On the distribution of betweenness centrality in random trees

Kevin Durant and Stephan Wagner[1]

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

Abstract

Betweenness centrality is a quantity that is frequently used to measure how 'central' a vertex v is. It is defined as the sum, over pairs of vertices other than v, of the proportions of shortest paths that pass through v. In this paper, we study the distribution of the betweenness centrality in random trees and related, subcritical graph families. Specifically, we prove that the betweenness centrality of the root vertex in a simply generated tree is usually of linear order, but has a mean of order $n^{3/2}$. We also show that a randomly chosen vertex typically also has linear-order betweenness centrality, and that the maximum betweenness centrality in a simply generated tree is of order n^2. We obtain limiting distributions for the betweenness centrality of the root vertex and of a randomly chosen vertex, as well as for the maximum betweenness centrality, and we also show that the centroid has positive probability in the limit to be the vertex of maximum betweenness centrality. Some similar results also hold for subcritical graph classes, which will be briefly discussed. Finally, we study random recursive trees and other families of increasing trees, where the situation is quite different: here, the root betweenness centrality is of quadratic order, as is the maximum betweenness centrality. The betweenness centrality of a random vertex, on the other hand, is again of linear order. Again, we also have limiting distributions upon suitable normalisation.

Keywords: betweenness centrality; random tree; simply generated tree; subcritical graph class; increasing tree; centroid
2010 MSC: 60C05; 05C05, 94C15

1. Introduction

Random models of graphs and trees are a subject of interest for a number of reasons—network scientists have a desire to identify the underlying processes

[1] This material is based upon work supported by the National Research Foundation of South Africa under grant number 96236.

that give rise to certain characteristics of real-life complex networks (graphs); combinatorialists are interested in enumerating graph-like structures and describing their shapes; and computer scientists commonly use certain kinds of trees in practice, making their properties relevant for the analysis of algorithms. This paper has ties of differing strengths to all three of these fields: the centrality measure we study is natural and widely used in network science; also, the second half of the paper is concerned with random increasing trees, which are linked to computer algorithms and data structures. However our main interest here is really combinatorial in nature, as we derive a number of results that characterise the betweenness centrality of vertices in simply generated and increasing trees. These results then help to colour the already rich understanding of the structural properties (height, degree distribution, path length, etc.) of random trees.

Let G be a graph; then the *betweenness centrality* (b.c.) of a given vertex r is defined as a sum over pairs $\{v, w\}$ of vertices other than r, counting for each pair the fraction $b_{vw}(r)$ of undirected shortest paths between them that pass through r:

$$b(r) = \sum_{\{v,w\}} b_{vw}(r),$$

where $0 \leq b_{vw}(r) \leq 1$. If $G = T$ is in fact a tree, then there is only one path between any two vertices, and $b(r)$ is the total number of paths that pass through r. In this case, the b.c. can be expressed in terms of the *branches* of T joined to r, which are the maximal subtrees of T not containing r:

$$b(r) = \sum_{i<j} |T_i||T_j|. \tag{1}$$

(Here T_i is a branch and $|T_i|$ its vertex count, or *size*.) This is precisely the number of ways to choose two unordered vertices from distinct branches of r,[2] which is a neatly phrased combinatorial problem. It is worth briefly noting here that the b.c. of any vertex in a graph is bounded from above by $\binom{n-1}{2}$. For more on betweenness centrality, we refer the reader to [1], [2, Section 7.7], or [3] for a practical application. A more mathematical survey is provided in [4]; [5] studies the betweenness centrality in real-world networks.

A key observation that follows from equation (1) will yield a limiting distribution for $b(r)$: if one of the branches of r (without loss of generality, T_1) is 'large', while the others combined contain a fixed number k of vertices, then $b(r)$ is dominated by paths between T_1 and the other branches: if the branch sizes are $n - k - 1 = n_1, n_2, \ldots, n_d$, so that $n_2 + \cdots + n_d = k$ remains fixed as the size n of the tree tends to infinity, then we have

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

[2] We will often refer to the branches of T joined to r as 'the branches of r.'

As it turns out, this observation applies, with high probability, to one of the random tree families we are interested in (simply generated trees).

The two broad families of random trees this paper is concerned with are simply generated (s.g.) and increasing trees, treated in Sections 2 and 4 respectively. Both types of trees are amenable to common methods based on generating functions (g.f.'s), but have markedly different combinatorial properties—the most significant being that an increasing tree is typically well balanced, so that its vertices are quite evenly distributed among the branches of its root vertex, whereas a s.g. tree is not. In addition to these two families, and as an extension to the study of s.g. trees, we also give some results derived for classes of subcritical graphs, which are tree-like in nature, in Section 3.

We follow a single course of analysis, and repeat it for each tree family: firstly, the moments of the b.c. of the root vertex are derived, and then a description of its limiting distribution is given. For increasing trees, we also describe the b.c. of the vertex with a given label. Secondly, a limiting distribution for the b.c. of a random vertex is obtained. Lastly, we consider the distribution of the maximum b.c. in a tree, along with the probability that the centroid vertex attains this maximum.[3] For s.g. trees, we rely on the continuum random tree and its connection to triangulations of the circle.

Our results can be summarised as follows: The kth moment of the b.c. of the root in a s.g. tree or subcritical graph of size n is $\Theta(n^{2k-(1/2)})$, however as $n \to \infty$, it is the *linearly* scaled b.c. of the root (or any randomly chosen vertex) which yields a limiting distribution, implying that vertices in a large s.g. tree typically have b.c. of $\Theta(n)$. In an increasing tree, the kth moment of the b.c. of any vertex with a fixed label is $\Theta(n^{2k})$, and its limiting distribution requires scaling by n^{-2}. The limiting distribution of a randomly chosen vertex, however, is once again obtained via a scaling factor of n^{-1}. The maximum b.c. in a s.g. tree or increasing tree is always $\Theta(n^2)$, and the probability that the centroid attains this maximum tends to a positive constant.

1.1. Notation

A brief word on the notation used throughout this paper: we use $\mathcal{T}_n$ to refer to those objects of size n in a class $\mathcal{T}$, and $[x^n]y(x)$ to denote the coefficient of x^n in a g.f. $y(x)$. When the vertices of a tree can be referred to directly (such as in labelled trees), $b(l)$ will refer to the b.c. of vertex l. Otherwise, $b(T)$ will be used to implicitly denote the b.c. of T's root.

2. Simply generated trees

If one couples a non-negative weight ϕ_i to each vertex in a rooted tree according to its out-degree i, and defines the weight $\omega(T)$ of the entire tree to be

[3] We have not investigated the maximum b.c. in subcritical graphs (or its relation to the centroid) here; this remains as possible future work.

the product of these weights, then the resulting class of trees can be counted using the g.f.

$$y(x) = \sum_{T \in \mathcal{T}} \omega(T) x^{|T|} = x\phi(y(x)), \tag{3}$$

where $\phi(u) = \sum_{i=0}^{\infty} \phi_i u^i$. Such a class is called *simply generated* (see [6], [7, Section VII.3] or [8, Section 1.2].[4] 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 [9, 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 s.g. trees has the characteristic property that its g.f. $y(x)$ 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:

$$y(x) = \tau - \gamma\sqrt{1 - \frac{x}{\rho}} + O\left(1 - \frac{x}{\rho}\right), \tag{4}$$

in which $y(\rho) = \tau$ and $\gamma = \sqrt{2\phi(\tau)/\phi''(\tau)}$. Because of this, many interesting properties of s.g. trees can be deduced almost mechanically using singularity analysis. The number of trees of size n, for example, is

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

The expected height of one of these trees is $\Theta(\sqrt{n})$, and the expected number of nodes at a fixed distance k from the root is only linear in k [7]. Another interesting result, considering that we are about to address the b.c. of the root vertex, is that the root of a s.g. tree is known to have up to three 'major' branches, with mean sizes of orders n, $\sqrt{n}$, and $\log n$ [10]. In light of this, one might expect that the b.c. 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 not only that this is the case, but also that the kth moment of the root's b.c. is $\Theta(n^{2k-(1/2)})$.

2.1. Moments of the betweenness centrality of the root

Theorem 1. *If $\mathcal{T}$ is a class of s.g. trees, then the mean b.c. of the root vertices in $\mathcal{T}_n$ is $\Theta(n^{3/2})$. More precisely:*

$$\mathbb{E}_n(b(T)) \sim \frac{\gamma^{-1}\tau}{2}\sqrt{\pi n^3}.$$

The proof of the above theorem is quick if one recalls that the b.c. of a vertex r is the number of ways to distinguish two unordered vertices from distinct branches of r, and notes that the g.f. of a 'pointed' tree—in which a vertex has been distinguished—is $\widehat{y}(x) = xy'(x)$.

[4] Because of the parallel between the recursive definition of s.g. trees and Galton-Watson branching processes, these trees are also referred to as Galton-Watson trees.

PROOF. The g.f. of trees in which two of the root's branches have been replaced with pointed branches encodes the total b.c. over the roots of trees of size n, and can be constructed explicitly:

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

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

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

Since $[x^n]H(x) = \sum_{\mathcal{T}_n} b(T)$, the result follows by computing $[x^n]H(x)/y_n$. $\square$

Considering that the b.c. of a vertex is bounded by $\binom{n-1}{2} = \Theta(n^2)$, and that it is certainly possible to construct trees—stars, for example—in which the root attains this bound, one can explain Theorem 1 intuitively by saying that the rather unlikely event (whose probability is only of order $n^{-1/2}$) of the root having two large branches, each with a number of vertices linear in n, dominates the asymptotic behaviour. By the same reasoning, one might then expect that the kth moment of $b(T)$ will be of order $n^{2k-(1/2)}$.

In deriving these higher-order moments, the following lemma will prove useful, both for s.g. trees and for subcritical graphs, the latter of which will be treated in the following section.

Lemma 1. *Let $\mathcal{C}$ be a 'tree-like' class, in that it is counted by a g.f. $c(x) = x\phi(f(x))$ 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 $f(x)$ of every tree with pointed branches—each of which may possibly distinguish multiple vertices, and which in total contain d distinguished vertices—yields a generating function whose dominant term is $\Theta((1-(x/\rho))^{-d+(m/2)})$.*

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

PROOF. The g.f. obtained after substitution is a linear combination of terms of the form

$$x\left(\prod_{i=1}^{m}\widehat{f}_{d_i}(x)\right)\phi^{(m)}(f(x)), \tag{5}$$

in which $\widehat{f}_{d_i}(x)$ is the g.f. of the ith substituted branch, which has d_i distinguished vertices:

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

where $\left\{ {j \atop l} \right\}$ denotes the Stirling numbers of the second kind. It is these branches that determine the overall asymptotic behaviour of the expression in (5), 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 (5) because $\sum_i d_i = d$ and $\phi(u)$ is analytic at $u = f(\rho)$. □

Theorem 2. *The kth moment of the b.c. of a root vertex in $\mathcal{T}_n$ is $\Theta(n^{2k-(1/2)})$, and satisfies, for $k \geq 1$,*

$$\mathbb{E}_n\big(b(T)^k\big) \sim \frac{\gamma^{-1}\tau}{2^{4k-3}}\binom{2k-2}{k-1}\sqrt{\pi n^{4k-1}}.$$

Proof. We are trying to derive the mean of the function $b(T)^k$, which can be expanded as

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

(where K is some constant that depends on k), since $b(T)^k$ involves k chances to choose a pair of branches. The mean of each of the sums in the above equation can be computed by interpreting the sum as a selection of $2k$ vertices from a number of branches, and then constructing the relevant g.f.; however Lemma 1 tells us that the term involving the fewest branches will have the greatest asymptotic order. With this in mind, we can simplify the g.f. that sums $b(T)^k$ over trees of size n to

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

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

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

Tree	$\phi(u)$	τ	ρ	γ	$\mathbb{E}_n(b(T))$	$\mathbb{V}_n(b(T))$
binary	$(1+u)^2$	1	1/4	2	$\sqrt{\pi n^3}/4$	$\sqrt{\pi n^7}/32$
plane	$(1-u)^{-1}$	1/2	1/4	1/2	$\sqrt{\pi n^3}/2$	$\sqrt{\pi n^7}/16$
labelled	$\exp(u)$	1	$1/e$	$\sqrt{2}$	$\sqrt{\pi n^3/8}$	$\sqrt{\pi n^7/512}$

Table 1: Lead-order asymptotics for the mean and variance of the b.c. of the root vertex in selected s.g. trees.

As in the proof of Theorem 1, the desired quantity is $[x^n]H_k(x)/y_n$, which one can extract using Theorem VI.1 of Flajolet and Sedgewick's comprehensive book [7]. □

The second moment of the b.c. of the root is asymptotically equivalent to $\gamma^{-1}\tau\sqrt{\pi n^7}/16$, and thus its variance is as well: $\mathbb{V}_n(b(T)) \sim \gamma^{-1}\tau\sqrt{\pi n^7}/16$. Table 1 gives some indicative values for a few common s.g. trees.

2.2. Limiting distribution of the betweenness centrality of the root

Although b.c.'s of order n^2 appear to dominate the moments of $b(T)$, the lagging factor of order $n^{-1/2}$ suggests that these events become increasingly rare as $n \to \infty$. In this small section, we show by symbolic construction that there is a limiting distribution for the *linearly* scaled b.c. of the root, $b(T)/n$. This implies that trees with one large root branch—of size linear in n—are sufficient to describe the distribution of $b(T)$ when n is large enough, which is in agreement with known results about the unbalanced nature of s.g. trees [10, 11].

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 vertices of trees in $\mathcal{L}_k$ have predictable, linear-order b.c., and in the limit $n \to \infty$, the classes $(\mathcal{L}_k)_n$ together describe $\mathcal{T}_n$.

Theorem 3. *The distribution of the linearly scaled b.c. of a root vertex in $\mathcal{T}_n$ converges weakly, as $n \to \infty$, to the discrete distribution defined by*

$$\mathbb{P}(k) = p_k = \rho^{k+1}[x^k]\phi'(y(x)).$$

Specifically, for fixed k and every $0 < \varepsilon < 1$:

$$\mathbb{P}_n(|(b(T)/n) - k| < \varepsilon) \xrightarrow[n\to\infty]{} p_k.$$

Proof. Firstly, we reiterate that the b.c. 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 vertices, then by equation (2) we have $b(T) = nk + O(k^2)$. Secondly, note that for large enough n, any two subclasses $\mathcal{L}_k$ and $\mathcal{L}_l$ are disjoint, since $(\mathcal{L}_k)_n \cap (\mathcal{L}_l)_n = \emptyset$ if $n > k+l+1$.

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 as n grows, and that the sum of these limiting probability masses is 1.

Begin by considering the g.f. $L_k(x)$ 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) vertices:

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

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 g.f., 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:

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

so that they describe a probability distribution. Thus the limiting distribution of $b(T)$ can be fully described using only the limit behaviour of the subclasses $\mathcal{L}_k$. □

It is also worth pointing out that an expansion of $\phi(u)$ around $u = \tau = y(\rho)$ gives $p_k \sim \gamma^{-1}\tau/\sqrt{\pi k^3}$, as $k \to \infty$, for any s.g. family of trees.

2.3. Limiting distribution of the betweenness centrality of a random vertex

The previous sections dealt specifically with the b.c. of the *root* vertex in s.g. trees, but the constructive idea of Section 2.2 can be used to obtain a limiting distribution for the b.c. of a *random* vertex as well. In the exceptional case of labelled trees (with $\phi(u) = \exp(u)$), all of the preceding results hold for non-root vertices automatically, because there is a natural mapping between unrooted and rooted labelled trees: each unrooted tree of size n gives rise to n rooted trees—one for each label—implying that iteration over the vertices of unrooted labelled trees is equivalent to iteration over the roots of rooted labelled trees. In general, however, this mapping does not hold for other s.g. trees. Still, we can show that like the root vertex, a randomly chosen vertex in a s.g. tree usually has b.c. of linear order.

Theorem 4. *The distribution of the linearly scaled b.c. of a randomly chosen vertex v in a s.g. tree $T \in \mathcal{T}_n$ converges weakly as $n \to \infty$ to the discrete distribution given by*

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

Specifically, for fixed k and every $0 < \varepsilon < 1$:

$$\mathbb{P}_n(|(b(v)/n) - k| < \varepsilon) \xrightarrow[n\to\infty]{} q_k.$$

The proof of Theorem 4 is similar to that of the corresponding result for root vertices, Theorem 3, except that in addition to its descendent branches, a non-root vertex 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 vertices among v's other branches.

Proof. Any vertex v with k descendants in a s.g. tree T of size n can be viewed as a leaf vertex of a rooted tree of size $n-k$ (its ancestral branch) to which a forest of size k (the descendent branches) has been grafted. If $(\mathcal{L}_k)_n$ is the resulting subclass of trees, its g.f. 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 g.f. $y(x,u)$ that marks the leaves of every tree with an auxiliary variable u, by taking the partial derivative of $y(x,u)$ with respect to u, and then setting $u=1$, yielding a g.f. that counts, for each tree, the possible points of attachment for our forest of size k. The entire g.f. of $\mathcal{L}_k$ is thus

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

in which $y(x,u) = x\phi(y(x,u)) + (u-1)\phi_0 x$ (c.f. [8, p. 84]), and the presence of ϕ_0^{-1} removes the weight that was assigned to the chosen leaf vertex, since a new weight will be assigned to it along with its grafted forest $\phi(y(x))$.

As in the proof of Theorem 3, v has b.c. $nk + O(k^2)$, and any two subclasses $(\mathcal{L}_k)_n$ and $(\mathcal{L}_l)_n$ $(k \neq l)$ are disjoint. To see that in the limit $n \to \infty$ a tree of size n with a distinguished vertex has probability q_k of belonging to $\mathcal{L}_k$, we need to express $L_k(x)$ 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, it follows that

$$\frac{d}{du}y(x,u)\bigg|_{u=1} = \phi_0 x(1 - x\phi'(y(x)))^{-1} \sim \phi_0\frac{\rho\gamma}{2\tau}\left(1 - \frac{x}{\rho}\right)^{-1/2}$$

as $x \to \rho$, with which $L_k(x)$ can be expressed, and the limiting probability q_k derived as

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

Note finally that the q_k sum to 1:

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

□

Tree	$\phi(u)$	τ	ρ	p_k	q_k
binary	$(1+u)^2$	1	$1/4$	$2^{-(2k+1)}\frac{1}{k+1}\binom{2k}{k}$	$4^{-(k+1)}\frac{1}{k+2}\binom{2k+2}{k+1}$
plane	$(1-u)^{-1}$	$1/2$	$1/4$	$4^{-(k+1)}\frac{1}{k+2}\binom{2k+2}{k+1}$	$2^{-(2k+1)}\frac{1}{k+1}\binom{2k}{k}$
labelled	$\exp(u)$	1	$1/e$	$e^{-(k+1)}\frac{(k+1)^{k-1}}{k!}$	$e^{-(k+1)}\frac{(k+1)^{k-1}}{k!}$

Table 2: The limiting probabilities p_k and q_k of a root and random vertex, respectively, in a s.g. tree of size n having b.c. that approaches nk.

It is once again worth pointing out that $q_k \sim (2\tau)^{-1}\gamma/\sqrt{\pi k^3}$ as $k \to \infty$ for any family of s.g. trees.

Table 2 lists values of the limiting probabilities $\mathbb{P}(k)$ for root and random vertices respectively, for some common trees. Observe that $p_k = q_k$ for labelled trees, as expected.

The final section on s.g. trees covers the b.c. of the centroid vertex and, more generally, the maximum b.c. in a tree. The motivation for considering the centroid is simple: vertices whose branch sizes are 'balanced' lead to high b.c.'s, and the centroid is in a sense the most balanced vertex in a tree.

2.4. Maximum betweenness centrality and the centroid

Together, the previous few sections have shown that the average b.c. of a root vertex in a s.g. tree is of order $n^{3/2}$, but that both the root and a randomly chosen vertex have 'typical' b.c. of only linear order. In contrast, the maximum b.c. is *always* of quadratic order, as we will now show. (In fact, we already know from Section 2.1 that root vertices with quadratic-order b.c., which are comparatively rare, dominate the root moments.)

A trivial lower bound for the maximum b.c. in a given tree of size n is $(n^2 - 2n)/4$. This can be shown by considering the *centroid* of the tree: the centroid consists of those vertices that minimise the total distance to all other vertices. Equivalently, one can define a centroid vertex as a vertex with the property that none of its branches contain more than half of the tree's vertices. It is well known that there is either a unique centroid vertex (in fact, this happens asymptotically almost surely in a random tree), which we will simply call the centroid, or two adjacent centroid vertices. In the latter case, removing the edge between the two centroid vertices must leave two components of exact size $n/2$. This was already shown by Jordan in 1869 [12]; see also [13, Chapter 4] or [14].

It is easy to see that the b.c. of a vertex decreases when vertices are transferred from one of its branches to another branch of greater or equal size. Therefore, the smallest possible b.c. of the centroid occurs when there are only two centroid branches whose sizes are $\lfloor (n-1)/2 \rfloor$ and $\lceil (n-1)/2 \rceil$. In this case, the b.c. is $\lfloor (n-1)^2/4 \rfloor \geq (n^2 - 2n)/4$. This provides a lower bound for the maximum b.c., as mentioned earlier.

Although a centroid vertex must necessarily have fairly large b.c., this does not imply that it is always the vertex where 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 b.c. of about $n^2/4$ in this case, while the centre of the star has about $5n^2/18$.

In spite of this counterexample, the centroid will play a major role in our analysis of the maximum b.c. As it turns out, the event that the centroid's b.c. is in fact the maximum has positive limiting probability, and we will also be able to show that the maximum b.c. of a random s.g. tree, once rescaled by a factor n^{-2}, has a limiting distribution. This limiting distribution—unlike the distribution of the b.c. of a randomly chosen vertex—is even independent of the specific class of s.g. trees. Before we give a rigorous argument, let us provide some intuition. To this end, let us review a connection to random triangulations of the circle that is due to Aldous [15], as well as some results of Meir and Moon [9] on centroid branches.

The limit object of s.g. trees is the celebrated continuum random tree (see the work of Aldous [11, 16, 17] and [8, Section 4.1.3]), and its dual (in some sense) is the random triangulation of a circle. This duality between trees and triangulations is best seen in the case of binary trees, where one is the plane dual of the other. Triangles in the limit correspond to vertices in the tree with three 'large' branches (of linear order); the lengths of the three arcs defined by a triangle correspond to the sizes of the branches.

The centroid corresponds to the triangle (almost surely, there is only one) that contains the centre of the circle. If we associate to a triangle with arc lengths a, b, c the weight $ab + bc + ca$, then this gives us (asymptotically up to a scaling factor n^2) the b.c. of the corresponding branching vertex. The maximum b.c. corresponds to the maximum weight of a triangle, and the distribution of this maximum is the limiting distribution of the b.c. We point out that a maximum indeed exists almost surely: it is easy to see that any triangle with a weight greater than that of the centroid triangle has to have a longer shortest arc than the centroid triangle, and there are at most finitely many such triangles.

Meir and Moon [9] showed, among other things, that the average b.c. of the centroid of a random s.g. tree is asymptotically equal to $(1 - (1/\sqrt{2}))n^2$ (they formulated it in terms of the probability that the path between two randomly chosen vertices contains the centroid). Note that $1 - (1/\sqrt{2}) \approx 0.293$. This result implies an asymptotic lower bound for the average maximum b.c., and it turns out that this is actually not far from the truth.

Let us now make these ideas more rigorous. For ease of presentation, we stick to the special case of labelled trees, but the same arguments apply (mutatis mutandis) also to other families of s.g. trees, and lead to the same result (in fact, with the same limiting distribution). Let us start with some technical preliminaries:

Lemma 2. *Fix ε with $0 < \varepsilon < \frac{1}{12}$. With probability tending to 1, there is no vertex in a random labelled tree with n vertices for which three of its branches all contain at least $n^{1-\varepsilon}$ vertices, and the rest of the tree (all except for those three branches) contains at least $n^{1-\varepsilon}$ vertices as well.*

Proof. This is achieved by means of the first moment method: we prove that the mean number of such vertices 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 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 simple consequence of Stirling's formula. Since the number of choices of n_1, n_2, n_3, m is $\Theta(n^3)$, we obtain that the total number of rooted trees with the property that three branches and the rest of the tree all have size at least $n^{1-\varepsilon}$ is

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

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

Lemma 3. *Fix constants $\alpha, \beta, \varepsilon$ with $0 < \alpha < \beta \leq \frac{1}{4}$ and $\varepsilon > 0$, and assume that n is sufficiently large. Let T be a tree with n vertices and a centroid vertex with three branches of size at least βn. If v is a non-centroid vertex with the property that all but at most $n^{1-\varepsilon}$ vertices belong to the three largest branches and the third-largest branch has at most αn vertices, then v has smaller b.c. than the centroid vertex.*

Proof. Recall that the b.c. of a vertex decreases when vertices are transferred from one of its branches to another branch of greater or equal size. This, together with the definition of a centroid, shows that the b.c. of the centroid is at least equal to

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

corresponding to three branches of sizes $\beta n, (n/2) - \beta n, n/2$ respectively. On the other hand, the b.c. of vertex v is at most

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

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

Lemma 4. *Fix a constant $\alpha > 0$. A tree T with n vertices has no more than $(1/\alpha) - 2$ vertices with at least three branches that each contain at least αn vertices.*

Proof. We call a vertex with three or more branches containing at least αn vertices a "big" vertex, while other vertices are called "small". Consider the tree R that is obtained as follows: take the tree consisting of all big vertices and the paths between them. Now suppress all small vertices, thereby reducing paths between large vertices that only contain small vertices to single edges.

Suppose that this tree has a total of r vertices, of which a_j have degree j. We note that vertices of degree 1 in this tree have to have two branches of size at least αn in T not containing any of the other vertices of R, while vertices of degree 2 in R have to have at least one such branch in T. This gives us a total of $2a_1 + a_2$ disjoint branches of at least αn vertices, so $2a_1 + a_2 \leq 1/\alpha$. On the other hand, since

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

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 to 4, we need the following result from the aforementioned papers of Aldous [15] and Meir and Moon [9]:

Lemma 5. *Let $(X_{1,n}, X_{2,n}, X_{3,n})$ denote the sizes of the three largest centroid branches of a random labelled tree with n vertices. Then the normalised random vector*

$$n^{-1}(X_{1,n}, X_{2,n}, X_{3,n})$$

converges in distribution to a vector with density $(12\pi)^{-1}(x_1x_2x_3)^{-3/2}$ on the set of triples (x_1, x_2, x_3) such that $0 < x_1, x_2, x_3 < 1/2$ and $x_1 + x_2 + x_3 = 1$.

Now we are ready for a formal proof of the following theorem that we alluded to earlier:

Theorem 5. *The maximum b.c. of a random labelled tree of size n, divided by n^2, converges weakly to a limiting distribution. The probability that the maximum b.c. is attained by the centroid tends to a positive constant.*

Proof. Consider the event that every vertex with maximum b.c. has at least three branches of size at least αn. Combining Lemmas 5, 3 and 2, we see that for $n > N_\alpha$, this event has a probability 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 vertices with three branches of size at least αn, of which there are, by Lemma 4, only a bounded number, and for which there are only a finite number of potential configurations with a nonzero limiting probability. Such a configuration can be seen as a labelled tree with $r \leq (1/\alpha) - 2$ vertices, no vertices of degree greater than 3, with edges representing birooted connecting trees (possibly empty, and the two roots may coincide), and vertices of degree $k < 3$ having $3 - k$ 'large' branches (rooted trees with at least αn vertices) attached to them, see Figure 9 for an example. Note that each of the

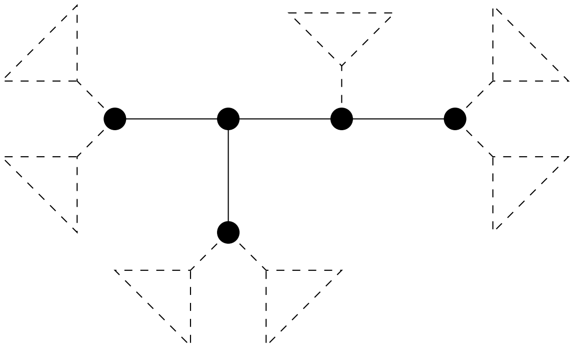

Figure 1: A configuration as described in the proof of Theorem 5. Edges represent (birooted) connecting trees, dashed triangles 'large' branches of size at least αn.

vertices may also have smaller branches with a total of at most $O(n^{1-\varepsilon})$ vertices. Let the sizes of the birooted trees and the sizes of the additional large branches be $x_1, x_2, \ldots, x_{r-1}$ and $y_1, y_2, \ldots, y_{r+2}$ respectively. Using the fact that there are $x_j^{x_j}$ possible birooted trees for each j and $y_j^{y_j-1}$ possible rooted trees for each j, we obtain an asymptotic expression for the number of possible trees corresponding to each configuration. We remark that there might actually be further vertices with three branches of size αn or more for a given configuration of r vertices inside the birooted connecting trees and large branches, which one can account for by means of an inclusion-exclusion argument.

In the end, one finds that the sizes of the connecting trees and large branches, scaled by a factor n, converge to a limiting distribution with an explicitly computable density for each configuration (as in Lemma 5). Since the b.c.'s of the vertices with three 'large' branches only depend on these sizes up to $O(n^{2-\varepsilon})$, we can infer the limiting distribution of the maximum b.c. of vertices with at least three branches of size αn or more, as well as a limiting probability that this maximum is attained by the centroid, for each fixed $\alpha > 0$. Letting α go to 0 now yields the desired result on the limiting distribution of the maximum b.c., and also shows that there must be a limiting probability for the centroid to attain the maximum b.c. To show that this probability is in fact positive, we can use a simple argument: Suppose that all three centroid branches have fewer than $((4/9) - \delta)n$ vertices (for some small $\delta > 0$), which happens with positive limiting probability by Lemma 5. Then the b.c. of the centroid is at least

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

which is obtained when the branches are as "unbalanced" as possible. On the other hand, every other vertex v has to have a branch of size at least $((5/9) + \delta)n$ (take the branch of v containing the centroid), and since with probability negligibly close to 1 v has at most 3 branches of size linear in n, an upped bound for its b.c. occurs when its second and third branches each contain $((4/9)-\delta)n/2$

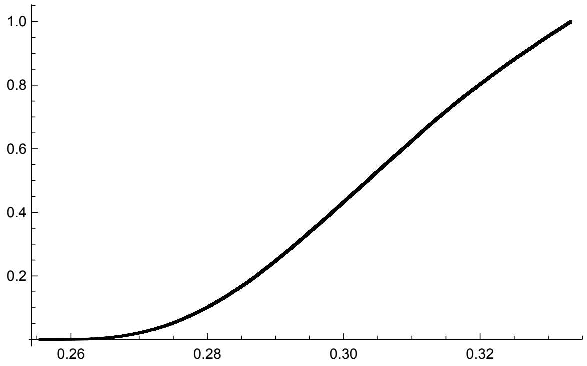

Figure 2: The cumulative distribution function of the limiting distribution of maximum b.c. in s.g. trees.

vertices:

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

which (for suitably small δ and then sufficiently large n) is strictly smaller than that of the centroid. This completes the proof. □

An argument similar to the final paragraph shows that the probability for the centroid to have maximum b.c. is strictly less than 1, and one can also show in the same fashion that the limiting random variable of the maximum b.c. has the interval $[1/4, 1/3]$ as its support.

Numerically, the average maximum is asymptotically equal to $0.303n^2$ (the numerical value of the constant was determined by Monte Carlo sampling; it might be possible to obtain an explicit expression for the constant, but this does not seem to be a trivial task). Moreover, the probability that the centroid is in fact also the vertex with maximum b.c. converges to a constant whose numerical value is 0.621. The limiting distribution function of the normalised maximum b.c. is shown in Figure 2. Just like the aforementioned constants, it was obtained by means of Monte Carlo sampling in view of the rather complicated nature of the limiting distribution.

One way to perform this Monte Carlo simulation is to first generate the centroid branch sizes according to the density given in Lemma 5; then one repeats this recursively for all branches. Once it is no longer possible to obtain vertices with b.c. greater than the current maximum (which must happen after a finite number of steps with probability 1), one can stop the process.

3. Subcritical graphs

There are classes of tree-like graphs that are in some ways similar to s.g. trees, called *subcritical graphs*. In particular, outerplanar, series-parallel, and cacti graphs are all special cases of subcritical graphs [18].[5]

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 vertices must be deleted before it becomes disconnected), every graph can be decomposed into its blocks, cut vertices (vertices whose removal disconnects the graph), and the induced edge set that links cut vertices to their incident blocks, leading to a bipartite tree known as the block-cut vertex tree. A class of graphs is called *block-stable* if it contains the two-vertex one-edge 'link' graph, and satisfies the property that a graph belongs to the class if and only if all of its blocks do as well.

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

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

where $B'(x)$ is the g.f. of the class $\mathcal{B}'$ of blocks with one removed vertex.

The 'subcriticality' property of subcritical graphs is a technical condition that requires the radii of convergence of $C(x)$ and $B(y)$, ρ and η respectively, to satisfy $C(\rho) < \eta$. This implies that $B'(y)$ 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 s.g. trees (see [18]). In particular, we have $\rho^{-1} = \exp(B'(\tau))B''(\tau)$ and

$$B'(y) = B'(\tau) + B''(\tau)(y-\tau) + O\big((y-\tau)^2\big), \tag{6}$$

$$C(x) = \tau - \mu\sqrt{1-\frac{x}{\rho}} + O\left(1-\frac{x}{\rho}\right), \tag{7}$$

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

Our goal is again to investigate the b.c. of the root vertex, however because we are considering labelled graphs in which any vertex can be distinguished as the root, *our results hold for a randomly chosen vertex as well.* The only real caveat when working with subcritical graphs is that the b.c. of a vertex v is no longer solely determined by paths between its branches (here, branches take the form of blocks with one vertex removed and subgraphs rooted to their remaining vertices, and have the g.f. $W(x) = B'(C(x))$). 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 $B \in \mathcal{B}'$. Because shortest paths within blocks are not necessarily unique, the contribution

[5] Although the vertices of subcritical graphs can be either labelled or unlabelled, we consider only the labelled case here.

of paths between the subgraphs of W to $b(C)$, the b.c. of the root vertex, is

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

where v, w is a pair of non-root vertices in B, such that $\sum b_{vw}(B) = b(B)$ is the b.c. of B's removed vertex with respect to paths contained within B, and C_v and C_w are the subgraphs rooted at v and w.

The full expression of the b.c. of a graph's root is then

$$\begin{aligned} b(C) &= \sum_{a<b} |W_a||W_b| + \sum_B \sum_{v<w} b_{vw}(B)|C_v||C_w| \\ &= b_1(C) + b_2(C), \end{aligned} \tag{8}$$

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

3.1. Moments of the betweenness centrality of the root

When deriving the moments of $b(C)$, we can handle the two terms in equation (8) individually. The contribution of $b_1(C)$ is identical, conceptually, to the b.c. of the root of a tree, so one need only count graphs with two distinguished vertices from distinct branches. A g.f. $\sum_{\mathcal{C}} b_2(C)x^{|C|}$ 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 B must be weighted by $b_{vw}(B)$. These observations lead to a relatively straightforward derivation of the expected b.c. of the root vertex.

Theorem 6. *Let $\mathcal{C}$ be the class of labelled subcritical graphs constructed from a block class $\mathcal{B}$. Then the root vertices in $\mathcal{C}_n$ have mean b.c. of order $n^{3/2}$:*

$$\mathbb{E}_n(b(C)) \sim K\sqrt{\pi n^3},$$

where

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

and $M(y) = \sum_{\mathcal{B}} b(B)y^{|B|}$ is the cumulative g.f. of $b(B)$ over blocks B in $\mathcal{B}$.

Proof. We desire the g.f. $H(x) = \sum_{\mathcal{C}} b(C)x^{|C|}$, which can be written as the sum of the corresponding g.f.'s for $b_1(C)$ and $b_2(C)$. The first of these two generating functions is

$$U_1(x) = \sum_{C\in\mathcal{C}} b_1(C)x^{|C|} = \frac{x^3}{2}W'(x)^2\exp(W(x)) = \frac{x^2}{2}W'(x)^2C(x).$$

From the expansions of $B'(y)$ and $C(x)$ given in equations (6) and (7), we can derive an expansion for $W(x)$:

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

so that $U_1(x)$ satisfies

$$U_1(x) \sim \frac{\tau}{2}\Big(\frac{\mu}{2}B''(\tau)\Big)^2\Big(1-\frac{x}{\rho}\Big)^{-1}. \tag{9}$$

The g.f. of $b_2(C)$ requires two stages of substitution, since we must first derive the g.f. $L(x)$ that describes branches that have had two vertices distinguished from their subgraphs. We will then have

$$U_2(x) = \sum_{C\in\mathcal{C}} b_2(C)x^{|C|} = xL(x)\exp(W(x)) = L(x)C(x).$$

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

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

where

$$M(y) = \sum_{B\in\mathcal{B}} b(B)y^{|B|} = m_2y^2 + m_3y^3 + \cdots .$$

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

$$U_2(x) \sim \frac{1}{\tau}\Big(\frac{\mu}{2}\Big)^2 M(\tau)\Big(1-\frac{x}{\rho}\Big)^{-1}. \tag{10}$$

Equations (9) and (10) imply that both kinds of paths contribute equally in order to the b.c. of the root vertex, and the expected b.c. of the root of a graph of size n is $[x^n](U_1(x)+U_2(x))/|\mathcal{C}_n|$. □

The higher-order moments of $b(C)$ are more interesting, because they involve the function

$$b(C)^k = (b_1(C)+b_2(C))^k = \sum_{j=0}^{k}\binom{k}{j} b_1(C)^{k-j}b_2(C)^j. \tag{11}$$

In the case of s.g. trees, $b(T)^k$ could be interpreted as a selection of $2k$ vertices from between 2 and $2k$ distinct branches, and we could restrict our calculation to the case of exactly 2 branches due to Lemma 1. This basic concept holds once again, for both $b_1(C)^k$ and $b_2(C)^k$, so that

$$b_1(C)^k \sim \sum_{a<b} |W_a|^k |W_b|^k,$$

$$b_2(C)^k \sim \sum_{B\in\mathcal{B}} \sum_{v<w} b_{vw}(B)^k |C_v|^k |C_w|^k.$$

Both of these terms lead to g.f.'s (of the form $\sum_{\mathcal{C}} b_i(C)^k x^{|C|}$) that are dominated by a term of order $(1-(x/\rho))^{-2k+1}$. The question, however, is whether the remaining terms in equation (11)—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 g.f. involving both $b_1(C)$ and $b_2(C)$ is three: some vertices 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 1 implies that the replacement of a branch or subgraph with one in which d vertices have been distinguished contributes $(1-(x/\rho))^{-d+(1/2)}$ to the final order of the g.f., which tells us that the 'mixed' terms of $b(C)^k$ grow at a slower rate than those involving only $b_1(C)$ or $b_2(C)$. This simplifies the asymptotic behaviour of $b(C)^k$ greatly:

$$\mathbb{E}_n\big(b(C)^k\big) \sim \mathbb{E}_n\big(b_1(C)^k\big) + \mathbb{E}_n\big(b_2(C)^k\big).$$

We find that the kth moment of the b.c. of the root vertex satisfies an expression that is very similar to the one derived for s.g. trees. The second moment is once again of order $n^{7/2}$, so that the variance of $b(C)$ is as well.

Theorem 7. *If $\mathcal{C}$ is a class of labelled subcritical graphs with block class $\mathcal{B}$, then the kth moment of the b.c. of a root vertex in $\mathcal{C}_n$ is $\Theta(n^{2k-(1/2)})$. Specifically, for $k \geq 1$:*

$$\mathbb{E}_n\big(b(C)^k\big) \sim K_k \sqrt{\pi n^{4k-1}},$$

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

Proof. The asymptotic behaviour of $H_k(x) = \sum_{\mathcal{C}} b(C)^k x^{|C|}$ is

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

in which

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

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

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

□

3.2. Limiting behaviour of the betweenness centrality of the root

Since the moments of the b.c. of the root vertex in a subcritical graph behave similarly to those found for s.g. trees, it is probably unsurprising that we can show that the majority of these root vertices (in subcritical graphs) have linear-order b.c., and that the balanced graphs which lead to quadratic-order b.c. become increasingly rare as $n \to \infty$.

To do so, we repeat the procedure of Section 2.2, defining unbalanced subclasses $\mathcal{L}_{k,m} \subset \mathcal{C}$ that not only have k non-root vertices outside their largest branch, but also have a dominant subgraph within that branch. This subgraph includes all but m of the large branch's vertices. If we let

$$\Lambda_{k,m} = \Big[[x^k] \exp(W(x))\Big] \Big[[x^m] B''(C(x))\Big]$$

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

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

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

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

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

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

Finally, the b.c. of the root of a graph C in subclass $(\mathcal{L}_{k,m})_n$ is of linear order, since there are linearly many of the two kinds of paths through the root: if k_i ($i = 2, \dots, \alpha$) and m_j ($j = 2, \dots, \beta$) are the minor branch and subgraph sizes respectively, we have

$$\begin{aligned} b(C) &\sim (n-k-m-1) \left(\sum_{i=2}^{\alpha} k_i + \sum_{j=2}^{\beta} b_{vw_j}(B) m_j \right) \\ &= nk + n \sum_{j=2}^{\beta} b_{vw_j}(B) m_j + O\big((k+m)^2\big). \end{aligned}$$

Noting that $0 \leq b_{vw_j}(B) \leq 1$, we have a linear bound on $b(C)$:

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

This gives us the following theorem, which is a qualitative analogue of Theorem 3, albeit less precise:

Theorem 8. *Let the graph C be randomly chosen from a labelled subcritical graph class $\mathcal{C}$. For every $\varepsilon > 0$, there exists a real number M such that*

$$\limsup_{n\to\infty} \mathbb{P}_n(b(C) > Mn) < \varepsilon.$$

In short, Theorem 8 says that $b(C_n)/n$ is bounded in probability: $b(C) = O_p(n)$.

If more information on the blocks of the specific class of subcritical graphs—and in particular their b.c.'s—is available, it is also possible to provide a more precise limit law, as for s.g. trees. We also remark again that the distribution is the same for a random vertex: as in the case of random labelled trees, every vertex of a random labelled subcritical graph has the same probability to be the root.

This brings to a close the first part of the paper, which dealt with s.g. trees and subcritical graphs. Both of these structures are characteristically unbalanced, or 'skinny', implying that their vertices will typically have b.c. that is linear in the size of the object. In the remainder of the paper we consider increasing trees, which, although superficially similar to s.g. trees (in terms of their global g.f.), have a markedly more balanced shape.

4. Increasing trees

An *increasing* tree is a rooted, labelled tree in which the labels along any path away from the root form an increasing sequence. Unlike labelled s.g. trees, in which labels are assigned somewhat arbitrarily, the labels 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 tree's fringes. In some sense this makes the investigation of a vertex's b.c. more satisfying than it was in the case of s.g. trees, because we can examine the b.c. $b(l)$ of each labelled vertex l individually.

The fact that the vertices of all increasing trees are labelled according to the order in which they were attached to the tree leads to a general form for their g.f.'s, somewhat like the g.f. for s.g. trees given in equation (3). Let the weight function $\phi(u) = \sum_0^\infty \phi_i u^i$ once again encode a sequence of non-negative out-degree weights $\{\phi_i\}$, such that $\phi_0 \neq 0$ and $\phi_i > 0$ for some $i \geq 2$. Then, recalling that the act of removing the vertex with the lowest label from every object in a class is represented by the differentiated g.f. $y'(x)$, the g.f. of a class of increasing trees $\mathcal{T}$[6] satisfies

$$y'(x) = \sum_{T\in\mathcal{T}} \frac{w(T)}{|T|!} x^{|T|} = \phi(y(x)), \tag{12}$$

where $w(T)$ is again the product of the weights assigned to T's vertices. Due to the fact that the generating functions of increasing trees satisfy differential

[6] Note that all generating functions in this section are *exponential* generating functions.

equations rather than functional equations, it is not always possible to carry out general analyses quite as thoroughly as it is for s.g. trees. Apart from the broad special case of increasing trees that have polynomial weight functions, it is usually necessary to specify ϕ in order to complete an application of singularity analysis to a parameter of interest in an increasing tree class [19].

Fortunately, there are a few particularly important varieties of increasing trees that have been well studied—namely recursive, d-ary recursive, and plane-oriented recursive trees (PORTs), and these special cases, along with the polynomial varieties mentioned above, tend to share important structural characteristics. For example, they have a mean path length of $\Theta(n \log n)$ [19], and the expected distance from the root of a randomly chosen vertex in one of these classes is $\Theta(\log n)$. The expected height of a tree from one of the three 'recursive' cases mentioned above is also $\Theta(\log n)$ [8], as opposed to the $\Theta(\sqrt{n})$ of s.g. trees. The weight functions that give rise to recursive trees, d-ary recursive trees, and PORTs are $\phi(u) = \exp(u)$, $(1+u)^d$, and $(1-u)^{-1}$ respectively.

Recursive trees, d-ary recursive trees and PORTs can also be obtained by means of a growth process (see [20]): in the simplest case of recursive trees, the process starts with a single vertex labelled 1 (the root), and at each step, vertex n is attached to one of the $n-1$ previous vertices, selected uniformly at random. PORTs and d-ary recursive trees can be obtained by a similar process, where the probabilities are however not uniform, but depend on the outdegrees.

With nothing but the known balanced nature of increasing trees to go on, one can perhaps anticipate that the kth moment of the b.c. of the root vertex in an increasing tree of size n will be of order n^2. This is indeed the case. However, instead of deriving first the mean and then the higher-order moments of the root vertex as we did in Sections 2 and 3, we consider immediately the more general problem of the kth moment of the b.c. of vertex l,[7] when l is fixed while $n \to \infty$. Once this analysis is complete, we make use of a recent result of Fuchs [21] to show that a randomly chosen vertex in a tree from one of the three commonly considered classes typically has linear-order b.c. Then in the final section of the paper, we consider the maximum b.c. in recursive trees specifically, and the probability that the centroid obtains this maximum.

4.1. Moments of the betweenness centrality of a vertex with a given label

To estimate a parameter of the lth vertex in an increasing tree, one first needs to describe the tree relative to vertex l. We do this here by fixing the subtree containing vertices 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 vertex in the subtree. The g.f. that models trees in this way is $y^{(l)}(x)$, since it 'disregards' the subtree containing the first l vertices, 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 g.f. satisfies

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

[7] That is, the vertex with label l.

We have

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

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

We phrase the following theorem in a relatively general way, framed by the two assumptions that allow us to extract the desired moments using singularity analysis. In particular, it covers the cases of recursive (with $r = \lambda = 1$), d-ary recursive ($r = d$, $\lambda = d-1$), and plane-oriented recursive trees ($r = 1$, $\lambda = 2$).

Theorem 9. *Let $\mathcal{T}$ be a class of increasing trees that has a g.f. $y(x)$ for which the following two assumptions hold:*

1. *For positive r and λ:*

$$y'(x) = \phi(y(x)) = (1-\lambda x)^{-r/\lambda}.$$

2. *For all $c > 0$, there is a constant $K_c(r,\lambda)$ (possibly 0 for large enough c) such that*

$$\phi^{(c)}(y(x)) \sim K_c(r,\lambda)(1-\lambda x)^{-c+((c-1)r/\lambda)}$$

Then for fixed $l > 0$, the kth moment of the b.c. of the vertex with label l in $\mathcal{T}_n$ is of order n^{2k}. Specifically, for $k \geq 1$:

$$\mathbb{E}_n(b(l)^k) \sim n^{2k}\frac{\Gamma(r/\lambda)}{\lambda^{l-1}2^k}\sum_{m=0}^{k}\binom{k}{m}\frac{(-1)^m}{\Gamma(l+2m-1+(r/\lambda))}D_l(m)$$

for some constants $D_l(m)$ that depend on $\mathcal{T}$ (detailed below in equation (14)*, in square brackets).*

Proof. As in Section 2, the b.c. function $b(l)$ can be interpreted symbolically as the act of choosing vertices from the branches of l. Unfortunately, there is no analogue of Lemma 1 that holds for increasing trees, and instead of reducing $b(l)^k$ to a selection of vertices from exactly *two* branches, we will have to consider all possible selections if we wish to accurately derive the constant factors present in the moments of $b(l)^k$. To make this computation a bit simpler, we reduce $b(l)$ to a form involving a sum $\widetilde{b}(l) = \sum_i |T_i|^2$ over single branches, instead of branch pairs:

$$\begin{aligned} b(l)^k = \left(\sum_{i<j}|T_i||T_j|\right)^k &= \frac{1}{2^k}\left(\left(\sum_i |T_i|\right)^2 - \sum_i |T_i|^2\right)^k \\ &= \frac{1}{2^k}\left((n-1)^2 - \widetilde{b}(l)\right)^k \\ &= \frac{1}{2^k}\sum_{m=0}^{k}\binom{k}{m}(-1)^m\widetilde{b}(l)^m(n-1)^{2(k-m)}. \qquad (13)\end{aligned}$$

The new function $\widetilde{b}(l)^m$ counts selections (with replacement) of $2m$ vertices from any number of branches, with the restriction that vertices 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 g.f. $\sum_{\mathcal{T}} \widetilde{b}(l)^m x^{|T|}$ can be constructed in a piecewise fashion, per composition.

Let l's ancestral branch, represented by the g.f. $A_l(x) = y^{(l)}(x)/y'(x)$, appear in i of the factors of $\widetilde{b}(l)^m$, with the remaining factors being distributed among c descendent branches according to the composition $a_1 + \cdots + a_c = m - i$. If $\widehat{A}_{l,i}(x)$ denotes the g.f. of an ancestral branch from which i vertices have been selected (with replacement), and $\widehat{y}_j(x)$ symbolises the selection of j vertices from a descendent branch, then the cumulative g.f. of $\widetilde{b}(l)^m$ is

$$\begin{aligned}\sum_{T\in\mathcal{T}} \widetilde{b}(l)^m \frac{x^{|T|-l}}{(|T|-l)!} &= \sum_{i=0}^{m} \binom{m}{i} \widehat{A}_{l,2i}(x) \sum_{c=0}^{m-i} \frac{1}{c!} \phi^{(c)}(y(x)) \\ &\quad \times \sum_{g_c(m-i)} \binom{m-i}{a_1,\ldots,a_c} \widehat{y}_{2a_1}(x) \cdots \widehat{y}_{2a_c}(x),\end{aligned}$$

where $g_c(m)$ enumerates the compositions of m into c parts, and the contribution to the sum over c from $c = 0$ vanishes unless $i = m$, in which case $\phi^{(0)}(y(x)) = y'(x)$ (that is, $K_0(r, \lambda) = 1$) and the last sum is 1.

Under the assumptions of the theorem, $\widehat{y}_j(x) \sim x^j y^{(j)}(x)$ (see the proof of Lemma 1). Furthermore, we also have

$$\begin{aligned} y^{(l)}(x) &= (1-\lambda x)^{-(l-1)-(r/\lambda)} \cdot \prod_{t=0}^{l-2} (r + t\lambda), \\ \widehat{y}_j(x) &\sim (1-\lambda x)^{-(j-1)-(r/\lambda)} \cdot \lambda^{-j} \prod_{t=0}^{j-2} (r + t\lambda), \\ \widehat{A}_{l,i}(x) &\sim (1-\lambda x)^{-(l+i-1)} \cdot (l-1)^{\overline{i}} \prod_{t=0}^{l-2} (r + t\lambda), \end{aligned}$$

where $(\cdot)^{\overline{i}}$ denotes the ith rising factorial power, and the asymptotic expressions hold as $x \to 1/\lambda$. These approximations, along with the second assumption, can be used to reduce the g.f. to (note the implicit nesting of the sums over i, c, and

$g_c(m-i)$):

$$
\begin{aligned}
\sum_{T\in\mathcal{T}} \widetilde{b}(l)^m \frac{x^{|T|-l}}{(|T|-l)!} &\sim (1-\lambda x)^{-(2m+l-1+(r/\lambda))} \\
&\quad\times \left[\lambda^{-2m}\left(\prod_{t=0}^{l-2}(r+t\lambda)\right)\sum_{i=0}^{m}\lambda^{2i}\binom{m}{i}(l-1)^{\overline{2i}}\right. \\
&\quad\left.\times \sum_{c=0}^{m-i}\frac{K_c(r,\lambda)}{c!}\sum_{g_c(m-i)}\binom{m-i}{a_1,\ldots,a_c}\prod_{j=1}^{c}\prod_{t=0}^{2a_j-2}(r+t\lambda)\right] \\
&= (1-\lambda x)^{-(2m+l-1+(r/\lambda))}\cdot D_l(m).
\end{aligned}
\tag{14}
$$

Of course the quantity we really seek is the sum of $b(l)^k$ over trees of size n, of which there are

$$n!\,[x^n]y(x) \sim \lambda^{n-1}n!\frac{n^{-2+(r/\lambda)}}{\Gamma(r/\lambda)}.$$

We have, from equation (13):

$$
\begin{aligned}
\mathbb{E}_n\big(b(l)^k\big) &= \frac{(n-l)!}{n!\,[x^n]y(x)}[x^{n-l}]\sum_{T\in\mathcal{T}} b(l)^k \frac{x^{|T|-l}}{(|T|-l)!} \\
&\sim n^{2k}\frac{\Gamma(r/\lambda)}{\lambda^{l-1}2^k}\sum_{m=0}^{k}\binom{k}{m}\frac{(-1)^m}{\Gamma(l+2m-1+(r/\lambda))}D_l(m).
\end{aligned}
$$

□

A few example values that were obtained using Theorem 9 are given in Table 3. In addition, we can be slightly more explicit when considering specific classes of increasing trees: For *recursive trees*, with $r=\lambda=1$,

$$K_c(r,\lambda)=1 \quad\text{and}\quad \prod_{t=0}^{l-2}(r+t\lambda)=(l-1)!;$$

for *plane-oriented recursive trees*, with $r=1$ and $\lambda=2$,

$$K_c(r,\lambda)=c! \quad\text{and}\quad \prod_{t=0}^{l-2}(r+t\lambda)=(2l-3)!!;$$

and for *d-ary recursive trees*, with $r=d$ and $\lambda=d-1$,

$$K_c(r,\lambda)=(d)^{\underline{c}} \quad\text{and}\quad \prod_{t=0}^{l-2}(r+t\lambda)=\prod_{t=1}^{l-1}(td-(t-1)),$$

where $(\cdot)^{\underline{c}}$ denotes the cth falling factorial power.

Tree	r	λ	$\mathbb{E}_n(b(1))/n^2$	$\mathbb{V}_n(b(1))/n^4$	$\mathbb{E}_n(b(2))/n^2$
recursive	1	1	1/4	1/96	1/4
PORT	1	2	1/3	4/315	1/5
binary	2	1	1/6	1/180	1/4

Table 3: Asymptotic expressions for the means and variances of the b.c.'s of some labelled vertices in increasing trees.

Although it is not possible to obtain the limiting distribution of a vertex's b.c. by a construction similar to that of Section 2.2, we do see that all the moments of the scaled random variable $b(l)/n^2$ converge to a limit:

$$\lim_{n\to\infty} \mathbb{E}_n\left(\frac{b(l)^k}{n^{2k}}\right) = c_{k,l}.$$

Since the b.c. of any vertex is trivially bounded by $\binom{n-1}{2}$, we automatically obtain $c_{k,l} \leq 2^{-k}$, which means that the g.f. of the constants $c_{k,l}$ converges in a neighbourhood of 0 and represents a moment g.f. This implies, in view of Theorem C.2 of [7], that $b(l)/n^2$ converges weakly to a distribution that is characterised by the moments c_k:

Theorem 10. *Under the assumptions of Theorem 9, the distribution of* $b(l)/n^2$ *converges weakly to a limiting distribution.*

4.2. Limiting behaviour of the betweenness centrality of a random vertex

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

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

We note that the expected number of vertices with a given number of descendants—referred to as the *subtree size profile* of a tree—has been recently studied for the case of increasing trees. In fact, the expected proportion of vertices with m descendants in a tree of size n has been given explicitly for the most interesting classes of increasing trees [21, Theorem 4.1], and from these expressions, limiting distributions for b.c. follow directly.

Theorem 11. *The distribution of the linearly scaled b.c. of a randomly chosen vertex* v *in an increasing tree of size* n *converges weakly to a limiting distribution as* $n \to \infty$. *For* $0 < \varepsilon < 1$, *we have,*

1. *for* recursive trees*,*
$$\mathbb{P}_n(|b(v)/n - m| < \varepsilon) \xrightarrow[n\to\infty]{} \frac{1}{(m+1)(m+2)};$$
2. *for* plane-oriented recursive trees*,*
$$\mathbb{P}_n(|b(v)/n - m| < \varepsilon) \xrightarrow[n\to\infty]{} \frac{2}{(2m+1)(2m+3)};$$
3. *and for* d-ary recursive trees*,*
$$\mathbb{P}_n(|b(v)/n - m| < \varepsilon) \xrightarrow[n\to\infty]{} \frac{d(d-1)}{((d-1)m+2d-1)((d-1)m+d)}.$$

Proof. We consider e.g. recursive trees. The expected number of vertices with m descendants (m is fixed) in a tree of size n is

$$s_n(m) = \frac{n}{(m+1)(m+2)},$$

and scaling by n, we obtain a limiting proportion:

$$s(m) = \lim_{n\to\infty} \frac{s_n(m)}{n} = \frac{1}{(m+1)(m+2)}.$$

Since the $s(m)$ sum to 1, and $\lim_{n\to\infty} b(v)/n = m$ for a vertex v with m descendants, the result follows in the same way as Theorem 3. □

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

4.3. 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 in this section is to show that the vertex of maximal b.c. in a recursive tree is unlikely to have a large number as its label. Specifically, if Q_n is a random variable over the label of this vertex, we wish to show that as the size of the tree tends to infinity, the probability distribution $\mathbb{P}(Q_n = l)$ converges weakly to a discrete limiting distribution.

Intuitively, this concentration property should hold, because the vertex of maximal b.c. cannot have any particularly large branches—including its ancestral branch—and thus is likely to have a large number of descendants. The chance of this being true of a vertex with label l should decrease exponentially as l increases, so we would expect $\mathbb{P}(Q_n = l)$ to decrease exponentially as well.

To be more specific, we have the following result:

Lemma 6. *The probability $\mathbb{P}(Q_n \geq L)$ that the maximum b.c. is attained by a vertex whose label is at least L can be bounded above as follows:*

$$\mathbb{P}(Q_n \geq L) < 16\Big(\frac{L}{3}+1\Big)\Big(\frac{3}{4}\Big)^L.$$

Proof. First of all, we note that a vertex l which has $m_l - 1$ descendants cannot possibly have maximal b.c. if $m_l < n/4$. To see why this assertion holds, recall from Section 2.4 that a lower bound on the maximum b.c. in a tree is given by the lower bound on the centroid's b.c., $n(n-2)/4$. Then note that the b.c. of vertex l in a tree of size $n \geq 2$ is at most

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

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

Such small subtrees, however, become more likely as l is increased, and in fact $\mathbb{P}_n(m_l \geq n/4) < (l-1)(3/4)^{l-1}$. This is also simple to prove: firstly, let $l > 1$, and recall that the tree can be viewed as a sequence of l subtrees, each one rooted to one of the first l vertices. The number of sequences in which the lth subtree is of size m_l is

$$\binom{n-l}{m_l-1}(m_l-1)!\binom{n-m_l-1}{l-2}(n-l-m_l+1)!,$$

because the number of ways to organise the remaining subtree sizes according to the composition $m_1 + \cdots + m_{l-1} = n - m_l$ is

$$\binom{n-l-m_l+1}{m_1-1,\ldots,m_{l-1}-1}(m_1-1)!\cdots(m_{l-1}-1)! = (n-l-m_l+1)!,$$

which is independent of the composition—of which there are $\binom{n-m_l-1}{l-2}$. Since there are $\binom{n-1}{l-1}(n-l)!$ sequences overall, the probability of l's subtree being of size m is

$$\mathbb{P}(m_l = m) = \binom{n-m-1}{l-2}\Big/\binom{n-1}{l-1}.$$

The result follows with a bit of algebra:

$$\begin{aligned}\mathbb{P}(m_l \geq n/4) &= \left[\binom{\lfloor 3n/4\rfloor - 1}{l-2} + \binom{\lfloor 3n/4\rfloor - 2}{l-2} + \cdots + \binom{l-2}{l-2}\right] \Big/ \binom{n-1}{l-1} \\ &< (\lfloor 3n/4\rfloor - l + 2)\binom{\lfloor 3n/4\rfloor}{l-2} \Big/ \binom{n-1}{l-1} \\ &= (l-1)\frac{(\lfloor 3n/4\rfloor)_{l-1}}{(n-1)_{l-1}} \\ &< (l-1)\left(\frac{\lfloor 3n/4\rfloor}{n}\right)^{l-1} \\ &\leq (l-1)\left(\frac{3}{4}\right)^{l-1}.\end{aligned}$$

Thus we have

$$\mathbb{P}(Q_n = l) \leq \mathbb{P}(m_l \geq n/4) < (l-1)\left(\frac{3}{4}\right)^{l-1},$$

and a bound on the tail probabilities follows immediately, for any n:

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

This completes the proof of the lemma. □

The upper bound on $\mathbb{P}_n(Q_n \geq L)$ is important firstly because it is *independent of* n, which means that regardless of the size of the tree, the probability of the maximum b.c. 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, $\mathbb{P}_n(Q_n < L)$ is positively bounded from below (independently of n).

Now we can follow a similar approach as in the proof of Theorem 5, and in fact the technical details are somewhat simpler. Before we formulate and prove our final result, let us consider the limit distribution of the root b.c. established in Theorem 10. A recursive tree decomposes naturally into the first root branch (that is, the branch containing label 2), and the rest. The number of trees of size n in which this branch has n_1 vertices 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 the size, each of the two pieces is again a uniformly random recursive tree.

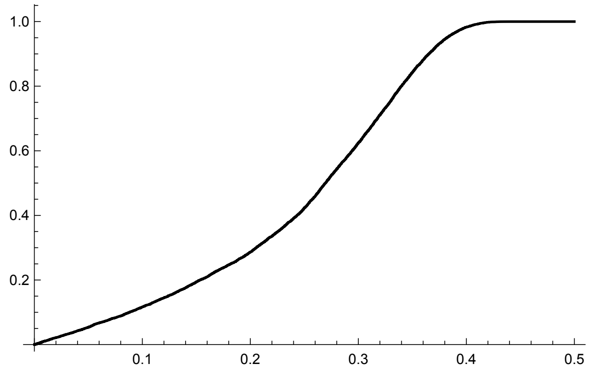

Figure 3: The cumulative distribution function of the limiting distribution of root b.c. in recursive trees.

If we let X be a random variable representing the limiting distribution of the root b.c., 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 for a plot of the distribution function).

A more general decomposition will yield the following theorem:

Theorem 12. *The maximum b.c. of a random recursive tree of size n, divided by n^2, converges weakly to a limiting distribution. The probability that the maximum b.c. is attained by the centroid tends to a positive constant, and the random variable Q_n giving the label of the vertex with maximum b.c. converges to a discrete limiting distribution.*

Proof. Instead of the maximum b.c. of an arbitrary vertex, we only consider the maximum among the first l vertices. By virtue of Lemma 6, we can then let l go to infinity.

If we fix the tree formed by the first l vertices (for which there are only finitely many possibilities), it decomposes naturally into l disjoint recursive trees. Let $n_1, n_2, \ldots, n_l$ be the sizes of these trees (n_j being the order of the tree rooted at j). Given all these sizes, there are

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

possible trees. This is independent of the values of $n_1, n_2, \ldots, n_l$ 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 b.c.'s, 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 b.c.'s of vertices $1, 2, \ldots, l$ are simple functionals of $U_1, U_2, \ldots, U_l$ and $X_1, X_2, \ldots, X_l$ (also depending on the shape of the tree formed by vertices $1, 2, \ldots, l$), so the theorem follows. □

With the help of a numerical simulation, we find that the expected maximum b.c. in a recursive tree is asymptotically equal to $0.35n^2$, and that the probability of the centroid vertex also being a vertex of maximal b.c. is roughly 0.87. In addition, it appears that the expected label of the vertex of maximal b.c. (breaking ties in favour of the vertex 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.[8]

Acknowledgment

The authors would like to thank an anonymous referee for thoroughly reading the manuscript and providing many useful comments.

[1] L. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.

[2] M. E. J. Newman, Networks: An Introduction, 1st Edition, Oxford University Press, 2010.

[3] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.

[4] S. Gago, J. C. Hurajová, T. Madaras, Betweenness centrality in graphs, in: Quantitative graph theory, Discrete Math. Appl. (Boca Raton), CRC Press, Boca Raton, FL, 2015, pp. 233–257.

[5] K.-I. Goh, E. Oh, H. Jeong, B. Kahng, D. Kim, Classification of scale-free networks, Proc. Natl. Acad. Sci. USA 99 (2002) 12583–12588.

[6] A. Meir, J. W. Moon, On the altitude of nodes in random trees, Canad. J. Math. 30 (1978) 997–1015.

[7] P. Flajolet, R. Sedgewick, Analytic Combinatorics, 1st Edition, Cambridge University Press, 2009.

[8] M. Drmota, Random Trees: An Interplay Between Combinatorics and Probability, 1st Edition, Springer, 2009.

[8] We also recover some interesting results of Moon [22], 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.

[9] A. Meir, J. W. Moon, On centroid branches of trees from certain families, Discrete Math. 250 (1–3) (2002) 153–170.

[10] A. Meir, J. W. Moon, On major and minor branches of rooted trees, Canad. J. Math. 39 (1987) 673–693.

[11] D. Aldous, The continuum random tree. I, Ann. Probab. 19 (1) (1991) 1–28.

[12] C. Jordan, Sur les assemblages des lignes, J. Reine Angew. Math. 70 (1869) 185–190.

[13] F. Harary, Graph theory, Addison-Wesley Publishing Co., Reading, Mass.-Menlo Park, Calif.-London, 1969.

[14] D. E. Knuth, The Art of Computer Programming, volume 1: Fundamental Algorithms, 3rd Edition, Addison-Wesley, 1997.

[15] D. Aldous, Recursive self-similarity for random trees, random triangulations and Brownian excursion, Ann. Probab. 22 (2) (1994) 527–545.

[16] D. Aldous, The continuum random tree. II. An overview, in: M. T. Barlow, N. H. Bingham (Eds.), Stochastic analysis, Cambridge University Press, 1991, pp. 23–70.

[17] D. Aldous, The continuum random tree. III, Ann. Probab. 21 (1) (1993) 248–289.

[18] M. Drmota, É. Fusy, M. Kang, V. Kraus, J. Rué, Asymptotic study of subcritical graph classes, SIAM J. Discrete Math. 25 (4) (2011) 1615–1651.

[19] F. Bergeron, P. Flajolet, B. Salvy, Varieties of increasing trees, in: J.-C. Raoult (Ed.), Lecture Notes in Computer Science, Vol. 581, Springer, 1992, pp. 24–48.

[20] A. Panholzer, H. Prodinger, Level of nodes in increasing trees revisited, Random Structures Algorithms 31 (2) (2007) 203–226.

[21] M. Fuchs, Limit theorems for subtree size profiles of increasing trees, Comb. Prob. Comp. 21 (3) (2012) 412–441.

[22] J. W. Moon, On the centroid of recursive trees, Austral. J. Comb. 25 (2002) 211–219.