INFERENCE FOR LARGE TREE-STRUCTURED DATA

KARTHIK BHARATH, PRABHANJAN KAMBADUR, DIPAK. K. DEY,
AND VEERABHADRAN BALADANDAYUTHAPANI

Abstract. We develop a parametric inferential framework for fully observed tree-structured data containing a large number of vertices using the distributional properties of the Continuum Random Tree (CRT) introduced by Aldous [1993]. Under a hypothesis testing context, we develop tests based on two equivalent characterizations of the CRT. In both cases, the Rayleigh distribution with a scale parameter belonging to the exponential family arises as a limiting distribution and consequently, the test statistics enjoy optimal statistical properties. We examine properties of the parametric families of distribution induced through the two approaches and perform detailed simulations evaluating the performance of the proposed tests. A secondary contribution is in the efficient simulation of large trees of a particular class used in this article, which is of independent interest.

Keywords: Conditioned Galton-Watson trees; Dyck path; Brownian excursion; Distinguishability of parametric families; Rayleigh distribution.

1. Introduction

The statistical analysis of tree-structured objects has received appreciable attention in recent years owing to the emergence of datasets wherein the underlying quantities of interest allow for tree-like representations. However, some central challenges have stymied the systematic development of tools for statistical inference: The non-Euclidean nature of the underlying space offers considerable challenges while developing probability models for fully observed trees; tree-structured data rarely contain the same number of vertices leading to issues in comparing trees of differing sizes; generating trees from a probability model for simulation purposes is not straightforward. Motivated by these issues, our approach in this article is based on the abstract notion of a Continuum Random Tree (CRT) from Aldous [1991a] and Aldous [1993] which arises as a continuous limit as the number of vertices grows without bound for a large class of random trees. Our objective is to investigate the utility in employing the CRT in developing asymptotic inferential tools on fully observed tree-structured data containing a large number of vertices. To this end, we confine our attention to finite, rooted trees: trees with a distinct vertex, referred to as the root, containing a finite number of vertices. These trees can be labelled or unlabelled, ordered or unordered, have positive branch lengths, unequal number of vertices and are referred to in combinatorial literature as simply generated trees, or equivalently (leaving aside some extreme cases) within the probability community as *Conditioned Galton-Watson trees* (CGW) obtained as the family tree of a Galton-Watson process conditioned on a given total number of vertices.

The CRT is an archetypal example of the weak convergence paradigm proposed by Aldous [1994b] based on an isometric ℓ_1 embedding of a tree with n vertices such that the graph distance between vertices are preserved. Such an embedding makes possible comparing trees for different n. Equipped with a probability measure on the vertices, a continuous weak limit as n tends to infinity is determined which then offers insight into structural and numerical properties of the original object which would otherwise be hard to ascertain. We propose to embed the statistical problem of interest on large trees in the continuous environment offered by the CRT and develop tools based on the two equivalent characterizations of the CRT leading to the definition of simple parametric probability models on trees.

The distribution of the CRT can be specified in four equivalent ways of which the following two will be of primary concern in this article (see Aldous [1991a] and Aldous [1993]): as a weak limit of family trees of critical Conditioned Galton-Watson processes; distributions of spanning subtrees which are analogous to finite dimensional distributions of a stochastic process. The first specification

is characterized by a continuous function on the vertex set, referred to as a *Dyck path*, obtained from the depth-first walk at unit speed on the tree which converges to a Brownian excursion. The second specification is based on the limiting distribution of family of subtrees referred to as Least Common Ancestor trees. These considerations on CGW tree-models for tree-structured data lead us to the primary focus of this article:

(i) Developing parametric families using the limiting distribution of a special class of subtrees of the CRT for hypothesis testing;
(ii) Using the Dyck path representation to define parametric families for hypothesis testing based on the Brownian excursion approximation;
(iii) Efficient simulation of CGW trees.

The parameter of interest is the variance of the offspring distribution generating the CGW trees. The parametric setup allows us to rigorously examine the theoretical properties of the family of distributions on trees and the proposed tests using conventional ideas such as Neyman-Pearson tests, UMP tests, exponential family etc. It will be seen that the induced family in (i) and (ii) is the Rayleigh family parameterized by a scale factor given by the variance of the offspring distribution of the CGW tree; the Rayleigh with a scale parameter is a member of the one-parameter exponential family enjoying various desirable statistical properties. While we restrict our attention to hypothesis tests, this article ought to be viewed a first step of a systematic program in developing statistical procedures on large trees using the CRT.

Another contribution of this article is in the efficient simulation of critical CGW trees which encompass a broad class of random trees including Catalan trees, Cayley trees, Binary trees, uniform random trees etc. We use the efficient algorithm for generating CGW trees proposed by Devroye [2012] for our simulations which has a universal linear expected run-time. As a consequence, we are able to generate a large number of CGW trees with thousands of vertices fairly quickly under a parallel distributed computing setup. The generation of CGW trees is, in general, not a trivial matter, and with our software, we provide means to simulate a broad class of trees from a critical Galton-Watson process which can be used as "ground-truth" for simulation experiments involving tree-structured data.

Methodology on tree-structured data has hitherto been characterized by nonparametric or algorithmic approaches (Wang and Marron [2007], Busch et al. [2009], Aydin et al. [2011], Shen et al. [2013], Wang et al. [2012], Aydin et al. [2009], Rosa et al. [2012] etc.). In relation to our approach, Shen et al. [2013] used FDA methods on Dyck paths using an aligning mechanism which led to the creation of some spurious tree-structures—for eg. negative branch lengths—while exploring modes of variation in the trees in a regression problem. A parametric route was taken by Steele [1987] wherein a one-parameter exponential family of distributions on labeled trees was proposed with the natural parameter representing the expected number of leaves (terminal vertices) in the trees. In similar vein, but with phylogenetics in mind, Aldous [1996] proposed a beta-splitting parametric model for cladograms; he noted the utility of a simple parametric model for phylogenetic tree construction. Motivated by the parametric approaches, we will consider a few one parameter family of distributions for testing statistical hypotheses on CGW trees considered in this article which are induced from the distributional properties of the CRT and Brownian excursion. CGW trees can be used as models for tree-structured data frequently encountered in many scientific settings. For instance, plane-rooted trees are considered in Busch et al. [2009] under the context of a protein classification problems; Shen et al. [2013], Aydin et al. [2009], Aylward and Bullitt [2002] modeled brain artery data as three dimensional trees embedded on the plane; also see Yang et al. [2005] and Tatikonda and Parthasarathy [2010] for datasets in the context of XML documents and secondary structure of RNA.

In section 2 we review the key ingredients of the CRT including CGW trees, Dyck paths and Least Common Ancestors trees. In section 3, we propose a parametric family induced by the spanning subtrees of the CRT, examine its properties, and propose one-sample and two-sample tests for distributions on trees. In section 4, we propose a parametric family based on the random projection of Dyck path representation and propose one and two-sample tests for distributions. In section 5, we generate CGW trees using the algorithm proposed by Devroye [2012] and verify the validity of the theoretical

results and performance of proposed tests. Section 6 discusses some salient aspects of our approach and comments on possible extensions. Proofs of results and details of simulations are relegated to the Appendix in section 7.

2. Preliminaries

2.1. **Representation of trees.** Consider a finite rooted tree τ_n as set of vertices $\mathcal{V}(\tau_n) = (\text{root}, v_1, \ldots, v_{n-1})$ and a set of edges $\mathcal{E}(\tau_n) = (e_1, \ldots, e_{n-1})$, represented as a point

$$\tau_n = \Big(\mathcal{V}(\tau_n), \mathcal{E}(\tau_n)\Big)$$

in the space $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$ where $\mathcal{T}_n$ is the set of all finite trees on n vertices. One way to compare different size trees is to embed $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$ as an element of the linear space ℓ_1, the Banach space of infinite sequences $x = (x_1, x_2, \ldots)$ such that $||x|| = \sum_i |x_i| < \infty$. Such an embedding makes possible comparison and scaling of trees consisting of different number of vertices in a natural way. Formally, suppose $d(v_1, v_2)$ is the distance between two vertices defined as the sum of edge lengths along the unique path from v_1 to v_2. The embedding of $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$ as a subset of ℓ_1 is the determination of points w_i for $1 \le i \le n$ in ℓ_1 such that $||w_i - w_j|| = d(v_i, v_j)$ for all $1 \le i, j \le n$. Then, the subset of ℓ_1 containing $w_1, \ldots, w_n$ and the connecting paths is referred to as the set representation of τ_n. In this article, however, we shall not directly employ the set representation; the use of Aldous' results obtained through the set representation, implies its indirect use. The formal definition the CRT is based on the set representation S and a probability measure μ on ℓ_1 connected to S through two technical conditions; see p. 253 in Aldous [1993]. The pair (S, μ) is then the CRT. We shall only be concerned with the CRT through its distributional properties.

2.2. **Conditioned Galton-Watson trees and random sampling.** Given a probability distribution $(\pi_k, k = 0, 1, \ldots)$ on the non-negative integers, or equivalently a random variable ξ with distribution π_k, we construct a *Galton-Watson* tree τ recursively starting with root and giving each node a number of children that is an independent copy of ξ; $P(\xi = k) = \pi_k$ for $k = 0, 1, \ldots$ is referred to as the offspring distribution and the out-degrees of the vertices are i.i.d. copies of ξ from π_k). As a consequence, ξ induces a unique distribution on τ as

$$P(\tau = t) = \prod_{v \in \mathcal{V}(t)} \pi_{o(v,t)},$$

where $o(v, t)$ is the out-degree or the number of children of vertex v in tree t.

If one wishes to model a set of tree-structured data using Galton-Watson trees, two issues arise at this point: Galton-Watson trees, with positive probability, can be infinite, whereas observed trees in practice are always finite; secondly, how could we ensure that the observed trees have been collected through random sampling? We shall address these issues by considering *Conditioned Galton-Watson* (CGW) trees; these are Galton-Watson trees conditioned on total progeny. That is, the distribution of a CGW tree τ_n conditioned to have n vertices is

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

Importantly, it is known, that for a fixed offspring distribution π_k, the corresponding CGW tree can be viewed a being picked according to a uniform distribution on certain types of tree with n vertices. For example, if we wish to choose a strictly binary tree (0 or two children only) with n vertices according to a uniform distribution on the space of n-vertex binary trees, then, we can equivalently construct a CGW tree with a offspring distribution 0.5 each for 0 and 2 children. We enumerate a few useful classes of trees for modeling purposes:

(i) *Ordered trees with unrestricted degree*: CGW trees with offspring distribution given by a Geometric distribution with success probability 1/2;
(ii) *Binary trees*: CGW trees with vertices containing 0,1 or 2 children with a Binomial distribution with 2 trials and success probability 1/2;

(iii) *Strict binary trees which are ordered*: CGW trees with vertices containing either 0 or 2 children with equal probability 1/2;
(iv) *Unary-binary trees which are ordered*: CGW trees with vertices containing 0, 1 or 2 children each with probability 1/3;
(v) *Unary-binary trees which are unordered and unlabelled*: CGW trees with vertices containing 0, 1 or 2 with probabilities $\pi_0 = \frac{1}{2+\sqrt{2}}$, $\pi_1 = \frac{\sqrt{2}}{2+\sqrt{2}}$ and $\pi_3 = \frac{1}{2+\sqrt{2}}$, respectively.
(vi) *m-ary trees*: CGW trees with vertices containing $0, 1, \ldots, m$ for $m > 3$ children with distribution given by a Binomial with m trials and success probability $1/m$ [1].

Ordered trees imply that they can be embedded on the plane and therefore possess a natural labelling mechanism. From a tree perspective, this implies that there is an order amongst the children at any given vertex. Note that the offspring distributions of the CGW trees considered are with unit mean implying that the Galton-Watson process generating the tree is critical. This is because conditioning on n makes the family of offspring distributions parameterized by a mean parameter identically distributed (see Kennedy [1975]). While it is conceivable that inference on such trees can be performed by a mere counting of the number of observed children at arbitrary, knowledge of distributions of local structural aspects like height, variations in branching structure and also information about branch lengths are not easy to obtain. In this article, we shall consider CGW trees and refer to them simply as trees.

2.3. **Dyck paths.** Any rooted ordered tree τ_n can be uniquely coded by a traversal of the tree; when the traversal is a depth-first walk, one can construct a function which is a bijection to the tree in the following manner: for ease of exposition, assume that the edges or branches of a tree τ_n with n vertices have length 1. For a fixed positive integer n, Dyck paths are lattice excursions of length $2n$, that is sequences $(d_j, 0 \leq j \leq 2n)$ where $d_0 = d_{2n} = 0$ and $d_j > 0$ with $d_{j+1} - d_j \in \{-1, +1\}$ for all $0 \leq j \leq 2n-1$. 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. Since it is clear that each edge will be crossed twice in this evolution, the total time needed to explore the tree is $2n$. For simplicity, suppose all edges are of unit length, the value $H_n(s)$ of a continuous function $H_n : [0, 2n] \to \mathbb{R}_{\geq 0}$ at time $s \in [0, 2n]$ is the distance (on the tree) between the position of the particle at time s and the root; figure 1, taken from Pitman [2006], offers a more intuitive description with edge lengths not all equal to one. Therefore, if τ_n is a tree of size n the sequence $(H_n(0), H_n(1), \cdots, H_n(2n))$ is its Dyck path of length $2n$. The representation of the Dyck path of a tree τ_n in terms of the distance d between its vertices is related to the function H_n as

$$H_n(s) = d(\text{root}, v), \tag{2.1}$$

where v is the vertex obtained during the depth-first walk such that the sum of the edges traversed till v is s. This discussion is formalized by the following proposition whose proof is straightforward.

Proposition 1. *The map $\tau_n \mapsto (H_n(0), H_n(1), \cdots, H_n(2n))$ is a bijection from the set of plane trees with size n to the set of all Dyck paths of length $2n$.*

The key result combining the two ideas is the following result in Aldous [1993]:

Theorem 1. *Let τ_n be a CGW tree conditioned to have n vertices with offspring distribution with mean 1 and variance $\sigma^2 \in (0, \infty)$. Let $H_n(k), 0 \leq k \leq 2n$ be the Dyck path associated with τ_n. Then, as $n \to \infty$ through the possible sizes of the unconditioned Galton-Watson tree,*

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

where B^{ex} is the standard Brownian excursion and $\Rightarrow$ implies weak convergence of processes in $C[0,1]$, the space of continuous functions on $[0,1]$, and $[\cdot]$ stands for the integer function.

[1] There is an identifiability issue for the m-ary trees with $m = 3$ since the variance is 2/3 which is the same as the variance for the unary-binary trees.

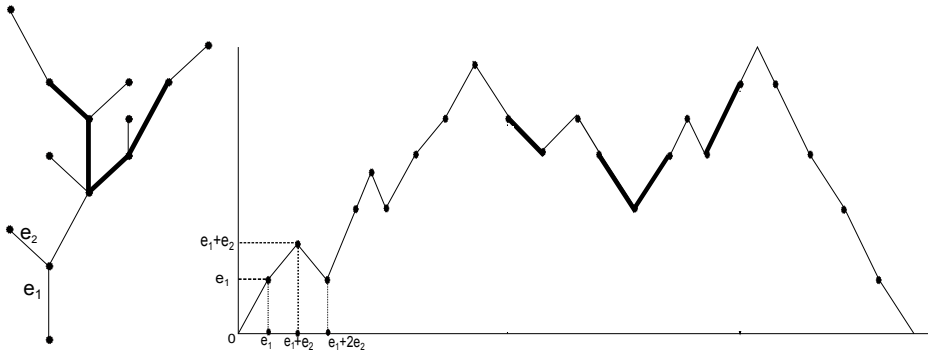

FIGURE 1. A tree with root at the bottom and its corresponding Dyck path. The x axis ranges from 0 to twice the sum of lengths of the edges; the Dyck path is constructed by traversing the tree in a depth-first manner at unit speed.

Roughly, the Brownian excursion in the limit is the 'Dyck path' of the CRT and the distribution of the CRT is specified by a careful construction from the excursion.

2.4. **Least Common Ancestor subtrees.** We define here the class of spanning subtrees which characterize the distribution of the CRT. For a CGW finite tree $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$, define its Least Common Ancestor (LCA) 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 the v_1 and v_2 from the root diverge or branch out. Now, the LCA tree corresponding to the subset B of vertices of τ_n is the tree, denoted as $LCA(\tau_n, B)$, containing 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 idea with $B = \{v_1, v_2, v_3, v_4\}$; the branch points are b_1 and b_2, and in order to preserve the distances from root to the vertices of B, the new edge from the branch points to elements of B are the sum of the edges along the path from the root to elements of B in the original tree. Now, for a tree τ_n randomly reorder

FIGURE 2. A tree on the left and its LCA tree on the right corresponding to vertices $\{v_1, v_2, v_3, v_4\}$. The LCA tree contains the root, the branch points b_1, b_2 and the set of vertices $\{v_1, v_2, v_3, v_4\}$.

the vertex set $\mathcal{V}(\tau_n)$ to obtain $(v_{n,1}, \ldots, v_{n,n})$. For a fixed $k < n$, consider the LCA tree of τ_n defined by $LCA(\tau_n, (v_{n,1}, \ldots, v_{k,n}))$; this is akin to picking $k < n$ vertices according to a uniform distribution on $\mathcal{V}(\tau_n)$. LCA trees have been used in the context of reconstruction of the trees; see Gronau and Moran [2007] for a phylogenetic applications and Aho et al. [1981] for related work in a computational context. Aldous showed that for each k, as $n \to \infty$, the random LCA trees $LCA(\tau_n, (v_{n,1}, \ldots, v_{k,n}))$ converge, as subsets of ℓ_1, to a limit tree $\mathcal{L}(k)$, which is strictly binary (each vertex has either 0 or 2

children) with k leaves or terminal vertices. The key point here is that Aldous proved that the family $(\mathcal{L}(k), k \geq 1)$ is a consistent class of subtrees which characterize the CRT (S, μ); in other words, for modeling purposes one can justifiably view the class $(\mathcal{L}(k), k \geq 1)$ as finite dimensional distributions of the CRT. One can then use the distribution of the limiting class to approximate the distributions of the LCA trees of CGW trees.

2.5. **A parametric family.** Throughout, we shall assume that we have a random sample of fully-observed trees $\tau_{n_i} = (\mathcal{V}(\tau_{n_i}, \mathcal{E}(\tau_{n_i})))$ for $i = 1, \ldots, N$ from some distribution on the product space $\otimes_{i=1}^{N} \mathcal{T}_{n_i} \times \mathbb{R}_{+}^{n_i - 1}$. The six types of CGW trees outlined in section 2.2 represent a very broad class for modeling purposes. The variances σ^2 of the distributions considered are, respectively, $2, \frac{1}{2}, 1, \frac{2}{3}, \frac{2}{2+\sqrt{2}}$ and $\frac{m-1}{m}$ for $m = 4, 5, \ldots$. Since the offspring distributions completely characterize the law on the CGW trees, it is conceivable that a class of probability distributions on the non-negative integers parameterized by their variance parameter σ^2 can be used for statistical purposes. Let $\mathcal{S} = \left\{2, \frac{1}{2}, 1, \frac{2}{3}, \frac{2}{2+\sqrt{2}}, \frac{m-1}{m}; m = 4, 5, \ldots\right\}$. Then, the class

$$\left\{\pi_{k,\sigma^2} : k = 0, 1, 2, \ldots; \sigma^2 \in \mathcal{S}\right\} \tag{2.2}$$

is a one-parameter class of probability models for fully observed finite rooted, ordered trees, with the constraints $\sum_{k=0}^{\infty} k\pi_{k,\sigma^2} = 1$ and $\sum_{k=0}^{\infty} k^2\pi_{k,\sigma^2} < \infty$, and set $\sigma^2 = \sum_{k=0}^{\infty} k^2\pi_{k,\sigma^2} - \left[\sum_{k=0}^{\infty} k\pi_{k,\sigma^2}\right]^2$. The first constraint implies that the Galton-Watson branching process generating the CGW tree is critical. An obvious shortcoming with the class in (2.2) is the absence of any branch-length information in the distribution—the probabilities purely reflect the topological structure or "shapes" of trees.

Since our primary interest is in developing hypothesis tests for parametric families, we recall the definition of distinguishable property of parametric families.

Definition 1. Suppose Θ is an index set and Θ_0 and Θ_1 are disjoint subsets of Θ such that $\Theta_0 \cup \Theta_1 = \Theta$. Denote by H_0 and H_1 the null and the alternative hypothesis that θ is a member of either Θ_0 or Θ_1. Then, the set of probability measures $\left\{P_\theta : \theta \in \Theta\right\}$ is distinguishable if

(i) $P_\theta \neq P_{\theta'}$ for all distinct $\theta, \theta' \in \Theta$;
(ii) There is at least one Borel set A such that $P_\theta(A) \neq P_{\theta'}(A)$ for $\theta \in H_0$ and $\theta' \in H_1$.

This is a crucial requirement while testing with parametric families; it will be shown that the parametric families induced by the LCA-tree based approach and Dyck path approach satisfy the above conditions.

3. Parametric family and test from LCA trees

In this section we consider a parametric family for finite rooted trees which may or may not be ordered; the canonical CGW tree contains n vertices and methodology is developed by constructing a subtree by choosing $k < n$ vertices according to a uniform distribution on the vertex set excluding the root; the root is always included in the subtree as its root. Recall that any tree τ_n is represented as $(\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$ with vertex set $\mathcal{V}(\tau_n) = \{\text{root}, v_1, \ldots, v_{n-1}\}$ and edge set $\mathcal{E}(\tau_n) = \{e_1, \ldots, e_{n-1}\}$; a similar representation holds for any subtree. From section 3, it is known that the family of subtrees $(\mathcal{L}(k), k \geq 1)$, arising as the limit of LCA trees, can be regarded as consistent "finite dimensional projections" of the CRT. For a random CGW tree τ_n with distribution π_{σ^2} for some σ^2 in $\mathcal{S}$, consider its $LCA(\tau_n, v_1, \ldots, v_k)$ defined earlier for $k < n$ vertices chosen randomly from $\mathcal{V}(\tau_n)$. Since $LCA(\tau_n, v_1, \ldots, v_k)$ converges in distribution to $\mathcal{L}(k)$ (see p. 251 in Aldous [1993]), the limit distribution inherits σ^2 too. Lemma 21 in Aldous [1993] provides the limit distribution and Theorem 3 proves the characterization of the CRT by $(\mathcal{L}(k), k \geq 1)$. We combine the two results into a single Lemma for our purposes.

Lemma 1. There exists a consistent family $(\mathcal{L}(k), k \geq 1)$ of strictly binary trees which define the CRT having the density

$$f(l(k)) = \left[\prod_{i=1}^{k-1} \frac{1}{2i-1}\right]^{-1} \frac{1}{2^{k-1}} s e^{\frac{-s^2}{2}},$$

where $s = e_1 + \cdots + e_{k-1}$ and $l(k)$ is a strict binary tree with k vertices.

We propose to use the density above for the LCA trees of large CGW trees τ_n with the dependence on σ^2 introduced in a natural manner: $1/\sigma$ will be the scale parameter; as a consequence, we have a parameterized class of distributions for $LCA(\tau_n, v_1, \ldots, v_k)$, given by

$$\left\{f_{\sigma^2} : \sigma^2 \in \mathcal{S}\right\}, \tag{3.1}$$

where, for the LCA tree with k vertices l_k,

$$f_{\sigma^2}(l_k) = \left[\prod_{i=1}^{k-1} \frac{1}{2i-1}\right]^{-1} \frac{1}{2^{k-1}} \sigma^2 s e^{\frac{-s^2\sigma^2}{2}}, \tag{3.2}$$

where l_k has edge set $\mathcal{E}(l_k) = (e_1, \cdots, e_{k-1})$. The vertex set manifests itself in the factor $2^{-(k-1)}$; see Aldous [1993] for details.

Remark 1. Note that the density in (3.2) uses information about the branch-length aspects of the tree; this is in contrast to the distribution in (2.2). Choosing k vertices from n according to a uniform distribution might appear to be restrictive. This can be relaxed in a simple manner as remarked in p.274 of Aldous [1993]. The word "consistent" in the Lemma refers to two properties: if an edge is removed from $\mathcal{L}(k)$, then the remainder tree is distributed as $\mathcal{L}(k-1)$; second, the labeling of the vertices are exchangeable. Upon ignoring the normalizing factors, when viewed as a density of the random variable representing the sum of the edges, the density in (3.2) is the Rayleigh density.

While class in (3.1) is a family of distributions on the LCA trees and consistent for the CRT, it is not immediately clear if the class can be extended to τ_n for every n; this is especially important while developing tests based on the LCA trees. Specifically, noting the obvious fact that $LCA(\tau_n, \mathcal{V}(\tau_n)) = \tau_n$, it is necessary that $f_{\sigma^2}(\cdot)$ defined on $LCA(\tau_n, B)$ for any $B \subset \mathcal{V}(\tau_n)$ can be extended to τ_n upon inserting vertices from $\mathcal{V}(\tau_n) - B$ to $LCA(\tau_n, B)$, while *retaining the interpretability of* σ^2. To formalize this ideas, recall that $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$ resides in $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$; its LCA corresponding to $B \subset \mathcal{V}(\tau_n)$, where $|B| = k$, lies in $\mathcal{T}_k \times \mathbb{R}_+^{k-1}$. Then, for all k and n with $k < n$, the class in (3.1) defined on $\mathcal{T}_k \times \mathbb{R}_+^{k-1}$ can be extended to $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$, or is *n-extendable*, if $f_{\sigma^2}(\cdot)$ on $LCA(\tau_n, B)$ can be recovered by marginalization over $f_{\sigma^2}(\tau_n)$ on $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$ for every σ^2. Denote by $\mathcal{P}_{\sigma^2}$ the law on CGW trees τ with $\sigma^2 \in \mathcal{S}$ corresponding to the density in (3.1). The following Propositions justifies the use of the class in (3.1) for defining a proper family on τ_n for every n, and its amenability for testing purposes.

Proposition 2. *The class* $\left\{f_{\sigma^2} : \sigma^2 \in \mathcal{S}\right\}$ *on* $\mathcal{T}_k \times \mathbb{R}_+^{k-1}$ *is n-extendable for every n.*

Proposition 3. *The parametric class of probability measures* $\left\{\mathcal{P}_{\sigma^2} : \sigma^2 \in \mathcal{S}\right\}$ *is distinguishable.*

Remark 2. The issue of extendability was considered by Shalizi and Rinaldo [2013] in the context of Exponential Random Graph Models, where they defined a notion of the class of distributions being *projective* for the exponential family of distributions. Without going into details of their work, it suffices here to note that the density in (3.2), when viewed as the density of s, is the Rayleigh distribution with a scale factor σ, which belongs to the one-parameter exponential family. Then, $s = e_1 + \cdots + e_{k-1}$ is the minimal sufficient statistic for σ^2 and is clearly *separable* as defined by Shalizi and Rinaldo [2013]—adding an edge increases the value of the sufficient statistic by an amount equaling the edge length. In fact, the *conditional volume factor* defined by them, in our case, is precisely the length, say e, of the newly added edge when viewed as the Lebesgue measure of the interval $[0, e]$ as a generalization of the definition of conditional volume factor. We remark here that the sufficient statistic s induces the

density on the tree and the density is hence with respect to the Lebesgue measure. This is in contrast to the density in the Exponential Random Graph Models which posses density with respect to the counting measure.

We now put to use the parametric class in (3.1) to the inferential problem of testing hypothesis on sets of random trees. The method involves choosing a subset of vertices (excluding the root) uniformly from the vertex set of each tree and constructing its corresponding LCA tree; this leads to a sample of LCA trees. The LCA trees are assumed to have been generated from the family in (3.1) parameterized by σ^2 as the number of vertices of the original trees approach infinity. The conclusions of the subsequent hypothesis test on the LCA trees are extended on to the fully observed trees using Proposition 2.

Theorem 2. *Suppose we have an independent sample of CGW trees $\tau_{n_i} = (\mathcal{V}(\tau_{n_i}, \mathcal{E}(\tau_{n_i})))$ for $i = 1, \ldots, N$ from a distribution π_{σ^2} on the product space $\otimes_{i=1}^{N} \mathcal{T}_{n_i} \times \mathbb{R}_+^{n_i - 1}$ with $\sigma^2 \in \mathcal{S}$. Let $\mathbf{K} = (K_1, \ldots, K_N)$, where $K_i \subset \mathcal{V}(\tau_{n_i})$ chosen according to a uniform distribution on $\mathcal{V}(\tau_{n_i})$ for each $i = 1, \ldots, N$; let $\#K_i$ denote the cardinality of set K_i and denote by $C_{\alpha,2N}$, the αth percentile of a Chi-square distribution with $2N$ degrees of freedom.*

(1) Given $\mathbf{K}$, define the critical function

$$\phi(\mathbf{K}, N, \alpha, \sigma_0^2) = \begin{cases} 1 & \text{if } \sigma_0^2 \sum_{i=1}^{N} s_i^2 < C_{\alpha,2N} \\ 0 & \text{if } \sigma_0^2 \sum_{i=1}^{N} s_i^2 > C_{\alpha,2N}, \end{cases}$$

where $s_i = e_1 + \cdots + e_{\#K_i - 1}$. Then, conditional on $\mathbf{K}$, for the pair of hypotheses $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 = \sigma_1^2$ where $\sigma_1^2 > \sigma_0^2$, the test given by $\phi(\mathbf{K}, N, \alpha, \sigma_0^2)$ is such that as $n_i \to \infty$ for each $i = 1, \ldots, N$, $E_\pi \phi(\mathbf{K}, N, \alpha, \sigma_0^2) \to \alpha$, and is the most powerful test for the pair of hypotheses.

(2) Given $\mathbf{K}$, the likelihood ratio test for testing $H_0 : \sigma^2 = \sigma_0^2$ vs $H_0 : \sigma^2 \neq \sigma_0^2$ is given by the critical function

$$\psi(\mathbf{K}, N, \alpha, \sigma_0^2) = \begin{cases} 1 & \text{if } \sigma_0^2 \sum_{i=1}^{N} s_i^2 < C_{\frac{\alpha}{2},2N} \quad \text{or} \quad \sigma_0^2 \sum_{i=1}^{N} s_i^2 > C_{1-\frac{\alpha}{2},2N}; \\ 0 & \text{otherwise}, \end{cases}$$

where as $n_i \to \infty$ for each $i = 1, \ldots, N$, $E_\pi \psi(\mathbf{K}, N, \alpha, \sigma_0^2) \to \alpha$ and all other quantities are as in part 1.

Theorem 3. *Suppose we have two independent samples of CGW trees $\tau_{n_i} = (\mathcal{V}(\tau_{n_i}), \mathcal{E}(\tau_{n_i}))$ for $i = 1, \ldots, N_1$, and $\eta_{m_j} = (\mathcal{V}(\eta_{m_j}), \mathcal{E}(\eta_{m_j}))$ for $j = 1, \ldots, N_2$, from distributions $\pi_{\sigma_1^2}$ and $\pi_{\sigma_2^2}$ on the product spaces $\otimes_{i=1}^{N_1} \mathcal{T}_{n_i} \times \mathbb{R}_+^{n_i - 1}$ and $\otimes_{j=1}^{N_2} \mathcal{T}_{m_j} \times \mathbb{R}_+^{m_j - 1}$, respectively, with $\sigma_i^2 \in \mathcal{S}$ for $i = 1, 2$. Let $\mathbf{K} = (K_1, \ldots, K_{N_1})$, where $K_i \subset \mathcal{V}(\tau_{n_i})$ is chosen according to a uniform distribution on $\mathcal{V}(\tau_{n_i})$ for $i = 1, \ldots, N_1$; in similar fashion let $\mathbf{L} = (L_1, \ldots, L_{N_2})$, where $L_j \subset \mathcal{V}(\eta_{m_j})$ is chosen according to a uniform distribution on $\mathcal{V}(\eta_{m_j})$. Given $\mathbf{K}$ and $\mathbf{L}$, as $n_i, m_j \to \infty$, for each i and j, the likelihood ratio test of asymptotic size α, for the pair of hypotheses $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_0 : \sigma_1^2 \neq \sigma_2^2$ is given by*

$$\phi(\mathbf{K}, \mathbf{L}, N_1, N_2, \alpha) = \begin{cases} 1 & \text{if } \frac{N_1 \sum_{i=1}^{N_2} r_i^2}{N_2 \sum_{i=1}^{N_1} s_i^2} < F_{\frac{\alpha}{2},2N_2,2N_1} \quad \text{or} \quad \frac{N_1 \sum_{i=1}^{N_2} r_i^2}{N_2 \sum_{i=1}^{N_1} s_i^2} > F_{1-\frac{\alpha}{2},2N_2,2N_1}; \\ 0 & \text{otherwise}, \end{cases}$$

where $s_i = e_1^\tau + \cdots + e_{\#K_i - 1}^\tau$ and $r_j = e_1^\eta + \cdots + e_{\#L_j - 1}^\eta$, for $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$, with e^τ and e^η representing generic elements of the edge sets $\mathcal{E}(\tau)$ and $\mathcal{E}(\eta)$ respectively; $F_{\alpha,a,b}$ denotes the αth percentile of an F distribution with a, b degrees of freedom.

Remark 3. Observe that the tests in Theorem 2 and 3 are finite sample tests in the sense that we do not let the sample size N_1 or N_2 tend to infinity. The class in (3.1) is a valid distributional class on a set of trees and whenever the data generating model is a CGW model, the test represents a useful tool to distinguish between a fairly general class of trees. It is evident now how crucial Proposition 2 on extendability is for inferential purposes.

4. Parametric family and test based on Dyck Path

In the section, we consider a parametric class and develop tests for trees which are ordered and can be embedded on the plane. The pertinent question behind the Dyck path representation of an ordered tree is this: suppose a CGW tree τ_n is distributed as a member of the class (2.2); what is the ramification of the bijective transformation $\tau_n \mapsto H_n$ on the class $\{\pi_{k,\sigma^2} : k = 0, 1, 2, \ldots; 0 < \sigma^2 < \infty\}$? If we propose to develop inferential tools on the space of Dyck paths, it is then required to establish the equivalence of statistical procedures, perhaps in the Le Cam sense, on $\{\pi_{k,\sigma^2} : k = 0, 1, 2, \ldots; 0 < \sigma^2 < \infty\}$ and the class resulting from the transformation. Indeed, this requires us to know exactly the induced class prior to establishing equivalence. The probabilistic structure of the Dyck path corresponding to an *arbitrary* CGW tree is not easily ascertained; only under the special case when the offspring distribution is the Geometric with success probability 1/2, is it known that the corresponding Dyck path can be modeled as a simple symmetric random walk conditioned on first return to 0 (see Aldous [1993]). This issue poses a serious difficulty if one wishes to establish some sort of equivalence between procedures on the two classes using the notion of a deficiency distance. However, weak equivalence of the procedure is easily established as consequence of the invariance principle in Theorem 1[2]: for a CGW tree τ_n, if $\{\mathbb{P}^n_{\sigma^2} : \sigma^2 \in \mathcal{S}\}$ is the experiment associated with its density (with respect to the counting measure) π^2_σ, then as $n \to \infty$ through the sizes of the unconditional CGW tree, $\{\mathbb{P}^n_{\sigma^2} : \sigma^2 \in \mathcal{S}\} \Rightarrow \{\mathcal{P}^{ex}_{\sigma^2} : \sigma^2 \in \mathcal{S}\}$, where $\mathcal{P}^{ex}$ is the law on the Brownian excursion. This is our motivation in using the weak convergence argument in developing statistical models on trees: we are able to circumvent the issue of proving equivalence since the limit process is a Brownian excursion regardless of the original offspring distribution. Conveniently though, the dependence on the offspring distribution arises through the variance parameter σ^2 as the scaling factor.

To recall, $\frac{2}{\sigma}B^{ex}$ is the limit of normalized Dyck paths which code CGW trees uniquely. Aldous' result connects the CRT to $\frac{2}{\sigma}B^{ex}$ in the following manner: Pick $U_1, \ldots, U_k$ uniformly from $[0, 1]$ and consider the order statistics $U_{1:k} < \cdots < U_{k:k}$. Set $V_i = \min_{U_{i:k} \le t \le U_{i+1:k}} \frac{2}{\sigma}B^{ex}(t)$. Draw an edge of length $\frac{2}{\sigma}B^{ex}(U_{1:k})$ and label one end as the root and the other end as U_1. Inductively, from $U_{i:k}$ move back a distance $\frac{2}{\sigma}B^{ex}(U_{i:k}) - V_i$ towards the root, draw a new edge of length $\frac{2}{\sigma}B^{ex}(U_{i+1:k}) - V_i$ and label the new endpoint U_{i+1}. Aldous then proved that the resulting binary tree on k vertices with $k - 1$ edges has the density given in (3.2). The implication of this construction is that the random tree constructed the Brownian excursion at k uniform random times leads to the class of consistent distributions given in (3.1) which characterize the CRT. This opens up the possibility of another family of parametric models for large CGW trees using the excursion.

For a tree τ_n, let $0 = U_{0:n} < U_{1:n} < \cdots < U_{n+1:n} = 1$ be uniform order statistics and let $V_i = \min_{U_{i:n} \le t \le U_{i+1:n}} \frac{2}{\sigma}B^{ex}(t)$. Now define the $2n + 2$ dimensional vector taking values in $\mathbb{R}^{2n+2}_+$ as

$$X_n = \Big(\frac{2}{\sigma}B^{ex}(U_{i:n}), \frac{2}{\sigma}B^{ex}(V_i)\Big).$$

Based on the construction above it can be seen that the distribution of the random vector X_n defines a distribution on the random tree constructed with n vertices. One way at looking at the density in (3.2) via the construction above is as the density of the random variable which is the total variation of the function obtained via a linear interpolation between points in X_n; using this approach it was

[2]Note that if we restrict ourselves to a finite set $\mathcal{S}$ for modeling purposes, then weak convergence of the procedures is equivalent to convergence in deficiency distance since the canonical Blackwell measure of the two experiments coincide.

shown in Theorem of Pitman [1999] that

$$X_n \stackrel{d}{=} \frac{\sigma(2\Gamma_{n+1})^{1/2}}{4}\Big(U_{i-1:n} - V_{i-1}, U_{i:n} - V_{i-1}; 1 \le i \le n+2 \Big| \cap_{i=1}^{n} (U_{i:n} > V_i)\Big),$$

where $U_{n+2:n} := 1$ and Γ_{n+1} is a Gamma random variable with shape $n+1$ and scale 1. While, in principle, it would be reasonable to define a parametric class, the distribution of X_n is not easy to compute.

4.1. **Test based on random projection.** Testing on trees with offspring variance $\sigma^2 \in \mathcal{S}$ is weakly equivalent to distinguishing between brownian excursions scaled by $\frac{2}{\sigma}$. We first construct a parametric class based on a random coordinate projection of B^{ex}; by this, we mean that we consider the family of distributions induced by the map $p_U : \frac{2}{\sigma}B^{ex} \mapsto \frac{2}{\sigma}B^{ex}(U)$, where U is chosen uniformly on $[0,1]$. In order for this approach to bear fruition, we first need to verify that the law on the Brownian excursion is completely determined by the law of $p_U(B^{ex})$ for a random U. This would then ensure that the resulting family based on random projections is distinguishable. In the tree-setting, the approach translates to the following scenario: For a tree τ_n with offspring variance σ^2, pick a vertex v from $\mathcal{V}(\tau_n)$ according to a uniform distribution, and use the distribution of $d(\text{root}, v)$ to define a parametric class. In the language of Dyck paths, we would be interested in the distribution of $H_n(s)$ for $0 \le s \le 2\sum_{i=1}^{n-1} e_i$ where s corresponds to the sum of the edges up to vertex v encountered during the depth-first walk.

Proposition 4. *For ordered CGW trees, the class of distributions* $\{r^U_{\sigma^2} : \sigma^2 \in \mathcal{S}\}$ *induced through* $p_U\left(\frac{2}{\sigma}B^{ex}\right)$ *is distinguishable.*

This immediately provides the following useful result:

Proposition 5. *On an ordered CGW tree* τ_n *with offspring variance* σ^2*, suppose* V *is a vertex chosen according to a uniform distribution on* $\mathcal{V}(\tau_n)$*. Then, the random variable*

$$n^{-1/2} d(\mathit{root}, V) \stackrel{d}{\to} W,$$

where W *is a Rayleigh distributed random variable with scale* $1/\sigma$*. Therefore,* $p_V\left(\frac{2}{\sigma}B^{ex}\right)$ *is Rayleigh distributed with scale* $1/\sigma$*.*

Remark 4. The question, Given a vertex v what is the distribution of $d(\text{root}, v)$?, is different to the one answered above, which is more meaningful for the following reason: the distance of a given vertex on tree is completely determined by the value of the Dyck path which was constructed using the depth-first walk; indeed, there are several ways to uniquely code a tree and the distance of a vertex from the root should not be dictated by the choice of a traversal. More importantly, if we wish to define a parametric class to distinguish between populations of trees, then this becomes a more pressing issue.

Remark 5. It is interesting to note that the Rayleigh distribution arises again as the limiting distribution—-this was the case in (3.2) when viewed as the density of $\sum_{i=1}^{k-1} e_i$. This is not a coincidence; in the interests of brevity, we refer to the intricate construction of the CRT and Corollary 22 in Aldous [1993] for an explanation of the connection. In the context of LCA trees, the Rayleigh density was used to define the density on an *entire* LCA tree; the tree functional of interest in this setup is, however, different. Suppose v is a vertex chosen randomly from τ_n and is a part of the subset of$B$ of $\mathcal{V}(\tau_n)$ chosen to construct $LCA(\tau_n, B)$. Note now that the distance from the root of v is preserved in $LCA(\tau_n, B)$. In the Dyck path approach the induced parametric family is based on this distance, whereas in the LCA-tree based approach the induced parametric family is based on the *sum* of all such distances in the $LCA(\tau_n, B)$ which additionally contains the branchpoints.

In the context of hypothesis testing on CGW trees, we are again under the exponential family framework with the Rayleigh distribution, but this time using the Dyck path approach. We state results omitting the proofs, as they are similar to the ones under the LCA approach.

Theorem 4. *Suppose we have an independent sample of ordered CGW trees $\tau_{n_i} = (\mathcal{V}(\tau_{n_i}), \mathcal{E}(\tau_{n_i})))$ for $i = 1, \ldots, N$ from a distribution π_{σ^2} on the product space $\otimes_{i=1}^{N} \mathcal{T}_{n_i} \times \mathbb{R}_+^{n_i - 1}$ with $\sigma^2 \in \mathcal{S}$. In each tree, choose a vertex V_i uniformly from $\mathcal{V}(\tau_{n_i}) = (root, v_1, \ldots, v_{n_i - 1})$ and record its distance from the root: $d(root, V_i)$; then let $d_i = n^{-1/2} d(root, V_i)$ for $i = 1, \ldots, N$.*

(1) Given $\mathbf{V} = (V_1, \ldots, V_N)$, define the critical function

$$\phi(\mathbf{V}, N, \alpha, \sigma_0^2) = \begin{cases} 1 & \text{if } \sigma_0^2 \sum_{i=1}^{N} d_i^2 < C_{\alpha, 2N} \\ 0 & \text{if } \sigma_0^2 \sum_{i=1}^{N} d_i^2 > C_{\alpha, 2N}. \end{cases}$$

Then, conditional on $\mathbf{V}$, for the pair of hypotheses $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 = \sigma_1^2$ where $\sigma_1^2 > \sigma_0^2$, the test given by $\phi(\mathbf{V}, N, \alpha, \sigma_0^2)$ is such that as $n_i \to \infty$ for each $i = 1, \ldots, N$, $E_\pi \phi(\mathbf{K}, N, \alpha, \sigma_0^2) \to \alpha$, and is the most powerful test for the pair of hypotheses.

(2) Given $\mathbf{V}$, the likelihood ratio test for testing $H_0 : \sigma^2 = \sigma_0^2$ vs $H_0 : \sigma^2 \neq \sigma_0^2$ is given by the critical function

$$\psi(\mathbf{K}, N, \alpha, \sigma_0^2) = \begin{cases} 1 & \text{if } \sigma_0^2 \sum_{i=1}^{N} d_i^2 < C_{\frac{\alpha}{2}, 2N} \quad \text{or} \quad \sigma_0^2 \sum_{i=1}^{N} d_i^2 > C_{1 - \frac{\alpha}{2}, 2N}; \\ 0 & \text{otherwise}, \end{cases}$$

where as $n_i \to \infty$ for each $i = 1, \ldots, N$, $E_\pi \psi(\mathbf{K}, N, \alpha, \sigma_0^2) \to \alpha$ and all other quantities are as in part 1.

Theorem 5. *Suppose we have two independent samples of ordered CGW trees $\tau_{n_i} = (\mathcal{V}(\tau_{n_i}), \mathcal{E}(\tau_{n_i}))$ for $i = 1, \ldots, N_1$, and $\eta_{m_j} = (\mathcal{V}(\eta_{m_j}), \mathcal{E}(\eta_{m_j}))$ for $j = 1, \ldots, N_2$, from distributions $\pi_{\sigma_1^2}$ and $\pi_{\sigma_2^2}$ on the product spaces $\otimes_{i=1}^{N_1} \mathcal{T}_{n_i} \times \mathbb{R}_+^{n_i - 1}$ and $\otimes_{j=1}^{N_2} \mathcal{T}_{m_j} \times \mathbb{R}_+^{m_j - 1}$, respectively, with $\sigma_i^2 \in \mathcal{S}$ for $i = 1, 2$. Given $\mathbf{V} = (V_1, \ldots, V_{N_1})$ and $\mathbf{W} = (W_1, \ldots, W_{N_2})$, let d_i and c_j be the normalized distances from the root for the chosen vertices as in Theorem 4, for $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$. Given $\mathbf{V}$ and $\mathbf{W}$, as $n_i, m_j \to \infty$, for each i and j, the likelihood ratio test of asymptotic size α, for the pair of hypotheses $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_0 : \sigma_1^2 \neq \sigma_2^2$ is given by*

$$\phi(\mathbf{V}, \mathbf{W}, N_1, N_2, \alpha) = \begin{cases} 1 & \text{if } \frac{N_1 \sum_{i=1}^{N_2} c_i^2}{N_2 \sum_{i=1}^{N_1} d_i^2} < F_{\frac{\alpha}{2}, 2N_2, 2N_1} \quad \text{or} \quad \frac{N_1 \sum_{i=1}^{N_2} c_i^2}{N_2 \sum_{i=1}^{N_1} d_i^2} > F_{1 - \frac{\alpha}{2}, 2N_2, 2N_1}; \\ 0 & \text{otherwise}, \end{cases}$$

where $F_{\alpha, a, b}$ denotes the αth percentile of an F distribution with a, b degrees of freedom.

5. Simulations

For a non-negative integer-valued random variable ξ with distribution π_k for $k = 0, 1, \ldots$, the construction of a Galton-Watson tree τ was explained in Section 2.2. However the construction of τ_n, the Galton-Watson tree conditioned to have n vertices, is not straightforward. Using a random-walk construction from n independent copies of ξ, it can seen that in order to generate a CGW τ_n, a necessary condition is to generate a vector $\Xi = (\xi_1, \ldots, \xi_n)$ such that $\sum_{i=1}^{n} \xi_i = n - 1$ and then determining a rotation of Ξ, i.e. a vector $(\xi_k, \xi_{k+1}, \ldots, \xi_n, \xi_1, \xi_2, \ldots, \xi_{k-1})$, with the property that the total number vertices of τ equals n. We use an efficient algorithm provided by Devroye [2012] with linear expected time to generate the CGW trees. This enables us to efficiently simulate a large number of CGW trees, each containing a large number of vertices—each tree is generated in expected linear time. We have made our C++ code is available at `www.github.com/pkambadu/DyckPaths`. Pseudo-code and description of the algorithm and shuffling can be found in the Appendix.

5.1. **Performance of LCA-based tests.** Computing details pertaining to the construction of an LCA tree following the CGW tree are elucidated in the Appendix. We report here the performance of the LCA test for distinguishing between two tree populations with offspring distributions π_{σ^2} where $\sigma^2 \in \mathcal{S}$. Recall from Theorem 2 that the critical function for the test was based on the statistic corresponding to the sum of edge lengths of the LCA tree constructed from a randomly chosen subset of the vertex set of the CGW tree. We generate CGW trees from π_{σ^2} for different values of σ^2 and present empirical rates of rejection for the test. We perform 10000 simulations at varying sample sizes of tree—note that Theorem 2 represents a finite sample test. Firstly, Figure 3 plots the histogram of the sum of the edges of the constructed LCA tree; as postulated, the Rayleigh distribution with scale $1/\sigma$ offers a good approximation. Tables 1 and 2 detail the performance of the one-sample most

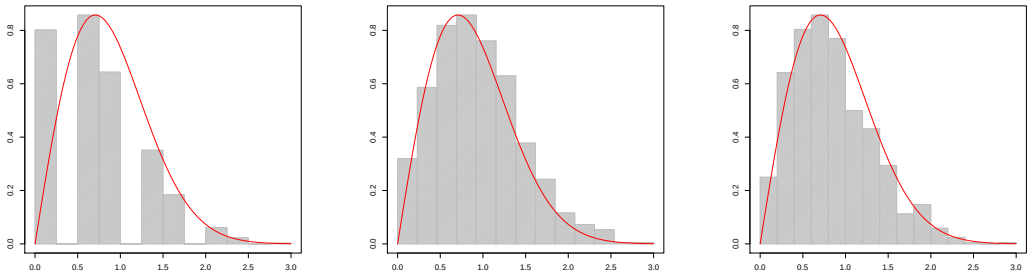

FIGURE 3. Histograms of $s = e_1 + \cdots + e_{k-1}$ from LCA trees of 10000 CGW trees from π_2 with number of vertices 10, 100 and 1000 (from left). LCA trees were constructed by choosing 40% of the vertices randomly from the vertex set of the CGW tree. Solid red curve is the Rayleigh density with scale $1/\sqrt{2}$.

powerful (MP) and Likelihood Ratio (LR) tests. Power calculation is performed only for the LR test since the MP is for simple hypotheses.

$H_0 : \pi = \pi_{\sigma^2}$	*Rejection rate for test of MP test*			*Rejection rate of LR test*		
	10	50	100	10	50	100
$\pi_{2/3}$	0.036	0.040	0.047	0.041	0.049	0.050
$\pi_{1/2}$	0.071	0.043	0.052	0.067	0.053	0.049
π_2	0.024	0.047	0.056	0.029	0.049	0.051

TABLE 1. Level of one sample UMP and LR tests under H_0 based on the LCA method for CGW trees with 1000 vertices each from different offspring distributions. LCA trees were constructed by choosing 30% of vertices from the vertex set of each tree at random.

For the two-sample LR test, we generate 1000-vertex CGW trees from different distributions at varying sample sizes and examine the simulation level of the test and its power; these are reported in Tables 3 and 4.

The tests based on the LCA, in general, appear to be performing well with low sample sizes as long as the number of vertices of the CGW trees are large, validating the use of the CRT approach. Since the LCA trees are, in a sense, finite dimensional distributions of the CRT, the tests based on them are able to distinguish between populations of trees quite efficiently.

$H_0 : \pi_{1/2}$ *vs* H_1 :	*Rejection rate for test of LR test*		
	10	50	100
$\pi_{2/3}$	0.948	0.964	0.991
π_1	0.924	0.967	0.999
π_2	0.919	0.957	0.987

TABLE 2. Rejection rate under alternative hypothesis when $H_0 : \pi_{\sigma^2} = \pi_{1/2}$ for the one sample LR test based on the LCA method for CGW trees with 1000 vertices each from different offspring distributions. LCA trees were constructed by choosing 30% of vertices from the vertex set of each tree at random.

$\pi_0 = \pi_1$ *under* H_0	*Level of LR test*		
	10	50	100
$\pi_{2/3}$	0.050	0.055	0.047
π_1	0.040	0.047	0.048
$\pi_{1/2}$	0.060	0.052	0.045
π_2	0.050	0.046	0.044

TABLE 3. Level of two-sample LR test under H_0 based on the LCA method for CGW trees from two distributions with 1000 vertices each from different offspring distributions. LCA trees were constructed by choosing 30% of vertices from the vertex set of each tree at random.

π_0 *vs* π_1 *under* H_1	*Rejection rate for LR test*		
	10	50	100
$\pi_{2/3}$ vs π_1	0.508	0.897	0.991
$\pi_{2/3}$ vs $\pi_{1/2}$	0.271	0.926	1.000
$\pi_{2/3}$ vs π_2	0.781	0.874	0.993
π_1 vs $\pi_{1/2}$	0.943	0.982	1.000
π_1 vs π_2	0.914	1.000	1.000
$\pi_{1/2}$ vs π_2	0.977	0.985	1.000

TABLE 4. Rejection rate of two-sample LR test under the alternative based on the LCA method for CGW trees from two distributions with 1000 vertices each from different offspring distributions. LCA trees were constructed by choosing 30% of vertices from the vertex set of each tree at random.

5.2. **Performance of Dyck path-based tests.** In this section we evaluate the performance of the tests based on the random projection method on normalized Dyck paths of CGW trees from Theorems 4 and 5. Figure 4 provides an illustration of CGW trees with offspring distributions $\pi_{1/2}$ and π_1 and their corresponding normalized Dyck paths. Recall that the test statistics were based on the distance from the root of a randomly chosen vertex which is equivalent to the value of the corresponding Dyck path at a randomly chosen point on the x-axis. Figure 5 below plots the histogram of this statistic scaled by $n^{-1/2}$ where n is the number of vertices of the tree for varying n. This offers verification of Proposition 5. We now examine the performance of the one-sample and two-sample tests based on the Dyck-path approach and Brownian excursion. Tables 5, 6,7 and 8 tabulate the results.

While the performance of the Dyck path-based tests are acceptable while verifying their attained level of significance, their power under small samples, say 10-20 trees, is quite poor; this can be seen from the power at sample size 10 in Table 8. But upon increasing the sample size to 100 or so, there

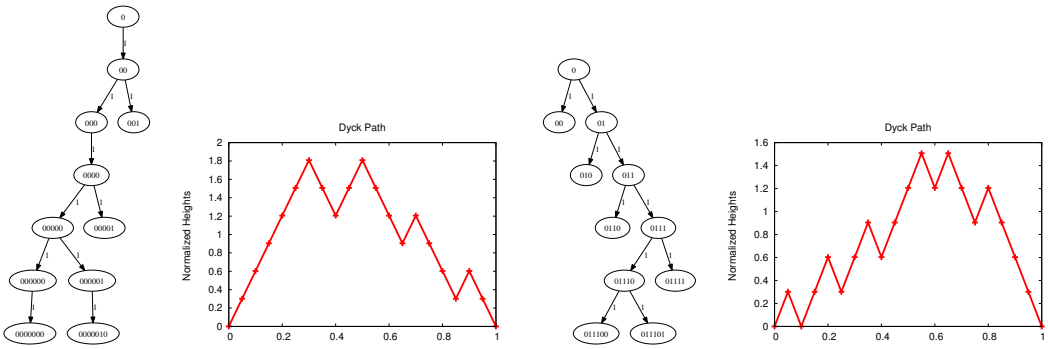

FIGURE 4. Left: 11-node CGW tree with offspring distribution $\pi_{1/2}$ and its corresponding normalized Dyck path. Right: 11-node CGW tree with offspring distribution π_1 and its normalized Dyck path.

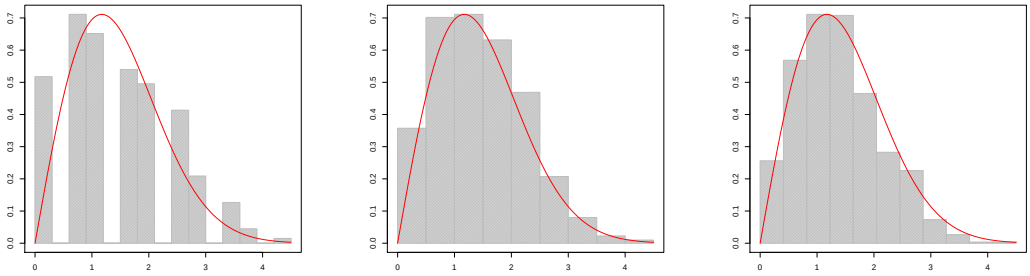

FIGURE 5. Histograms of $n^{-1/2}d(\text{root}, V)$ where V is vertex chosen at random on 10000 CGW trees from $\pi_{1/2}$ with number of vertices $n = 10$, 100 and 1000 (from left). Solid red curve is the Rayleigh density with scale $\sqrt{2}$.

$H_0 : \pi = \pi_{\sigma^2}$	*Rejection rate for test of MP test*			*Rejection rate of LR test*		
	10	50	100	10	50	100
$\pi_{2/3}$	0.043	0.050	0.049	0.041	0.044	0.052
π_1	0.051	0.049	0.046	0.061	0.047	0.049
$\pi_{1/2}$	0.046	0.054	0.049	0.057	0.043	0.049
π_2	0.047	0.051	0.051	0.049	0.051	0.050

TABLE 5. Level of one sample UMP and LR tests under H_0 based on the Dyck path method for CGW trees with 1000 vertices each from different offspring distributions.

is marked improvement in the distinguishing power. The fact that the test is based on one randomly chosen vertex, as opposed to LCA subtrees, is reflected in the its minimal utility in small sample sizes.

6. Discussion

Aldous' papers on the CRT and variants (see, in addition, Aldous [1994a, 1991b]) provide useful distributional results and connections to common stochastic processes, which in principle can be harnessed in developing asymptotic statistical tools. The circumscription of our considerations to

$H_0 : \pi_{1/2}$ *vs* H_1 :	*Rejection rate for test of LR test*		
	10	50	100
$\pi_{2/3}$	0.262	0.873	0.998
π_1	0.197	0.764	0.989
π_1	0.441	0.777	0.979
π_2	0.631	0.816	1.000

TABLE 6. Rejection rate under alternative hypothesis when $H_0 : \pi_{\sigma^2} = \pi_{1/2}$ for the one sample LR test based on the Dyck path method for CGW trees with 1000 vertices each from different offspring distributions.

$\pi_0 = \pi_1$ *under* H_0	*Level of LR test*		
	10	50	100
$\pi_{2/3}$	0.056	0.056	0.047
π_1	0.052	0.046	0.053
$\pi_{1/2}$	0.057	0.051	0.054
π_2	0.050	0.043	0.052

TABLE 7. Level of two-sample LR test under H_0 based on the Dyck path method for CGW trees from two distributions with 1000 vertices each from different offspring distributions.

π_0 *vs* π_1 *under* H_1	*Rejection rate for LR test*		
	10	50	100
$\pi_{2/3}$ vs π_1	0.258	0.833	0.994
$\pi_{2/3}$ vs $\pi_{1/2}$	0.169	0.371	0.996
$\pi_{2/3}$ vs π_2	0.628	0.999	1.000
π_1 vs $\pi_{1/2}$	0.308	0.939	1.000
π_1 vs π_2	0.235	0.859	0.994
$\pi_{1/2}$ vs π_2	0.817	0.993	1.000

TABLE 8. Rejection rate of two-sample LR test under the alternative based on the Dyck path method for CGW trees from two distributions with 1000 vertices each from different offspring distributions.

hypothesis testing for distributions is not a shortcoming of the CRT-based approach; rather, as mentioned in the introduction, this article represents a first step in developing inferential procedures on large tree-structured data. One immediate extension to this work is to approximate distributions of local tree-functionals by corresponding Brownian excursion functional. For example, the Wiener index of a tree, popular in phylogenetics and chemistry, is exactly $\frac{2}{n}A_n$ where A_n is the area under the curve of the Dyck path of a tree with n vertices. The distribution of A_n can be approximated by the distribution of Brownian excursion area which is well-know (albeit difficult to compute); see Janson [2012] and references therein for details. Other tree functionals whose limit distributions as Brownian excursion functionals are known include height of a tree, which is the number of generations before extension, maximal distance between a pair of vertices etc. Preliminary investigations based on simulations appear promising.

The 'weak convergence paradigm' set forth by Aldous wherein properties (global and local) of random combinatorial objects likes trees, triangulations, planar maps etc. are studied through continuous approximations offers a fertile ground for development of statistical methodologies on such objects;

see Aldous [1994b]. This is an investigative route well worth pursuing in today's data centric climate wherein complex data structures, modeled in a combinatorial fashion, is prevalent.

On a modeling note, our statistical characterizations of the trees using CRT admits a full likelihood based inference using frequentist and Bayesian techniques. For the latter, a particular context of importance might be regression models that enable modeling of covariate effects on tree-structured responses. For instance, the offspring variance σ^2 can be modeled as a function of covariates through parametric and non-parametric prior specifications. This might be particularly appealing in applied contexts where evaluations of systematic variations induced by the covariates are of prime interest. We leave these tasks for future consideration.

7. Appendix

*Proof. P*roposition 2. First observe that the density in (3.2) implies that the edge lengths are exchangeable: $s = \sum_{i=1}^{k-1} e_i$ is invariant to permutations of e_i. This implies that the actual labeling to the vertices and the edges has no relevance to the distribution. For ease of notation, let

$$\mathcal{C}_k = \mathcal{T}_k \times \mathbb{R}_+^{k-1}.$$

For a tree, $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$, with n vertices, we shall consider its LCA tree, $LCA(\tau_n, v_1, \ldots, v_k)$, constructed from k vertices, chosen uniformly, with $k < n$. The question of extendability is basically a question of whether models specified in terms of joint distributions over a class of index sets is *projective* as defined in eq. 13, p. 92 of Kallenberg [1997]—this is the basis of our definition of n-extendability. The probability kernel p_k from $\mathcal{C}_1 \times \cdots \times \mathcal{C}_{k-1}$ to $\mathcal{C}_k$ is defined in terms of the conditional density (ignoring the normalizing factors)

$$f_{\sigma^2}\Big(LCA(\tau_n, v_1, \ldots, v_k) = \cdot \Big| LCA(\tau_n, v_1, \ldots, v_{k-1})\Big),$$

which is obtained as

$$\begin{aligned} f(LCA(\tau_n, v_1, \ldots, v_k) | LCA(\tau_n, v_1, \ldots, v_{k-1})) &= \frac{f(LCA(\tau_n, v_1, \ldots, v_k))}{f(LCA(\tau_n, v_1, \ldots, v_{k-1}))} \\ &= \frac{s'}{s} e^{-\frac{(s'^2 - s^2)\sigma^2}{2}}, \end{aligned}$$

with $s = e_1 + \cdots + e_{k-1}$ and $s' = s + e_k$. By induction on k, we can extend the existence of the probability kernel p_n to $\mathcal{C}_n$ with conditional density

$$f(LCA(\tau_n, v_1, \ldots, v_n) | LCA(\tau_n, v_1, \ldots, v_{n-1})) = \frac{s'}{s} e^{-\frac{(s'^2 - s^2)\sigma^2}{2}},$$

where $s = e_1 + \cdots + e_{n-1}$ and $s' = s + e_n$. By Theorem 5.17 in Kallenberg [1997], we can assert the existence of the tree τ_n with distribution $p_1 \otimes \cdots \otimes p_n$; in other words, the distribution on τ_n can be defined via the conditional densities as

$$\begin{aligned} f(\tau_n) = f(LCA(\tau_n, v_1)) f(LCA(\tau_n, v_2) | LCA(\tau_n, v_1)) f(LCA(\tau_n, v_3) | LCA(\tau_n, v_1, v_2)) \\ \ldots f(LCA(\tau_n, v_n) | LCA(\tau_n, v_1, \ldots, v_{n-1})). \end{aligned}$$

Straightforward computation with the conditional densities verifies this fact. If $B = \{v_1, \ldots, v_k\}$ and $\mathcal{V}(\tau_n) - B$ is the set difference, what should be noted is that

$$\sum_{\tau \in \mathcal{V}(\tau_n) - B} f_{\sigma^2}(\tau) = \int_{e_k > 0} \int_{e_{k+1} > 0} \ldots \int_{e_{n-1} > 0} f_{\sigma^2}(\tau) \quad de_k \ldots de_{n-1}.$$

The density on the tree is induced by the density of s with respect to the Lebesgue measure. □

Proof. Proposition 3. Suppose B is a Borel subset of $\mathcal{C}_n = \mathcal{T}_n \otimes \mathbb{R}_+^{n-1}$. Suppose we define a relation $\sim$ on subsets B_1 and B_2 of $\mathcal{C}_n$ as $B_1 \sim B_2$ if they contain *all* trees with n vertices; by this we mean that the "shape" of the tree is disregarded and imply that all trees with n vertices are equivalent. Note that $\sim$ is an equivalence relation and the generates the quotient class $\mathcal{C}_n^{\sim} = (t_n, (e_1, \ldots, e_{n-1}))$ with $(e_1, \ldots, e_{n-1}) \in \mathbb{R}_+^{n-1}$ and t_n is the canonical tree with n vertices. The Borel sets of $C_n^{\sim}$ are the usual

open rectangles generating the Euclidean space $\mathbb{R}_+^{n-1}$. Note that the law P_{σ^2} assigns different mass to distinct elements in $\mathcal{C}^\sim$. We are hence interested primarily in Borel subsets of $C^\sim$ and will restrict our examination of distinguishability to this equivalence class.

Suppose the null hypothesis H_0 is that $\sigma^2 \in \mathcal{S}_0$ and the alternative H_1 is that $\sigma^2 \in \mathcal{S}_1$ where $\mathcal{S}_0 \cup \mathcal{S}_1 = \mathcal{S}$ and $\mathcal{S}_0 \cap \mathcal{S}_1 = \emptyset$. Note that

$$\mathcal{S} = \left\{2, \frac{1}{2}, 1, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \dots\right\}$$

is a countable set and consequently, so are $\mathcal{S}_0$ and $\mathcal{S}_1$. Furthermore the distribution function associate with P_{σ^2}, for a tree τ_n, corresponding to the continuous density $f_{\sigma^2}(\cdot)$

$$F_{\sigma^2}(x_1, \dots, x_{n-1}) = \int_0^{x_1} \dots \int_0^{x_n} f_{\sigma^2}(\tau) \quad de_1 \dots de_n$$

is continuous for each vector $(x_1, \dots, x_n)$ representing edge lengths. Therefore by part (i) of Theorem 1 in Rao [2000], the proof is complete. □

Proof. Theorem 2.

(1) *Simple hypotheses*:

The key observation here is that conditional on $\mathbf{K}$, $\mathbf{s} = (s_1, \dots, s_N)$ where $s_i = e_1 + \dots + e_{k_i}$, is a vector of independent Rayleigh distributed random variables with scale $1/\sigma$. Let us first perform some calculations in the limit as $n_i \to \infty$ for each $i = 1, \dots, N$. Suppose $W_1, \dots, W_N$ are i.i.d $\mathbb{R}_+$ -valued random variables from a Rayleigh distribution with scale $1/\sigma$ and density

$$f_\sigma^2(w) = \sigma^2 w e^{-\frac{w^2\sigma^2}{2}},$$

leading to the likelihood

$$L_{\sigma^2} = (\sigma^2)^N \exp\left[-\frac{\sigma^2}{2}\sum_{i=1}^N w_i^2\right] \prod_{i=1}^N w_i.$$

Consider

$$\Lambda = \frac{L_{\sigma_1^2}}{L_{\sigma_0^2}} \propto \exp\left[\frac{\sum_{i=1}^N w_i^2(\sigma_1^2 - \sigma_0^2)}{2}\right].$$

By the Neyman-Pearson Lemma (see, p. 60 Lehmann and Romano [2005]), the most powerful test for testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 = \sigma_1^2$, where $\sigma_1^2 > \sigma_0^2$ is given by the rejection region

$$\{(w_1, \dots, w_n) : \Lambda > C_\alpha\}$$

for a suitable value C_α such that $P(\Lambda > C_\alpha) = \alpha$ with P denoting the law corresponding to the Rayleigh density. Note now that $\Lambda > C_\alpha$ if and only if $\sum_{i=1}^N w_i^2 < C_\alpha$. It is easy to ascertain that $\sigma^2 \sum_{i=1}^N W_i^2$ follows a Chi-square distribution with $2N$ degrees of freedom; therefore the rejection region defined as

$$\{(w_1, \dots, w_n) : \sigma_0^2 \sum_{i=1}^N W_i^2 < C_{\alpha,2N}\},$$

where $C_{\alpha,2N}$ is chosen such that $P(\chi_{2N} > C_{\alpha,2N}) = \alpha$ with χ_{2N} denoting a Chi-square random variable with $2N$ degrees of freedom, is of size α. The power function of the test is

$$\theta(N, \alpha, \sigma^2) = P\Big(\sigma^2 \sum_{i=1}^N W_i^2 < C_{\alpha,2N}\Big), \tag{7.1}$$

with $\theta(N, \alpha, \sigma_0^2) = \alpha$. For each $i = 1, \ldots, N$, from discussion earlier, we know that $LCA(\tau_{n_i j}, v_1, \ldots, v_{k_i} | K_i = k_i)$ converges in distribution to $\mathcal{L}(k_i)$, as $n_i \to \infty$, with density

$$f_{\sigma^2}(l(k_i)) := f_{\sigma^2}(e_1, \ldots, e_{k_i - 1}) = \left[\prod_{j=1}^{k_i - 1} \frac{1}{2j-1} \right]^{-1} \frac{1}{2^{k_i - 1}} \sigma^2 s_i e^{-\frac{\sigma^2 s_i^2}{2}}.$$

Note now that a calculation of likelihood ratio for $\mathcal{L}(k_i)$ using the density above leads to the exact ratio as Λ. Therefore it is now easy to see that as $n_i \to \infty$, for every $i = 1, \ldots, N$, and for every $\mathbf{k}$,

$$E_\pi \phi(\mathbf{K}, N, \alpha, \sigma^2) \to \theta(N, \alpha, \sigma^2) \quad \forall \sigma^2 \in \mathcal{S},$$

ensured by the extendability of the class proved in Proposition 2; quite naturally then,

$$E_\pi \phi(\mathbf{K}, N, \alpha, \sigma_0^2) \to \theta(N, \alpha, \sigma_0^2) = \alpha.$$

(2) *Composite hypothesis*:

If $W_i, i = 1, \ldots, N$ are i.i.d. Rayleigh distributed random variables with scale $1/\sigma$, the it is easy to determine that the MLE of σ^2 is $\hat{\sigma^2} = \frac{2N}{\sum_{i=1}^N W_i^2}$. The likelihood ratio test, then, is to reject H_0 if and only if

$$\begin{aligned} & \frac{(\sigma_0^2)^N e^{-\frac{\sigma_0^2 \sum_{i=1}^N w_i^2}{2}}}{(\hat{\sigma^2})^N e^{-N}} < \beta \\ \iff & \left[t e^{1-t} \right]^N < \beta, \end{aligned}$$

where $t = \frac{\sigma_0^2 \sum_{i=1}^N w_i^2}{2}$ for a suitable β. Observe that the function $g(t) = te^{1-t}$ for $t > 0$; g is increasing for $t < 1$ and decreasing for $t > 1$. Therefore, the likelihood ratio test is equivalent to rejecting H_0 if and only if

$$\sigma_0^2 \sum_{i=1}^N w_i^2 < \beta_1 \quad \text{or} \quad \sigma_0^2 \sum_{i=1}^N w_i^2 > \beta_2;$$

then, β_1 and β_2 are determined as in part 1 for the Neyman-Pearson test. Using identical arguments with power function and weak convergence, as in part 1, the proof is complete.

□

Proof. Theorem 3. We will work again in the limiting scenario of the Rayleigh densities. In the interests of brevity, we refer the reader to the proof of Theorem 2 for arguments concerning the trees—they follow along identical lines. Suppose $X_1, \ldots, X_{N_1}$ are i.i.d. from a Rayleigh distribution with scale $1/\sigma_1$ and $Y_1, \ldots, Y_{N_1}$ are i.i.d. from a Rayleigh distribution with scale $1/\sigma_2$. Under $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{2(N_1 + N_2)}{\sum_{i=1}^{N_1} x_i^2 + \sum_{i=1}^{N_2} y_i^2}.$$

The maximum likelihood estimates, in general, of σ_1^2 and σ_2^2 are, respectively,

$$\hat{\sigma_1}^2 = \frac{2N_1}{\sum_{i=1}^{N_1} x_i^2} \quad \text{and} \quad \hat{\sigma_2}^2 = \frac{2N_2}{\sum_{i=1}^{N_2} y_i^2}.$$

Then likelihood ratio is

$$\begin{aligned}\Lambda &= \frac{(\hat{\sigma}^2)^{N_1+N_2}}{(\hat{\sigma_1}^2)^{N_1}(\hat{\sigma_2}^2)^{N_2}} \\ &= \frac{\left(\frac{N_1}{N_2}+1\right)^{N_1+N_2}}{\left(\frac{N_1}{N_2}\right)^{N_1}} \left(1+\frac{\sum_{i=1}^{N_2} y_i^2}{\sum_{i=1}^{N_1} x_i^2}\right)^{-N_1} \left(1+\frac{\sum_{i=1}^{N_1} x_i^2}{\sum_{i=1}^{N_2} y_i^2}\right)^{-N_2}.\end{aligned}$$

Consider now function $g(t) = t^{N_2}(1+t)^{-(N_1+N_2)}$ for $t > 0$. Observe that $g(0) = 0$ and

$$g'(t) = \frac{t^{N_2-1}}{(1+t)^{N_1+N_2}}\left[N_2 - \frac{t(N_1+N_2)}{(1+t)}\right],$$

which is positive (negative) for $t > (<)\frac{N_2}{N_1}$, implying that $g(t)$ is increasing (decreasing) for $t > (<)\frac{N_2}{N_1}$. Setting $t = \frac{\sum_{i=1}^{N_1} x_i^2}{\sum_{i=1}^{N_2} y_i^2}$, it is the case that $t \sim \frac{N_2}{N_1} F_{2N_2,2N_1}$ under H_0 where $\sigma_1^2 = \sigma_2^2$. □

Proof. Proposition 4. First, we need to establish that law on the Brownian excursion is completely determined by the law of $p_u(B^{ex})$ for a random u on $[0,1]$. For this we use a result from Cuesta-Albertos et al. [2006]. It can be checked that $m_k = \int ||x||^k ne(de)$ for $k \in \mathbb{N}$, the moments of the scaled excursion are finite, where ne is the normalized Ito measure of positive excursion of linear Brownian motion—this is true for every $\sigma^2 \in \mathcal{S}$. Now suppose ne_1 and ne_2 are two excursion measures on $\mathbb{R}_+$ and ne_1^u and ne_2^u are the randomly projected excursion measures corresponding to the projection p_u where u is uniform in $[0,1]$. Consider the set

$$\mathbb{E}_u := \{x : ne_1^u(x) = ne_2^u(x)\}.$$

Since ne is atomless it is the case that $ne(\mathbb{E}_u) > 0$ for every u. Using the so-called Carleman condition (see, for instance, p. 19 Shohat and Tamarkin [1943]) and Theorem 2.8 from Cuesta-Albertos et al. [2006], we can claim $ne_1 = ne_2$. The implication of this is that the distribution of the excursion is fully determined by just one random projection p_u for a uniform u in $[0,1]$.

Before proving distinguishability, we need to ascertain the density $r^u_{\sigma^2}$ of $p_U\left(\frac{2}{\sigma}B^{ex}\right) = \frac{2}{\sigma}B^{ex}(U)$ where U is uniform on $[0,1]$. Note that the density of the Brownian excursion $\frac{2}{\sigma}B^{ex}$ at time $t \in (0,1)$ is given by (see Takács [1991])

$$f(t,x) = \frac{x^2\sigma^3}{4\sqrt{2\pi t^3(1-t)^3}} e^{\frac{-x^2\sigma^2}{8t(1-t)}}, \quad x > 0. \tag{7.2}$$

This implies that

$$\begin{aligned}r^u_{\sigma^2}(x) &= \int_0^\infty f(s,x)ds \\ &= \sigma^2 x e^{-\frac{1}{2}x^2\sigma^2},\end{aligned}$$

which is a Rayleigh density with scale $\frac{1}{\sigma}$, with continuous distribution function

$$R^u_{\sigma^2}(x) = 1 - e^{-\frac{x^2\sigma^2}{2}}.$$

Bearing in mind that $r^u_{\sigma^2}$ completely determines the distribution of $\frac{2}{\sigma}B^{ex}$, from Theorem 1 in Rao [2000] we have distinguishability. We have omitted details regarding the Borel sets, as detailed in the proof of Proposition 3, in the interests of brevity. □

Proof. Proposition 5. Let H_n be the Dyck path corresponding to τ_n. Then, $d(\text{root}, V)$ is distributed as $H_n(2nV)$. Since for $0 \le s \le 1$, $n^{-1/2}H_n(2ns)$ converges weakly in $C[0,1]$ to $B^{ex}(s)$, we can claim that $n^{-1/2}d(\text{root}, v) \xrightarrow{d} B^{ex}(v)$ on the set $\{V = v\}$. Using (7.2), we can ascertain the unconditional density of $B^{ex}(V)$ as

$$r(x) = \int_0^\infty f(s,x)ds,$$

since V is uniform on $[0,1]$. Note that the map $B^{ex} \mapsto B^{ex}(V)$ is a one-dimensional random coordinate projection, and is clearly continuous on $C[0,1]$ with respect to the uniform norm. Using the continuous mapping theorem (see Billingsley [1968]), $d(\text{root}, V) \xrightarrow{d} \frac{2}{\sigma} B^{ex}(V)$, which follows a Rayleigh distribution as described in the proof of Proposition 4. □

7.1. **Computing notes.** In this section, we provide details on the parallel and high-performance simulation platform that we used for our experiments; this software has been open-sourced on GitHub (*www.github.com/pkambadu/DyckPaths*) under a BSD-style license. Our implementation is written in C++ and makes use of the Boost Graph library (Schling [2011]) to represent trees, the Boost options library (Schling [2011]) to parse command line options, the Boost random library (Schling [2011]) to generate various distributions, and OpenMP for parallelism and therefore, its dependencies. Our code can be compiled and run on any operating system that has a C++ compiler (with or without OpenMP support) as long as the above mentioned Boost libraries have also been installed; we have tested our implementation on Darwin 10.7 using GCC 4.2.1 and Ubuntu Linux 2.6.31-23-server using GCC 4.4.1.

Generating CGW trees with given offspring distribution
In order to generate a CGW tree τ_n with offspring distribution π_k based on the algorithm in Devroye [2012], it is required to generate a vector $\Xi = (\xi_1, \ldots, \xi_n)$ where ξ_i are independent copies from π; subsequently, we are required to rotate Ξ to ensure that $\sum_{i=1}^{n} \xi_i = n - 1$. We shall describe the construction of the CGW tree with unit edge lengths. Such a setup implies that (see Devroye [2012]):

(1) Ξ is a multinomial random vector with success probabilities determined by π_k;
(2) Elements of Ξ are bounded between 0 and $n - 1$.

Once the vector Ξ has been generated, it is then required to shuffle it to ensure that $\sum_{i=1}^{n} \xi_i = n - 1$. The first n_0 entries of Ξ contain 0, the next n_1 entries contain 1's, and so on. We first impart random structure to the CGW tree represented by Ξ by a random shuffling or permutation of Ξ. We then need to rotate Ξ to ensure that a Depth First Search (DFS) traversal will cover all the n nodes. As an example, suppose following the shuffling we are left with $\Xi = [0, 0, 1, 2]$. Our DFS based construction algorithm would assign $0(\psi[0])$ children to the root node, thereby terminating the tree generation. For this Ξ to be valid for our tree construction, we have to rotate to get $\Xi = [1, 2, 0, 0]$. The index $i, 1 \leq i \leq n$ at which Ξ has to be rotated is given in Devroye [2012]. Given a properly constructed, shuffled and rotated Ξ, construction of the CGW tree is achieved by a DFS based algorithm that is best illustrated through the use of an example. Consider $\Xi = [2, 1, 0, 3, 0, 0, 0]$; when augmented with the index information, $\psi = \left[\frac{1}{2}, \frac{2}{1}, \frac{3}{0}, \frac{4}{3}, \frac{5}{0}, \frac{6}{0}, \frac{7}{0}\right]$; here, the numerator denotes the node-ID and the denominator denotes the number of children (out-degree) of the node. We start by considering node 1 as the root of the CGW tree; in our example, node 1 has an out-degree of 2. Therefore, we mark nodes 2 and 3 as the children of 1 and connect them in our tree. As we explore in DFS-order, we next consider node 2, which has 1 child; as the next unmarked node is 4, we connect 4 to be 2's child. Next, we explore 4, which has 3 children; therefore, we allocate $5, 6, 7$ as 4's children and connect them. Next we explore nodes $5, 6$, and 7, each of which has 0 children before returning to node 3, which also has 0 children. This completes our tree construction, which is shown in Figure 6. Notice that in addition to

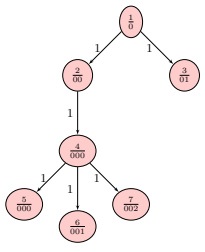

FIGURE 6. The graph constructed for $\psi = \left[\frac{1}{2}, \frac{2}{1}, \frac{3}{0}, \frac{4}{3}, \frac{5}{0}, \frac{6}{0}, \frac{7}{0}\right]$ using DFS-visit with the DFS-label for each node starting from the root (0);

constructing the tree, a DFS-label is generated for every node starting from the root (0); the left-most child appends a 0 to it's parent's DFS-label to form it's own label, it's right sibling appends a 1, and so on. This DFS-label is useful in finding the LCA of two nodes in a CGW tree.

Constructing LCA trees

Let $T = (V, E)$ represent our CGW tree, with vertex set V and edge set E. Let B represent a subset of vertices of V that includes the root. Given (T, B), the goal of the LCA algorithm is to construct a minimum path tree T_{new} that contains all the vertices of B and some additional LCA vertices; that is, we augment B with additional LCA vertices. Notice that in the worst case, after running our LCA algorithm, $B = V$. The LCA of two nodes (v_1, v_2) in a tree is a node that is the lowest among the common ancestors of v_1 and v_2; the caveat is that the LCA of two nodes can be one of the nodes themselves as each node is it's own ancestor. The algorithm to compute the LCA is rather simple:

(1) Let V_{new} represent the list of vertices B plus the LCA vertices that augment B — initialize this list to $V_{new} = B$;
(2) For each pair of vertices $(v_1, v_2) \in B$, compute v_{LCA}, the LCA of (v_1, v_2) and add it to V_{new}; there are $\binom{\#B}{2}$ such vertex pairs;
(3) Construct the LCA tree T_{new} by joining the vertices in V_{new} using the edge information in T; specifically, when connecting vertices that originally had an intermediate vertex between them in T, augment the new edge to include the weights of the edges that were skipped in T.

We now turn our attention to efficient computation of $v_{LCA} = LCA(v_1, v_2)$. Notice that we label each of the vertices in T with their DFS-label (see Figure 6). This DFS-label can be used directly to determine the LCA; the LCA of (v_1, v_2) is the longest common prefix of the labels of v_1 and v_2. For example, consider the nodes 3 and 5 in Figure 6, which have the labels '$01'$ and '$000'$, respectively. The longest common prefix is '$0'$, which points to vertex 0, which also is the LCA of 3 and 5. As we store the DFS-label of each node succinctly as a string, we are able to quickly find the LCA using the `std::mismatch` algorithm, which returns the first position of mismatch in the two DFS-labels.

Parallel execution

The basic control structure of our simulations is: (a) generate a large number of CGW trees; (b) compute local statistics on each CGW tree; and (c) combine the local statistics to make inferences. As mentioned earlier, generating a single tree is expensive and may potentially incur many failed attempts before success. Therefore, we parallelize the simulation framework by parallelizing step (b) above using OpenMP; that is, multiple trials of the experiments are run simultaneously when possible and combined with care to ensure consistency. Given that most of the computing hardware has inherent parallelism in the form of multi-cores and multi-sockets, our approach results in linear speedups (w.r.t number of computational resources) in throughput. Notice that parallelizing step (a) is hard both because of the sequential dependency in generating Ξ from the multinomial distribution and because our current random number generators are not thread-safe. However, as we conduct thousands of experiments, we are able to fully utilize clusters with similar processor counts; that is, parallelizing step (a) is not necessary.

References

A Aho, Y Sagiv, T G Szymanski, and J D Ullman. Inferring a Tree from Lowest Common Ancestors with an Application to the Optimization of Relational Expressions. *SIAM Journal of Computation*, 10:405–421, 1981.

D Aldous. The Continuum Random Tree I. *Annals of Probability*, 19:1–28, 1991a.

D Aldous. Asymptotic Fringe Distributions for General Families of Random Trees. *Annals of Applied Probability*, 1:228–266, 1991b.

D Aldous. The Continuum Random Tree III. *Annals of Probability*, 21:248–289, 1993.

D Aldous. Recursive Self-similarity for Random Trees, Random Triangulations and Brownian Excursion. *Annals of Probability*, 22:527–545, 1994a.

D Aldous. Triangulating the Circle, at Random. *The American Mathematical Monthly*, 101:223–233, 1994b.

D Aldous. Probability Distributions on Cladograms. *In: Random Discrete Structures, ed. David Aldous and R. Pemantle, IMA Volumes Math. Appl*, 76:1–18, 1996.

B Aydin, G Pataki, H Wang, E Bullitt, and J S Marron. A Principal Component Analysis for Trees. *Annals of Applied Statistics*, 3:1597–1615, 2009.

B Aydin, G Pataki, H Wang, A Ladha, E Bullitt, and J S Marron. Visualizing the Structure of Large Trees. *Electronic Journal of Statistics*, 5:405–420, 2011.

S Aylward and E Bullitt. Initialization, Noise, Singularities and Scale in Height Ridge Traversal for Tubular Object Centerline Extraction. *IEEE Transactions of Medical Imaging*, 21:61–75, 2002.

P Billingsley. *Weak Convergence of Probability Measures*. John Wiley and Sons, New York, 1968.

J R Busch, P A Ferrari, A G Flesia, R Fraiman, S P Grynberg, and F Leonardi. Testing statistical hypothesis on random trees and applications to the protein classification problem. *Annals of Applied Statistics*, 3:542–563, 2009.

J A Cuesta-Albertos, R Fraiman, and T Ransford. Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of Brazilian Mathematical Society, New Series*, 37:1–25, 2006.

L Devroye. Simulating Size-constrained Galton-Watson Trees. *SIAM Journal of Computing*, 41:1–11, 2012.

I Gronau and S Moran. Neighbor Joining Algorithms for Inferring Phylogenies via LCA Distances. *Journal of Computational Biology*, 14:1–15, 2007.

S Janson. Simply generated trees, conditioned GaltonWatson trees, random allocations and condensation. *Probability Surveys*, 9:103–252, 2012.

O Kallenberg. *Foundations of Modern Probability Theory, First edition* . Springer-Verlag, New York, 1997.

D P Kennedy. The Galton-Watson Process Conditioned on the Total Progeny. *Journal of Applied Probability*, 12:800–806, 1975.

E L Lehmann and J P Romano. *Testing Statistical Hypotheses, 2nd Edition*. Springer, New York, 2005.

J Pitman. Brownian Motion, Bridge, Excursion, and Meander Characterized by Sampling at Independent Uniform Times. *Electronic Journal of Probability*, 4:1–33, 1999.

J Pitman. *Combinatorial Stochastic Processes*. Springer-Verlag, New York, 2006.

M M Rao. *Stochastic Processes: Inference Theory*. Kluwer, Dordrecht, Netherlends, 2000.

P S La Rosa, B Shands, E Deych, Y Zhou, E Sodergen, G Weinstock, and W D Shannon. Statistical Object Data Analysis on Taxonomic Trees from Human Microbiome Data. *PLoS ONE*, 7:1–12, 2012.

Boris Schling. *The boost C++ libraries*. Xml Press, 2011.

C R Shalizi and A Rinaldo. Consistency under Sampling of Exponential Random Graph Models. *Annals of Statistics*, 41:508–535, 2013.

D Shen, H Shen, S Bhamidi, Y Munoz-Maldonado, Y Kim, and J S Marron. Functional Data Analysis of Tree Structured Objects. *Journal of Computational and Graphical Statistics*, DOI:10.1080/10618600.2013.786943, 2013.

J A Shohat and J D Tamarkin. *The Problem on Moments*. 1943.

J M Steele. 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, 1987.

L Takács. A Bernoulli Excursion and Its Various Applications. *Advances in Applied Probability*, 23: 557–585, 1991.

S Tatikonda and S Parthasarathy. Hashing tree-structured data: Methods and applications. In *IEEE 26th International Conference on Data Engineering, Long Beach*, 2010.

H Wang and J S Marron. Object oriented data analysis: Sets of trees. *Annals of Statistics*, 35: 1849–2311, 2007.

Y Wang, J S Marron, B Aydin, A Ladha, E Bullitt, and H Wang. A Nonparametric Regression Model with Tree-structured Response. *Journal of the American Statistical Association*, 107:1272–1285,

2012.

R Yang, P Kalnis, and A K H Tung. Similarity evaluation on tree-structured data. In *SIGMOD, Baltimore*, 2005.

Department of Statistics, The Ohio State University,, 1958 Neil Ave, Columbus, OH 43210, USA
E-mail address: karthikbharath@gmail.com

IBM T. J. Watson Research Center,, Yorktown Heights, NY 10598, USA
E-mail address: pkambadu@us.ibm.com

Department of Statistics, University of Connecticut,, 215 Glenbrook Road, Storrs, CT 06269, USA
E-mail address: dipak.dey@uconn.edu

Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center,, 1515 Holcombe Blvd, Houston, TX 77030, USA
E-mail address: veera@mdanderson.org