), New York: Wiley. [1738]

Hartigan, J. (1975), Clustering Algorithms, New York: Wiley. [1741]

Haulk, C. (2011), "Aspects of Exchangeable Partitions and Trees," Ph.D. dissertation, University of California, Berkeley. [1740,1741]

Heller, K. A., and Graharamani, Z. (2005), "Bayesian Hierarchical Clustering," in ICMIL, pp. 297–304. [1733]

Hu, Y., Boyd-Garber, J. L., Daume, H., and Ying, Z. I. (2013), "Binary to Bushy: Bayesian Hierarchical Clustering with the Beta Coalescent," in NIPS, pp. 1079–1087. [1733]

Janson, S. (2012), "Simply Generated Trees, Conditioned Galton–Watson Trees, Random allocations and Condensation," Probability Surveys, 9, 103–252. [1737]

Just, N. (2014), "Improving Tumour Heterogeneity MRI Assessment with Histograms," British Journal of Cancer, 111, 2205–2213. [1734,1741]

Kolbe, D. L., and Eddy, S. R. (2009), "Local RNA Structure Alignment with Incomplete Sequence," Bioinformatics, 25, 1236–1243. [1733]

Lanfear, R. (2010), "The Local-Clock Permutation Test: A Simple Test to Compare Rates of Molecular Evolution on Phylogenetic Trees," Evolution, 65, 606–611. [1736]

Le, S. Y., Nussinov, R., and Maizel, J. V. (1989), "Tree Graphs of RNA Secondary Structures and their Comparisons," Computers and Biomedical Research, 22, 461–473. [1733]

Neal, R. (2001), "Defining Priors for Distributions Using Dirichlet Diffusion Trees," Technical Report No. 0104, Department of Statistics, University of Toronto. [1733]

Orbanz, P., and Roy, D. M. (2015), "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures," IEEE Transactions Pattern Analysis and Machine Intelligence, 37, 437–461. [1734]

Pitman, J. (1999), "Brownian Motion, Bridge, Excursion, and Meander Characterized by Sampling at Independent Uniform Times," Electronic Journal of Probability, 4, 1–33. [1743]

——— (2006), Combinatorial Stochastic Processes, New York: Springer-Verlag. [1739]

Schmitt, W. R., and Waterman, M. S. (1994), "Linear Trees and RNA Secondary Structure," Discrete Applied Mathematics, 51, 317–323. [1733]

Shen, D., Shen, H., Bhamidi, S., Munoz-Maldonado, Y., Kim, Y., and Marron, J. S. (2014), "Functional Data Analysis of Tree Structured Objects," Journal of Computational and Graphical Statistics, 23, 418–438. [1733,1739]

Steele, J. M. (1987), "Gibbs’ Measures on Combinatorial Objects and The Central Limit Theorem For an Exponential Family of Random Trees," Probability in the Engineering and Informational Sciences, 1, 47–59. [1733]

Tatikonda, S., and Parthasarathy, S. (2010), "Hashing Tree-Structured Data: Methods and Applications," in IEEE 26th International Conference on Data Engineering, Long Beach, pp. 429–440. [1733]

Teh, Y. W., Daume, H., and Roy, D. M. (2008), "Bayesian Agglomerative Clustering with Coalescents," in NIPS, pp. 1473–1480. [1733]

Wang, Y., Marron, J. S., Aydin, B., Ladha, A., Bullitt, E., and Wang, H. (2012), "A Nonparametric Regression Model With Tree-Structured Response," Journal of the American Statistical Association, 107, 1272–1285. [1733]

Yang, R., Kalnis, P., and Tung, A. K. H. (2005), "Similarity Evaluation on Tree-Structured Data," in SIGMOD, Baltimore, pp. 754–765. [1733]