Advances in Applied Mathematics 96 (2018) 39–75

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Recovering a tree from the lengths of subtrees spanned by a randomly chosen sequence of leaves

Steven N. Evans [a,*,1], Daniel Lanoue [b]

[a] Department of Statistics, University of California, 367 Evans Hall #3860, Berkeley, CA 94720-3860, USA
[b] Department of Mathematics, University of California, 970 Evans Hall #3840, Berkeley, CA 94720-3840, USA

A R T I C L E   I N F O

Article history:
Received 31 August 2016
Received in revised form 10 January 2018
Accepted 14 January 2018

MSC:
05C05
05C60
05C80

Keywords:
Tree reconstruction
Graph isomorphism
Phylogenetic diversity
Random tree

A B S T R A C T

Given an edge-weighted tree $\mathbf{T}$ with n leaves, sample the leaves uniformly at random without replacement and let W_k, $2 \leq k \leq n$, be the length of the subtree spanned by the first k leaves. We consider the question, "Can $\mathbf{T}$ be identified (up to isomorphism) by the joint probability distribution of the random vector $(W_2, \ldots, W_n)$?" We show that if $\mathbf{T}$ is known *a priori* to belong to one of various families of edge-weighted trees, then the answer is, "Yes." These families include the edge-weighted trees with edge-weights in general position, the ultrametric edge-weighted trees, and certain families with equal weights on all edges such as $(k+1)$-valent and rooted k-ary trees for $k \geq 2$ and caterpillars.

* Corresponding author.
E-mail addresses: evans@stat.berkeley.edu (S.N. Evans), dlanoue@math.berkeley.edu (D. Lanoue).
[1] SNE supported in part by NSF grants DMS-0907630 and DMS-1512933, and NIH grant 1R01GM109454-01.

https://doi.org/10.1016/j.aam.2018.01.001
0196-8858/

1. Introduction

1.1. Background and motivation

What features of an edge-weighted tree identify it uniquely up to isomorphism, perhaps within some class of such trees? Here an *edge-weighted tree* is a connected, acyclic finite graph $\mathbf{T}$ with vertex set $\mathbf{V}(\mathbf{T})$ and edge set $\mathbf{E}(\mathbf{T})$ which is equipped with a function $\mathbf{W}_{\mathbf{T}} : \mathbf{E}(\mathbf{T}) \to \mathbb{R}_{++} := (0, \infty)$. The value of $\mathbf{W}_{\mathbf{T}}(e)$ for an edge $e \in \mathbf{E}(\mathbf{T})$ is called the *weight* or the *length* of e. Two such trees $\mathbf{T}'$ and $\mathbf{T}''$ are isomorphic if there is a bijection $\sigma : \mathbf{V}(\mathbf{T}') \to \mathbf{V}(\mathbf{T}'')$ such that:

- $\{u, v\} \in \mathbf{E}(\mathbf{T}')$ if and only if $\{\sigma(u), \sigma(v)\} \in \mathbf{E}(\mathbf{T}'')$,
- $\mathbf{W}_{\mathbf{T}'}(\{u, v\}) = \mathbf{W}_{\mathbf{T}''}(\{\sigma(u), \sigma(v)\})$ for all $\{u, v\} \in \mathbf{E}(\mathbf{T}')$.

The question above is, more formally, one of asking for a given class of edge-weighted trees $\mathbb{T}$ about the possible sets $\mathbb{U}$ and functions $\Phi : \mathbb{T} \to \mathbb{U}$ such that for all $\mathbf{T}', \mathbf{T}'' \in \mathbb{T}$ we have $\Phi(\mathbf{T}') = \Phi(\mathbf{T}'')$ if and only if $\mathbf{T}'$ and $\mathbf{T}''$ are isomorphic. If the class $\mathbb{T}$ consists of edge-weighted trees for which all edges have length 1 (we will call such objects *combinatorial trees* for the sake of emphasis), then determining whether two trees in $\mathbb{T}$ are isomorphic is just a particular case of the standard graph isomorphism problem. The general graph isomorphism problem has been the subject of a large amount of work in combinatorics and computer science – [33] already speaks of the "graph isomorphism disease" – and, in particular, there are many results on reconstructing the isomorphism type of a graph from the isomorphism types of subgraphs of various sorts (see, for example, the review [6]). There is also a substantial volume of somewhat parallel research on graph isomorphism in computational chemistry (see, for example, [10] for a review). There seems to be considerably less work on determining isomorphism (in the obvious sense) of edge-weighted graphs; of course, in order for two edge-weighted graphs to be isomorphic the underlying combinatorial graphs must be isomorphic, but this does not imply that the best way for checking that two edge-weighted graphs are isomorphic proceeds by first determining whether the underlying combinatorial graphs are isomorphic and then somehow testing whether some isomorphism of the combinatorial graphs is still an isomorphism when the edge-weights are considered.

We begin with a discussion of previous results that address various aspects of the problem of determining when two edge-weighted or combinatorial trees are isomorphic.

A result in [3] gives the following criterion for a bijection $\sigma : \mathbf{V}(\mathbf{T}') \to \mathbf{V}(\mathbf{T}'')$, where $\mathbf{T}'$ and $\mathbf{T}''$ are combinatorial trees, to be an isomorphism: if $v_0, v_1, \ldots, v_m$ is any sequence from $\mathbf{V}(\mathbf{T}') \sqcup \mathbf{V}(\mathbf{T}'')$ (here $\sqcup$ denotes a disjoint union) such that $v_0 = v_m$ and

$$\{v_i, v_j\} \in \mathbf{E}(\mathbf{T}') \sqcup \mathbf{E}(\mathbf{T}'') \sqcup \{\{u, \sigma(u)\} : u \in \mathbf{V}(\mathbf{T}')\} \iff i - j \equiv \pm 1 \mod m,$$

then $m = 4$.