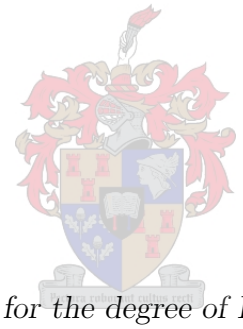
Centrality in Random Trees

by

Kevin Durant



Dissertation presented for the degree of Doctor of Philosophy in Mathematics in the Faculty of Science at Stellenbosch University

Supervisor: Prof. S. Wagner

December 2017

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2017

Abstract

Centrality in Random Trees

K. Durant

Department of Mathematical Sciences
Stellenbosch University
Private Bag X1, Matieland 7602, South Africa

Dissertation: PhD (Mathematics)

December 2017

We consider two notions of centrality—namely, the betweenness centrality of a node and whether or not it is a centroid—in families of simply generated and increasing trees. Both of these concepts are defined in terms of paths within a tree: the betweenness centrality of a node v is the sum, over pairs of nodes, of the proportions of shortest paths that pass through v; and v is a centroid (there can be at most two) if it minimises the sum of the distances to the other nodes in the tree.

We find that betweenness centrality in a large, random simply generated tree is generally linear in the size n of the tree, and that due to the tall, thin nature of simply generated trees, the probability of a random node having quadratic-order betweenness centrality decreases as n increases. This leads to a kth moment of order $n^{2k-(1/2)}$ for the betweenness centrality of a root node, even though a limiting distribution arises upon linearly rescaling the betweenness centrality. The class of labelled subcritical graphs, which are tree-like in structure, behave similarly.

Betweenness centrality in a random increasing tree is also usually linear, except for nodes near to the root of the tree, which typically have centralities of order n^2. The kth moment of the betweenness centrality of any node with a fixed label is thus of order n^{2k}, but once again the distribution of the betweenness centrality of a random node converges to a limit when scaled by $1/n$.

To complement known results involving centroid nodes in simply generated trees, we also derive limiting distributions, along with limits of moments, for the depth, label, and subtree size of the centroid nearest to the root in an increasing tree. The first two of these distributions are concentrated around the root, while the latter is a combination of a point measure at 1 and a decreasing density on $[1/2, 1)$.

In addition, we show that the distributions of the maximum betweenness centrality in simply generated and increasing trees converge, upon suitable rescalings, to limiting distributions, and that the probability of the centroid attaining maximal betweenness centrality tends in both cases to a limiting constant.

Uittreksel

Sentraliteit in Stogastiese Bome

Centrality in Random Trees

K. Durant

Departement van Wiskundige Wetenskappe
Universiteit Stellenbosch
Privaatsak X1, Matieland 7602, Suid-Afrika

Proefskrif: PhD (Wiskunde)

Desember 2017

Ons beskou twee konsepte van sentraliteit—naamlik, die tussensentraliteit van 'n punt en of dit 'n sentroïed is of nie—in families van eenvoudig gegenereerde en toenemende bome. Albei hierdie konsepte word gedefinieer in terme van paaie binne 'n boom: die tussensentraliteit van 'n punt v is die som, oor pare punte, van die verhouding van kortste paaie wat deur v beweeg; en v is 'n sentroïed indien dit die som van die afstande tot ander punte in die boom minimeer.

Ons vind dat tussensentraliteit in 'n groot, lukrake eenvoudig gegenereerde boom is oor die algemeen lineêr in verhouding tot die grootte n van die boom, en as gevolg van die lang, dun aard van eenvoudig gegenereerde bome, sal die waarskynlikheid dat 'n ewekansige punt kwadratiese-orde tussensentraliteit sal hê verminder soos wat n vermeerder word. Dit lei tot 'n $k$de moment van orde $n^{2k-(1/2)}$, alhoewel daar 'n limietverdeling ontstaan sodra die tussensentraliteit lineêr herskaal word.

Tussensentraliteit in 'n lukrake toenemende boom is ook gewoonlik lineêr, behalwe vir punte naby aan die wortel van die boom, wat tipies sentraliteite het van orde n^2. Die $k$de moment van die tussensentraliteit van enige punt met 'n vaste kode is dus van orde n^{2k}, maar weereens sal die verspreiding van die tussensentraliteit van 'n ewekansige punt konvergeer tot 'n limiet wanneer daar geskaal word met $1/n$.

Om by bekende resultate wat sentroïed punte in eenvoudig gegenereerde bome insluit aan te vul, lei ons ook limietverdelings af, saam met limiete van momente, vir die diepte, kode, en subboomgrootte van die sentroïed naaste aan die wortel in 'n toenemende boom. Die eerste twee van hierdie verdelings is gekonsentreer rondom die wortel, terwyl die laasgenoemde is 'n kombinasie van 'n puntmaat by 1 en 'n dalende digtheid op $[1/2, 1)$.

Daarbenewens wys ons dat die verdelings van die maksimum tussensentraliteit in eenvoudig gegenereerde en toenemende bome konvergeer, op geskikte skalerings, na limietverdelings, en die waarskynlikheid dat die sentroïed ook 'n maksimale tussensentraliteit bereik neig in albei gevalle tot limietkonstantes.

Acknowledgements

By far the person who has had the greatest influence on this work, and to whom I am the most indebted, is Stephan Wagner. His patient, insightful guidance, and his deep understanding of and enthusiasm for the topic at hand, have made putting together this thesis under his supervision a rewarding and (if I'm allowed to say so) thoroughly enjoyable experience. The advice of several anonymous reviewers along the way, and the generosity of the analytic combinatorics community in general, have also been of great value.

There are many others who I consider myself lucky to have had in my life over the last few years as well; and while I am genuinely grateful to all of them, a few enduring names deserve special mention: my parents David and Deborah and the rest of my family; my good friends Jon Ambler and Travis Myburgh; and, for far more than this, God.

I am also grateful to Stellenbosch University and VASTech Ltd. for the many opportunities they have afforded me—this research having been funded by the latter of the two. In particular, I appreciate the encouragement of Marius and Charlotte Ackerman as well as the support of Gavin Gray.

Contents

Chapter 1

Introduction

For a structure that is, at first sight, quite simple and natural—even elegant—graph-theoretic trees are a remarkably rich source of mathematical problems. While many of these are due to their ubiquity in the applied sciences, where they are used as efficient data structures or to encode natural processes, others (such as the ones we consider here) are more theoretical. Our current interest is primarily combinatorial, and here one sees that the subtleties involved in defining trees concretely give rise to a wealth of interesting questions simply regarding the structure of the tree itself—to say nothing of those that involve encodings or bijections with other counting objects.

In what follows, we concern ourselves with questions of centrality in trees—distinct but complementary notions of a 'centre'. To be more specific: we address both betweenness centrality and the concept of the centroid in simply generated and (very simple) increasing trees. Of the four intersections, only the properties of the centroid in simply generated trees have so far been studied in any generality (although the centroid has also been considered in recursive trees, which are a certain kind of increasing tree).

The results we present are rather varied in scope. Although our aim, for the most part, is to describe moments and limiting distributions for parameters of interest, one quickly realises that there is by no means a shortage of these: for example, when considering betweenness centrality in a random tree, we may ask about the root, a specifically labelled node, a random node, or even the node at which the maximum is obtained. A study of centroid nodes, on the other hand, calls for somewhat different parameters, since most trees have only one (and in rare situations, a second) centroid.

Nonetheless, these considerations can be broken up quite neatly: in Chapters 2 and 3 we study the betweenness centrality parameters mentioned above, along with the probability that the centroid and the node of maximal betweenness centrality coincide, in simply generated and increasing trees respectively.[1] The chapter on simply generated trees also touches on betweenness centrality in subcritical graph families, which share many structural characteristics with simply generated trees. In Chapter 4, we consider both the depth and label of the centroid in increasing trees, as well as the size of the subtree rooted at the centroid node (equivalently, the

[1]The results of these chapters have been presented previously in Durant and Wagner (2016, 2017).

size of its ancestral branch).

Broadly, our results are as follows: the betweenness centrality of a randomly chosen node in a large tree—simply generated or increasing—is typically linear in the size n of the tree. This holds for the root node in a simply generated tree or subcritical graph, too. The moments of betweenness-related parameters reveal a different picture, however: the root of a simply generated tree (which is indicative of a random node) or subcritical graph has a betweenness centrality whose kth moment is $\Theta(n^{2k-(1/2)})$. That of any node with a fixed label in an increasing tree, on the other hand, is $\Theta(n^{2k})$. The intuitive reasons for this will be made clear in the coming sections, but we elucidate slightly by saying that whereas it is rare—but not impossible—for a node in a simply generated tree to have betweenness centrality of quadratic order, this is quite normal for nodes with small (i.e., fixed relative to n) labels in an increasing tree. We also consider the maximum betweenness centrality of a tree's nodes, and show that in both classes of trees, the distribution of the maximum converges to a limiting distribution once rescaled by $1/n^2$. This limiting behaviour allows us, in a way, to link our two notions of centrality to one another: it turns out that the probability that the centroid also attains maximal betweenness centrality converges, as $n \to \infty$, to constants close to 0.62 and 0.87 in labelled (simply generated) and recursive (increasing) trees respectively.

Finally, limits of the distributions and moments of the depth, label, and subtree size of the centroid in an increasing tree are derived. Consequences of these derivations are, e.g., that the expected depth of the centroid in random plane-oriented, recursive, or binary increasing trees, respectively, tends to 1/2, 1, and 2; the mean label tends to 7/4, 5/2, and 4; and the expected proportion of the tree accounted for by the centroid's ancestral branch approaches, roughly, 0.13, 0.24, and 0.38. A noticeable trend is that the root is further from the centroid in a binary increasing family than in any other type of increasing tree, and in fact it will follow from the limiting distribution of the centroid's label that the probability of the root being the centroid tends to 0.59, 0.31, and 0, respectively, in the three families mentioned above.

The remainder of this chapter contains descriptions of the chosen tree models and centrality measures, although formal definitions of the tree classes are left to their respective chapters. Of note are Sections 1.1.3 and 1.2.3, since together they sketch the known results regarding the behaviour of the centroid in simply generated trees—to which we have little to add.

1.1 Random tree models

Our general object of interest is a family of trees $\mathcal{T}$, and, usually, the subset $\mathcal{T}_n \subset \mathcal{T}$ of trees made up of n nodes. A tree $T \in \mathcal{T}_n$ is said to have *size* n, denoted by $|T| = n$.[2] A probabilistic model for any given tree parameter then arises quite naturally if one considers trees drawn randomly from $\mathcal{T}_n$. In the most obvious case, this is done in a uniformly random manner, so that each tree is equally likely to be

[2]The variable n will be reserved for the size of a tree of interest throughout this thesis—even when not pointed out explicitly, as in "betweenness centrality of order n^2."

chosen, but both simply generated and increasing trees allow for models in which trees are weighted relative to one another.

Our focus is necessarily on trees in which a single node has been distinguished as the root, because this forms part of the definitions of both simply generated and increasing trees (however the distinction between rooted and unrooted trees—when it is sensible to make it—is usually not particularly important). And on this note, there is one more piece of terminology that should be introduced, since it will be used freely throughout the next few chapters: the *branches* of a tree T at node v are the maximal subtrees of T that do not contain v.[3] When dealing with a specific node v in a rooted tree, the members of the branch of v containing the root are *ancestral* nodes, while members of the remaining branches are *descendants*. Direct ancestors and descendants are called *parents* and *children*, respectively.

1.1.1 Increasing trees

Increasing trees are rooted, labelled trees in which paths away from the root are labelled in increasing order. Their variety stems from a relative weighting scheme, as alluded to above: each family of increasing trees is defined by a set of weights (which may be zero) that are assigned to nodes according to their out-degrees (that is, their number of children), and the weight of a tree is the product of those of its nodes.

One of the most interesting aspects of increasing trees is that there is an important subclass of trees, called *very simple increasing trees*, that can be characterised by a probabilistic growth process: begin with the root node 1, and repeatedly attach nodes to the existing tree according to certain probabilistic rules, determined by the family's out-degree weights. The simplest such family is that of recursive trees, in which each new node is attached uniformly at random to an existing one. Other common families include plane-oriented and binary increasing trees.

In terms of structure, very simple increasing trees are first and foremost distinguished by a height distribution that is concentrated around a mean of order $\log n$ (Drmota, 2009, Chapter 6). This implies that the sizes of the branches in a random tree of size n are well balanced (recall, e.g., that a strict binary tree of size n has height at least $\log n$). In fact somewhat more than this is known: the depths of the nodes in a very simple increasing tree follow a normal distribution with both mean and variance of order $\log n$, and the expected path length of a tree (the sum of the distances from the root to all other nodes) is $\Theta(n \log n)$ (Bergeron *et al.*, 1992).

1.1.2 Simply generated trees

The class of simply generated trees is also one of rooted trees in which each node is weighted according to its out-degree, but without the additional restriction that the labels along any path away from the root form an increasing sequence. Indeed, a family of simply generated trees need not even be labelled. By design, two of the most common combinatorial trees can be seen as families of simply generated trees:

[3] We will use the shorthand "branches of v" for the branches of T at v, along with "branches of T" for the branches of T at its root node.

unlabelled plane—or Catalan—trees, which are enumerated by the Catalan numbers; and non-plane labelled—or Cayley—trees, of which there are n^{n-1} of size n.[4]

Like very simple increasing trees, simply generated trees can be viewed from the perspective of a probabilistic growth process—in this case, that of a phylogenetic (or 'family') tree. Each node gives rise to a number of children, in accordance with a set of relative out-degree weights, and each child is once again the root of a simply generated tree from the given family. This process is the reason that simply generated trees are often referred to as Galton-Watson trees—a correspondence that is concisely described by Aldous (1991*b*).

Unlike increasing trees, simply generated trees are characteristically thin: for example, Meir and Moon (1987) have shown that a typical simply generated tree has up to three branches of interest: the first has height of order $\sqrt{n}$ and size of order n; the second, height and size of orders $\log n$ and $\sqrt{n}$ respectively; and the third has constant-order height, and size of order $\log n$. Another way of stating this thinness is to say that for $h(n) = o(\sqrt{n})$ such that $h(n)$ tends to infinity with n, it is likely that there is a *unique* path of length $h(n)$ from the root that can be extended to order $\sqrt{n}$ (Aldous, 1991*a*).

1.1.3 The continuum random tree

One of the remarkable aspects of simply generated trees is that a certain limiting object appears with high probability when one considers ever-larger random trees. This limit, called the *compact continuum random tree*, was introduced by Aldous (1991*a*), and can be defined in a number of different ways. A precise probabilistic definition is not of any particular use to us here, so we instead give a brief description in terms of its relation to Brownian excursion (that is, Brownian motion conditioned to be 0 at its start and end points, and positive in-between): the continuum random tree is the rescaled infinite tree whose depth-first search distribution is Brownian excursion of duration 2.

The key concepts underlying this link are relatively intuitive: consider random-walk excursions in which positive and negative steps are equally likely, and let R be such an excursion of length $2n$. If positive steps represent movement within a tree from a node to its first unvisited child, and negative steps represent movement towards the root, then R traces out the depth-first search process of a unique rooted, ordered tree of size n. Scaling step width and height by $1/n$ and $\sqrt{n}$ respectively, and letting $n \to \infty$, the random trees constructed in this way converge to a family of infinite trees whose depth-first search process is Brownian excursion of duration 2.

The finite bijection we have described corresponds specifically to the case of unlabelled plane trees in which a node with i children is assigned the weight $\phi_i = 2^{-i}$, since this is the probability of a random walk generating positive steps on i consecutive visits to height h (and thus a node of out-degree i at depth h). A remarkable property of the continuum random tree, however, is that the Brownian excursion distribution holds *regardless* of the family of simply generated trees—the only effect that a change of family has is to scale the excursion function by a factor

[4] We should point out that non-plane unlabelled trees do *not* fall into this class of trees, since their generating functions involve Pólya operators of the form $\phi(y(z), y(z^2), \dots)$. The fourth combination, plane labelled trees, are a simply generated family.

$1/\sigma = \sqrt{\phi(\tau)/\phi''(\tau)}/\tau$ (these variables will be introduced in Chapter 2; see also Aldous (1991*b*, Theorem 2)). Stated another way: all simply generated trees share the same limiting object.

One of the strengths of the continuum random tree is that one can often rephrase questions about the limiting behaviour of finite trees as questions about the continuum random tree itself. For example, the 'thin' shape of simply generated trees carries through to their limiting object, and as such, known results concerning the distances between nodes in a large simply generated tree can also be deduced from the continuum random tree's relation to Brownian excursion. Deductions made in the opposite direction are possible as well: the fact that the probabilistic model for rooted labelled trees is unchanged when a random node is chosen as a new root implies that the continuum random tree is also invariant under random re-rooting (Aldous, 1991*b*).

As a precursor to Section 1.2.3, let us briefly state that our chief interest in the continuum random tree is due to the fact that the bijection described above can be extended to include a third process—random triangulations of the circle (Aldous, 1994*a*,*b*). In terms of node centrality, the triangulation perspective is particularly interesting, because the triangle in which the centre of the circle is contained corresponds to the branchpoint that arises at the centroid of a tree.

1.2 Centrality measures

The term 'centrality' as we use it here simply refers to the idea that certain nodes are nearer to a graph's central point than others, where the idea of a 'centre' is sometimes based on intuitive, or even aesthetic, properties. Measures of a node's centrality are often interpreted in an applied sense—especially in the network science community—as an indication of how 'important' that node is to the graph.

There are various concrete definitions of centrality, each giving rise to a different measure. The simplest is arguably that of degree centrality, in which a node's centrality is nothing but its degree (and this problem has of course been studied for classes of random trees: see Bergeron *et al.* (1992) and Flajolet and Sedgewick (2009, Section VII.3.2)). The two measures we consider here are the most common path-based measures: betweenness and closeness centrality (Freeman, 1978), although we choose to approach the latter from the point of view of the centroid of a tree. There are more complex examples of centrality as well, the most notable being those based on random walks: Katz and eigenvector centrality, and even PageRank, which forms (or perhaps, formed) the core of Google's search algorithm.

1.2.1 Betweenness centrality

Let G be a graph; then the *betweenness centrality* of a node v is the sum over pairs $\{u, w\}$ of nodes other than v that counts for each pair the fraction $b_{uw}(v)$ of undirected shortest paths between them that pass through v:

$$b(v) = \sum_{\{u,w\}} b_{uw}(v),$$

where $0 \leq b_{uw}(v) \leq 1$. If $G = T$ is in fact a tree, then there is only one path between any two nodes, and $b(v)$ is the total number of paths that pass through v. In this case, the betweenness centrality can be expressed in terms of the sizes of v's branches T_i:

$$b(v) = \sum_{i<j} |T_i||T_j|. \tag{1.1}$$

This is precisely the number of ways to choose two unordered nodes from distinct branches of v. We also briefly note that the betweenness centrality of any node in a graph of size n is bounded from above by $\binom{n-1}{2}$.

The notion of betweenness centrality was introduced by Freeman (1977), and subsequently presented as part of a trio of basic centrality measures (Freeman, 1978), the other two being degree centrality and closeness centrality. For more on betweenness centrality in the context of graphs, we refer the reader to Newman (2010, Section 7.7). More mathematical treatments are also available (Gago *et al.*, 2015). We also highlight two practical applications of betweenness centrality to real-world networks (graphs): the first to the problem of 'community detection' (Girvan and Newman, 2002), and the second as a tool to classify networks (Goh *et al.*, 2002).

1.2.2 Centroid nodes

A more classical way, perhaps, of measuring the centrality of a node in a tree (or graph) is in terms of its distances to other nodes. Two similar sets of 'central' nodes are of immediate interest: those for which the maximum distance to any other node (often called the *eccentricity*) is minimised—known as *centres*—and those which minimise the average distance to another node—called *centroids*. We will focus on the latter.

In the network science literature, the average distance from a node v to the other nodes in a graph is generally referred to as the (inverse) *closeness centrality* of v, and, like betweenness centrality, it appears in a few interesting practical applications, such as the identification of the source of a rumour in a network (Shah and Zaman, 2011). From this perspective, a network scientist might find it fitting to think of our results on the probability that a centroid also has maximum betweenness centrality (Sections 2.3 and 3.5) in terms of the coincidence of maximum betweenness and closeness centrality instead. The 'centroid' terminology we have used, however, is far more natural when dealing with trees, and we will continue to use it exclusively.

On that note, the definition we have given for a centroid node is not the most commonly presented one: a node in a tree is usually defined as a centroid if each of its branches contains at most half of the tree's nodes. The two definitions are equivalent (Zelinka, 1968), but this latter property is generally more useful when it comes to analyses—our own included. Another indispensable fact, due to Jordan (1869), is that every tree has either one or two centroids. In the latter case, the size n of the tree must be even, the centroids are adjacent, and the largest branch of each has exactly $n/2$ nodes.

Combinatorially, there are also a number of interesting results regarding centroids of random trees, especially when it comes to simply generated trees, and we give an account of some of the most noteworthy of these below. Although work has been done to investigate the behaviour of the centroid in increasing trees, this has

so far been restricted to recursive trees. We give an overview of those results in the introduction to Chapter 4.

1.2.3 The centroid in simply generated trees

For now, let $\mathcal{Y}_n$ denote a family of simply generated trees of size n. To avoid possible confusion, we remark first that most combinatorial results deal with the *nearest* centroid, either to the root or the node in question. This is nothing more than a small technicality though, since the probability that a random tree $T \in \mathcal{Y}_n$ has two centroids (when n is even) decreases as $1/\sqrt{n}$ (Meir and Moon, 2002).

As we have already mentioned, the scale of the distances in simply generated trees is inevitably of order $\sqrt{n}$: in particular, the height of T, the depth of a random node, the distance between two random nodes, and, notably, the distance from the centroid to a random node, are all $\Theta(\sqrt{n})$ (see Flajolet and Sedgewick (2009, Section VII.10.2) and Moon (1985)). In fact far more than this is known: Aldous, to complement his treatment of the continuum random tree as the limiting object of all simply generated trees, has shown that in a manner somewhat similar to that in which a tree and its branches can be viewed recursively—as a number of independent random trees attached to a root—the branches of the centroid, in the limit $n \to \infty$, behave like random trees themselves, albeit conditioned on a certain size distribution. A pleasing property of the centroid-based decomposition is that it gives rise to multiple large branches, and thus remains 'visible' on the macro scale. The root-branch decomposition, on the other hand, gives rise to a dominant branch of order n, and secondary and tertiary branches of orders $\sqrt{n}$ and $\log n$ respectively (Meir and Moon, 1987). As such the root, when viewed as a branching point, becomes less and less apparent as $n \to \infty$. The result of Aldous that is of the most interest to us is the following:

Theorem 1.1 (Aldous (1994*a*, Theorem 4))**.** *Let T_1, T_2, and T_3 be the three largest branches, randomly ordered, of the centroid in a tree $T \in \mathcal{Y}_n$. Then as $n \to \infty$, the sum of the sizes of these three branches is asymptotic to n. In particular, there is convergence in distribution of $(|T_1|, |T_2|, |T_3|)/n$ to the continuous distribution with support $\{0 < x_1, x_2, x_3 < 1/2, x_1 + x_2 + x_3 = 1\}$ and density:*

$$f(x_1, x_2, x_3) = \frac{1}{12\pi}(x_1 x_2 x_3)^{-3/2}.$$

In the limit each branch, scaled by $1/|T_i|$, is an independent copy of the continuum random tree.

Not only will the centroid of the limiting object almost surely have exactly three branches, the distribution of the sizes of these branches is known explicitly.

There is another, similar result worth mentioning here: if one chooses three random nodes from a large tree, then with high probability there will be a unique node v such that each of the chosen nodes lies in a distinct branch of v. (The alternative—that one of the nodes lies on the path between the other two—has probability $\Theta(1/\sqrt{n})$.) Furthermore, an analogue of Theorem 1.1 holds for these branches as well. Interestingly, the probability that v is also the centroid of the tree tends to a value near 0.121.

A few refinements of these results were given by Meir and Moon (2002), who, for example, showed that the expected sizes of the root branch and two largest descendent branches of the centroid converge to the rough values 0.414, 0.438, and 0.146 respectively.

Our final comment on the centroid of a simply generated tree will only be applied in Section 2.3, but we mention it here for the sake of completeness. The bijection between random walk excursions and random trees—which carries through to Brownian excursion and the continuum random tree—can be extended to a three-way mapping by considering random triangulations of the regular n-gon. The link between trees and triangulations can be most clearly seen in the case of a random unlabelled plane tree: such a tree can be mapped bijectively to a binary tree, the internal nodes of the binary tree each give rise to three branches, and these branch-points lead to the triangles of the triangulation. The size of a branch corresponds to the number of nodes contained in the segment of the n-gon marked off by an edge of a triangle (see Aldous (1994*a*,*b*) for more information). In the limit $n \to \infty$, one obtains, in a sense, random triangulations of a circle. From this perspective, the centroid is 'seen' as the triangle containing the centre of the circle, and the case of two centroids occurs when an edge of some triangle is a diameter.[5]

The elegance of this bijection is most apparent when one considers Theorem 1.1, since it implies that one can first generate the triangle corresponding to the centroid, and then, considering each of the three segments independently, continue generating 'centroid' triangles recursively. On the other hand, this also provides an intuitive way of thinking about the branchpoints resulting from choices of three random nodes: since each such branchpoint corresponds to a triangle, the branchpoint can be viewed as the recursive centroid of some earlier centroid's branch, a certain number of steps removed from the tree's actual centroid.

We finish this introduction with two comparative figures that will hopefully provide the reader with a feeling for both the characteristic shapes of simply generated and increasing trees, as well as the manner in which centrality varies within them.

[5]The fact that triangulations are counted by the Catalan numbers, combined with Stirling's formula, provides a simple argument for the $1/\sqrt{n}$ order of the probability of this event.

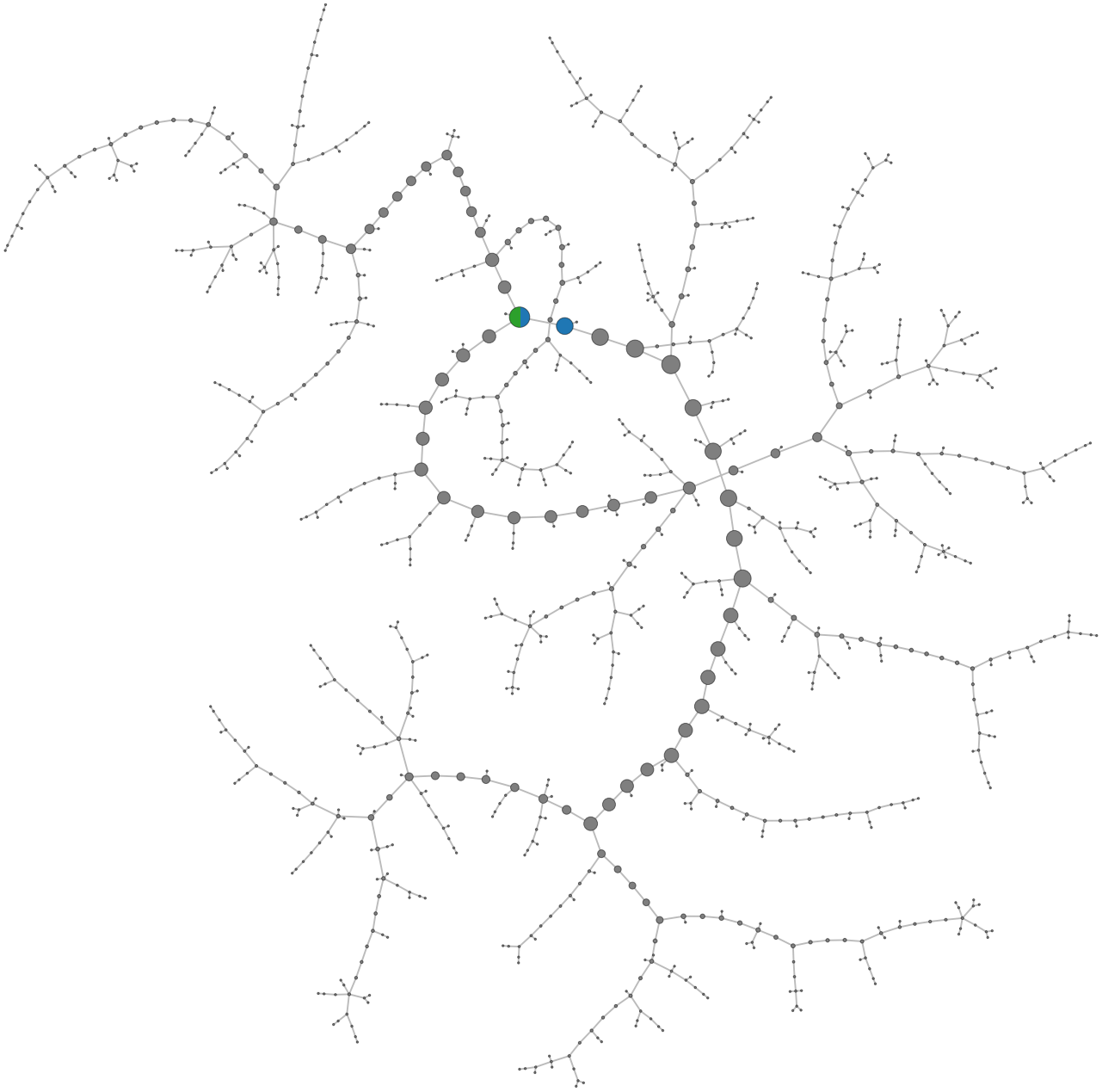


Figure 1.1: A random non-plane labelled (Cayley) tree of size $n = 1000$. Nodes are scaled according to their betweenness centralities, and those that are centroids or have maximal betweenness centrality are coloured blue and green respectively. A few important features are apparent: there are three large, spine-like branches that extend outwards from one of the centroid nodes (which coincides, in this case, with the node of maximum betweenness centrality), and it is along these spines that the nodes of largest betweenness centrality are located. Many of these nodes—of which there are $O(\sqrt{n}\,)$—must have betweenness centralities that are quadratic in n, and indeed, the probability of such quadratic values scales as $1/\sqrt{n}$. The remaining nodes all have noticeably smaller betweenness centralities.

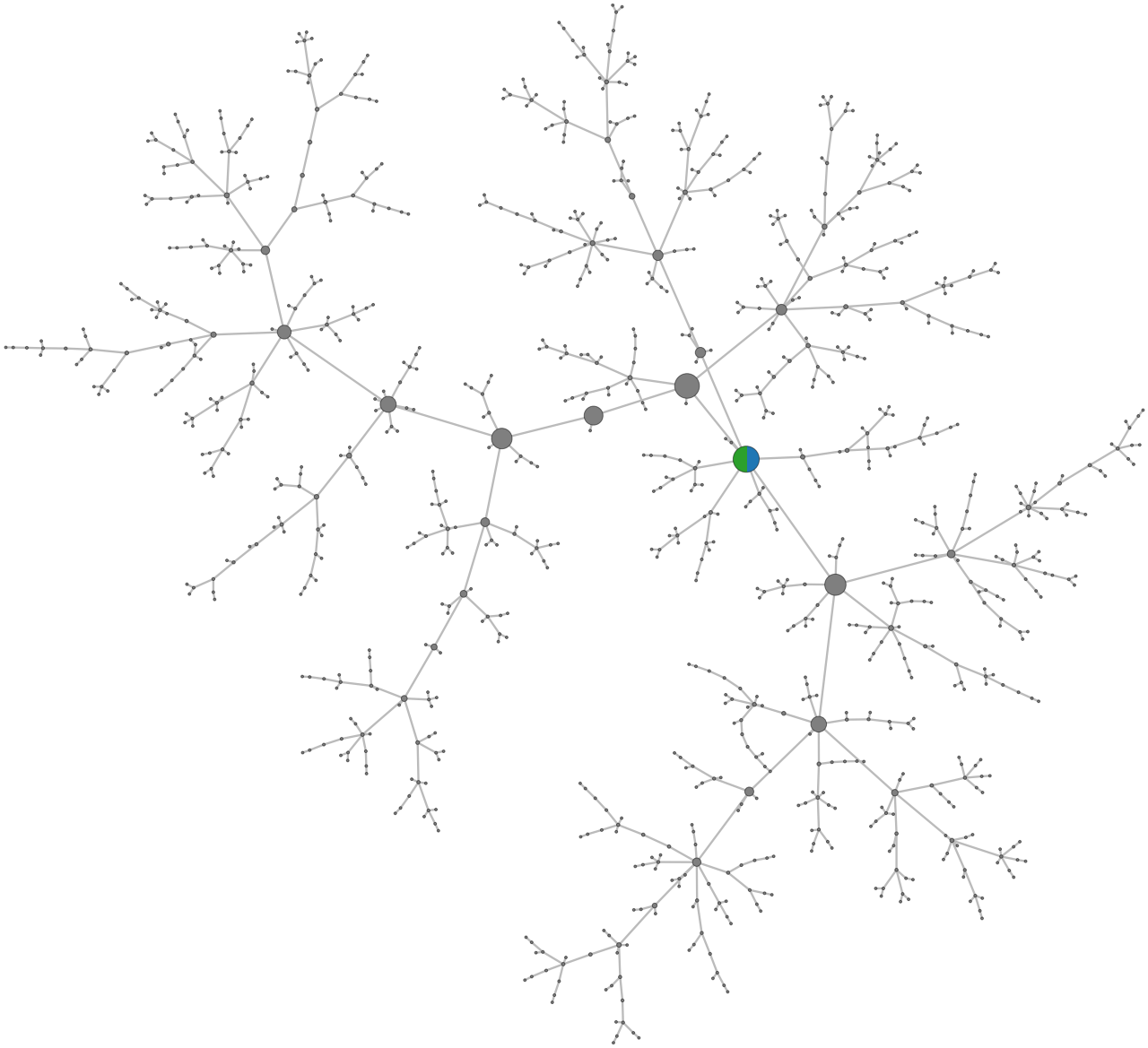


Figure 1.2: A random recursive tree with $n = 1000$ nodes. Once again the size of a node represents its betweenness centrality, and the centroid and node of maximum betweenness centrality (which again coincide) are depicted in green and blue respectively. In this particular example, the label of the centroid is 2, and the root is the (tiny) node of degree 4 positioned directly below it. Although it is not visible here, the ten largest nodes all have labels less than 20, however, apart from such small-labelled nodes, which lie close to the root, betweenness centrality is for the most part linear in n. (This concentration of large centralities around the root is in contrast to simply generated trees, where the root plays no significant role.) Finally, as one would expect, the shape of this tree is rather more balanced than that of Figure 1.1.

Chapter 2

Betweenness Centrality in Simply Generated Trees

2.1 Introduction and preliminaries

Our first class of trees, the simply generated families, are counted by generating functions with the interesting property that they satisfy analytic expansions in powers of 1/2—that is, expansions of the form:

$$y(x) = a_0 + a_1\sqrt{1-\frac{x}{\rho}} + a_2\left(1-\frac{x}{\rho}\right) + \cdots .$$

We mention this here (it will be covered again below) for two reasons: there is a tree-like class of graphs, known as subcritical graphs, that possesses this property as well; and this so-called square-root expansion determines the overarching shape of both classes. Most notably, a large, random simply generated tree or subcritical graph is 'thin', in that it typically has one root branch that contains most of the tree's nodes.

The unbalanced nature of the root's branches has a clear implication on its betweenness centrality: paths through the root more often than not connect nodes in the largest branch to the rest of the tree. Having already defined betweenness centrality in Section 1.2.1, we can state this idea more plainly with the help of equation (1.1): if one of the branches of a node v (without loss of generality, T_1) is large, while the remaining branches together contain a relatively small number k of nodes, then $b(v)$ is dominated by paths between T_1 and the other branches. If the branch sizes are $n_1, \ldots, n_d$, and $n_2 + \cdots + n_d = k$ remains fixed as the size of the tree tends to infinity, then we have:

$$b(v) = (n-k-1)\sum_{i=2}^{d} n_i + \sum_{1<i<j} n_i n_j = nk + O\left(k^2\right). \tag{2.1}$$

We will see that this imbalance has a determining effect on the distribution of the betweenness centrality of the majority of nodes in both simply generated trees and subcritical graphs. In particular, the linearly rescaled betweenness centrality of the root of a simply generated tree converges to an explicit limiting distribution as the size n of the tree increases. Its kth moment, however, is of order $n^{2k-(1/2)}$. A similar, linearly rescaled, limiting distribution is also found for the betweenness centrality of a random node in a simply generated tree, and the existence of a limiting distribution for the quadratically rescaled maximum betweenness centrality in a random non-plane labelled tree is proved. Along with this, it is shown that the probability that the maximum is attained by the centroid tends to a constant (approximately 0.62) as $n \to \infty$. Finally, the chapter concludes with a relatively brief look at subcritical graphs: the kth moment of the betweenness centrality of the root is again of order $n^{2k-(1/2)}$, and we can show that as n grows, non-linear betweenness centralities become increasingly rare.

2.1.1 Simply generated trees

This section contains a small overview of the main analytic properties of simply generated families; for a more thorough treatment the reader is referred to, e.g., Meir and Moon (1978, 1987), Flajolet and Sedgewick (2009, Section VII.3), or Drmota (2009, Section 1.2).

A family of simply generated trees is defined concretely by coupling a non-negative weight ϕ_i to each node in a rooted tree according to its out-degree i, and then letting the weight $\omega(T)$ of the tree be the product of the per-node weights. The resulting family of trees can be counted using the generating function:

$$y(x) = \sum_{T \in \mathcal{T}} \omega(T) x^{|T|} = x\phi(y(x)), \tag{2.2}$$

in which $\phi(u) = \sum_i \phi_i u^i$. In particular, one recovers the classes of binary, plane, and labelled trees via the weight functions $\phi(u) = (1+u)^2$, $(1-u)^{-1}$, and $\exp(u)$ respectively.

Under a few technical conditions on $\phi(u)$ (see Meir and Moon (2002, Theorem 2.1)), including the existence of a unique positive solution τ of $\phi(\tau) = \tau\phi'(\tau)$ within the radius of convergence of ϕ, every class of simply generated trees has the characteristic property that its generating function $y(x)$ has a dominant singularity at $x = \rho$, determined by $\rho = \tau/\phi(\tau) = 1/\phi'(\tau)$. Furthermore, $y(x)$ satisfies a square-root expansion around this singularity:

$$y(x) = \tau - \frac{\tau\sqrt{2}}{\sigma}\sqrt{1 - \frac{x}{\rho}} + O\left(1 - \frac{x}{\rho}\right), \tag{2.3}$$

in which $y(\rho) = \tau$ and $\sigma = \tau\sqrt{\phi''(\tau)/\phi(\tau)}$.[1] Because of this, many interesting properties of simply generated trees can be deduced almost mechanically using singularity analysis. The total weight of trees of size n, for example, is:

$$y_n = [x^n]y(x) \sim \frac{\tau\rho^{-n}}{\sigma\sqrt{2\pi n^3}}.$$

[1] This σ is the standard deviation of the corresponding Galton-Watson process, as pointed out by Aldous (1991*b*).

The expected height of one of these trees is $\Theta(\sqrt{n}\,)$, and the expected number of nodes at a fixed distance h from the root is only linear in h (Flajolet and Sedgewick, 2009). Another interesting result, considering that we are about to address the betweenness centrality of the root node, is that the root of a simply generated tree is known to have up to three 'major' branches, with mean sizes of orders n, $\sqrt{n}$, and $\log n$ (Meir and Moon, 1987). In light of this, one might expect that the betweenness centrality of the root will be dominated by paths between the two largest branches, of which there are $\Theta(n^{3/2})$. In the following section, we show that this view is only partially complete, and that the kth moment of the root's betweenness centrality is in fact $\Theta(n^{2k-(1/2)})$.

2.2 The betweenness centrality of the root

Let $\mathcal{B}(T)$ denote the betweenness centrality of the root node in a simply generated tree T. Instead of thinking of $\mathcal{B}(T)$ as the number of paths through the root, it will be useful to view it as the number of ways to choose two nodes from distinct branches of T. Then analytically, this provides us with a clear way forward, since the act of distinguishing a node in a tree that has the generating function $y(x)$ corresponds to the 'pointing' operation $y(x) \to xy'(x)$. To begin with, we use this fact to derive the moments of $\mathcal{B}(T)$.

2.2.1 Moments of the betweenness centrality of the root

Theorem 2.1. *The expected betweenness centrality of the root node in a simply generated tree of size n is $\Theta(n^{3/2})$, satisfying:*

$$\mathrm{E}_n(\mathcal{B}) \sim \frac{\sigma}{4}\sqrt{2\pi n^3}.$$

Proof. The generating function of trees in which two of the root's branches have been replaced with pointed branches counts all the paths through roots in $\mathcal{T}_n$, and can be constructed explicitly:

$$\begin{aligned} H(x) = \sum_{T\in\mathcal{T}} \mathcal{B}(T)\omega(T)x^{|T|} &= x\sum_{i\geq 2}\phi_i\binom{i}{2}x^2y'(x)^2y(x)^{i-2} \\ &= \frac{x^3}{2}y'(x)^2\phi''(y(x)). \end{aligned}$$

Taking advantage of the square-root expansion of $y(x)$ at $x=\rho$, and the fact that $\phi(u)$ is analytic at $u=\tau$, the asymptotic form of $H(x)$ is:

$$\begin{aligned} H(x) &\sim \frac{\rho}{4}\left(\frac{\tau}{\sigma}\right)^2\phi''(\tau)\left(1-\frac{x}{\rho}\right)^{-1} \\ &= \frac{\tau}{4}\left(1-\frac{x}{\rho}\right)^{-1}. \end{aligned}$$

Since $[x^n]H(x) = \sum_T \mathcal{B}(T)\omega(T)$, the result follows as $[x^n]H(x)/y_n$, which one can extract using Theorem VI.1 of Flajolet and Sedgewick (2009). □

As explained above, root betweenness centralities of order $n^{3/2}$ are by no means surprising if one considers the typical branch structure of a simply generated tree. However it is certainly possible to construct non-typical trees—stars, for example—in which the root obtains the quadratic upper bound $\mathcal{B}(T) = \binom{n-1}{2}$, so one can also explain Theorem 2.1 by saying that although it is unlikely (of order $n^{-1/2}$) for the root to have two large branches[2], this event nonetheless dominates the asymptotic behaviour of its betweenness centrality. By this reasoning, one might anticipate that the kth moment of $\mathcal{B}(T)$ will be of order $n^{2k-(1/2)}$.

In deriving these higher-order moments, the following lemma will prove useful, both for simply generated trees and for subcritical graphs, which will be treated in Section 2.4.

Lemma 2.1. *Let $\mathcal{C}$ be a 'tree-like' family, in that it is counted by a generating function $c(x) = x\phi(f(x))$ such that both $c(x)$ and $f(x)$ permit square-root expansions around a common singularity $x = \rho$, and $\phi(u)$ is analytic at $u = f(\rho)$. Then the substitution of m 'branches' $f(x)$ with pointed branches—each of which may possibly distinguish multiple nodes, and which in total contain d distinguished nodes—yields a generating function whose dominant term is $\Theta((1-(x/\rho))^{-d+(m/2)})$.*

It follows from this lemma that when choosing d nodes from a simply generated tree, the resulting asymptotic behaviour depends only on the configuration that affects the fewest branches.

Proof. The generating function obtained after the substitution described above is a linear combination of terms of the form:

$$x\left(\prod_{i=1}^{m} \widehat{f}_{d_i}(x)\right)\phi^{(m)}(f(x)), \tag{2.4}$$

in which $\widehat{f}_{d_i}(x)$ is the generating function of the ith substituted branch, which has d_i distinguished nodes:

$$\widehat{f}_{d_i}(x) = x\frac{d}{dx}\widehat{f}_{d_i-1}(x) = \sum_{l=1}^{d_i} \left\{ {d_i \atop l} \right\} x^l f^{(l)}(x).$$

Here, $\left\{ {j \atop l} \right\}$ denotes the Stirling numbers of the second kind. It is these branches that determine the overall asymptotic behaviour of the expression in (2.4), since $f(x)$ permits a square-root expansion. Specifically, $f^{(l)}(x)$ is of order $(1-(x/\rho))^{-l+(1/2)}$, and:

$$\widehat{f}_{d_i}(x) \sim x^{d_i} f^{(d_i)}(x) \sim K_{d_i}\left(1-\frac{x}{\rho}\right)^{-d_i+(1/2)}$$

for some constant K_{d_i}. The result follows from equation (2.4) because $\sum_i d_i = d$ and $\phi(u)$ is analytic at $u = f(\rho)$. □

[2] A counting exercise (using, e.g., binary plane trees) shows that the probability of this event is $\Theta(n^{-1/2})$. Alternatively, one can argue heuristically from the fact that the probability of choosing three random nodes and having one of them lie on the path between the other two is of order $n^{-1/2}$ (Aldous, 1994*a*).

Theorem 2.2. *The kth moment of the betweenness centrality of the root node in a simply generated tree of size n is $\Theta(n^{2k-(1/2)})$, and satisfies, for $k \geq 1$:*

$$\mathrm{E}_n(\mathcal{B}^k) \sim \frac{\sigma}{2^{4k-2}} \binom{2k-2}{k-1} \sqrt{2\pi n^{4k-1}}.$$

Proof. We are trying to derive the mean of the function $\mathcal{B}(T)^k$, which can be expanded as:

$$\mathcal{B}(T)^k = \left(\sum_{i<j} |T_i||T_j| \right)^k = \sum_{i<j} |T_i|^k |T_j|^k + \cdots + K \sum_{i_1<\cdots<i_{2k}} |T_{i_1}| \cdots |T_{i_{2k}}|,$$

(where K is some constant that depends on k), since $\mathcal{B}(T)^k$ involves k chances to choose a pair of branches. Each of the sums in the above equation can be interpreted as a selection of $2k$ nodes from a number of branches, and their means can be computed by constructing the corresponding generating functions; however Lemma 2.1 tells us that the term involving the fewest branches will have the greatest asymptotic order. With this in mind, we can simplify the generating function that sums $\mathcal{B}(T)^k$ over trees of size n to:

$$H_k(x) = \sum_{T \in \mathcal{T}} \mathcal{B}(T)^k \omega(T) x^{|T|} \sim \sum_{T \in \mathcal{T}} \left(\sum_{i<j} |T_i|^k |T_j|^k \right) \omega(T) x^{|T|}.$$

This counts, for every tree, the number of ways to choose two branches and distinguish k (not necessarily distinct) nodes in each, and is represented symbolically as:

$$\begin{aligned} H_k(x) &\sim \frac{x^{2k+1}}{2} y^{(k)}(x)^2 \phi''(y(x)) \\ &\sim \tau \left(\frac{(2k-2)!}{2^{2k-1}(k-1)!} \right)^2 \left(1 - \frac{x}{\rho} \right)^{-2k+1}. \end{aligned}$$

As in the proof of Theorem 2.1, the desired quantity is $[x^n]H_k(x)/y_n$. □

The second moment of the betweenness centrality of the root is of a greater asymptotic order than the mean, and thus the variance is as well:

$$\mathrm{V}_n(\mathcal{B}) \sim \frac{\sigma}{32} \sqrt{2\pi n^7}.$$

Table 2.1 gives some indicative values for a few common simply generated families.

2.2.2 A limiting distribution for the root node

Although betweenness centralities of order n^2 appear to dominate the moments of $\mathcal{B}(T)$, the fact that the probability of such large values occurring is $\Theta(n^{-1/2})$ implies that these events become increasingly rare as $n \to \infty$. In this section we make this idea more rigorous by showing that there is a limiting distribution for the *linearly* scaled betweenness centrality of the root, $\mathcal{B}(T)/n$. Stated differently: trees with one

Tree	$\phi(u)$	τ	ρ	σ	$\mathrm{E}_n(\mathcal{B})$	$\mathrm{V}_n(\mathcal{B})$
binary	$(1+u)^2$	1	$1/4$	$1/\sqrt{2}$	$\sqrt{\pi n^3}/4$	$\sqrt{\pi n^7}/32$
plane	$(1-u)^{-1}$	$1/2$	$1/4$	$\sqrt{2}$	$\sqrt{\pi n^3}/2$	$\sqrt{\pi n^7}/16$
labelled	$\exp(u)$	1	$1/e$	1	$\sqrt{2\pi n^3}/4$	$\sqrt{2\pi n^7}/32$

Table 2.1: Lead-order asymptotics for the mean and variance of the betweenness centrality of the root node in selected families of simply generated trees.

large root branch—of size linear in n—are sufficient to describe the distribution of $\mathcal{B}(T)$ when n is large enough, which is in agreement with other known results about the unbalanced nature of simply generated trees.

To prove this, we define subclasses of trees $\mathcal{L}_k \subset \mathcal{T}$ in such a way that the trees in $\mathcal{L}_k$ have one dominant branch, along with a few small branches of total size k. Formally, $(\mathcal{L}_k)_n$ consists of trees of $\mathcal{T}_n$ with one distinguished branch of size $n-k-1$. (Note that a tree may thus *a priori* belong to more than one subclass.) For fixed k, the root nodes of trees in $\mathcal{L}_k$ have predictable, linear-order betweenness centrality, and in the limit $n \to \infty$, the classes $(\mathcal{L}_k)_n$ together describe $\mathcal{T}_n$.

Theorem 2.3. *The linearly scaled betweenness centrality of the root node in a random tree of size n, $\mathcal{B}(\mathcal{T}_n)/n$, converges in distribution to the discrete random variable $\mathcal{B}_\star$ supported by $\mathbb{Z}_{\geq 0}$ and with mass function:*

$$\mathrm{P}(\mathcal{B}_\star = k) = \rho^{k+1}[x^k]\phi'(y(x)) \sim \sigma\Big(2\pi k^3\Big)^{-1/2},$$

in which the asymptotic expression holds as $k \to \infty$.

Proof. Firstly, we reiterate that the betweenness centrality of the root of a tree $T \in (\mathcal{L}_k)_n$ is of linear order for large n and constant k: if the root has a branch of size $n-k-1$, while the other branches contain k nodes, then by equation (2.1) we have $\mathcal{B}(T) = nk + O(k^2)$. Secondly, note that $(\mathcal{L}_k)_n \cap (\mathcal{L}_l)_n = \emptyset$ if $n > k+l+1$, so that for large enough n, any two subclasses $\mathcal{L}_k$ and $\mathcal{L}_l$ are disjoint. Finally, one must show that the probability of a random tree $T \in \mathcal{T}_n$ belonging to $(\mathcal{L}_k)_n$ tends to the constant probability $p_k = \mathrm{P}(\mathcal{B}_\star = k)$ as n grows, and that the sum of these limiting probability masses is 1.

We begin by considering a generating function $L_k(x)$ that counts the trees of a subclass $\mathcal{L}_k$ according to their sizes: it must account for a single branch of variable size (and its i possible points of attachment), as well as the $[x^k]y(x)^{i-1}$ configurations of the remaining (non-root) nodes:

$$\begin{aligned} L_k(x) &= x^{k+1}y(x)\sum_{i\geq 1} i\phi_i[x^k]y(x)^{i-1} \\ &= x^{k+1}y(x)[x^k]\phi'(y(x)). \end{aligned}$$

(Note that the maximum root degree of a tree in $\mathcal{L}_k$ is $k+1$, accounted for by the fact that $[x^k]y(x)^{i-1} = 0$ whenever $i-1 > k$.) From this generating function, it is evident that the probability of a tree belonging to $\mathcal{L}_k$ tends to:

$$p_k = \lim_{n\to\infty} \frac{[x^n]L_k(x)}{y_n} = \rho^{k+1}[x^k]\phi'(y(x)).$$

The sum of these constants is indeed 1:

$$\sum_{k\geq 0} p_k = \rho\phi'(y(\rho)) = 1.$$

Thus the limiting distribution of $\mathcal{B}(T)$ can be fully described using only the limiting behaviour of the subclasses $\mathcal{L}_k$. Specifically, for fixed $k \geq 0$ and every $0 < \varepsilon < 1$:

$$\mathrm{P}_n(|(\mathcal{B}/n) - k| < \varepsilon) \xrightarrow[n\to\infty]{} p_k.$$

The asymptotic form of p_k follows from an expansion of $\phi(u)$ around $u = \tau = y(\rho)$. □

2.2.3 A limiting distribution for random nodes

The previous sections dealt specifically with the betweenness centrality of the *root* node in a simply generated tree, but the constructive idea of Section 2.2.2 can be used to obtain a limiting distribution for the betweenness centrality of a random node as well. In the exceptional case of labelled trees (with $\phi(u) = \exp(u)$), all of the preceding results hold for non-root nodes automatically, because each unrooted tree of size n gives rise to exactly n distinct rooted trees, implying that iteration over the nodes of unrooted labelled trees is equivalent to iteration over the roots of rooted labelled trees. In general, however, such a mapping does not hold for other simply generated trees. Still, we can show that like the root node, a randomly chosen node usually has betweenness centrality of linear order. Let the random variable $\mathcal{R}(T)$ denote the betweenness centrality of a random node in T, so that $\mathrm{P}_n(\mathcal{R} = k)$ is the proportion of nodes in $\mathcal{T}_n$ that have betweenness centrality k.

Theorem 2.4. *The linearly scaled betweenness centrality of a randomly chosen node in a simply generated tree of size n, $\mathcal{R}(\mathcal{T}_n)/n$, converges in distribution to the discrete random variable $\mathcal{R}_\star$ with support $\mathbb{Z}_{\geq 0}$ and mass function:*

$$\mathrm{P}(\mathcal{R}_\star = k) = \frac{\rho^{k+1}}{\tau}[x^{k+1}]y(x) \sim \frac{4}{\sigma}\Big(2\pi k^3\Big)^{-1/2},$$

where the asymptotic expression holds for $k \to \infty$.

The proof of Theorem 2.4 is mostly similar to that of Theorem 2.3, the corresponding result for root nodes, except that in addition to its descendent branches, a non-root node v also has an 'ancestral' branch that contains the root. The idea is to let this ancestral branch be large, and to share a fixed number k of nodes among v's other branches.

Proof. Any node v with k descendants in a tree $T \in \mathcal{T}_n$ can be viewed as a leaf node of a rooted tree of size $n - k$ (its ancestral branch) to which a forest of size k has been grafted. If $(\mathcal{M}_k)_n$ is the resulting subclass of trees, its generating function must account for the $[x^k]\phi(y(x))$ configurations of the smaller branches, as well as the selection of a leaf from a tree of size $n - k$. The latter part can be derived from a bivariate generating function $y(x, u)$ that marks the leaves of every tree with an auxiliary variable u (see Drmota (2009, Section 3.2.1) or, more generally, Flajolet

and Sedgewick (2009, Chapter 3)), since taking the partial derivative of $y(x,u)$ with respect to u and then setting $u=1$ yields a generating function that counts, for each tree, the possible points of attachment for our forest of size k. The entire generating function of $\mathcal{M}_k$ is thus:

$$M_k(x) = \Big([x^k]\phi(y(x))\Big)x^k \times \frac{1}{\phi_0}\frac{d}{du}\big[y(x,u)\big]_{u=1},$$

in which $y(x,u) = x\phi(y(x,u)) + (u-1)\phi_0 x$. The presence of ϕ_0^{-1} in the above equation removes the weight that was assigned to the chosen leaf node, since a new weight will be assigned to it along with the grafted forest $\phi(y(x))$.

As in the proof of Theorem 2.3, the node of interest has betweenness centrality $nk + O(k^2)$. Furthermore, for $k \neq l$, any two subclasses $(\mathcal{M}_k)_n$ and $(\mathcal{M}_l)_n$ are disjoint. To see that, in the limit $n \to \infty$, a tree with a distinguished node has probability $q_k = \mathrm{P}(\mathcal{R}_\star = k)$ of belonging to $\mathcal{M}_k$, we need to express $M_k(x)$ asymptotically. Quickly note that by differentiating $y(x) = x\phi(y(x))$, we have $(1 - x\phi'(y(x)))^{-1} = xy'(x)y(x)^{-1}$. With this in mind:

$$\frac{d}{du}\big[y(x,u)\big]_{u=1} = \phi_0 x(1 - x\phi'(y(x)))^{-1} \sim \phi_0 \frac{\rho}{\sigma\sqrt{2}}\left(1 - \frac{x}{\rho}\right)^{-1/2},$$

as $x \to \rho$. This grants us the desired expression for $M_k(x)$, with which the limiting probability q_k can be derived:

$$q_k = \lim_{n\to\infty} \frac{[x^n]L_k(x)}{ny_n} = \frac{\rho^{k+1}}{\tau}[x^{k+1}]y(x).$$

Note finally that the q_k sum to 1:

$$\sum_{k\geq 0} q_k = \frac{1}{\tau}\sum_{k\geq 0}\rho^{k+1}[x^{k+1}]y(x) = \frac{1}{\tau}y(\rho) = 1.$$

Altogether, we have, for fixed k and every $0 < \varepsilon < 1$:

$$\mathrm{P}_n(|(\mathcal{R}/n) - k| < \varepsilon) \xrightarrow[n\to\infty]{} q_k. \qquad \square$$

Table 2.2 lists values of the limiting probabilities for root and random nodes respectively, for some common trees. Observe that these probabilities are equal for the family of labelled trees, as expected.

The final section on simply generated trees covers the betweenness centrality of the centroid node and, more generally, the maximum betweenness centrality in a tree. Since centroids are the other notion of centrality of most interest to us, this next section—along with a similar one to be found in Chapter 3—is notable for its intersection of the two ideas.

2.3 Maximum betweenness centrality and the centroid

So far we have shown that although betweenness centrality in random simply generated trees is for the most part of linear order, the average betweenness centrality

Tree	$\phi(u)$	σ	$\mathrm{P}(\mathcal{B}_\star = k)$	$\mathrm{P}(\mathcal{R}_\star = k)$
binary	$(1+u)^2$	$1/\sqrt{2}$	$2^{-(2k+1)}\frac{1}{k+1}\binom{2k}{k}$	$4^{-(k+1)}\frac{1}{k+2}\binom{2k+2}{k+1}$
plane	$(1-u)^{-1}$	$\sqrt{2}$	$4^{-(k+1)}\frac{1}{k+2}\binom{2k+2}{k+1}$	$2^{-(2k+1)}\frac{1}{k+1}\binom{2k}{k}$
labelled	$\exp(u)$	1	$e^{-(k+1)}\frac{(k+1)^{k-1}}{k!}$	$e^{-(k+1)}\frac{(k+1)^{k-1}}{k!}$

Table 2.2: The limiting probabilities of a root and random node, respectively, having betweenness centrality that approaches nk (in a simply generated tree of size n).

of the root is $\Theta(n^{3/2})$, being dominated, along with all the other moments of $\mathcal{B}(T)$, by quadratic-order values. One expects the moments of a random node to behave similarly. This section—which establishes the existence of a limiting distribution for the maximum betweenness centrality—begins with a small addition to these results, by demonstrating that the maximum in any simply generated tree is *always* of order n^2.

Firstly, a trivial lower bound for the maximum is $(n^2 - 2n)/4$, which follows if one considers the centroid of the tree. We know that nodes whose branch sizes are 'balanced' lead to large betweenness centralities, and that the centroid is in a sense the most balanced node of all (recall from Section 1.2.2 that the centroid minimises the total distance to all other nodes, and (equivalently) that none of its branches contain more than half the nodes of the tree). By noting that the betweenness centrality of a node decreases when nodes are moved from one of its branches to another branch of greater or equal size, we see that the smallest possible betweenness centrality of a centroid occurs when it has only two branches whose sizes are $\lfloor (n-1)/2 \rfloor$ and $\lceil (n-1)/2 \rceil$. In this case, the betweenness centrality is $\lfloor (n-1)^2/4 \rfloor \geq (n^2 - 2n)/4$; and since every tree has a centroid (and in the limit, almost surely only one), this gives the above-mentioned—quadratic—lower bound for the maximum betweenness centrality.

Although a centroid node must necessarily have fairly large betweenness centrality, this does not imply that it is always the node at which the maximum is attained. As a counterexample, consider a star of size $n/3$ with a path of length $2n/3$ attached to it: the centroid has a betweenness centrality of about $n^2/4$, while that of the centre of the star is roughly $5n^2/18$. In spite of this counterexample, the centroid will play a major role in our analysis of maximum betweenness centrality. As it turns out, the event that the centroid's betweenness centrality is in fact the maximum has positive limiting probability, and we will also be able to show that the maximum in a random simply generated tree of size n, once rescaled by a factor n^{-2}, has a limiting distribution. This limiting distribution—unlike that of the betweenness centrality of a randomly chosen node—is in fact independent of the specific family of simply generated trees.

Recall from Section 1.2.3 that the limiting object of any simply generated tree is the continuum random tree, and that its dual (in some sense) is the random triangulation of the circle with unit circumference. Triangles, in the limit, correspond to nodes of the tree with three large branches (of linear order), and the lengths of the arcs described by a triangle correspond to the sizes of these branches. The centroid, as we know, is represented by the triangle that contains the centre of the circle.

If we assign the weight $ab+bc+ac$ to a triangle with arc lengths a, b, and c, then

this gives us (asymptotically, and subject to a scaling factor n^2) the betweenness centrality of the corresponding branchpoint. The maximum betweenness centrality corresponds to the maximum weight of a triangle, and, in the limit, the distribution of this maximum weight is also the distribution of the maximum betweenness centrality. Note that a maximum weight exists almost surely, since any triangle with a weight greater than that of the centroid's has to have a longer shortest arc than the centroid triangle,[3] and there are at most finitely many such triangles.

We should also point out that Meir and Moon (2002) showed, among other things, that the average betweenness centrality of the centroid of a random simply generated tree is asymptotically equal to $(1-(1/\sqrt{2}))n^2 \approx 0.293n^2$, formulating their result in terms of the probability that the path between two randomly chosen nodes contains the centroid. This implies an asymptotic lower bound for the expected maximum betweenness centrality, and as an estimate, this bound is actually not far from the truth.

The remainder of this section presents the above-mentioned ideas more rigorously, starting with a few technical lemmas that will be required in the proof of the main theorem. For ease of presentation we stick to the special case of (non-plane) labelled trees, but similar arguments apply to the other families of simply generated trees as well—and lead to the same result.

Lemma 2.2. *Fix ε such that $0 < \varepsilon < 1/12$. In a random labelled tree of size n, the probability that there is no node that has three branches that each contain at least $n^{1-\varepsilon}$ nodes, and whose remaining branches together have $n^{1-\varepsilon}$ nodes as well, tends to 1 as $n \to \infty$.*

Proof. This is achieved by means of the first moment method: we prove that the mean number of such nodes tends to zero by counting all rooted trees whose root has the stated property. Let n_1, n_2, n_3 and $m = n - n_1 - n_2 - n_3$ be the sizes of the three branches and the remaining tree respectively. Each of them is a rooted labelled tree, so that the total number of possible trees is:

$$\binom{n}{n_1, n_2, n_3, m} n_1^{n_1-1} n_2^{n_2-1} n_3^{n_3-1} m^{m-1} = \Theta\Big(n^{n+(1/2)} n_1^{-3/2} n_2^{-3/2} n_3^{-3/2} m^{-3/2}\Big),$$

the asymptotic estimate being a consequence of Stirling's formula. Since the number of choices of n_1, n_2, n_3, and m is $\Theta(n^3)$, the total number of rooted trees with the property that three of the root's branches and the rest of the tree all have sizes at least $n^{1-\varepsilon}$ is:

$$O\Big(n^{n+(7/2)}\Big(n^{-3(1-\varepsilon)/2}\Big)^4\Big) = O\Big(n^{n-(5/2)+6\varepsilon}\Big).$$

Noting that the number of unrooted labelled trees is n^{n-2}, we find that the average number of nodes with the property given in the lemma is $O(n^{6\varepsilon-(1/2)})$, which completes the proof. □

[3] Let the centroid and non-centroid triangles have arc lengths a_1, b_1, c_1, and a_2, b_2, c_2, respectively, and assume (without loss of generality) that the second triangle lies in the segment corresponding to a_1, and that the arc lengths are labelled such that $b_1 \geq c_1$ and $a_2 \geq b_2 \geq c_2$. With the triangles' weights written in the form $a(1-a)+bc$, the fact that $a_2 \geq 1 - a_1 > 1/2$ implies $a_1(1-a_1) \geq a_2(1-a_2)$. We also have $((1/2)-c_2)c_2 > b_2c_2$, and $b_1c_1 > ((1/2)-c_1)c_1$; so were the non-centroid triangle to have the greater weight, $b_2c_2 \geq b_1c_1$ would imply $((1/2)-c_2)c_2 > ((1/2)-c_1)c_1$. Since both $c_2 < (1/2) - c_2$ and $c_1 \leq (1-a_1)/2 \leq (1/2) - c_2$, we have $c_2 > c_1$.

Lemma 2.3. *Fix constants α, β, and ε such that $0 < \alpha < \beta \leq 1/4$ and $\varepsilon > 0$. Let T be a tree of size n (with n sufficiently large) in which the centroid node has three branches of size at least βn. If v is a non-centroid node with the property that all but at most $n^{1-\varepsilon}$ nodes belong to its three largest branches, and whose third-largest branch contains at most αn nodes, then v has smaller betweenness centrality than the centroid.*

Proof. Recall that the betweenness centrality of a node decreases when nodes are transferred from one of its branches to another branch of equal or greater size. This, together with the fact that each of a centroid's branches contains at most $n/2$ nodes, implies that a lower bound for the betweenness centrality of the centroid occurs when its three largest branches have sizes $n/2$, $(n/2) - \beta n$, and βn, and is:

$$\frac{1 + 2\beta - 4\beta^2}{4} n^2.$$

On the other hand, node v must have a branch that contains at least $n/2$ nodes, so using similar reasoning one finds an upper bound for its betweenness centrality:

$$\frac{1 + 2\alpha - 4\alpha^2}{4} n^2 + O\Big(n^{2-\varepsilon}\Big).$$

Since $\alpha < \beta$ and the function $x \mapsto (1 + 2x - 4x^2)/4$ is increasing, the lemma follows immediately. □

Lemma 2.4. *Fix $\alpha > 0$. A tree T of size n has at most $(1/\alpha) - 2$ nodes that have three or more branches of size at least αn.*

Proof. We will call nodes with at least three branches of size αn or larger 'big' nodes, and other nodes 'small'. Consider the tree R that is obtained as follows: take the tree consisting of all big nodes and the paths between them, and then remove all small nodes, thereby reducing paths between big nodes that only contain small nodes to single edges. Suppose that this tree has a total of r nodes, of which a_j have degree j. We note that nodes of degree 1 in this tree have to have two branches in T that each contain at least αn small nodes, but no big nodes; nodes of degree 2 in R have to have at least one such branch in T. This implies a total of $2a_1 + a_2$ disjoint branches of at least αn nodes, so that $2a_1 + a_2 \leq 1/\alpha$. On the other hand, since:

$$\sum_{k \geq 1} a_k = r \quad \text{and} \quad \sum_{k \geq 1} k a_k = 2(r - 1),$$

we have:

$$\frac{1}{\alpha} \geq 2a_1 + a_2 \geq \sum_{k \geq 1} (3 - k) a_k = r + 2,$$

which proves the statement. □

In addition to Lemmas 2.2 to 2.4, we will need a result of Aldous (1994*a*) that was previously introduced as Theorem 1.1. It states that the limiting density of the sizes of the three largest (rescaled) branches of the centroid is given by:

$$f(x_1, x_2, x_3) = \frac{1}{12\pi} (x_1 x_2 x_3)^{-3/2}, \tag{2.5}$$

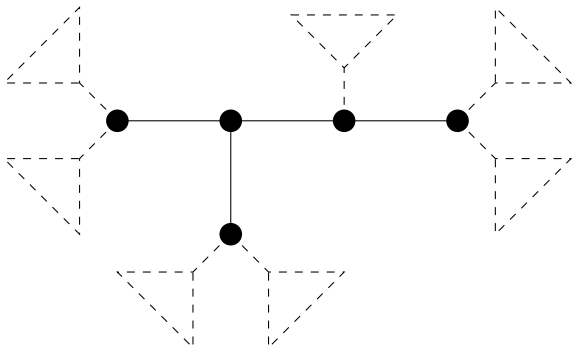


Figure 2.1: A configuration as described in the proof of Theorem 2.5. Edges represent (birooted) connecting trees, and dashed triangles branches of size at least αn.

where $0 < x_1, x_2, x_3 < 1/2$ and $x_1 + x_2 + x_3 = 1$. With these preliminaries, we are ready for a formal proof of the theorem that was alluded to earlier.

Theorem 2.5. *The maximum betweenness centrality of a random labelled tree of size n, divided by n^2, converges weakly to a limiting distribution. The probability that the maximum betweenness centrality is attained by the centroid tends to a positive constant.*

Proof. Consider the event that every node with maximum betweenness centrality has three branches each containing at least αn nodes, for some $\alpha > 0$. By Theorem 1.1, the probability that the centroid has three such branches tends to 1 as $\alpha \to 0$, and when it does, Lemma 2.3 implies that all nodes at which the maximum is attained must have either four linear-order branches, or three branches of size at least αn. Lemma 2.2 accounts for the diminishing probability of the former. Concisely, we may say that for $n > N(\alpha)$, the probability that every node with maximum betweenness centrality has three branches of size αn or larger is bounded below by $1 - f(\alpha)$, where $f(\alpha)$ is a function that goes to zero as α does.

So for fixed $\alpha > 0$, we can focus on nodes with three branches of at least αn nodes each, of which there are, by Lemma 2.4, only a bounded number. These nodes can only be configured in a finite number of ways: each configuration can be seen as a labelled tree with $r \leq (1/\alpha) - 2$ nodes, in which there are no nodes of degree greater than 3, nodes of degree $k < 3$ have $3 - k$ large branches attached to them (rooted trees with at least αn nodes), and edges represent birooted trees (possibly empty, or with coinciding roots). See Figure 2.3 for an example. Note that each of the nodes may also have smaller branches with a total of at most $O(n^{1-\varepsilon})$ nodes.

Let the sizes of the birooted trees and the sizes of the additional large branches be $x_1, \ldots, x_{r-1}$ and $y_1, \ldots, y_{r+2}$ respectively. Using the fact that there are $x_i^{x_i}$ possible birooted trees for each i and $y_l^{y_l - 1}$ possible rooted trees for each l, we obtain an asymptotic expression for the number of possible trees corresponding to each configuration. We should point out that since we have placed no further constraints on the x_i and y_l, there might actually be nodes with three branches of size αn or larger inside the birooted connecting trees or peripheral branches, however one can account for these cases by means of an inclusion-exclusion argument.

All in all one finds, for each configuration, that the sizes of the connecting trees and large branches, scaled by $1/n$, converge to a limiting distribution with a computable density (as in equation (2.5)). Since the betweenness centralities of the nodes with three large branches only depend on these sizes up to $O(n^{2-\varepsilon})$, we can infer a limiting distribution for the maximum betweenness centrality of nodes with at least three branches of size αn or larger, as well as a limiting probability that this maximum is attained by the centroid, for each fixed $\alpha > 0$. Letting α go to zero now yields the desired result on the limiting distribution of the maximum betweenness centrality, and also shows that there must be a limiting probability for the centroid to attain the maximum betweenness centrality. To show that this probability is in fact positive, we can use a straightforward argument: suppose that all three centroid branches have fewer than $((4/9) - \delta)n$ nodes (for some small $\delta > 0$), which happens with positive limiting probability by Theorem 1.1. Then the betweenness centrality of the centroid is at least:

$$\left(\left(\frac{4}{9} - \delta\right)^2 + 2\left(\frac{4}{9} - \delta\right)\left(\frac{1}{9} + 2\delta\right) + o(1)\right)n^2 = \left(\frac{8}{27} + \frac{2\delta}{3} - 3\delta^2 + o(1)\right)n^2.$$

On the other hand, every other node v has to have a branch of size at least $((5/9) + \delta)n$ (consider the branch containing the centroid); and since, with probability approaching 1, node v has at most three branches of size linear in n, an upper bound for its betweenness centrality occurs when its second and third branches each contain $((4/9) - \delta)n/2$ nodes:

$$\left(\left(\frac{2}{9} - \frac{\delta}{2}\right)^2 + 2\left(\frac{2}{9} - \frac{\delta}{2}\right)\left(\frac{5}{9} + \delta\right) + o(1)\right)n^2 = \left(\frac{8}{27} - \frac{\delta}{3} - \frac{3\delta^2}{4} + o(1)\right)n^2.$$

This betweenness centrality is—for small δ and sufficiently large n—strictly smaller than that of the centroid. □

An argument similar to the final paragraph, but comparing the betweenness centrality of the centroid to that of the centroid of one of its largest branches, shows that the probability that the centroid has maximum betweenness centrality is strictly less than 1. Also in this fashion, one can show that the limiting random variable of the rescaled maximum betweenness centrality has the interval $[1/4, 1/3]$ as its support.

Numerically, the average maximum is asymptotically equal to $0.303n^2$, the (approximate) constant being determined by Monte Carlo sampling. Moreover, the probability that the centroid is in fact also the node with maximum betweenness centrality converges to a constant close to 0.621. The limiting distribution of the (normalised) maximum betweenness centrality, which can also be obtained by Monte Carlo sampling, is shown in Figure 2.2. One way to perform this simulation is to first generate the centroid's branch sizes according to equation (2.5), and then repeat this recursively within each branch. Once it is no longer possible to generate nodes with betweenness centrality greater than the current maximum (which almost surely happens within a finite number of steps), one can stop the process.

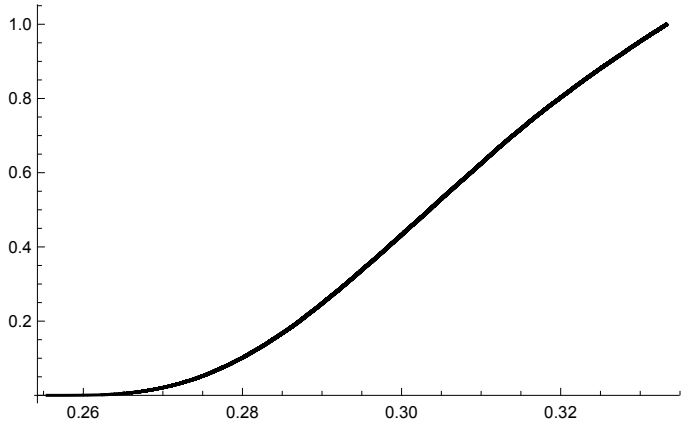


Figure 2.2: The cumulative distribution function of the limiting distribution of rescaled maximum betweenness centrality in simply generated trees.

2.4 Betweenness centrality in subcritical graphs

There are certain families of graphs that are analytically quite similar to simply generated trees, and which, symbolically, can be seen as an extension of them. These are the so-called subcritical graphs, of which outerplanar, series-parallel, and cacti graphs are special cases. And although graphs are not the main focus of this thesis, it seems natural when considering betweenness centrality in simply generated trees to ask whether similar behaviour is witnessed in families of subcritical graphs. Because these graphs are tree-like in structure, this is in fact the case.

We give a brief definition of subcritical graphs here; they have been described in more detail by, e.g., Drmota *et al.* (2011). Note that in general, the nodes of subcritical graphs can be either labelled or unlabelled, however we will consider only the labelled case—as with non-plane trees, the functional equations for the unlabelled case involve certain Pólya operators.

The key concept is the block decomposition of a graph: by defining the blocks of a graph to be its maximal 2-connected subgraphs (a graph is k-connected if at least k of its nodes must be deleted before it becomes disconnected), every graph can be decomposed into blocks, cut nodes (nodes whose removal disconnects the graph), and the induced edge set that links cut nodes to their incident blocks. This leads to a bipartite tree. A family of graphs is called *block-stable* if it contains the two-node one-edge 'link' graph, and satisfies the property that a graph belongs to the family if and only if all of its blocks do as well.

Let $\mathcal{C}$ be a block-stable family of rooted, labelled, *connected* graphs whose blocks form the class $\mathcal{A}$. Then the bipartite block decomposition described above implies a symbolic definition of $\mathcal{C}$: start with a root node, and graft a set of blocks to it by removing a node from each block and fitting the detached edges to the root. Then graft sets of blocks to every newly added node in the same way, and continue. The generating function that counts graphs of $\mathcal{C}$ according to their size captures this construction:

$$C(x) = x \exp(A'(C(x))),$$

where $A'(x)$ is the (exponential) generating function of the class $\mathcal{A}'$ of blocks with

one removed node.

The 'subcriticality' property of subcritical graphs is a technical condition that requires the radii of convergence of $C(x)$ and $A'(y)$, ρ and η respectively, to satisfy $C(\rho) < \eta$. This implies that $A'(y)$ is analytic at $y = \tau = C(\rho)$, and that $C(x)$ permits a square-root expansion around its singularity $x = \rho$, much like in the case of simply generated trees (see Drmota *et al.*, 2011). In particular, we have $1/\rho = \exp(A'(\tau))A''(\tau)$ and:

$$\begin{aligned} A'(y) &= A'(\tau) + A''(\tau)(y-\tau) + O\Big((y-\tau)^2\Big), \\ C(x) &= \tau - \mu\sqrt{1-\frac{x}{\rho}} + O\Big(1-\frac{x}{\rho}\Big), \end{aligned} \tag{2.6}$$

in which $\mu = \sqrt{2/(A''(\tau)^2 + A^{(3)}(\tau))}$.

Our goal is again to investigate the betweenness centrality of the root of a graph $C \in \mathcal{C}$, which will be denoted by $\mathcal{B}(C)$ as before.[4] However it should be noted that, because we are considering labelled graphs in which any node can be distinguished as the root, our results hold for a randomly chosen node as well.

The only real caveat when working with subcritical graphs is that the betweenness centrality of a node v is no longer solely determined by paths between its branches (here, branches take the form of blocks with one node removed and subgraphs rooted to their remaining nodes, and will be denoted by the generating function $W(x) = A'(C(x))$). In addition to the usual inter-branch paths, we must also consider shortest paths between subgraphs of each branch's root block, since it may be the case that these pass through v.

Consider one of the root's branches W, along with its root block $A \in \mathcal{A}'$. Because shortest paths within blocks are not necessarily unique, the contribution of paths between the subgraphs of W to $\mathcal{B}(C)$, the betweenness centrality of the root node, is:

$$\sum_{v<w} b_{vw}(A)|C_v||C_w|,$$

where v, w is a pair of non-root nodes in A, so that $\sum b_{vw}(A) = \mathcal{B}(A)$ is the betweenness centrality of A's removed node with respect to paths contained within A; and C_v and C_w are the subgraphs rooted at v and w.

A full expression for the betweenness centrality of a graph's root is thus:

$$\begin{aligned} \mathcal{B}(C) &= \sum_{a<b} |W_a||W_b| + \sum_{A}\sum_{v<w} b_{vw}(A)|C_v||C_w| \\ &= B_1(C) + B_2(C), \end{aligned}$$

the first sum being over all pairs of root branches and the second sum being over all root blocks.

2.4.1 Moments of the betweenness centrality of the root

When deriving the moments of $\mathcal{B}(C)$, we can handle the two terms in equation (2.7) separately. The contribution of $B_1(C)$ is identical, conceptually, to the betweenness

[4] We will also abuse this notation slightly by writing $\mathcal{B}(A)$ for the betweenness centrality of the distinguished (and removed) node of a block $A \in \mathcal{A}'$.

centrality of the root of a tree, so one need only count graphs with two distinguished nodes from distinct branches. A generating function $\sum_C B_2(C)x^{|C|}/|C|!$ for the second term can be derived in essentially the same way, as long as we note that every path between two subgraphs C_v and C_w rooted to a block A must be weighted by $b_{vw}(A)$. These observations lead to a relatively straightforward derivation of the average betweenness centrality of the root node.

Theorem 2.6. *The expected betweenness centrality of the root node in a subcritical graph of size n is $\Theta(n^{3/2})$, satisfying:*

$$\mathrm{E}_n(\mathcal{B}) \sim K\sqrt{\pi n^3},$$

where:

$$K = \frac{\mu}{2}\left(\frac{\tau}{2}A''(\tau)^2 + \frac{1}{\tau}M(\tau)\right),$$

and $M(y) = \sum_A \mathcal{B}(A)y^{|A|}/|A|!$ is the cumulative generating function of $\mathcal{B}(A)$ over blocks $A \in \mathcal{A}'$.

Proof. We desire the generating function $H(x) = \sum_C \mathcal{B}(C)x^{|C|}/|C|!$, which can be written as the sum of the corresponding generating functions for $B_1(C)$ and $B_2(C)$. The first of these is:

$$U_1(x) = \sum_{C\in\mathcal{C}} B_1(C)\frac{x^{|C|}}{|C|!} = \frac{x^3}{2}W'(x)^2\exp(W(x)) = \frac{x^2}{2}W'(x)^2C(x).$$

From the expansions of $A'(y)$ and $C(x)$ given in equation (2.6) we can derive a similar expansion for $W(x)$:

$$W(x) = A'(\tau) - \mu A''(\tau)\sqrt{1-\frac{x}{\rho}} + O\left(1-\frac{x}{\rho}\right),$$

so that $U_1(x)$ satisfies:

$$U_1(x) \sim \frac{\tau}{2}\left(\frac{\mu}{2}A''(\tau)\right)^2\left(1-\frac{x}{\rho}\right)^{-1}. \tag{2.7}$$

The generating function of $B_2(C)$ requires two stages of substitution, since we must first derive a generating function $L(x)$ that describes branches that have had two nodes distinguished from their subgraphs. We will then have:

$$U_2(x) = \sum_{C\in\mathcal{C}} B_2(C)\frac{x^{|C|}}{|C|!} = xL(x)\exp(W(x)) = L(x)C(x).$$

To obtain $L(x)$, recall that the paths between subgraphs of a branch's root block must be weighted; then:

$$\begin{aligned} L(x) &= \sum_{A\in\mathcal{A}'}\sum_{v<w} b_{vw}(A)\frac{C(x)^{|A|-2}(xC'(x))^2}{|A|!} \\ &= (xC'(x))^2\sum_{A\in\mathcal{A}'}\mathcal{B}(A)\frac{C(x)^{|A|-2}}{|A|!} \\ &= M(C(x))\frac{(xC'(x))^2}{C(x)^2}, \end{aligned}$$

where:

$$M(y) = \sum_{A \in \mathcal{A}'} \mathcal{B}(A) \frac{y^{|A|}}{|A|!} = m_2 y^2 + m_3 y^3 + \cdots .$$

We remark that $M(y)$ has the same (or possibly even greater) radius of convergence as $A'(y)$, since $\mathcal{B}(A)$ can be bounded trivially by $|A|^2$. Noting that $C(x)^{-1}$ also permits a square-root expansion around $x = \rho$, beginning $(1/\tau)+\cdots$, the asymptotic form of the second generating function is:

$$U_2(x) \sim \frac{1}{\tau}\left(\frac{\mu}{2}\right)^2 M(\tau)\left(1 - \frac{x}{\rho}\right)^{-1}. \tag{2.8}$$

Equations (2.7) and (2.8) imply that both kinds of paths contribute equally in order to the betweenness centrality of the root node, and the expected betweenness centrality of the root of a graph of size n is $[x^n](U_1(x) + U_2(x))/[x^n]C(x)$. □

The higher-order moments of $\mathcal{B}(C)$ are more interesting, because they involve the function:

$$\mathcal{B}(C)^k = (B_1(C) + B_2(C))^k = \sum_{j=0}^{k} \binom{k}{j} B_1(C)^{k-j} B_2(C)^j. \tag{2.9}$$

In the case of simply generated trees, $\mathcal{B}(T)^k$ could be interpreted as a selection of $2k$ nodes from between 2 and $2k$ distinct branches, and we could restrict our calculation to the case of exactly two branches due to Lemma 2.1. This basic concept holds once again—for both $B_1(C)^k$ and $B_2(C)^k$—so that:

$$B_1(C)^k \sim \sum_{a<b} |W_a|^k |W_b|^k,$$
$$B_2(C)^k \sim \sum_{A} \sum_{v<w} b_{vw}(A)^k |C_v|^k |C_w|^k.$$

Both of these terms lead to generating functions (of the form $\sum_C B_i(C)^k x^{|C|}/|C|!$) that are dominated by a term of order $(1 - (x/\rho))^{-2k+1}$. The question, however, is whether the remaining terms in equation (2.9)—which involve a product of powers of $B_1(C)$ and $B_2(C)$—are of lower or equal order. Note that the smallest number of substitutions of branches and subgraphs with pointed structures that can be made when constructing a generating function involving both $B_1(C)$ and $B_2(C)$ is three: some nodes must be chosen from at least two branches, and the rest from at least two subgraphs. At best, subgraphs from one of the pointed branches could be affected, leading to three substitutions. Lemma 2.1 implies that the replacement of a branch or subgraph with one in which d nodes have been distinguished contributes $(1 - (x/\rho))^{-d+(1/2)}$ to the final order of the generating function, which tells us that the 'mixed' terms of $\mathcal{B}(C)^k$ grow at a slower rate than those involving only $B_1(C)$ or $B_2(C)$. This simplifies the asymptotic behaviour of $\mathcal{B}(C)^k$ greatly:

$$\mathrm{E}_n(\mathcal{B}^k) \sim \mathrm{E}_n(B_1^k) + \mathrm{E}_n(B_2^k).$$

We find that the kth moment of the betweenness centrality of the root node satisfies an expression that is very similar to the one derived for simply generated trees. The second moment is once again of order $n^{7/2}$, so that the variance of $\mathcal{B}(C)$ is as well.

Theorem 2.7. *The kth moment of the betweenness centrality of the root node in a subcritical graph of size n is $\Theta(n^{2k-(1/2)})$, and satisfies, for $k \geq 1$:*

$$\mathrm{E}_n(\mathcal{B}^k) \sim K_k \sqrt{\pi n^{4k-1}},$$

for a constant K_k that depends on $\mathcal{C}$.

Proof. The asymptotic behaviour of $H_k(x) = \sum_C \mathcal{B}(C)^k x^{|C|}/|C|!$ is:

$$H_k(x) \sim \frac{\tau}{2}\left(\frac{\mu(2k-3)!!}{2^k}A''(\tau)\right)^2\left(1-\frac{x}{\rho}\right)^{-2k+1} + \frac{1}{\tau}\left(\frac{\mu(2k-3)!!}{2^k}\right)^2 M_k(\tau)\left(1-\frac{x}{\rho}\right)^{-2k+1},$$

in which:

$$M_k(y) = \sum_{A\in\mathcal{A}'}\sum_{v<w} b_{vw}(A)^k \frac{y^{|A|}}{|A|!}.$$

The desired moment is $[x^n]H_k(x)/[x^n]C(x)$, so the theorem follows with:

$$K_k = \frac{\mu}{2^{4k-3}}\binom{2k-2}{k-1}\left(\frac{\tau}{2}A''(\tau)^2 + \frac{1}{\tau}M_k(\tau)\right). \qquad \square$$

2.4.2 Limiting behaviour of the betweenness centrality of the root

Since the moments of the betweenness centrality of the root in a random subcritical graph are similar to those found for simply generated trees, it is probably unsurprising that we can show that the majority of these root nodes (in subcritical graphs) have linear-order betweenness centrality, and that the balanced graphs which lead to quadratic-order betweenness centrality become increasingly rare as $n \to \infty$.

To do so, we repeat the procedure of Section 2.2.2, defining unbalanced subclasses $\mathcal{L}_{k,m} \subset \mathcal{C}$ that not only have k non-root nodes outside their largest branch, but also have a dominant subgraph within that branch. This subgraph includes all but m of the large branch's nodes. If we let:

$$H(k,m) = \Big[[x^k]\exp(W(x))\Big]\Big[[x^m]A''(C(x))\Big]$$

be the number of ways in which the minor branches and subgraphs can be configured, then the generating function of $\mathcal{L}_{k,m}$ can be written as:

$$L_{k,m}(x) = H(k,m)x^{k+m+1}C(x).$$

From this generating function, the limiting probability of a random graph C belonging to $\mathcal{L}_{k,m}$ is seen to be a function of k and m:

$$\lim_{n\to\infty} \mathrm{P}_n(C \in (\mathcal{L}_{k,m})_n) = H(k,m)\rho^{k+m+1}.$$

As expected, these proportions account for the entire limiting distribution:

$$\sum_{k\geq 0}\sum_{m\geq 0}\lim_{n\to\infty}\mathrm{P}_n(C\in(\mathcal{L}_{k,m})_n) = \rho\exp(W(\rho))A''(C(\rho)) = 1.$$

Finally, the betweenness centrality of the root of a graph $C \in (\mathcal{L}_{k,m})_n$ is of linear order, since there are linearly many of the two kinds of paths through the root: if k_i and m_j are the minor branch and subgraph sizes respectively, with, say, $i = 2, \ldots, \alpha$ and $j = 2, \ldots, \beta$, then:

$$\begin{aligned}\mathcal{B}(C) &\sim (n-k-m-1)\left(\sum_{i=2}^{\alpha} k_i + \sum_{j=2}^{\beta} b_{vw_j}(A)m_j\right) \\ &= nk + n\sum_{j=2}^{\beta} b_{vw_j}(A)m_j + O\Big((k+m)^2\Big), \end{aligned} \tag{2.10}$$

in which A denotes the root block of the dominant branch. Noting that $0 \leq b_{vw_j}(A) \leq 1$, we have a linear bound on $\mathcal{B}(C)$:

$$k \leq \lim_{n\to\infty} \frac{\mathcal{B}(C)}{n} \leq k + m.$$

This gives us the following theorem, which is a qualitative analogue of Theorem 2.3, albeit less precise:

Theorem 2.8. *The betweenness centrality of the root in a random subcritical graph is bounded in probability:*

$$\mathcal{B}(\mathcal{C}_n) = O_p(n).$$

Specifically, if C is randomly chosen from a labelled subcritical graph family $\mathcal{C}$, then for every $\varepsilon > 0$ there exists a real number M such that:

$$\limsup_{n\to\infty} \mathrm{P}_n(\mathcal{B}(C) > Mn) < \varepsilon.$$

If more information on the blocks of the specific family of subcritical graphs—and in particular their betweenness centralities—is available, it is also possible to provide a more precise limit law, as for simply generated trees. We also remark again that the distribution is the same for a random node: as in the case of random labelled trees, every node of a random labelled subcritical graph has the same probability to be the root.

2.4.3 Cacti graphs

To illustrate the results of this section more plainly, we can apply them to the special case of cacti graphs, which are defined using polygonal blocks. Specifically, $\mathcal{A}$ consists of the two-node link graph and all (labelled) unoriented, convex polygons of size 3 or more, and the derived class $\mathcal{A}'$ is counted by the generating function $\mathcal{A}'(y) = y + y^2(1-y)^{-1}/2$, which has radius of convergence $\eta = 1$. The radius of convergence ρ of $C(x)$, and the constant $\tau = C(\rho)$, are roughly 0.239 and 0.456 respectively (see Drmota *et al.*, 2011, Section 9.1).

The derivation of the generating function $M(y) = \sum_A \mathcal{B}(A)y^{|A|}/|A|!$ that appears in Theorem 2.6 is relatively straightforward: if a derived polygon $A \in \mathcal{A}'$ has $u+1$ nodes—including its root—then every shortest path within A consists of at most $\lfloor (u+1)/2 \rfloor = \lceil u/2 \rceil$ edges. All of these shortest paths are unique except when $u+1$

is even, in which case paths of length $(u+1)/2$ can be oriented in two different ways. There are $l-1$ paths of length l that pass through the root (because such a path contains $l-1$ non-terminal nodes), so the betweenness centrality of the root of A is given by:

$$\begin{aligned}\mathcal{B}(A) = B_u &= \left(\sum_{l=2}^{\lceil \frac{u}{2} \rceil} l - 1\right) - \frac{1}{2}\left(\frac{u+1}{2} - 1\right) I[u \text{ odd}] \\ &= \begin{cases} \frac{1}{8}u(u-2) & \text{if } u \text{ is even,} \\ \frac{1}{8}(u-1)^2 & \text{if } u \text{ is odd.} \end{cases}\end{aligned}$$

Here we have used I to denote the use of Iverson's notation, which in this case yields 1 for odd u and 0 otherwise. Letting $a_u = [y^u]A'(y) = 1/2$ for $u > 1$, the generating function $M(y)$ satisfies:

$$\begin{aligned}M(y) = \sum_{u\geq 2} a_u B_u y^u &= \frac{1}{16}\sum_{u \text{ even}} u(u-2)y^u + \frac{1}{16}\sum_{u \text{ odd}} (u-1)^2 y^u \\ &= \frac{y^3}{16}\frac{d}{dy}\left[\frac{1}{y}\frac{d}{dy}\left(1-y^2\right)^{-1}\right] + \frac{y^2}{16}\frac{d}{dy}\left[y\frac{d}{dy}\left(1-y^2\right)^{-1}\right] \\ &= \frac{y^3}{4}\left(1-y^2\right)^{-2} + \frac{y^4+y^5}{2}\left(1-y^2\right)^{-3}.\end{aligned}$$

Immediately, we see that $M(\tau) \approx 0.101$, and that the constant of Theorem 2.6—which describes, asymptotically, the mean betweenness centrality of the root—is $K \approx 0.281$. The constant K_k of Theorem 2.7 is similar, except that it refers to the generating function $M_k(y) = \sum_A \sum_{v,w} b_{vw}(A)^k y^{|A|}/|A|!$, which requires that proportions of paths through the root be raised to the kth power. Of course this change affects only those proportions $b_{vw}(A)$ that are not equal to 1, which means the only paths that are affected are those of length $(u+1)/2$ in blocks where $u+1$ is even. For these paths, $b_{vw}(A) = 1/2$ becomes $1/2^k$, so that the altered betweenness centrality of the root is:

$$\begin{aligned}\sum_{v<w} b_{vw}(A)^k = B_{u,k} &= \left(\sum_{l=2}^{\lceil \frac{u}{2} \rceil} l - 1\right) - \left(1 - \frac{1}{2^k}\right)\left(\frac{u+1}{2} - 1\right) I[u \text{ odd}] \\ &= \begin{cases} \frac{1}{8}u(u-2) & \text{if } u \text{ is even,} \\ \frac{1}{8}(u-1)(u-3) + \frac{1}{2^{k+1}}(u-1) & \text{if } u \text{ is odd.} \end{cases}\end{aligned}$$

In the same manner as above, this leads to the generating function:

$$M_k(y) = \sum_{u\geq 2} a_u B_{u,k} y^u = \frac{y^3}{2^{k+1}}\left(1-y^2\right)^{-2} + \frac{y^4+y^5}{2}\left(1-y^2\right)^{-3}.$$

Using the definition given in the proof of Theorem 2.7, we can derive the constant associated with the kth moment of the betweenness centrality of the root node:

$$K_k \approx \frac{1}{2^{4k}}\binom{2k-2}{k-1}\left(4.209 + \frac{0.566}{2^k}\right).$$

Finally, we can consider the limiting distribution that arises for the betweenness centrality of the root. Recall equation (2.10), in which k counts the nodes outside the largest branch, and m_j the nodes of the jth subgraph attached to the largest branch's root block. Assume that this root block A has $u+1$ nodes, and that its subgraphs have sizes $m_1, m_2, \ldots, m_u$, with $m_2 + m_3 + \cdots + m_u = m$ fixed. Although we paid them little attention in the previous section, these subgraph sizes are crucial to the asymptotic betweenness centrality of the root as n tends to infinity.

Our goal is to describe the exact set of graphs that lead to a specific, linear betweenness centrality—in particular, graphs for which $\mathcal{B}(C)/n \sim k + t$, where:

$$t = \sum_{j=2}^{u} b_{vw_j}(A) m_j.$$

Here v is the root of the largest subgraph, w_j is the root of the subgraph of size m_j, and u corresponds to the β of equation (2.10). If v is positioned directly across from the root of A (there is one such position for odd u and two for even u), then none of the shortest paths between its subgraph's nodes and the nodes of other subgraphs pass through the root, implying $t = 0$. On the other hand, if v lies $l < \lceil u/2 \rceil$ edges away from the root, there are $\lceil u/2 \rceil - l$ nodes on the 'far' side of the root that are of interest. We need to take into consideration the size u of A, the $2\lceil u/2 \rceil - 2$ possible positions for v, and the sizes $m_2, m_3, \ldots, m_{\lceil u/2 \rceil - l + 1}$ of the subgraphs whose paths through v contribute to the betweenness centrality of the root of A.

To count the configurations of the dominant branch that add nt to the overall asymptotic betweenness centrality, define $J(m,t)$, where:

$$\begin{aligned} J(m,0) &= [x^m]\left(\sum_{u \text{ odd}} a_u C(x)^{u-1} + 2 \sum_{u \text{ even}} a_u C(x)^{u-1} \right) \\ &= [x^m] \frac{C(x)}{2} \left(\frac{1}{1 - C(x)} + \frac{1}{1 - C(x)^2} \right), \end{aligned}$$

and, for $t > 0$:

$$\begin{aligned} J(m,t) = &\sum_{u \text{ even}} \Big[[x^u] A'(x) \Big] \sum_{l=1}^{\lceil u/2 \rceil - 1} 2 \Big[[x^t] C(x)^{\lceil u/2 \rceil - l} \Big] \Big[[x^{m-t}] C(x)^{\lceil u/2 \rceil + l - 1} \Big] \\ &+ \sum_{u \text{ odd}} \Big[[x^u] A'(x) \Big] \sum_{l=1}^{\lceil u/2 \rceil - 1} 2 \sum_{i \geq 1} \Big[[x^i] C(x) \Big] \Big[[x^{t-(i/2)}] C(x)^{\lceil u/2 \rceil - l - 1} \Big] \\ &\qquad \times \Big[[x^{m-t-(i/2)}] C(x)^{\lceil u/2 \rceil + l - 1} \Big]. \end{aligned}$$

In the sum over odd u, the variable i counts the nodes of the furthest subgraph, which leads to paths of length $(u+1)/2$.

The coefficient $H(k,m)$ of Section 2.4.2 can then be extended to one that accounts for the constants k and m, as well as a betweenness centrality for the root that satisfies $\mathcal{B}(C)/n \sim k + t$:

$$H(k,m,t) = \Big[[x^k] \exp(W(x)) \Big] J(m,t).$$

The total probability of the subset of graphs whose roots have betweenness centralities that satisfy $\mathcal{B}(C)/n \sim r$ thus tends to:

$$p_r = \sum_{k=0}^{r} \sum_{m \geq r-k} H(k, m, r-k)\rho^{k+m+1},$$

and, omitting the details, these probability masses define a limiting distribution, as described in the following theorem.

Theorem 2.9. *The linearly scaled betweenness centrality of the root node in a random cacti graph of size n, $\mathcal{B}(\mathcal{C}_n)/n$, converges in distribution to the discrete random variable with support $\mathbb{Z}_{\geq 0}$ and mass function $r \mapsto p_r$.*

This brings to a close our second chapter, which has dealt with simply generated trees and subcritical graphs. Both of these structures are characteristically unbalanced, or 'thin', implying that their nodes will typically have betweenness centrality that is linear in the size of the object. In the next chapter we consider betweenness centrality in increasing trees, which, although superficially similar to simply generated trees (in terms of their global generating function), have a markedly more balanced shape.

Chapter 3

Betweenness Centrality in Increasing Trees

3.1 A brief summary of results

This short chapter retraces, for the most part, the steps of the previous one—but with a focus on the class of increasing instead of simply generated trees. Our results are only partially similar: the betweenness centrality of a random node in a very simple increasing tree indeed converges to a limiting distribution when rescaled (linearly) by the size of the tree; however, for any node with a fixed label (including the root), the kth moment is of order n^{2k}, and a limiting distribution exists only when the betweenness centrality is rescaled quadratically. To complement the analogous result for labelled trees in Chapter 2, the maximum betweenness centrality in a recursive tree, divided by n^2, converges to a limiting distribution, and the probability that the centroid attains this maximum approaches a constant (roughly 0.87).

3.2 Increasing trees

An increasing tree is a rooted, labelled tree in which the labels along any path leading away from the root form an increasing sequence. This ordering constraint, innocuous as it may seem, gives rise to families of trees that are markedly different from those that were dealt with in the previous chapter. When it comes to practical matters, these differences make increasing trees perhaps the more familiar of the two classes, especially in terms of their shape: distances in an increasing tree of size n are usually of order $\log n$.

There are smaller, conceptual distinctions as well: increasing trees are necessarily labelled, but unlike labelled simply generated trees, in which labels are somewhat arbitrarily assigned, those in an increasing tree are quite significant. The root is

always given the label 1, and one can expect the largest labels to be found close to the fringes of a tree. In some sense, this will make the investigation of betweenness centrality more satisfying than it was in the case of simply generated trees, because we can study the betweenness centrality $\mathcal{B}_l(T)$ of each labelled node l individually.

The fact that the nodes of an increasing tree are labelled according to the order in which they are attached to the tree implies a kind of generating function that is reminiscent of the one given for simply generated trees (equation (2.2)), but with two important differences: firstly, whereas a simply generated tree could be associated with an exponential or ordinary generating function according to whether or not it was labelled, the generating functions we use for increasing trees—throughout this chapter and the next—are all exponential. Secondly, these generating functions satisfy differential equations, as opposed to the functional equations of the previous chapter.

Let the characteristic weight function $\phi(u) = \sum_i \phi_i u^i$ once again encode a sequence of non-negative out-degree weights $\{\phi_i\}$, such that $\phi_i \neq 0$ and $\phi_i > 0$ for some $i \geq 2$. Then, recalling that the act of removing the node with the lowest label from every object in a class is represented by the differential operator $y'(x)$, the generating function for the class of increasing trees $\mathcal{T}$ satisfies:

$$y'(x) = \sum_{T \in \mathcal{T}} \omega(T) \frac{x^{|T|-1}}{(|T|-1)!} = \phi(y(x)), \tag{3.1}$$

where $\omega(T)$ is again the product of the weights assigned to T's nodes. Due to the fact that the generating functions of increasing trees satisfy differential equations, it is not always possible to carry out general analyses quite as thoroughly as it is for simply generated trees. Apart from the broad special case of increasing trees with polynomial weight functions, it is often necessary to specify ϕ in order to complete an application of singularity analysis to a parameter of interest (see Bergeron *et al.* (1992) for several illustrative cases).

3.2.1 Very simple increasing trees

Fortunately, there are a few particularly important varieties of increasing trees that can be characterised in a number of useful ways, and which also share important structural characteristics. These are general recursive, plane-oriented[1], and d-ary increasing trees.

Lemma 3.1 (Panholzer and Prodinger (2007, Lemma 5))**.** *Let $\mathcal{T}$ be a family of increasing trees; then $\mathcal{T}$ is a family of* very simple *increasing trees if the following (equivalent) properties hold:*

- *the total weight of trees of size n, denoted by y_n, satisfies $y_{n+1}/y_n = c_1 n + c_2$ for certain $c_1, c_2 \in \mathbb{R}$;*
- *repeatedly pruning the node with the largest label from a random tree yields another, smaller, random tree; and*

[1]Plane-oriented trees are also known as heap-ordered trees.

- *trees can be constructed by way of a probabilistic growth process.*

Alternatively, families of very simple increasing trees can be identified by their characteristic functions, which correspond to one of the following:

- *general recursive* trees:
$$\phi(u) = \exp(c_1 u), \text{ with } c_1 > 0;$$
- *general plane-oriented* trees:
$$\phi(u) = (1 + c_2 u)^{1+c_1/c_2}, \text{ with } c_2 < 0 \text{ and } c_1/c_2 < -1;$$
- *general d-ary increasing* trees:
$$\phi(u) = (1 + c_2 u)^{1+c_1/c_2}, \text{ with } c_2 > 0 \text{ and } c_1/c_2 \in \mathbb{Z}_{>0}.$$

(Note that in the above characteristic functions we have implicitly assumed that the weight assigned to a leaf node is 1, since $[u^0]\phi(u) = 1$ in all three of them.) We have called these families 'general' because each one has a more standard form, corresponding to certain fixed values of c_1 and c_2. These are:

- *recursive* trees ($c_1 = 1$):
$$\phi(u) = \exp(u) \implies y(x) = -\log(1 - x);$$
- *plane-oriented* trees ($c_2 = -1$ and $c_1 = 2$):
$$\phi(u) = (1 - u)^{-1} \implies y(x) = 1 - \sqrt{1 - 2x};$$
- *d-ary increasing* trees ($c_2 = 1$ and $c_1 = d - 1 \in \mathbb{Z}_{>0}$):
$$\phi(u) = (1 + u)^d \implies y(x) = -1 + (1 - (d - 1)x)^{-1/(d-1)}.$$

Throughout the majority of the next two chapters we will work with the more general forms, only referring to specific cases for the sake of intuition or examples. The notable exception is Section 3.5, in which we focus on recursive trees exclusively—though similar results will hold for other very simple families.

The probabilistic growth processes mentioned in Lemma 3.1 were also introduced briefly in Section 1.1.1, but perhaps deserve more attention here. In the simplest case of recursive trees, the process starts with a root node—always labelled 1—and at each step, node n is attached to one of the $n - 1$ previous nodes, uniformly at random. As implied above, the tree obtained after the nth step is random (relative to its weight) in $\mathcal{T}_n$. Clearly, the number of recursive trees of size n satisfies $y_n = (n - 1)y_{n-1} = (n - 1)!$.

The processes for plane-oriented and d-ary increasing trees are similar, but with attachment probabilities that depend on the out-degrees of the existing nodes. Firstly, in the plane-oriented case, a node with m children is viewed as having $m + 1$ distinct attachment points. Since there are a total of $2n - 1$ such points in a tree of size n, the number of plane-oriented trees is $y_n = (2n - 3)y_{n-1} = (2n - 3)!!$. Secondly, the defining characteristic of a d-ary tree is that each of its nodes starts with d attachment points, so that $y_n = ((d - 1)n + 1)y_{n-1}$. Note that these counts all abide by the general rule $y_n = (c_1 n + c_2)y_{n-1}$ described above.

3.2.2 The common form of the derived generating function

The case-specific expressions we have mentioned for the generating function $y(x)$ are all quite familiar and amenable to analysis; and although this is a trait shared by the whole class of very simple increasing trees, it will be sufficient, for our purposes, to work with the derived generating function $y'(x)$, which has a *common*, manageable form for all families:

$$y'(x) = (1 - c_1 x)^{-(1+(c_2/c_1))}, \text{ with } c_1 \neq 0. \tag{3.2}$$

This expression is not only a necessary property of very simple families, but a sufficient one as well. To see this, note that it is usual to assume that $y_n \geq 0$ for $n > 0$, along with $y_0 = 0$. Since $y_n = \prod_{j=1}^{n-1}(c_1 j + c_2)$, this is the case only if $c_1 > 0$ and $1 + c_2/c_1 > 0$, which corresponds to general recursive and plane-oriented trees when $c_2 = 0$ and $c_2 < 0$, respectively.

Another typical assumption is that $\phi_i = [u^i]\phi(u) \geq 0$ for $i > 0$, with $\phi_0 > 0$ and $\phi_i > 0$ for some $i \geq 2$. This can also be applied here, since by integrating $y'(x)$ when $c_2 > 0$ (the constant is determined using $y_0 = 0$) it follows that $y'(x) = \phi(y(x)) = (1 + c_2 y(x))^{1+(c_1/c_2)}$. This characteristic function satisfies the assumption only if $c_1/c_2 \in \mathbb{Z}_{\geq 0}$, although the case $c_1/c_2 = 0$ is generally excluded to avoid the family of 'path' trees.

Thus it is possible to derive results that are specific to very simple increasing trees by working with the derived form (3.2) and assuming the non-negativity of the y_n and ϕ_i. That being said, our results will remain unaffected if we drop the constraints on y_n and ϕ_i, so in this chapter and the next, we define α and assume that:

$$y'(x) = (1 - c_1 x)^{-\alpha}, \text{ with } \alpha = 1 + \frac{c_2}{c_1} > 0.$$

In particular, we can write $y_n = c_1^{n-1}\alpha^{\overline{n-1}}$. Recursive, plane-oriented, and d-ary increasing trees now correspond to $\alpha = 1$, $\alpha = 1/2$, and $\alpha = d/(d-1)$ respectively (binary increasing trees imply $\alpha = 2$).

We remarked at the beginning of the chapter that distances in increasing trees are typically of order $\log n$—to be more specific, it is known that the mean path length of a family of increasing trees of size n is $\Theta(n \log n)$ (Bergeron *et al.*, 1992), and that the expected distance from the root of a randomly chosen node in one of these families is $\Theta(\log n)$. The expected height of a tree, in particular, is also $\Theta(\log n)$, as opposed to the $\Theta(\sqrt{n})$ of simply generated trees (Drmota, 2009).

With nothing but this balanced nature of increasing trees to go on (none of the branches is inordinately large), one can perhaps anticipate that the kth moment of the betweenness centrality of the root node will be of order n^2. This is indeed the case. Instead of deriving first the mean and then the higher-order moments of the root node—as we did in Sections 2.2.1 and 2.4.1—we consider immediately the more general problem of the kth moment of the betweenness centrality of the node with label l, when l is fixed while $n \to \infty$.

Once this analysis is complete, we make use of a recent result of Fuchs (2012) to show that a randomly chosen node in a very simple increasing tree typically has linear-order betweenness centrality. Then in the final section of the chapter, we

consider the maximum betweenness centrality in recursive trees specifically, and the probability that the centroid obtains this maximum.

3.3 Moments of the betweenness centrality of a node

To estimate a parameter of the lth node in an increasing tree, one first needs to describe the tree in relation to that node. We do this here by fixing the subtree containing nodes 1 to l and noting that the rest of the tree is simply a sequence of l forests, each one the descendent branches of a node in the subtree. The generating function that models trees in this way is $y^{(l)}(x)$, since it 'disregards' the subtree containing the first l nodes, so that although their possible configurations are still counted, they no longer contribute to the overall size of the tree.

Take for example the class of recursive trees, whose generating function satisfies:

$$y'(x) = \exp(y(x)) = (1-x)^{-1}.$$

We have:

$$y^{(l)}(x) = (l-1)!\,(1-x)^{-l} = (l-1)!\,y'(x)^l;$$

and since we know that the descendent branches of node l are counted by $y'(x)$, this tells us that l's ancestral branch—which contains the root—has the generating function $(l-1)!\,y'(x)^{l-1}$. In general, the generating function of the ancestral branch of node l is $y^{(l)}(x)/y'(x)$.

Theorem 3.1. *The kth moment of the betweenness centrality of the node with label l in a very simple increasing tree of size n is of order n^{2k}. Specifically, for $k, l \geq 1$:*

$$\mathrm{E}_n(\mathcal{B}_l^k) \sim n^{2k} \frac{\Gamma(\alpha)}{c_1^{l-1} 2^k} \sum_{m=0}^{k} \binom{k}{m} \frac{(-1)^m}{\Gamma(\alpha + l + 2m - 1)} D_l(m),$$

for some constants $D_l(m)$ that depend on $\mathcal{T}$:

$$D_l(m) = c_1^{l-1} \alpha^{\overline{l-1}} \sum_{i=0}^{m} \binom{m}{i} (l-1)^{\overline{2i}} \sum_{r=0}^{m-i} \frac{\alpha}{r!} \left(\prod_{j=1}^{r-2} (1 - j(\alpha - 1)) \right) \times \sum_{\mathcal{Q}_r(m-i)} \binom{m-i}{a_1, \ldots, a_r} \prod_{j=1}^{r} \alpha^{\overline{2a_j - 1}}.$$

(Here $\mathcal{Q}_r(m)$ enumerates the compositions of the integer m into r parts.)

Proof. As in Section 2.2.1, the betweenness centrality $\mathcal{B}_l(T)$ can be interpreted symbolically as the act of choosing nodes from the branches of l. Unfortunately, there is no analogue of Lemma 2.1 that holds for increasing trees, and instead of reducing $\mathcal{B}_l(T)^k$ to a selection of nodes from exactly *two* branches, we will have to consider all possible selections if we wish to accurately derive the constant factors present in the kth moment. To make this computation a bit simpler, we reduce $\mathcal{B}_l(T)$ to a

form involving a sum $B_l(T) = \sum_i |T_i|^2$ over single branches, as opposed to branch pairs:

$$\begin{aligned}\mathcal{B}_l(T)^k = \left(\sum_{i<j} |T_i||T_j|\right)^k &= \frac{1}{2^k}\left(\left(\sum_i |T_i|\right)^2 - \sum_i |T_i|^2\right)^k \\ &= \frac{1}{2^k}\Big((n-1)^2 - B_l(T)\Big)^k \\ &= \frac{1}{2^k}\sum_{m=0}^{k}\binom{k}{m}(-1)^m B_l(T)^m (n-1)^{2(k-m)}. \end{aligned} \tag{3.3}$$

The new function $B_l(T)^m$ counts selections (with replacement) of $2m$ nodes from any number of branches, with the restriction that nodes are chosen two at a time. More specifically, since all labelled branches (whether ordered or unordered) can be numbered deterministically, every selection can be regarded as a composition of the integer m. This means that the generating function $\sum_T B_l(T)^m x^{|T|}$ can be constructed in a piecewise fashion, per composition.

Let l's ancestral branch, which is represented by the generating function $A_l(x) = y^{(l)}(x)/y'(x)$, appear in i of the factors of $B_l(T)^m$, with the remaining factors being distributed among r descendent branches according to the composition $a_1+\cdots+a_r = m-i$. If $\widehat{A}_{l,i}(x)$ denotes the generating function of an ancestral branch from which i nodes have been selected (with replacement), and $\widehat{y}_j(x)$ symbolises the selection of j nodes from a descendent branch, then the cumulative generating function of $B_l(T)^m$ is:

$$\begin{aligned}\sum_{T\in\mathcal{T}} B_l(T)^m \frac{x^{|T|-l}}{(|T|-l)!} = \sum_{i=0}^{m}\binom{m}{i}\widehat{A}_{l,2i}(x)\sum_{r=0}^{m-i}\frac{1}{r!}\phi^{(r)}(y(x)) \\ \times \sum_{\mathcal{Q}_r(m-i)}\binom{m-i}{a_1,\ldots,a_r}\widehat{y}_{2a_1}(x)\cdots\widehat{y}_{2a_r}(x),\end{aligned}$$

where $\mathcal{Q}_r(m)$ enumerates the compositions of m into r parts, and the contribution to the sum over r from $r=0$ vanishes unless $i=m$, in which case $\phi^{(0)}(y(x)) = y'(x)$ and the last sum is 1.

Due to the form of a very simple increasing tree's generating function, $\widehat{y}_j(x) \sim x^j y^{(j)}(x)$ (see the proof of Lemma 2.1). Furthermore, we also have:

$$\begin{aligned} y^{(l)}(x) &= c_1^{l-1}\alpha^{\overline{l-1}}(1-c_1x)^{-(\alpha+l-1)}, \\ \widehat{y}_j(x) &\sim \frac{\alpha^{\overline{j-1}}}{c_1}(1-c_1x)^{-(\alpha+j-1)}, \\ \widehat{A}_{l,i}(x) &\sim c_1^{l-1}\alpha^{\overline{l-1}}(l-1)^{\overline{i}}(1-c_1x)^{-(l+i-1)}, \\ \phi^{(r)}(y(x)) &= c_1^r\alpha(1-c_1x)^{(r-1)\alpha-r}\cdot\prod_{j=1}^{r-2}(1-j(\alpha-1)), \end{aligned}$$

where the asymptotic expressions hold as $x \to 1/c_1$, and the final derivative follows by solving $\phi'(y(x)) = y''(x)/y'(x)$ and differentiating both sides repeatedly. These

Tree	α	$\mathrm{E}_n(\mathcal{B}_1)/n^2$	$\mathrm{V}_n(\mathcal{B}_1)/n^4$	$\mathrm{E}_n(\mathcal{B}_2)/n^2$
recursive	1	1/4	1/96	1/4
plane-oriented	1/2	1/3	4/315	1/5
binary increasing	2	1/6	1/180	1/4

Table 3.1: Asymptotic expressions for the means and variances of the betweenness centralities of some labelled nodes in very simple increasing trees.

approximations can be used to reduce the generating function to an asymptotic form (note the implicit nesting of the sums over i, r, and $\mathcal{Q}_r(m-i)$):

$$\begin{aligned}\sum_{T\in\mathcal{T}} B_l(T)^m \frac{x^{|T|-l}}{(|T|-l)!} &\sim (1-c_1x)^{-(\alpha+2m+l-1)} \\ &\times \left[c_1^{l-1}\alpha^{\overline{l-1}}\sum_{i=0}^{m}\binom{m}{i}(l-1)^{\overline{2i}}\right. \\ &\times \left.\sum_{r=0}^{m-i}\frac{\alpha}{r!}\left(\prod_{j=1}^{r-2}(1-j(\alpha-1))\right)\sum_{\mathcal{Q}_r(m-i)}\binom{m-i}{a_1,\ldots,a_r}\prod_{j=1}^{r}\alpha^{\overline{2a_j-1}}\right] \\ &= (1-\lambda x)^{-(2m+l-1+(r/\lambda))}\cdot D_l(m).\end{aligned}$$

Of course the quantity we really seek is the sum of $\mathcal{B}_l(T)^k$ over trees of size n, of which there are:

$$n!\,[x^n]y(x) \sim c_1^{n-1}n!\frac{n^{\alpha-2}}{\Gamma(\alpha)}.$$

We have, from equation (3.3):

$$\begin{aligned}\mathrm{E}_n(\mathcal{B}_l^k) &= \frac{(n-l)!}{n!\,[x^n]y(x)}[x^{n-l}]\sum_{T\in\mathcal{T}}\mathcal{B}_l(T)^k\frac{x^{|T|-l}}{(|T|-l)!} \\ &\sim n^{2k}\frac{\Gamma(\alpha)}{c_1^{l-1}2^k}\sum_{m=0}^{k}\binom{k}{m}\frac{(-1)^m}{\Gamma(\alpha+2m+l-1)}D_l(m). \qquad \square\end{aligned}$$

A few illustrative values that were obtained using Theorem 3.1 are given in Table 3.1. Although it is not possible to obtain the limiting distribution of a node's betweenness centrality by a construction similar to that of Section 2.2.2, we do see that all the moments of the scaled random variable $\mathcal{B}_l(T)/n^2$ converge to a limit:

$$\lim_{n\to\infty}\mathrm{E}_n(\mathcal{B}_l^k/n^{2k}) = s_{k,l}.$$

Since the betweenness centrality of any node is trivially bounded by $\binom{n-1}{2}$, we automatically obtain $s_{k,l} \le 2^{-k}$, which means that the generating function of the constants $s_{k,l}$ converges in a neighbourhood of 0 and represents a moment generating function. This implies, in light of a result that can be found in the book of Flajolet and Sedgewick (2009, Theorem C.2), that $\mathcal{B}_l(T)/n^2$ converges weakly to a distribution that is characterised by the moments $s_{k,l}$:

Theorem 3.2. *If $\mathcal{T}$ is a family of very simple increasing trees, then the distribution of $\mathcal{B}_l(\mathcal{T}_n)/n^2$ converges weakly to a limiting distribution.*

3.4 A limiting distribution for random nodes

Because increasing trees are generally well balanced, the majority of nodes in any given one will lie near its fringes. These extremal nodes have few descendants, which implies that their betweenness centralities will be relatively small—linear in the size of the tree. So in contrast with the quadratic betweenness centrality that arises by fixing a label l and letting n tend to infinity, we would expect the distribution of a randomly chosen node in an increasing tree to be dominated by linear-order values.

To show that this is indeed the case, one can count nodes with a fixed number of descendants in a subclass of trees of size n, because the proportion of nodes in $\mathcal{T}_n$ that have m descendants is an approximation of the probability that a randomly chosen node has betweenness centrality of roughly nm. Letting $n \to \infty$ makes this approximation more accurate, and yields the limiting distribution of the betweenness centrality of a randomly chosen node.

We note that the number of nodes with a given number of descendants—referred to as the *subtree size profile* of a tree—has recently been studied for various families of increasing trees. In fact, the expected proportion of nodes with $m-1$ descendants (each forming a rooted subtree of size m) has been given explicitly for the most interesting families (see Fuchs, 2012, Section 3 and Theorem 4.1). Letting $U_m(T)$ denote the number of subtrees of size m in a random tree, we perform the derivation in a more general way here.

Lemma 3.2. *For $1 \leq m < n$, the expected number of subtrees of size m in a random very simple increasing tree of size n is given by:*

$$\mathrm{E}_n(U_m) = \frac{\alpha(\alpha+n-1)}{(\alpha+m)(\alpha+m-1)}.$$

Proof. Firstly, note that $\mathrm{E}_n(U_m) = \sum_l \mathrm{P}_n(S_l = m)$, where S_l is the size of the subtree rooted at l. To form a tree of size n in which l has $m-1$ descendants, begin with a tree of l nodes, and consider the ways in which the remaining $n-l$ nodes can be attached: there are currently $lc_1 + c_2 = c_1(\alpha + l - 1)$ attachment points[2], of which $c_1\alpha$ belong to l, and the remaining $c_1(l-1)$ do not. Continuing iteratively, and accounting for labels, it follows that:

$$\begin{aligned}
\mathrm{P}_n(S_l = m) &= \binom{n-l}{m-1} \frac{\alpha^{\overline{m-1}}(l-1)^{\overline{n-l-m+1}}}{(\alpha+l-1)^{\overline{n-l}}} \\
&= \binom{n-l}{m-1} \frac{\Gamma(\alpha+m-1)\Gamma(n-m)}{\Gamma(\alpha+n-1)} \frac{\Gamma(\alpha+l-1)}{\Gamma(\alpha)\Gamma(l-1)} \\
&= \binom{n-l}{m-1} \mathrm{B}(n-m, \alpha+m-1) \binom{\alpha+l-2}{l-2} \alpha. \qquad (3.4)
\end{aligned}$$

[2]This is technically only correct when each tree is assigned a weight of 1 (corresponding to $c_1 = 1$ in the case of recursive trees and $c_2 = 1$ otherwise). The quantity $lc_1 + c_2$ is in fact the sum of the weight-adjusting factors (either always c_1 or always c_2) that would result from each of the possible attachments, and is thus proportional to the number of attachment points—but need not be an integer.

Summing over possible labels (the root is omitted since $m < n$) yields:

$$\begin{aligned}\mathrm{E}_n(U_m) &= \sum_{l=2}^{n-m+1} \mathrm{P}_n(S_l = m) \\ &= \alpha\,\mathrm{B}(n-m, \alpha+m-1) \sum_{l=2}^{n-m+1} \binom{\alpha+l-2}{l-2}\binom{n-l}{m-1} \\ &= \alpha\,\mathrm{B}(n-m, \alpha+m-1)\binom{\alpha+n-1}{n-m-1}.\end{aligned}$$

in which the final step is due to the Chu-Vandermonde identity, once the numerators of the binomial coefficients have been converted to constants:

$$\sum_{l=0}^{n-m-1} \binom{\alpha+l}{l}\binom{n-l-2}{m-1} = (-1)^{n-m-1} \sum_{l=0}^{n-m-1} \binom{-\alpha-1}{l}\binom{-m}{n-m-1-l}.$$

The stated result is obtained after simplifying. □

Mirroring Section 2.2.3, let $\mathcal{W}(T)$ denote the betweenness centrality of a random node in T. The limiting behaviour of $\mathcal{W}(T)$ is detailed in the following theorem—which shares the form of Theorem 2.4.

Theorem 3.3. *The linearly scaled betweenness centrality of a randomly chosen node in a very simple increasing tree of size n, $\mathcal{W}(\mathcal{T}_n)/n$, converges in distribution to the discrete random variable $\mathcal{W}_\star$ with support $\mathbb{Z}_{\geq 0}$ and mass function:*

$$\mathrm{P}(\mathcal{W}_\star = m) = \frac{\alpha}{(\alpha+m)(\alpha+m+1)}.$$

Proof. By Lemma 3.2, the expected number of nodes with m descendants (m is fixed) in a tree of size n is:

$$\mathrm{E}_n(U_{m+1}) = \frac{\alpha(\alpha+n-1)}{(\alpha+m)(\alpha+m+1)},$$

and scaling by n, we obtain a limiting proportion:

$$s(m) = \lim_{n\to\infty} \frac{\mathrm{E}_n(U_{m+1})}{n} = \frac{\alpha}{(\alpha+m)(\alpha+m+1)}.$$

Since the $s(m)$ sum to 1, and the betweenness centrality of a node v with m descendants, divided by n, tends to m, the result follows in the same way as Theorem 2.3. □

Corollary 3.1. *With $0 < \varepsilon < 1$, we have, for* recursive trees*:*

$$\mathrm{P}_n(|\mathcal{W}/n - m| < \varepsilon) \xrightarrow[n\to\infty]{} \frac{1}{(m+1)(m+2)}.$$

For plane-oriented trees*:*

$$\mathrm{P}_n(|\mathcal{W}/n - m| < \varepsilon) \xrightarrow[n\to\infty]{} \frac{2}{(2m+1)(2m+3)}.$$

For d-ary increasing trees*:*

$$\mathrm{P}_n(|\mathcal{W}/n - m| < \varepsilon) \xrightarrow[n\to\infty]{} \frac{d(d-1)}{((d-1)m+d)((d-1)m+2d-1)}.$$

The idea that a node near to the fringes of an increasing tree must have small betweenness centrality is intuitive, and from it, one can reason that a node with a large label—which is likely to be far from the root—should have small betweenness centrality as well. In the next section, we derive an explicit bound on the probability, in a recursive tree, that a node with a given label can attain a significantly large betweenness centrality. This bound allows us to numerically determine the expected maximum betweenness centrality in a random recursive tree, as well as the probability that the centroid has maximal betweenness centrality.

3.5 Maximum betweenness centrality and the centroid

For the rest of this chapter, we focus on recursive trees, although analogous statements can be obtained for other varieties of increasing trees in the same manner.

Our first goal is to show that the node of maximal betweenness centrality in a recursive tree is unlikely to have a large label. Specifically, if $Q(T)$ is a random variable over the label of this node, we wish to show that as the size of the tree tends to infinity, the probability distribution $\mathrm{P}_n(Q = l)$ converges weakly to a discrete limiting distribution.

Intuitively, this concentration property should hold, because the node of maximal betweenness centrality cannot have any particularly large branches—specifically, its ancestral branch cannot be disproportionately large—and thus is likely to have a large number of descendants. The chance of this being true of a node with label l should decrease exponentially as l increases, so we would expect $\mathrm{P}_n(Q = l)$ to decrease exponentially as well. To be more specific, we have the following result:

Lemma 3.3. *The probability* $\mathrm{P}_n(Q \geq L)$ *that the maximum betweenness centrality is attained by a node whose label is at least* L *can be bounded above as follows:*

$$\mathrm{P}_n(Q \geq L) < 16\left(\frac{L-1}{3} + 1\right)\left(\frac{3}{4}\right)^{L-1}.$$

Proof. First of all, we note that in a tree of size $n \geq 2$, a node l which has $m-1$ descendants cannot possibly have maximal betweenness centrality if $m < n/4$. To see why this assertion holds, recall from Section 2.3 that a lower bound on the maximum betweenness centrality in a tree is given by the lower bound on the centroid's betweenness centrality, $n(n-2)/4$. Furthermore, the betweenness centrality of node l is at most:

$$\begin{aligned}(n-m)(m-1) + \binom{m-1}{2} &= m(n-2) - \frac{1}{2}\left(m^2 - 3m - 2\right) - n \\ &\leq m(n-2),\end{aligned}$$

which is strictly less than $n(n-2)/4$ whenever $m < n/4$.

Such small subtrees, however, become ever more likely as l is increased, and in fact $\mathrm{P}_n(S_l \geq n/4) < (l-1)(3/4)^{l-2}$ for $l > 1$, where, as in the proof of Lemma 3.2, $S_l(T)$ yields the size of the subtree rooted at l. This is also simple to prove: equation (3.4) reduces, in the case of recursive trees, to:

$$\mathrm{P}_n(S_l = m) = \binom{n-m-1}{l-2} \Big/ \binom{n-1}{l-1},$$

from which the result follows algebraically:

$$\begin{aligned}
\mathrm{P}_n(S_l \geq n/4) &= \left[\binom{\lfloor 3n/4 \rfloor - 1}{l-2} + \binom{\lfloor 3n/4 \rfloor - 2}{l-2} + \cdots + \binom{l-2}{l-2}\right] \Big/ \binom{n-1}{l-1} \\
&< (\lfloor 3n/4 \rfloor - l + 2) \binom{\lfloor 3n/4 \rfloor - 1}{l-2} \Big/ \binom{n-1}{l-1} \\
&= (l-1) \frac{(\lfloor 3n/4 \rfloor - l + 2)}{n-l+1} \frac{(\lfloor 3n/4 \rfloor - 1)^{\underline{l-2}}}{(n-1)^{\underline{l-2}}} \\
&< (l-1) \left(\frac{\lfloor 3n/4 \rfloor}{n}\right)^{l-2} \\
&\leq (l-1) \left(\frac{3}{4}\right)^{l-2},
\end{aligned}$$

as long as $n \geq 4$. Thus we have:

$$\mathrm{P}_n(Q = l) \leq \mathrm{P}_n(S_l \geq n/4) < (l-1)\left(\frac{3}{4}\right)^{l-2},$$

and a bound on the tail probabilities follows immediately, for any $n \geq 4$:

$$\begin{aligned}
\mathrm{P}_n(Q \geq L) &= \sum_{l \geq L} \mathrm{P}_n(Q = l) \\
&< \sum_{l \geq L-1} l \left(\frac{3}{4}\right)^{l-1} = 16\left(\frac{L-1}{3} + 1\right)\left(\frac{3}{4}\right)^{L-1}.
\end{aligned}$$

Although it is of no real significance, the bound holds for trees of size $n < 4$ as well: in these cases the probability is trivially 0 for $L > 3$, and the bound is greater than 1 whenever $1 \leq L \leq 3$. This completes the proof of the lemma. □

The upper bound on $\mathrm{P}_n(Q \geq L)$ is important firstly because it is uniform in n, which means that regardless of the size of the tree, the probability of the maximum betweenness centrality being attained at a label L or greater is bounded from above; and secondly because it approaches 0 as $L \to \infty$. Conversely, this means that for any reasonably large finite tree, $\mathrm{P}_n(Q < L)$ is positively bounded from below (independently of n).

Now we can take an approach similar to that which was used in the proof of Theorem 2.5—however the technical details here are somewhat simpler. Before we formulate and prove this result, consider the limiting distribution of the root betweenness centrality established in Theorem 3.2: a recursive tree decomposes

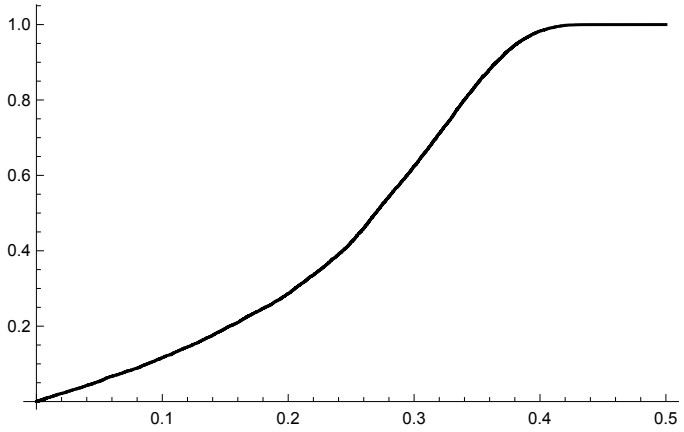


Figure 3.1: The cumulative distribution function of the limiting distribution of root betweenness centrality in recursive trees.

naturally into the first root branch (that is, the branch containing label 2), and the rest of the tree. The number of trees of size n in which this branch has n_1 nodes is:

$$\binom{n-2}{n_1-1}(n_1-1)!\,(n-n_1-1)! = (n-2)!,$$

implying that the size of the first branch is uniformly distributed. Conditioned on this size, each of the two subtrees is again a uniformly random recursive tree. If we let X be a random variable representing the limiting distribution of the root's betweenness centrality, then we obtain from this decomposition that:

$$X \overset{(d)}{=} U^2\widetilde{X} + U(1-U),$$

where U follows a uniform distribution on $[0,1]$ and $\widetilde{X}$ follows the same distribution as X and is independent of U. Making use of the 'smoothing' influence of the uniform distribution, one can use this decomposition also to prove that X is continuous (see Figure 3.1 for a plot of the distribution function). A slightly more general decomposition yields the following theorem:

Theorem 3.4. *The maximum betweenness centrality of a random recursive tree of size n, divided by n^2, converges weakly to a limiting distribution. The probability that the maximum betweenness centrality is attained by the centroid tends to a positive constant, and the random variable $Q(\mathcal{T}_n)$ giving the label of the node with maximum betweenness centrality converges to a discrete limiting distribution.*

Proof. Instead of the maximum betweenness centrality of an arbitrary node, we only consider the maximum among the first l nodes. By virtue of Lemma 3.3, we can then let l go to infinity.

If we fix the tree formed by the first l nodes (for which there are only finitely many possibilities), it decomposes naturally into l disjoint recursive trees, each rooted at a node $1,\dots,l$. Let $n_1, n_2, \dots, n_l$ be the sizes of these subtrees; then there are:

$$\binom{n-l}{n_1-1, n_2-1, \dots, n_l-1} \cdot (n_1-1)!\,(n_2-1)!\cdots(n_l-1)! = (n-l)!$$

possible trees, which is independent of the values of $n_1, \ldots, n_l$, and also of the shape of the tree formed by the first l labels. Therefore, the vector formed by the sizes of these l trees converges, upon normalisation by a factor n^{-1}, to a uniformly random composition $(U_1, U_2, \ldots, U_l)$ of 1. The root betweenness centralities of the l trees converge—again upon suitable normalisation—to random variables $X_1, X_2, \ldots, X_l$ that all follow the same limiting distribution (described earlier). The normalised limits of the betweenness centralities of nodes 1 to l are simple functionals of $U_1, \ldots, U_l$ and $X_1, \ldots, X_l$ (also depending on the shape of the tree formed by nodes $1, \ldots, l$), so the theorem follows. □

With the help of a numerical simulation, we find that the expected maximum betweenness centrality in a recursive tree is asymptotically equal to $0.35n^2$, and that the probability of the centroid node also being a node of maximal betweenness centrality is roughly 0.87. In addition, it appears that the expected label of the node of maximal betweenness centrality (breaking ties in favour of the node with the smaller label if necessary, although this occurs with asymptotic probability 0) is 2.57, and that its mean distance from the root is 1.03.[3]

[3]Such a simulation also recovers some interesting results of Moon (2002), which state that the expected label of the centroid of a recursive tree is 5/2, and that its mean distance from the root is 1. These results will be generalised in the next chapter.

Chapter 4

The Centroid in Increasing Trees

4.1 Introduction and known results

Our final chapter is concerned solely with the nearest centroid of a random increasing tree (the known limiting behaviour of a simply generated tree's centroid having been discussed in Section 1.2.3). To begin with, note that since a tree can only have one or two centroid nodes, the parameters of interest that arise in this context are quite different from those of the previous two chapters. Instead of random variables that range over centrality values, we are now more concerned with parameters more familiar to the tree itself: depth, label, and subtree size. In essence, we want to know where the centroid lies.

As it happens, this question can be answered both precisely and in some generality by looking at the behaviour of the above parameters in the limit $n \to \infty$. (Finite results are within reach for specific families (see below), but do not seem to be available in general.) In short, we derive limits for the distributions of the depth, label, and ancestral branch size of the centroid in a random very simple increasing tree, as well as the limits of the moments of these distributions. The mean depth and label are, in the limit, particularly plain: in the notation of Chapter 3 (in which plane-oriented, recursive, and binary increasing trees correspond to $\alpha = 1/2$, $\alpha = 1$, and $\alpha = 2$ respectively) the expected depth of the centroid tends to α and the expected label to $1 + 3\alpha/2$. Of course the limiting distributions yield more specialised values as well.

4.1.1 The centroid in recursive trees

The depth and label of the centroid were actually the subject of a similar (topically) study by Moon (2002), who, considering the special case of recursive trees, obtained explicit formulae for their means: if $M = \lfloor (n-1)/2 \rfloor$, then the expected depth of the nearest centroid in a recursive tree of size n is $M/(n-M)$, and as $n \to \infty$, the probability that this depth is at least h tends to $(\ln 2)^h/h!$. Furthermore, the expected label of the centroid is:

$$\frac{1}{2} + \frac{n(n+1)}{2(n-M)(n+1-M)},$$

and the probability that the centroid and root coincide tends to $1 - \ln 2$. Moon also gave the limiting probability for any specific label—see Corollary 4.4 below. Finally, it was shown that the probability that the ancestral branch of the centroid has B nodes is:

$$\frac{n}{(n-B)(n-B+1)} \left(1 - \sum_{b=\lceil (n+1)/2 \rceil}^{\lfloor n-B-1 \rfloor} 1/h \right),$$

and that the mean of the proportion of the tree accounted for by this branch approaches $(\ln 2)^2/2$.

Since we are interested in the locality of the centroid in large, general trees here, all of the stated asymptotic expressions will reappear as special cases of results in this chapter.

4.2 The depth of the centroid

The starting point for our results involving the centroid is the observation that in a tree of size n, the node with label k is on the path between the root and the centroid (always considering the nearer if there are two) if and only if k has at least $\lfloor n/2 \rfloor$ descendent nodes (Moon, 2002). Let $\Lambda_k(1/2)$ mark the occurrence of this event. Our first goal will be to derive a closed form for the probability $\mathrm{P}_n(\Lambda_k(1/2))$, which, due to its reliance on the label k, will be obtained via the exponential generating function $y^{(k)}(x)$. By combining this closed form with a known probability generating function for the depth of node k, we can describe the event that k is both at depth h and on the path from the root to the centroid, and by marginalising over k, we arrive at a generating function that yields, in the limit $n \to \infty$, the behaviour of the depth of the (nearest) centroid node.

4.2.1 The probability of a node appearing on the centroid path

There is, of course, a natural extension of $\Lambda_k(1/2)$ to the event $\Lambda_k(\sigma)$ that node k has at least $\lfloor \sigma n \rfloor$ descendants, where $1/2 \leq \sigma < 1$; and in fact the closed form we desire for $\mathrm{P}_n(\Lambda_k(1/2))$ is simply a special case of a similar expression for the more general probability $\mathrm{P}_n(\Lambda_k(\sigma))$. Although we have no need for it in determining the depth and label of the centroid, this more general version will in fact be required in Section 4.4, when considering the size of the centroid branch containing the root;

and since the two derivations are essentially identical, we deal with the variable case immediately.

Consider the exponential generating function $y^{(k)}(x)$, which counts trees (of size at least k) as if the nodes 1 through k were 'size-less'. We have:

$$\begin{aligned} y^{(k)}(x) &= c_1^{k-1}\alpha^{\overline{k-1}}(1-c_1x)^{-(\alpha+k-1)} \\ &= y_k \cdot y'(x) \cdot (1-c_1x)^{-(k-1)}, \end{aligned}$$

where the three terms can be interpreted as accounting for the configurations of the first k nodes, the subtree rooted at node k, and the remaining subtrees, respectively. With this in mind, the number (more accurately, the total weight) of trees of size n in which k has m descendent nodes is:

$$y_k\binom{n-k}{m}(m!\,[x^m]y'(x))\Big((n-k-m)!\,[x^{n-k-m}](1-c_1x)^{-(k-1)}\Big),$$

and the proportion of trees in which k's descendants number at least $\lfloor \sigma n \rfloor$ is, for $k > 1$:

$$\begin{aligned} \mathrm{P}_n(\Lambda_k(\sigma)) &= \frac{y_k(n-k)!}{y_n}\sum_{m=\lfloor \sigma n\rfloor}^{n-k}([x^m]y'(x))\Big([x^{n-k-m}](1-c_1x)^{-(k-1)}\Big) \\ &= \sum_{m=\lfloor \sigma n\rfloor}^{n-k}\binom{\alpha+m-1}{m}\binom{n-m-2}{k-2}\Big/\binom{\alpha+n-2}{n-k}, \end{aligned} \tag{4.1}$$

where we have used the fact that $y_n = (n-k)!\,[x^{n-k}]y^{(k)}(x)$.

As long as the label k is small relative to the tree's size n, a more explicit (and usable) expression for $\mathrm{P}_n(\Lambda_k(\sigma))$ holds:

Theorem 4.1. *For $1 < k < \lceil (1-\sigma)n \rceil$ such that $k = o(n^{1/4})$, the probability $\mathrm{P}_n(\Lambda_k(\sigma))$ that the node with label k has at least $\lfloor \sigma n \rfloor$ descendants satisfies:*

$$\mathrm{P}_n(\Lambda_k(\sigma)) = I_{1-\sigma}(k-1,\alpha)\left(1+O\left(\frac{k^2}{\sqrt{n}}\right)\right), \tag{4.2}$$

where the error term is uniform in σ over subsets of the form $[1/2, 1-\delta)$, for $0 < \delta < 1/2$; and:

$$I_x(a,b) = \frac{\mathrm{B}(x;a,b)}{\mathrm{B}(a,b)}$$

is the regularised incomplete beta function.

Proof. The main step in going from (4.1) to (4.2) is an application of the Euler-Maclaurin formula; but first, note that as n grows:

$$\begin{aligned} \mathrm{P}_n(\Lambda_k(\sigma)) &= \frac{\displaystyle\sum_{m=\lfloor \sigma n\rfloor}^{n-k}\frac{m^{\alpha-1}}{\Gamma(\alpha)}\left(1+O\left(\frac{1}{m}\right)\right)\binom{n-m-2}{k-2}}{\dfrac{(n-k)^{\alpha+k-2}}{\Gamma(\alpha+k-1)}\left(1+\dfrac{(\alpha+k-1)^{\underline{2}}}{2(n-k)}+O\left(\dfrac{k^4}{(n-k)^2}\right)\right)} \\ &= \frac{n^{-(\alpha+k-2)}}{\mathrm{B}(k-1,\alpha)}\sum_{m=k}^{\lceil (1-\sigma)n)\rceil}(n-m)^{\alpha-1}(m-2)^{\underline{k-2}}\left(1+O\left(\frac{k^2}{n}\right)\right). \end{aligned}$$

Splitting the sum at an intermediate value $n^{1-\varepsilon}$, where $0 < \varepsilon < 1$, reveals that the contribution from smaller values of m is minimal:

$$n^{-(\alpha+k-2)} \sum_{m=k}^{\lceil n^{1-\varepsilon} \rceil - 1} (n-m)^{\alpha-1} (m-2)^{\underline{k-2}} = O\Big(n^{-(k-1)\varepsilon}\Big). \tag{4.3}$$

It suffices now to apply the Euler-Maclaurin formula to the dominant portion of the sum:

$$\begin{aligned} & n^{-(\alpha+k-2)} \sum_{m=\lceil n^{1-\varepsilon} \rceil}^{\lceil (1-\sigma)n \rceil} (n-m)^{\alpha-1} m^{k-2} \left(1 + O\left(\frac{k^2}{m}\right)\right) \\ &= n^{-(\alpha+k-2)} \int_{\lceil n^{1-\varepsilon} \rceil}^{\lceil (1-\sigma)n \rceil} (n-u)^{\alpha-1} u^{k-2}\, du \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right) + O\left(\frac{1}{n}\right) \\ &= \int_{n^{-\varepsilon}+O(1/n)}^{1-\sigma+O(1/n)} t^{k-2}(1-t)^{\alpha-1}\, dt \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right). \end{aligned} \tag{4.4}$$

Note that in absorbing the error term $O(1/n)$ we have implicitly treated σ as a constant with respect to n, and as such, the uniformity of the error term only holds as long as σ is not allowed to tend arbitrarily close to 1. Also, as long as $\varepsilon \geq 1/k$, the term $k^2/n^{1-\varepsilon}$ is not smaller in order than that of equation (4.3), and the contribution of the first portion of the sum can be ignored. This is the case for all $k > 1$ when $1/2 \leq \varepsilon < 1$.

As n grows, the bounds of the integral in (4.4) approach 0 and $1-\sigma$ at the following rates:

$$\int_0^{n^{-\varepsilon}+O(1/n)} t^{k-2}(1-t)^{\alpha-1}\, dt = O\Big(n^{-(k-1)\varepsilon}\Big),$$

and

$$\int_{1-\sigma}^{1-\sigma+O(1/n)} t^{k-2}(1-t)^{\alpha-1}\, dt = O\left(\frac{1}{n}\right).$$

Noting that these terms are also of orders less than $k^2/n^{1-\varepsilon}$ when $1/2 \leq \varepsilon < 1$, we see that the probability of node k having at least $\lfloor \sigma n \rfloor$ descendants can be written, for $1 < k = o(n^{(1-\varepsilon)/2})$, as:

$$\begin{aligned} \mathrm{P}_n(\Lambda_k(\sigma)) &= \frac{1}{\mathrm{B}(k-1,\alpha)} \int_0^{1-\sigma} t^{k-2}(1-t)^{\alpha-1}\, dt \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right) \\ &= I_{1-\sigma}(k-1,\alpha) \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right). \end{aligned}$$

□

We mention also that an alternative representation of $\mathrm{P}_n(\Lambda_k(\sigma))$ can be obtained using the binomial theorem, since

$$\begin{aligned} \int_0^{1-\sigma} t^{k-2}(1-t)^{\alpha-1}\, dt &= \mathrm{B}(k-1,\alpha) - \int_{1-\sigma}^{1} t^{k-2}(1-t)^{\alpha-1}\, dt \\ &= \mathrm{B}(k-1,\alpha) - \sum_{l=0}^{k-2} \binom{k-2}{l} \frac{(-1)^l}{l+\alpha} \sigma^{l+\alpha}. \end{aligned}$$

Corollary 4.1. *For $1 < k = o(n^{1/4})$, the probability that node k is on the path from the root to the (nearest) centroid node satisfies:*

$$\mathrm{P}_n(\Lambda_k(1/2)) = I_{1/2}(k-1,\alpha)\left(1 + O\left(\frac{k^2}{\sqrt{n}}\right)\right).$$

Finally, we give limiting probabilities for the event that node k is on the path to the centroid in the two simplest families of very simple increasing trees, for which the incomplete beta function can be easily simplified:

Corollary 4.2. *For* recursive trees*:*

$$\lim_{n\to\infty} \mathrm{P}_n(\Lambda_k(1/2)) = I_{1/2}(k-1,1) = 2^{-(k-1)}.$$

For binary increasing trees*:*

$$\lim_{n\to\infty} \mathrm{P}_n(\Lambda_k(1/2)) = I_{1/2}(k-1,2) = (k+1)2^{-k}.$$

4.2.2 A uniform bound on the path probability

In addition to the asymptotic expression for $\mathrm{P}_n(\Lambda_k(\sigma))$ given above, we can show that the probability of a specific node k appearing on the path to the centroid not only vanishes for large k, but does so exponentially in k and uniformly over n. In particular:

$$\mathrm{P}_n(\Lambda_k(1/2)) \leq \frac{\alpha^{\overline{k-1}}}{(k-1)!}2^{-(k-2)}. \tag{4.5}$$

It is this fact, in combination with Corollary 4.1, that will allow us to derive limiting distributions for events that depend on $\mathrm{P}_n(\Lambda_k(1/2))$.

Once again a more general form of this result will be required later, in Section 4.4, so to avoid a repeated derivation, we give the version for variable σ here.

Lemma 4.1. *For $1 \leq k$ and $1/2 \leq \sigma < 1$, the probability that node k has at least $\lfloor \sigma n \rfloor$ descendants in a tree of size $n \geq 3$ is subject to an upper bound that decreases exponentially with k:*

$$\mathrm{P}_n(\Lambda_k(\sigma)) \leq \frac{3}{\sigma}\frac{\alpha^{\overline{k-1}}}{(k-1)!}(1-\sigma)^{k-1}. \tag{4.6}$$

Proof. Firstly, take note of the following inequality involving binomial coefficients: if $\alpha \in \mathbb{R}_{\geq 0}$ and $m \leq n \in \mathbb{Z}_{\geq 0}$, then:

$$\begin{aligned}\binom{m-1+\alpha}{m} &= \frac{(\alpha+m-1)^{\underline{m}}}{m!} \\ &\leq \frac{(\alpha+n)\cdots(\alpha+m)}{n\cdots m}\frac{(\alpha+m-1)^{\underline{m}}}{m!} \\ &= \frac{n+\alpha}{m}\binom{n-1+\alpha}{n}.\end{aligned} \tag{4.7}$$

Since $\mathrm{P}_n(\Lambda_1(\sigma)) = 1$, and $\mathrm{P}_n(\Lambda_k(\sigma)) = 0$ whenever $k > \lceil(1-\sigma)n\rceil$, we need only consider $1 < k \leq \lceil(1-\sigma)n\rceil$. In this case:

$$\begin{aligned}
\mathrm{P}_n(\Lambda_k(\sigma)) &= \sum_{m=\lfloor\sigma n\rfloor}^{n-k} \binom{m-1+\alpha}{m}\binom{n-m-2}{k-2}\Big/\binom{n-2+\alpha}{n-k} \\
&\leq \frac{n-k+\alpha}{\lfloor\sigma n\rfloor}\binom{n-k-1+\alpha}{n-k}\binom{\lceil(1-\sigma)n\rceil-1}{k-1}\Big/\binom{n-2+\alpha}{n-k} \\
&= \frac{n-k+\alpha}{\lfloor\sigma n\rfloor}\frac{(\lceil(1-\sigma)n\rceil-1)^{\underline{k-1}}}{(k-1)!}\frac{\Gamma(n-k+\alpha)}{\Gamma(n-1+\alpha)}\frac{\Gamma(k-1+\alpha)}{\Gamma(\alpha)} \\
&= \frac{\alpha^{\overline{k-1}}}{(k-1)!}\frac{n-k+\alpha}{\lfloor\sigma n\rfloor}\frac{(\lceil(1-\sigma)n\rceil-1)^{\underline{k-1}}}{(n-2+\alpha)^{\underline{k-1}}} \\
&\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!}\frac{n-k+\alpha}{n-2}\frac{n(n-2)\cdots(n-2k+4)}{(n-2+\alpha)\cdots(n-k+\alpha)}\frac{(1-\sigma)^{k-1}}{\sigma} \\
&\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!}\frac{n}{n-2}\frac{(1-\sigma)^{k-1}}{\sigma},
\end{aligned}$$

which yields the stated bound whenever $n \geq 3$. In the specific case $\sigma = 1/2$, the final two lines are slightly stronger, resulting in equation (4.5). □

4.2.3 A limiting distribution for the depth of the centroid

Let $\mathcal{D}(T)$ denote the depth of the centroid—that is, the number of edges on the path from the root to the centroid node (the nearer if there are two)—in a random tree T. As mentioned earlier, $\mathcal{D}(T) \geq h$ if and only if there is a vertex at depth h that is on this path. Breaking this event down per label, we may say that the depth of the centroid is at least h if and only if for some label k, node k has both depth h and is present on the path to the centroid. Since these per-label events are disjoint:

$$\mathrm{P}_n(\mathcal{D} \geq h) = \sum_{k\geq 1} \mathrm{P}(D_k = h)\,\mathrm{P}_n(\Lambda_k(1/2)),$$

in which D_k is a random variable over the possible depths $h \in \mathbb{Z}_{\geq 0}$ of node k. This random variable has a known probability generating function (Panholzer and Prodinger, 2007):

$$\sum_{h\geq 0} \mathrm{P}(D_k = h)v^h = \prod_{j=0}^{k-2}\frac{\alpha v + j}{\alpha + j} = \frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}, \tag{4.8}$$

which is independent of n, implying that $\mathrm{P}_n(D_k = h) = \mathrm{P}(D_k = h)$ is as well. Combining these two expressions yields a (complementary cumulative) probability generating function for the depth of the centroid:

$$\begin{aligned}
C_n(v) = \sum_{h\geq 0}\mathrm{P}_n(\mathcal{D}\geq h)v^h &= \sum_{k\geq 1}\mathrm{P}_n(\Lambda_k(1/2))\sum_{h\geq 0}\mathrm{P}(D_k = h)v^h \\
&= \sum_{k\geq 1}\mathrm{P}_n(\Lambda_k(1/2))\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}.
\end{aligned} \tag{4.9}$$

Our goal in this section, then, is to apply the asymptotic form of $\mathrm{P}_n(\Lambda_k(1/2))$ given in Corollary 4.1 to find the limit of this generating function, and then simply extract the desired probabilities as coefficients.

Theorem 4.2. *The depth $\mathcal{D}(\mathcal{T}_n)$ of the centroid node in a random tree of size n converges in probability to the discrete random variable $\mathcal{D}_\star$ supported by $\mathbb{Z}_{\geq 0}$ and with cumulative distribution function:*

$$\mathrm{P}(\mathcal{D}_\star \geq h) = \left(\frac{\alpha}{\alpha-1}\right)^h \left(1 - 2^{1-\alpha} \sum_{j=0}^{h-1} \frac{((\alpha-1)\ln 2)^j}{j!}\right)$$

and mass function:

$$\mathrm{P}(\mathcal{D}_\star = h) = \frac{\alpha^h}{(\alpha-1)^{h+1}} \left[2^{1-\alpha}\left(\sum_{j=0}^{h-1} \frac{((\alpha-1)\ln 2)^j}{j!} + \frac{\alpha\big((\alpha-1)\ln 2\big)^h}{h!}\right) - 1\right].$$

Proof. Letting n be large, and fixing $K = \lfloor n^{1/4-\varepsilon} \rfloor$ for an arbitrarily small $\varepsilon > 0$, Corollary 4.1 and equation (4.9) above imply:

$$C_n(v) = 1 + \sum_{k=2}^{K} I_{1/2}(k-1,\alpha)\left(1 + O\left(\frac{k^2}{\sqrt{n}}\right)\right)\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}} + \sum_{k=K+1}^{\lceil n/2 \rceil} \mathrm{P}_n(\Lambda_k(1/2))\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}.$$

As $n \to \infty$, assuming $|v| < 1$, the second sum tends to 0:

$$\begin{aligned}
\sum_{k=K+1}^{\lceil n/2 \rceil} \left|\mathrm{P}_n(\Lambda_k(1/2))\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}\right| &\leq \sum_{k=K+1}^{\lceil n/2 \rceil} \frac{\alpha^{\overline{k-1}}}{(k-1)!} 2^{-(k-2)} \\
&\leq 2 \sum_{k=K}^{\lceil n/2 \rceil - 1} \binom{k-1+\alpha}{k} 2^{-k} \\
&\leq n \frac{\frac{n}{2}+\alpha}{\lfloor n^{1/4-\varepsilon} \rfloor} \binom{\frac{n}{2}-1+\alpha}{\frac{n}{2}} 2^{-\lfloor n^{1/4-\varepsilon} \rfloor} \\
&= O\left(\frac{n^{\alpha+3/4+\varepsilon}}{2^{\lfloor n^{1/4-\varepsilon} \rfloor}}\right) \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

where we have made use of the upper bound on $\mathrm{P}_n(\Lambda_k(1/2))$ given in equation (4.5), as well as the binomial inequality (4.7). Similarly, the tail which replaces this sum is negligible:

$$\begin{aligned}
\sum_{k>K} \left|I_{1/2}(k-1,\alpha)\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}\right| &\leq \sum_{k>K} \mathrm{B}(k-1,\alpha)^{-1} \int_0^{1/2} t^{k-2}(1-t)^{\alpha-1}\,dt \\
&\leq \alpha \sum_{k>K} \binom{\alpha+k-2}{k-2} 2^{-(k-1)} \\
&= O\Big(K^\alpha 2^{-K}\Big) \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

Thus, for values of v within the unit circle, the pointwise limit of $C_n(v)$ is:

$$\begin{aligned} C(v) = \lim_{n\to\infty} C_n(v) &= 1 + \sum_{k\geq 2} I_{1/2}(k-1,\alpha)\frac{\Gamma(\alpha v + k - 1)\Gamma(\alpha)}{\Gamma(\alpha v)\Gamma(\alpha + k - 1)} \\ &= 1 + \sum_{k\geq 2} \frac{\Gamma(\alpha v + k - 1)}{\Gamma(\alpha v)\Gamma(k-1)} \int_0^{1/2} t^{k-2}(1-t)^{\alpha-1}\,dt \\ &= 1 + \alpha v \int_0^{1/2} (1-t)^{\alpha-1} \sum_{k\geq 2} \binom{\alpha v + k - 2}{k-2} t^{k-2}\,dt \\ &= 1 + \alpha v \int_0^{1/2} (1-t)^{-\alpha(v-1)-2}\,dt \\ &= 1 + \frac{\alpha v}{\alpha(v-1)+1}\Big(2^{\alpha(v-1)+1} - 1\Big). \end{aligned} \tag{4.10}$$

Considering now the probability generating function involving the mass $P_n(\mathcal{D} = h)$, say $A_n(v)$, we see that this too converges:

$$\begin{aligned} A_n(v) = \sum_{h\geq 0} \mathrm{P}_n(\mathcal{D} = h)v^h = C_n(v) - \frac{C_n(v) - 1}{v} \\ \to C(v) - \frac{C(v) - 1}{v} \\ = 1 + \frac{\alpha(v-1)}{\alpha(v-1)+1}\Big(2^{\alpha(v-1)+1} - 1\Big) = A(v). \end{aligned} \tag{4.11}$$

Since pointwise convergence of probability generating functions (here $A_n(v) \to A(v)$) implies convergence in probability of their distributions (Flajolet and Sedgewick, 2009, Theorem IX.1), we have the stated theorem via $\mathrm{P}_n(\mathcal{D} \geq h) \to [v^h]C(v)$. The limiting mass function follows naturally. □

The singular case $\alpha = 1$, corresponding to recursive trees, is a simple consequence of equation (4.10):[1]

Corollary 4.3 (Moon (2002))**.** *For* recursive trees*:*

$$\lim_{n\to\infty} \mathrm{P}_n(\mathcal{D} \geq h) = [v^h]2^v = (\ln 2)^h/h!.$$

4.2.4 Moments of the depth distribution

To close our discussion of the centroid's depth, we consider the moments of $\mathcal{D}(T)$ as the size of a tree tends to infinity. More specifically, with $C_n(v)$ and $A_n(v)$ as they were in the proof of Theorem 4.2 (along with their respective limits), we are interested in:

$$\lim_{n\to\infty} \mathrm{E}_n(\mathcal{D}^{\underline{m}}) = \lim_{n\to\infty} \sum_{h\geq m} h^{\underline{m}}\,\mathrm{P}_n(\mathcal{D} = h) = \lim_{n\to\infty} A_n^{(m)}(1).$$

[1] Alternatively, singular cases such as this—which will appear throughout this chapter, and usually in reference to recursive trees—can be derived as limits of the more general cases.

We show firstly that the moments of $\mathrm{P}_n(\mathcal{D} = h)$ converge to those of its limiting distribution $\mathrm{P}(\mathcal{D}_\star = h)$, and then, instead of dealing with $A_n(v)$ directly, derive the moments' asymptotic behaviour using the limiting generating function $A(v)$.

Lemma 4.2. *The moments of the distribution of the centroid's depth $\mathcal{D}(\mathcal{T}_n)$ converge to those of $\mathcal{D}_\star$, i.e.:*

$$\lim_{n\to\infty} \mathrm{E}_n(\mathcal{D}^m) = \mathrm{E}(\mathcal{D}_\star^m).$$

Proof. This follows from Lemma 4.1 and Lebesgue's dominated convergence theorem, which states that if (f_n) is a sequence of real-valued functions, and g a function such that $|f_n| \le g$ for all n, then if $\int g < \infty$, one has $\lim_{n\to\infty} \int f_n = \int \lim_{n\to\infty} f_n$.

For our purposes, let $f_n(h) = h^{\underline{m}}\,\mathrm{P}_n(\mathcal{D} = h)$. The bound on $\mathrm{P}_n(\Lambda_k(1/2))$ given in equation (4.5) then leads to a similar one (also uniform over n) on $\mathrm{P}_n(\mathcal{D} = h)$, as follows (recall that $[v^h]C_n(v) = \mathrm{P}_n(\mathcal{D} \ge h)$):

$$\begin{aligned}
\mathrm{P}_n(\mathcal{D} = h) \le [v^h]C_n(v) &= [v^h] \sum_{k\ge 1} \mathrm{P}_n(\Lambda_k(1/2)) \frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}} \\
&\le [v^h] \sum_{k\ge 1} \binom{\alpha v + k - 2}{k-1} 2^{-(k-2)} \\
&= [v^h] 2^{1+\alpha v} = 2\frac{(\alpha \ln 2)^h}{h!}.
\end{aligned}$$

Since the range of the random variable $\mathcal{D}$ is countable, the integrals in Lebesgue's dominated convergence theorem become sums, and we are left to show that:

$$2\sum_{h\ge 0} h^{\underline{m}} \frac{(\alpha \ln 2)^h}{h!} = (\alpha \ln 2)^m\, 2^{\alpha+1} < \infty.$$

Thus the factorial moments of the distributions $\mathrm{P}_n(\mathcal{D} = h)$ converge to those of $\mathrm{P}(\mathcal{D}_\star = h)$, and since the usual higher-order moments $\mathrm{E}(\mathcal{D}_\star^m)$ are (finite) linear combinations of the factorial moments, convergence holds for them as well. □

With convergence established, all that remains is to compute the moments of $\mathcal{D}_\star$ by making use of its probability generating function $A(v) = \sum_{h\ge 0} \mathrm{P}(\mathcal{D}_\star = h) v^h$, since $\mathrm{E}(\mathcal{D}_\star^{\underline{m}}) = A^{(m)}(1)$.

Theorem 4.3. *The limit of the mth factorial moment of the centroid's depth $\mathcal{D}(\mathcal{T}_n)$ is given by:*

$$\mathrm{E}(\mathcal{D}_\star^{\underline{m}}) = m\alpha^m \left(2\sum_{j=0}^{m-2} \binom{m-1}{j} (-1)^j j! \,(\ln 2)^{m-1-j} + (-1)^{m-1}(m-1)! \right).$$

In particular, the limits of its mean and variance are:

$$\begin{aligned}
\mathrm{E}(\mathcal{D}_\star) &= \alpha, \\
\mathrm{V}(\mathcal{D}_\star) &= \alpha^2(4\ln 2 - 3) + \alpha.
\end{aligned}$$

Proof. The calculation can be simplified slightly by writing the derivative of the expression in equation (4.11) as:

$$\begin{aligned} A^{(m)}(v) &= \frac{d^m}{dv^m}\Big[\alpha(v-1)\cdot\big(1+\alpha(v-1)\big)^{-1}\cdot\Big(2^{\alpha(v-1)+1}-1\Big)\Big] \\ &= \frac{d^m}{dv^m}[a(v)\cdot b(v)\cdot c(v)] \\ &= \sum_{i+j+k=m}\binom{m}{i,j,k}a^{(i)}(v)\,b^{(j)}(v)\,c^{(k)}(v). \end{aligned}$$

Since we are interested in $v=1$, note that $a'(1)=\alpha$, but $a^{(i)}(1)=0$ when $i\neq 1$. Furthermore, $b^{(j)}(1)=(-\alpha)^j j!$ for $j\geq 0$, and $c^{(k)}(1)=2(\alpha\ln 2)^k$ for $k>0$, whereas $c(1)=1$. This leads to:

$$\begin{aligned} \mathrm{E}(\mathcal{D}_\star^{\underline{m}}) &= m\alpha\sum_{j+k=m-1}\binom{m-1}{j,k}b^{(j)}(1)\,c^{(k)}(1) \\ &= m\alpha^m\left(2\sum_{j=0}^{m-2}\binom{m-1}{j}(-1)^j j!\,(\ln 2)^{m-1-j}+(-1)^{m-1}(m-1)!\right). \end{aligned}$$

The mean is computed more simply as $C(1)-1=\alpha$, and the second factorial moment is $\mathrm{E}(\mathcal{D}_\star^{\underline{2}})=2\alpha^2(2\ln 2-1)$. □

Finally, we make two small remarks: that the limits of the mean and variance of the depth of the centroid in a random recursive tree (1 and $4\ln 2-2$ respectively) were previously given by Moon (2002); and that Theorem 4.3 implies that the mean and variance of the centroid's depth are greatest (in the limit) in the case of binary increasing trees ($\alpha=2$).

4.3 The label of the centroid

Our second task regarding the centroid of an increasing tree is to describe its label, which we will denote by $\mathcal{L}(T)$.

Since we will have no need for the general event $\Lambda_k(\sigma)$ throughout this section, let us adopt the shorthand $\Lambda_k=\Lambda_k(1/2)$ to denote the presence of label k on the path between the root and centroid nodes. A key observation is that the event $\mathcal{L}(T)=k$ can be expressed in terms of the presence—or lack thereof—of nodes $k,k+1,\ldots$ on this path, namely:

$$\begin{aligned} \mathrm{P}_n(\mathcal{L}=k) = \mathrm{P}_n(\Lambda_k) &- \mathrm{P}_n(\Lambda_k\cap\Lambda_{k+1}) \\ &- \mathrm{P}_n(\Lambda_k\cap\overline{\Lambda}_{k+1}\cap\Lambda_{k+2}) \\ &- \cdots. \end{aligned} \tag{4.12}$$

Here $\overline{\Lambda}_l$ is the complement of Λ_l, i.e., it is the event that node l is *not* on the path to the centroid. Equation (4.12) simply states that the centroid has label k if and only if k is on the path to the centroid, but none of the nodes $k+1,k+2,\ldots$ are.

One can write a similar expression for the probability that the centroid's label is at least k:

$$\begin{aligned} \mathrm{P}_n(\mathcal{L} \geq k) = \mathrm{P}_n(\Lambda_k) &+ \mathrm{P}_n(\overline{\Lambda}_k \cap \Lambda_{k+1}) \\ &+ \mathrm{P}_n(\overline{\Lambda}_k \cap \overline{\Lambda}_{k+1} \cap \Lambda_{k+2}) \\ &+ \cdots . \end{aligned} \tag{4.13}$$

The composite event $\Lambda_k \cap \overline{\Lambda}_{k+1} \cap \cdots \cap \overline{\Lambda}_{k+j-1} \cap \Lambda_{k+j}$ in the first of the two equations requires that k and $k+j$ appear on the path to the centroid, but none of $k+1, \ldots, k+j-1$ do; and this occurs if and only if $k+j$ is on the path and has node k as its parent. (This is a simpler condition than the one required by equation (4.13), which would be that node $k+j$ is on the path and its parent is any one of the nodes $1, \ldots, k-1$.)

Let $A_l(T)$ be the random variable that yields node l's parent, which if we consider the increasing tree's probabilistic growth process, is the node l was 'attached' to. Then we are interested in $\mathrm{P}_n(\Lambda_{k+j} \cap (A_{k+j} = k))$, for fixed k and j, as $n \to \infty$. Because the size of the subtree consisting of node $k+j$ and its descendants is independent of the node to which $k+j$ was attached, we have:

$$\mathrm{P}_n(\Lambda_{k+j} \cap (A_{k+j} = k)) = \mathrm{P}_n(\Lambda_{k+j})\,\mathrm{P}(A_{k+j} = k), \tag{4.14}$$

where the second probability in the product is independent of the size n of the tree. Since, by Corollary 4.1, we already know the asymptotic behaviour of $\mathrm{P}_n(\Lambda_{k+j})$, we are left to derive an expression for the probability that node $k+j$ attaches to node k. Although similar expressions have been derived before (see Dobrow and Smythe (1996, Theorem 5) or Kuba and Wagner (2010)), we do so here by making straightforward use of generating functions.

4.3.1 The probability of a specific attachment

To compute $\mathrm{P}(A_{k+j} = k)$, we simply count the (weighted) trees of size $k+j$ in which k is the parent of $k+j$. From a symbolic perspective, this requires counting trees of size $k+j$ in which one of k's descendent branches has been replaced with a branch of size 1, and the label $k+j$ has been constrained to appear in this singleton branch. Both replacements and label constraints (on the smallest or largest labels in the structure) can be encoded using generating functions (see, e.g., Flajolet and Sedgewick, 2009, Sections II.6.1 and II.6.3).

Lemma 4.3 (Dobrow and Smythe (1996))**.** *For $k, j \in \mathbb{Z}_{>0}$, the probability that the parent of node $k+j$ has label k is given by:*

$$\mathrm{P}(A_{k+j} = k) = \frac{\alpha k^{\overline{j-1}}}{(\alpha + k - 1)^{\overline{j}}},$$

which does not depend on the size of the tree.

Proof. Consider again the exponential generating function of a tree whose first k nodes do not contribute to its size:

$$B(x) = y^{(k)}(x) = y_k \cdot y'(x) \cdot (1 - c_1 x)^{-(k-1)},$$

in which $y'(x)$ represents the subtree attached to (but excluding) node k. Replacing one of this subtree's branches with a singleton branch, and restricting the label $k+j$ to appear in this new branch, yields the generating function:

$$\int_0^x \frac{d}{dt} t \cdot \phi'(y(t))\, dt = \int_0^x \phi'(y(t))\, dt,$$

which represents the desired structure of the subtree attached to node k. Thus the generating function of a tree whose first k nodes have been 'deleted', but whose largest node is attached to k, is:

$$\begin{aligned} A(x) &= y_k \int_0^x \frac{d}{du}\left[\int_0^u \phi'(y(t))\, dt\right](1-c_1u)^{-(k-1)}\, du \\ &= y_k \int_0^x \phi'(y(u))(1-c_1u)^{-(k-1)}\, du \\ &= y_k c_1 \alpha \int_0^x (1-c_1u)^{-k}\, du \\ &= y_k \frac{\alpha}{k-1}\Big((1-c_1x)^{-(k-1)} - 1\Big), \end{aligned}$$

where

$$\frac{d}{dx} y'(x) = \frac{d}{dx}\phi(y(x)) = \phi'(y(x))y'(x)$$

has implied that

$$\phi'(y(x)) = \frac{y''(x)}{y'(x)} = c_1\alpha(1-c_1x)^{-1}.$$

Since the exponential generating functions $A(x)$ and $B(x)$ represent constrained and unconstrained trees respectively, the ratio of their jth coefficients is the probability that k is the parent of $k+j$:

$$\begin{aligned} \mathrm{P}(A_{k+j} = k) &= \frac{j!\,[x^j]A(x)}{j!\,[x^j]B(x)} \\ &= \frac{\alpha}{k-1}\frac{[x^j](1-c_1x)^{-(k-1)}}{[x^j](1-c_1x)^{-(\alpha+k-1)}} \\ &= \frac{\alpha}{k-1}\binom{k+j-2}{j}\Big/\binom{\alpha+k+j-2}{j} \\ &= \frac{\alpha}{k-1}\prod_{l=0}^{j-1}\frac{k+l-1}{\alpha+k+l-1} = \frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}}. \end{aligned}$$

□

Immediately, we see that in the case of recursive trees ($\alpha = 1$), the probability of a particular attachment is $\mathrm{P}(A_{k+j} = k) = 1/(k+j-1)$, which agrees with the family's known growth process.

4.3.2 A limiting distribution for the label of the centroid

Following on from equations (4.12) and (4.14), we can write the probability of the centroid assuming a certain label k in terms of the events Λ_k and $A_{k+j} = k$ (for

which we have closed-form expressions):

$$\mathrm{P}_n(\mathcal{L}=k)=\mathrm{P}_n(\Lambda_k)-\sum_{j\geq 1}\mathrm{P}_n(\Lambda_{k+j})\,\mathrm{P}_n(A_{k+j}=k).$$

As a consequence, we arrive at the desired result of this section:

Theorem 4.4. *The label $\mathcal{L}(\mathcal{T}_n)$ of the centroid node in a random tree of size n converges in probability to a discrete random variable $\mathcal{L}_\star$ supported by $\mathbb{Z}_{\geq 0}$ and with mass function:*

$$\mathrm{P}(\mathcal{L}_\star=k)=\begin{cases}1-\frac{\alpha}{\alpha-1}\left(1-2^{-(\alpha-1)}\right) & \text{if } k=1,\\ I_{1/2}(k-1,\alpha)-\frac{\alpha}{\alpha-1}I_{1/2}(k,\alpha-1) & \text{otherwise.}\end{cases}$$

An alternative form, which holds for $k\geq 1$, is given by:

$$\mathrm{P}(\mathcal{L}_\star=k)=-\frac{1}{\alpha-1}I_{1/2}(k,\alpha)+\left(1+\frac{\alpha}{\alpha-1}\right)\binom{\alpha+k-2}{k-1}2^{-(\alpha+k-1)}. \tag{4.15}$$

Proof. Recalling the asymptotic expression for $\mathrm{P}_n(\Lambda_k)$ (Corollary 4.1), assume now that n is large, and fix J so that $k+J=\lfloor n^{1/4-\varepsilon}\rfloor$, for some arbitrarily small, positive ε. Then:

$$\begin{aligned}\mathrm{P}_n(\mathcal{L}=k)&=\mathrm{P}_n(\Lambda_k)-\sum_{j=1}^{\lceil n/2\rceil-k}\mathrm{P}_n(\Lambda_{k+j})\,\mathrm{P}_n(A_{k+j}=k)\\&=\mathrm{P}_n(\Lambda_k)-\sum_{j=1}^{J}I_{1/2}(k+j-1,\alpha)\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}}\left(1+O\left(\frac{(k+j)^2}{\sqrt{n}}\right)\right)\\&\quad-\sum_{j=J+1}^{\lceil n/2\rceil-k}\mathrm{P}_n(\Lambda_{k+j})\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}},\end{aligned}$$

As in the proof of Theorem 4.2—which dealt with the depth of the centroid—the upper bound for $\mathrm{P}_n(\Lambda_k)$ given in equation (4.5) implies that the sum over larger labels vanishes as n grows:

$$\begin{aligned}\sum_{j=J+1}^{\lceil n/2\rceil-k}\mathrm{P}_n(\Lambda_{k+j})\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}}&\leq\sum_{j=J+1}^{\lceil n/2\rceil-k}\frac{\alpha^{\overline{k+j-1}}}{(k+j-1)!}\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}}2^{-(k+j-2)}\\&\leq\frac{\alpha^{\overline{k-1}}}{(k-1)!}\sum_{j=J+1}^{\lceil n/2\rceil-k}\frac{\alpha}{k+j-1}2^{-(k+j-2)}\\&\leq\frac{\alpha^{\overline{k-1}}}{(k-1)!}\frac{\alpha n}{\lfloor n^{1/4-\varepsilon}\rfloor}2^{-\lfloor n^{1/4-\varepsilon}\rfloor}\xrightarrow[n\to\infty]{}0.\end{aligned}$$

Also, the extension of the first sum to an infinite one is permissible, since:

$$\begin{aligned}\sum_{j>J} I_{1/2}(k+j-1,\alpha)\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}} &= \sum_{j>J}\frac{\alpha\Gamma(\alpha+k-1)}{\Gamma(\alpha)\Gamma(k)}\int_0^{1/2} t^{k+j-2}(1-t)^{\alpha-1}\,dt\\ &\le \alpha\binom{\alpha+k-2}{k-1}\sum_{j>J}2^{-(k+j-1)}\\ &= O\Big(2^{-(k+J-1)}\Big)\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Letting $n\to\infty$, and assuming $k>1$, we obtain the limiting probability $\mathrm{P}(\mathcal{L}_\star=k)$:

$$\begin{aligned}\mathrm{P}(\mathcal{L}_\star=k) &= \lim_{n\to\infty}\mathrm{P}_n(\mathcal{L}=k)\\ &= I_{1/2}(k-1,\alpha)-\sum_{j\ge1} I_{1/2}(k+j-1,\alpha)\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}}\\ &= I_{1/2}(k-1,\alpha)-\alpha\frac{\Gamma(\alpha+k-1)}{\Gamma(k)\Gamma(\alpha)}\int_0^{1/2}(1-t)^{\alpha-1}\sum_{j\ge1}t^{k+j-2}\,dt \qquad (4.16)\\ &= I_{1/2}(k-1,\alpha)-\frac{\alpha}{\alpha-1}\,\mathrm{B}(k,\alpha-1)^{-1}\int_0^{1/2}t^{k-1}(1-t)^{\alpha-2}\,dt\\ &= I_{1/2}(k-1,\alpha)-\frac{\alpha}{\alpha-1}I_{1/2}(k,\alpha-1).\end{aligned}$$

When $k=1$, the first incomplete beta function is replaced by 1. The consolidated form given in equation (4.15) is due to the following property of the incomplete beta function:

$$\begin{aligned}I_x(a-1,b)-\frac{x^{a-1}(1-x)^b}{(a-1)\,\mathrm{B}(a-1,b)} &= I_x(a,b)\\ &= I_x(a,b-1)+\frac{x^a(1-x)^{b-1}}{(b-1)\,\mathrm{B}(a,b-1)},\end{aligned}$$ □

Repeating (4.16) for $\alpha=1$ yields:

Corollary 4.4 (Moon (2002))**.** *For* recursive trees*:*

$$\lim_{n\to\infty}\mathrm{P}_n(\mathcal{L}=k)=2^{-(k-1)}-\sum_{j\ge k}\frac{2^{-j}}{j}.$$

In particular (for recursive trees), $\lim_n \mathrm{P}_n(\mathcal{L}=1)=1-\ln 2$.

4.3.3 Moments of the label distribution

Just as we did when dealing with the depth of the centroid, we can apply Lebesgue's dominated convergence theorem to prove convergence of the moments of $\mathrm{P}_n(\mathcal{L}=k)$ to those of $\mathrm{P}(\mathcal{L}_\star=k)$, and then derive their limits using the more convenient form of the limiting distribution. In the present case of the centroid's label, however, the proof of convergence is almost immediate.

Lemma 4.4. *The moments of the distribution of the centroid's label $\mathcal{L}(\mathcal{T}_n)$ converge to those of $\mathcal{L}_\star$, i.e.:*

$$\lim_{n\to\infty} \mathrm{E}_n(\mathcal{L}^m) = \mathrm{E}(\mathcal{L}_\star^m).$$

Proof. The line of argument is the same as that which was used for Lemma 4.2: we must find a uniform bound $g(k)$ for $k^{\underline{m}}\,\mathrm{P}_n(\mathcal{L}=k)$ such that $\sum_k k^{\underline{m}}\,g(k)$ converges. Once again by equation (4.5):

$$\begin{aligned}
\sum_{k\geq 1} k^{\underline{m}}\,\mathrm{P}_n(\mathcal{L}=k) &\leq \sum_{k\geq m} k^{\underline{m}}\,\mathrm{P}_n(\Lambda_k)\\
&\leq \sum_{k\geq m} \frac{k^{\underline{m}}\,\alpha^{\overline{k-1}}}{(k-1)!}\,2^{-(k-2)}\\
&= 2^{-(m-2)} \sum_{k\geq m-1} \binom{\alpha+k-1}{k} (k+1)^{\underline{m}}\,2^{-(k-m+1)}\\
&= 2^{-(m-2)}\,\frac{d^m}{du^m}\left[u(1-u)^{-\alpha}\right]_{u=1/2} < \infty. \qquad \square
\end{aligned}$$

Theorem 4.5. *The limit of the mth factorial moment of the centroid's label $\mathcal{L}(\mathcal{T}_n)$ is given by:*

$$\mathrm{E}(\mathcal{L}_\star^{\underline{m}}) = \frac{4m^2+2\alpha m+\alpha-2}{m+1}\,\alpha^{\overline{m-1}}.$$

In particular, the limits of its mean and variance are:

$$\begin{aligned}
\mathrm{E}(\mathcal{L}_\star) &= 1+\frac{3}{2}\alpha,\\
\mathrm{V}(\mathcal{L}_\star) &= -\frac{7}{12}\alpha^2+\frac{19}{6}\alpha.
\end{aligned}$$

Proof. The factorial moments of the limiting distribution can be computed directly using equation (4.15):

$$\begin{aligned}
\sum_{k\geq 1} k^{\underline{m}}\,\mathrm{P}(\mathcal{L}_\star=k) &= -\frac{1}{\alpha-1}\sum_{k\geq 1} k^{\underline{m}}\,I_{1/2}(k,\alpha)\\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right)\sum_{k\geq 1}\binom{\alpha+k-2}{k-1} k^{\underline{m}}\,2^{-(\alpha+k-1)}\\
&= -\frac{1}{\alpha-1}\int_0^{1/2}(1-t)^{\alpha-1}\sum_{k\geq 1}\binom{\alpha+k-1}{k-1}\alpha k^{\underline{m}}\,t^{k-1}\,dt\\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right)\sum_{k\geq 1}\binom{\alpha+k-2}{k-1} k^{\underline{m}}\,2^{-(\alpha+k-1)}\\
&= -\frac{\alpha}{\alpha-1}\int_0^{1/2}(1-t)^{\alpha-1}t^{m-1}\,\frac{d^m}{dt^m}\left[t(1-t)^{-(\alpha+1)}\right]dt\\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right)2^{-(\alpha+m-1)}\,\frac{d^m}{dt^m}\left[t(1-t)^{-\alpha}\right]_{t=1/2}
\end{aligned}$$

$$= -\frac{\alpha}{\alpha-1}\int_0^{1/2}\sum_{i=0}^{1}\binom{m}{i}(\alpha+1)^{\overline{m-i}}\,t^{m-i}(1-t)^{-(m+2-i)}\,dt + \left(1+\frac{\alpha}{\alpha-1}\right)\left(\alpha^{\overline{m}}+m\alpha^{\overline{m-1}}\right).$$

Noting that the integrals within the sum are all of a common, solvable, form:

$$\int_0^{1/2} t^m(1-t)^{-(m+2)}\,dt = \frac{1}{m+1}\left[\left(\frac{t}{1-t}\right)^{m+1}\right]_{t=0}^{1/2} = \frac{1}{m+1},$$

the mth factorial moment reduces to:

$$\begin{aligned}\sum_{k\geq 1} k^{\underline{m}}\,\mathrm{P}(\mathcal{L}_\star = k) &= -\frac{\alpha}{\alpha-1}\left(\frac{(\alpha+1)^{\overline{m}}}{m+1}+\frac{m(\alpha+1)^{\overline{m-1}}}{m}\right)\\ &\quad+\left(1+\frac{\alpha}{\alpha-1}\right)\left(\alpha^{\overline{m}}+m\alpha^{\overline{m-1}}\right)\\ &= \left(1+\frac{\alpha}{\alpha-1}\right)m\alpha^{\overline{m-1}}+2\alpha^{\overline{m}}-\frac{1}{\alpha-1}\frac{\alpha^{\overline{m+1}}}{m+1}.\end{aligned}$$

□

Once again, the fact that the expected label of the centroid in a random recursive tree tends to 5/2 was first proved by Moon (2002). And lastly, but perhaps not unexpectedly, it follows from Theorem 4.5 that binary increasing trees ($\alpha = 2$) lead to the greatest eventual mean and variance.

4.4 The size of the centroid's root branch

Our third and final set of results involving the centroid of an increasing tree can be contrasted with the case of simply generated trees in a way that the results of the previous two sections could not: whereas the depth and label (where applicable) of the centroid in a simply generated tree are relatively uninformative (because roots or specifically labelled nodes in simply generated trees are, for the most part, no different from randomly selected nodes), the number and sizes of its centroid branches are strikingly well-determined. We have already mentioned that almost all simply generated trees of size n have three large centroid branches that together contain most of the tree's nodes (see Section 1.2.3), and in fact Meir and Moon (2002) have proved, among other things, that the size of the centroid's ancestral branch, divided by n, tends to $\sqrt{2}-1 \approx 0.414$ as $n \to \infty$ (independently of the specific family of simply generated trees). Our main goal in this section is an analogue of this result, but for increasing trees; however we will phrase it (analogously) in terms of the size of the subtree rooted at the centroid.

Note that in a way, the ancestral branch is the most interesting of the centroid's branches, because its other, descendent branches behave mostly (in particular, their number and sizes do) like those of the root branches of a random increasing tree—albeit under the extra condition that no one branch contains more than $\lfloor n/2 \rfloor$ nodes.

Let $\mathcal{S}(T)$ denote the size of the subtree containing the centroid and all of its descendent branches, and $\mathrm{P}_n(\mathcal{S} = m = \lfloor \theta n \rfloor)$ the relevant probability distribution. Since the ancestral branch contains at most $\lfloor n/2 \rfloor$ nodes, the ranges of m and θ

are $\{\lceil n/2\rceil, \ldots, n\}$ and $[1/2, 1]$ respectively, with $m = 1n$ characterising the case in which the root and centroid coincide (for which Theorem 4.4 already provides a limiting probability).

4.4.1 A preliminary equation

The event that the centroid's subtree is made up of m nodes (where $m \geq n/2$) can be decomposed into a pair of simpler events: firstly, that the tree contains a subtree of size m (there can be at most one); and secondly, given the presence of such a subtree, that its root is the centroid. This second event can be stated more explicitly as the case, in a tree of size m, that the root is the only node with at least $\lfloor n/2\rfloor$ descendants.

It is here that we will make use of $\mathrm{P}_n(\Lambda_k(\sigma))$, which was introduced in Section 4.2.1 as a generalisation of the 'path' probability $\mathrm{P}_n(\Lambda_k(1/2))$. Letting $X_m(T)$ mark the existence of a subtree of size m in a tree T, the probability that the centroid's subtree contains exactly m nodes can be expressed as:

$$\begin{aligned}\mathrm{P}_n(\mathcal{S} = m) &= \mathrm{P}_n(X_m)\Big(1 - \mathrm{P}_m\Big(\textstyle\bigcup_{j\geq 2} \Lambda_j(\frac{n}{2m})\Big)\Big) \\ &= \mathrm{P}_n(X_m)\left(1 - \sum_{j=2}^{m} \mathrm{P}_m(A_j = 1 \cap \Lambda_j(\tfrac{n}{2m}))\right) \\ &= \mathrm{P}_n(X_m) \cdot (1 - A_m(\tfrac{n}{2m})), \end{aligned} \tag{4.17}$$

where A_j gives the label of node j's parent, so that $A_j = 1$ characterises the root's children. The probabilities that appear within the sum refer to disjoint events, since at most one of the subtree's root branches can contain $\lfloor n/2\rfloor + 1$ nodes. We note—as we did for equation (4.14)—that the size of the subtree rooted at j is independent of the node j was attached to, so that:

$$\mathrm{P}_m(A_j = 1 \cap \Lambda_j(\tfrac{n}{2m})) = \mathrm{P}(A_j = 1)\,\mathrm{P}_m(\Lambda_j(\tfrac{n}{2m})).$$

Consider the first probability $\mathrm{P}_n(X_m)$ in equation (4.17), of the event that a tree of n nodes contains a subtree of size m. We assume here that $n/2 \leq m < n$. Since there can be at most one, this probability can be rephrased as the expected number of such subtrees—a problem that we dealt with briefly in Section 3.4. Specifically, Lemma 3.2 tells us that in a tree of size n, the expected number of subtrees of size m is:

$$\sum_{k=2}^{n-m+1} \mathrm{P}_n(S_k = m) = \frac{\alpha(\alpha + n - 1)}{(\alpha + m)(\alpha + m - 1)}, \tag{4.18}$$

where, as before, S_k denotes the size of the subtree rooted at k. As long as $m \geq n/2$, we have $\mathrm{P}_n(X_m) = \sum_k \mathrm{P}_n(S_k = m)$.

4.4.2 The probability that the root of a subtree is the centroid

The second probability in equation (4.17), denoted by $1 - A_m(n/(2m))$, accounts for the cases in which the root of a subtree of size m is the centroid of the entire

tree, where $m/n = \theta$ for some fixed θ. We have:

$$A_m(\tfrac{n}{2m}) = \sum_{j=2}^{m} \mathrm{P}(A_j = 1)\, \mathrm{P}_m(\Lambda_j(\tfrac{n}{2m})), \tag{4.19}$$

in which both of the terms contained within the sum are manageable—the first by Lemma 4.3:

$$\mathrm{P}(A_j = 1) = \frac{\alpha(j-2)!}{\alpha^{\overline{j-1}}} = \alpha\, \mathrm{B}(j-1, \alpha),$$

and the second due to Theorem 4.1 and Lemma 4.1, which provide an asymptotic form and an upper bound respectively.

Lemma 4.5. *For $n/2 \le m < n$ in a tree of size n, the probability that the root of a subtree of size m is* not *the centroid of the tree satisfies, for all $0 < \varepsilon < 1/2$:*

$$A_m(\tfrac{n}{2m}) = \frac{\alpha}{\alpha - 1}\Big(1 - (\tfrac{n}{2m})^{\alpha-1}\Big) + O(n^{-\varepsilon}).$$

Proof. The sum given in equation (4.19) can be split at a value small enough for Theorem 4.1 to be applied, say $J = \lfloor m^{1/4-\varepsilon/2} \rfloor$ so that $J^2/\sqrt{m} = O(m^{-\varepsilon})$:

$$\begin{aligned} A_m(\tfrac{n}{2m}) = &\sum_{j=2}^{J} \mathrm{P}(A_j = 1) I_{1-\frac{n}{2m}}(j-1, \alpha)\left(1 + O\left(\frac{j^2}{\sqrt{m}}\right)\right) \\ &+ \sum_{j=J+1}^{m} \mathrm{P}(A_j = 1)\, \mathrm{P}_m(\Lambda_j(\tfrac{n}{2m})). \end{aligned}$$

Applying the bound of Lemma 4.1 affirms that the second sum is small for large values of J:

$$\begin{aligned} \sum_{j=J+1}^{m} \mathrm{P}(A_j = 1)\, \mathrm{P}_m(\Lambda_j(\tfrac{n}{2m})) &\le 6\alpha \frac{m}{n} \sum_{j=J+1}^{m} \mathrm{B}(j-1, \alpha) \frac{\alpha^{\overline{j-1}}}{(j-1)!} \left(1 - \frac{n}{2m}\right)^{j-1} \\ &\le 6\alpha \frac{m}{n} \sum_{j=J+1}^{m} \left(1 - \frac{n}{2m}\right)^{j-1} \\ &< 12\alpha \left(\frac{m}{n}\right)^2 \left(1 - \frac{n}{2m}\right)^{J} \xrightarrow[m\to\infty]{} 0, \end{aligned}$$

Similarly, extending the first sum to an infinite one has an effect that vanishes as m and n grow:

$$\begin{aligned} \sum_{j>J} \mathrm{P}(A_j = 1) I_{1-\frac{n}{2m}}(j-1, \alpha) &= \alpha \sum_{j>J} \int_0^{1-\frac{n}{2m}} t^{j-2}(1-t)^{\alpha-1}\, dt \\ &\le \alpha \sum_{j>J} \left(1 - \frac{n}{2m}\right)^{j-1} \xrightarrow[m\to\infty]{} 0. \end{aligned}$$

Combined, these two substitutions give the asymptotic behaviour of $A_m(n/(2m))$:

$$\begin{aligned} A_m(\tfrac{n}{2m}) &= \alpha \sum_{j=2}^{J} \int_0^{1-\frac{n}{2m}} t^{j-2}(1-t)^{\alpha-1}\,dt\left(1+O\left(\frac{J^2}{\sqrt{m}}\right)\right) + O\left(\left(1-\frac{n}{2m}\right)^J\right) \\ &= \alpha \int_0^{1-\frac{n}{2m}} (1-t)^{\alpha-2}\,dt(1+O(m^{-\varepsilon})) \\ &= \frac{\alpha}{\alpha-1}\left(1-\left(\frac{n}{2m}\right)^{\alpha-1}\right) + O(n^{-\varepsilon}), \end{aligned}$$

which should be compared to the case $m = n$ as given in Theorem 4.4. □

When dealing with recursive trees, the final step is slightly different, resulting in:

Corollary 4.5. *For* recursive trees*:*

$$A_m(\tfrac{n}{2m}) \sim \ln\left(\tfrac{2m}{n}\right).$$

4.4.3 The distribution of the size of the centroid's subtree

Now that $\mathrm{P}_n(X_m)$ is known explicitly (equation (4.18)), and $A_m(n/(2m))$ asymptotically, we are ready to derive an expression for the distribution of $\mathcal{S}(T)$. In the light of equation (4.17), which states that:

$$\mathrm{P}_n(\mathcal{S} = m) = \mathrm{P}_n(X_m) \cdot (1 - A_m(\tfrac{n}{2m})),$$

we have the main theorem of this section:

Lemma 4.6. *For $n/2 \le m < n$ and any $0 < \varepsilon < 1/2$, the probability that the centroid has $m-1$ descendent nodes is given by:*

$$\mathrm{P}_n(\mathcal{S} = m) = \frac{4}{n}\frac{\alpha}{\alpha-1}\left(\alpha\left(\frac{n}{2m}\right)^{\alpha+1} - \left(\frac{n}{2m}\right)^2\right) + O\left(n^{-1-\varepsilon}\right).$$

Proof. The result is simply an application of equation (4.18) and Lemma 4.5 to the above expression:

$$\begin{aligned} \mathrm{P}_n(\mathcal{S} = m) &= \left(1 - \frac{\alpha}{\alpha-1}\left(1-\left(\frac{n}{2m}\right)^{\alpha-1}\right) + O(n^{-\varepsilon})\right)\frac{\alpha(\alpha+n-1)}{(\alpha+m)(\alpha+m-1)} \\ &= \left(1 - \frac{\alpha}{\alpha-1}\left(1-\left(\frac{n}{2m}\right)^{\alpha-1}\right)\right)\frac{\alpha}{n}\left(\frac{n}{m}\right)^2 + O\left(n^{-1-\varepsilon}\right) \\ &= \frac{4}{n}\frac{\alpha}{\alpha-1}\left(\alpha\left(\frac{n}{2m}\right)^{\alpha+1} - \left(\frac{n}{2m}\right)^2\right) + O\left(n^{-1-\varepsilon}\right). \end{aligned}$$

□

Combined with the special case $m = n$, which is covered by Theorem 4.4, this asymptotically describes the size of the centroid's subtree, and thus the size $n - m$ of its root branch as well.

As with the distributions of the centroid's depth and label, we can also show convergence to a limiting distribution—however in this case, although the finite probability distributions are discrete, the limiting distribution is a mixture of a continuous distribution with support $[1/2, 1)$ and a point measure at 1. The notion of convergence is also slightly different, in that it is weaker than those of the previous two sections.

Theorem 4.6. *The proportion $\mathcal{S}(\mathcal{T}_n)/n$ of nodes accounted for by the subtree consisting of the centroid and all of its descendants in a random tree of size n converges in distribution to the random variable $\mathcal{S}_\star$, defined on $[1/2, 1)$ by the density:*

$$f(\theta) = 4\frac{\alpha}{\alpha-1}\Big(\alpha(2\theta)^{-(\alpha+1)} - (2\theta)^{-2}\Big),$$

and at the boundary $\theta = 1$ by the point measure:

$$\mathrm{P}(\mathcal{S}_\star = 1) = 1 - \frac{\alpha}{\alpha-1}\Big(1 - 2^{-(\alpha-1)}\Big).$$

Proof. Consider the cumulative distribution function arising from Lemma 4.6:

$$\begin{aligned}\mathrm{P}_n(\mathcal{S} \le \sigma n) &= \frac{4}{n}\frac{\alpha}{\alpha-1}\sum_{m=\lceil n/2\rceil}^{\lfloor \sigma n\rfloor}\left(\alpha\left(\frac{n}{2m}\right)^{\alpha+1} - \left(\frac{n}{2m}\right)^{2}\right) + O(n^{-\varepsilon}) \\ &= 4\frac{\alpha}{\alpha-1}\int_{1/2}^{\sigma}\Big(\alpha 2\theta)^{-(\alpha+1)} - (2\theta)^{-2}\Big)\,d\theta + O(n^{-\varepsilon}).\end{aligned}$$

Note that the error term—which traces back to Theorem 4.1 via Lemma 4.5—is uniform in σ over subsets of the form $[1/2, 1-\delta)$. And since each point of continuity (there is discontinuity at 1) is contained in such a subset, this makes explicit (for $\sigma < 1$) the convergence of $\mathrm{P}_n(\mathcal{S} \le \sigma n)$ to the continuous distribution with the stated density. The point measure simply corresponds to $\mathrm{P}(\mathcal{L}_\star = 1)$. □

Once again, the result for recursive trees differs slightly:

Corollary 4.6. *For* recursive trees*:*

$$f(\theta) = \frac{1 - \ln(2\theta)}{\theta^2} \quad \textit{and} \quad \mathrm{P}(\mathcal{S}_\star = 1) = 1 - \ln 2.$$

4.4.4 Moments of the subtree's size distribution

Finally, we can detail the limiting behaviour of the moments of $\mathcal{S}(T)$, and in particular, the expected size of the centroid's subtree. This is simply mechanical, since our random variables have bounded support, so that convergence in distribution implies convergence of moments.

Theorem 4.7. *The moments of the distribution of the proportion $\mathcal{S}(\mathcal{T}_n)/n$ of the tree accounted for by the centroid and its descendants converge to those of $\mathcal{S}_\star$, i.e.:*

$$\lim_{n\to\infty} \mathrm{E}_n((\mathcal{S}/n)^r) = \mathrm{E}(\mathcal{S}_\star^r).$$

The limit of the rth moment satisfies:

$$\mathrm{E}(\mathcal{S}_\star^r) = \mathrm{P}(\mathcal{S}_\star = 1) + \frac{\alpha}{\alpha-1}\left(\frac{\alpha}{\alpha-r}\left(2^{-(r-1)} - 2^{-(\alpha-1)}\right) - \frac{1}{r-1}\left(1 - 2^{-(r-1)}\right)\right).$$

In particular, the limit of its mean is:

$$\mathrm{E}(\mathcal{S}_\star) = 1 + \frac{\alpha}{(\alpha-1)^2}\left(1 - 2^{-(\alpha-1)} - (\alpha-1)\ln 2\right).$$

Proof. For $\alpha \notin \{1, r\}$ and $r \in \mathbb{Z}_{>0}$, and with $f(\theta)$ as in Theorem 4.6, we have:

$$\begin{aligned}
\mathrm{E}(\mathcal{S}_\star^r) &= \mathrm{P}(\mathcal{S}_\star = 1) + \int_{1/2}^{1} \theta^r f(\theta)\, d\theta \\
&= \mathrm{P}(\mathcal{S}_\star = 1) + 2^{-(r-1)}\frac{\alpha}{\alpha-1}\int_1^2 \mu^r\left(\alpha\mu^{-(\alpha+1)} - \mu^{-2}\right) d\mu \\
&= \mathrm{P}(\mathcal{S}_\star = 1) + 2^{-(r-1)}\frac{\alpha}{\alpha-1}\left[-\frac{\alpha}{\alpha-r}\mu^{r-\alpha} - \frac{1}{r-1}\mu^{r-1}\right]_1^2 \\
&= \mathrm{P}(\mathcal{S}_\star = 1) + \frac{\alpha}{\alpha-1}\left(\frac{\alpha}{\alpha-r}\left(2^{-(r-1)} - 2^{-(\alpha-1)}\right) - \frac{1}{r-1}\left(1 - 2^{-(r-1)}\right)\right). \square
\end{aligned}$$

For plane-oriented and binary increasing trees ($\alpha = 1/2$ and $\alpha = 2$), this yields means of $3 - 2\sqrt{2} + \ln 2 \approx 0.86$ and $2 - 2\ln 2 \approx 0.61$ respectively. It is worth noting that the proof's requirement that $\alpha \neq r$ is weak for two reasons: because the standard definition of very simple increasing trees deals with the range $0 < \alpha \leq 2$, and because singular cases such as these can be seen as limits of the above function, or, in the case of recursive trees, be derived directly from Corollaries 4.4 and 4.6. For instance:

$$\begin{aligned}
\mathrm{E}(\mathcal{S}_\star)|_{\alpha=1} &= 1 - \frac{1}{2}(\ln 2)^2 \approx 0.76, \\
\mathrm{E}(\mathcal{S}_\star^r)|_{\alpha=1} &= 1 + \frac{r}{(r-1)^2}\left(1 - 2^{-(r-1)}\right) - \frac{r}{r-1}\ln 2, \\
\mathrm{E}(\mathcal{S}_\star^2)|_{\alpha=2} &= 2\ln 2 - 1.
\end{aligned}$$

The limit of the mean in the case of recursive trees was given by Moon (2002). The asymptotic variances for plane-oriented, recursive, and binary increasing trees are approximately 0.03, 0.04, and 0.01 respectively.

4.5 Concluding remarks

Altogether, the results of this chapter paint a reasonably consistent picture of the behaviour of the (nearest) centroid in a large, random very simple increasing tree: one expects the centroid to lie, on average, within two edges of the root (for the usual case $0 < \alpha \leq 2$), and its root branch to account for a significant portion of the entire tree.

That being said, there are still a number of related questions that could be raised, either in connection with this chapter or the thesis as a whole. For example, we might investigate other parameters of the centroid—in the simplest case, its

degree—or ask to what extent the above behaviour generalises to the entire class of increasing trees, which do not necessarily satisfy the properties of Lemma 3.1.

Furthermore, one could attempt to characterise the distribution of closeness centrality in a random tree in a way similar to that in which we have handled betweenness centrality in Chapters 2 and 3; or, as an alternative definition of a tree's most 'central' node, consider its centres (i.e., nodes with minimal eccentricity). Finally, there are classes of random trees other than simply generated and increasing trees that we have not mentioned at all here—the most notable being the various families of search trees (Drmota, 2009, Section 1.4).

We offer no real guidance on questions such as these here, other than to say that given the relatively straightforward treatment of the more well-known centrality measures and random tree models of the previous chapters, one might expect (or at least hope) that several of these problems will also be amenable to the tools and methods of analytic combinatorics.

List of References

Aldous, D. (1991*a*). The continuum random tree. I. *Ann. Probab.*, vol. 19, no. 1, pp. 1–28.

Aldous, D. (1991*b*). The continuum random tree. II. An overview. In: Barlow, M.T. and Bingham, N.H. (eds.), *Stochastic Analysis*, pp. 23–70. Cambridge University Press.

Aldous, D. (1994*a*). Recursive self-similarity for random trees, random triangulations and Brownian excursion. *Ann. Probab.*, vol. 22, no. 2, pp. 527–545.

Aldous, D. (1994*b*). Triangulating the circle, at random. *Am. Math. Mon.*, vol. 101, no. 3, pp. 223–233.

Bergeron, F., Flajolet, P. and Salvy, B. (1992). Varieties of increasing trees. In: Raoult, J.-C. (ed.), *Lecture Notes in Computer Science*, vol. 581, pp. 24–48. Springer.

Dobrow, R.P. and Smythe, R.T. (1996). Poisson approximations for functionals of random trees. *Random Struct. Algorithms*, vol. 9, no. 1–2, pp. 79–92.

Drmota, M. (2009). *Random Trees: An Interplay Between Combinatorics and Probability*. 1st edn. Springer, Vienna. ISBN 9783211753552.

Drmota, M., Fusy, É., Kang, M., Kraus, V. and Rué, J. (2011). Asymptotic study of subcritical graph classes. *SIAM J. Discrete Math.*, vol. 25, no. 4, pp. 1615–1651.

Durant, K. and Wagner, S. (2016). Betweenness centrality in random trees. In: Fill, J.A. and Ward, M.D. (eds.), *2016 Proceedings of the Thirteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pp. 66–79. Society for Industrial and Applied Mathematics.

Durant, K. and Wagner, S. (2017). On the distribution of betweenness centrality in random trees. *Theor. Comput. Sci.*, vol. 699, pp. 33–52.

Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. 1st edn. Cambridge University Press, New York. ISBN 9780521898065.

Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociom.*, vol. 40, no. 1, pp. 35–41.

Freeman, L.C. (1978). Centrality in social networks: conceptual clarification. *Soc. Netw.*, vol. 1, no. 3, pp. 215–239.

Fuchs, M. (2012). Limit theorems for subtree size profiles of increasing trees. *Comb. Probab. Comput.*, vol. 21, no. 3, pp. 412–441.

Gago, S., Hurajová, J.C. and Madaras, T. (2015). Betweenness centrality in graphs. In: Dehmer, M. and Emmert-Streib, F. (eds.), *Quantitative graph theory*, Discrete Math. Appl., pp. 233–257. CRC Press.

Girvan, M. and Newman, M.E.J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826.

Goh, K.-I., Oh, E., Jeong, H., Kahng, B. and Kim, D. (2002). Classification of scale-free networks. *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 20, pp. 12583–12588.

Jordan, C. (1869). Sur les assemblages des lignes. *J. Reine Angew. Math.*, vol. 70, pp. 185–190.

Kuba, M. and Wagner, S. (2010). On the distribution of depths in increasing trees. *Electron. J. Comb.*, vol. 17, no. R137.

Meir, A. and Moon, J.W. (1978). On the altitude of nodes in random trees. *Can. J. Math.*, vol. 30, pp. 997–1015.

Meir, A. and Moon, J.W. (1987). On major and minor branches of rooted trees. *Can. J. Math.*, vol. 39, pp. 673–693.

Meir, A. and Moon, J.W. (2002). On centroid branches of trees from certain families. *Discrete Math.*, vol. 250, no. 1–3, pp. 153–170.

Moon, J.W. (1985). On the expected distance from the centroid of a tree. *Ars Comb.*, vol. 20, pp. 263–276.

Moon, J.W. (2002). On the centroid of recursive trees. *Australas. J. Comb.*, vol. 25, pp. 211–219.

Newman, M.E.J. (2010). *Networks: An Introduction*. 1st edn. Oxford University Press, New York. ISBN 9780199206650.

Panholzer, A. and Prodinger, H. (2007). Level of nodes in increasing trees revisited. *Random Struct. Algorithms*, vol. 31, no. 2, pp. 203–226.

Shah, D. and Zaman, T. (2011). Rumors in a network: who's the culprit? *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5163–5181.

Zelinka, B. (1968). Medians and peripherians of trees. *Arch. Math.*, vol. 4, no. 2, pp. 87–95.