# Betweenness Centrality in Random Trees*

Kevin Durant and Stephan Wagner†

**Abstract**

Betweenness centrality is a quantity that is frequently used in graph theory 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. The latter is the same for all simply generated families, and can be obtained by means of the continuum random tree. In this context, 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, where the situation is quite different: here, the root betweenness centrality is of quadratic order. Again, we also have a limiting distribution upon suitable normalisation.

## 1 Introduction

In graph theory, one of the variety of centrality measures used to assign importance to vertices is betweenness centrality, which gives an indication of the number of shortest paths a vertex appears on. For a given vertex $r$ in an undirected graph $G = (V, E)$, the betweenness centrality (BC) is defined as a sum over subsets $\{v, w\}$ of $V \setminus \{r\}$, counting for each pair of vertices the fraction of undirected shortest paths between them that pass through $r$:

$$g(r) = \sum_{\{v,w\}} g_{vw}(r)\,,$$

where $0 \leq g_{vw}(r) \leq 1$. If $G$ is a tree, paths are unique (so $g_{vw}(r)$ is either 0 or 1), and $g(r)$ counts the number of shortest paths through $r$. For a graph of $n$ vertices, the betweenness centrality of any vertex is bounded by the number of possible vertex pairs, and is thus at most quadratic in order.

The origins of betweenness centrality in sociology date back almost 40 years [8]. It is often used to compare the connectivity of vertices in large complex and social networks: a good general reference is Newman's book [17, Section 7.7]. In particular, betweenness centrality can be used to identify vertices (or edges) through which much information flows, and this property was leveraged to create one of the first useful vertex clustering algorithms [10]. See also [11] for a study of the betweenness centrality in real-world networks.

There are, of course, more mathematical considerations of the topic as well [9]. Here, we restrict the discussion to tree-like graphs, which are more amenable to analytic methods, and yet still particularly valuable, especially in the analysis of algorithms.

In the following sections we derive, mostly with the aid of analytic combinatorics and singularity analysis, various moments and limiting distributions of betweenness centrality in simply generated trees, subcritical graphs, and recursive trees (Sections 2, 3, and 4 respectively). We obtain moments for the betweenness centrality of the root

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

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

vertex in all three cases, and limiting distributions when dealing with the two classes of trees. Simply generated trees and subcritical graphs—similar in that their generating functions permit square-root expansions—typically possess an unbalanced shape, so that in both cases the betweenness centrality of the root vertex exhibits a mean and variance of orders $n^{3/2}$ and $n^{7/2}$ respectively. Recursive trees, on the other hand, are of a more balanced nature, yielding a mean and variance of orders $n^2$ and $n^4$ respectively. Limiting distributions for root betweenness centrality (that is, the betweenness centrality of the root vertex) are obtained by considering the linearly scaled function $g(r)/n$ in the case of simply generated trees, and $g(r)/n^2$ in the case of recursive trees.

Furthermore, for simply generated trees, we consider vertices other than the root, and obtain a limiting distribution for the betweenness centrality of a randomly chosen vertex. Although the betweenness centrality of a vertex is typically of linear order, the maximum betweenness centrality of the vertices in such trees is of quadratic order; we derive its limiting distribution as well.

The concept of the centroid of a tree will play a major role in this context: a centroid vertex is a vertex for which the sum (or average) of all distances to the other vertices attains its minimum. Such a vertex is therefore "most central" in another sense. For trees, with high probability there is only one centroid vertex, and we will also show that the centroid has positive limiting probability of also being the vertex with maximum betweenness centrality.

## 2 Simply generated trees

To begin with, we consider simply generated families of trees (also known as Galton-Watson trees; for an overview, see Drmota's book [5, Section 1.2.7]). These are families of labelled/unlabelled rooted trees whose vertex out-degrees can be restricted to a certain set $\Omega$, and assigned nonnegative real weights $\{\, \phi_\omega \mid \omega \in \Omega \,\}$. The classes of binary, plane (Catalan), and rooted labelled trees are all simply generated. One of the many known properties of simply generated trees (see [5, Chapters 3 and 4]) is that they tend to contain a few disproportionately large branches, leading to a tall, unbalanced structure. Whereas a perfectly balanced tree of size $n$ would have a height of order $\log n$, the mean height of a simply generated tree is of order $\sqrt{n}$, with the $k$th moment being of order $n^{k/2}$. The sum of the distances of all vertices from the root—known as the path length of a rooted tree—has mean order $n^{3/2}$ in simply generated trees, implying that a random vertex in the tree is, on average, at a distance of order $\sqrt{n}$ from the root. In fact, the cumulative generating function of path length for simply generated trees is markedly similar to that of root betweenness centrality, which we shall derive shortly.

From the symbolic viewpoint of analytic combinatorics, simply generated trees can be represented as a collection of branches attached to a root vertex, and because of this, a generating function for the sequence $(y_n)$, where $y_n$ is the (weighted) number of trees of size $n$ in a simply generated family $\mathcal{T}$, takes on the general form $y(z) = z\phi(y(z))$. Here $\phi(u) = \sum \phi_\omega u^\omega$ encodes the possible out-degrees of the trees' vertices, and for technical reasons, is assumed to be analytic at $u = 0$ and non-periodic.

Under the above conditions, there is a unique solution $\tau$ of the equation $\phi(\tau) - \tau\phi'(\tau) = 0$ within the radius of convergence of $\phi$, and by considering $y(z)$ as the inverse of the function $z = y/\phi(y)$, it can be shown [5, 7, 14] that the generating function $y(z)$ has a singularity at the point $z = \rho$, determined by $\rho = \tau/\phi(\tau) = 1/\phi'(\tau)$. Furthermore, $y(z)$ satisfies a square-root expansion about this singularity:

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

in which $y(\rho) = \tau$ and

$$\gamma = \sqrt{\frac{2\phi(\tau)}{\phi''(\tau)}}.$$

This implies that for a simply generated family of trees $\mathcal{T}$, the number of trees of size $n$ is of order $\rho^{-n}n^{-3/2}$:

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

Turning to the question of betweenness centrality in simply generated trees, we consider in particular the root vertex. Since the betweenness centrality of a vertex counts the shortest paths that pass through it, the betweenness centrality of the root is given by the number of ways in which unordered pairs of vertices can be chosen from different branches of the tree. For a tree $T \in \mathcal{T}$ with root $r$, we adopt the implicit notation $g(T) \equiv g(r)$, and have the equality

$$g(T) = \sum_{i<j} |T_i||T_j|\,, \tag{2.1}$$

in which $T_1, T_2, \ldots, T_\omega$ are the branches of $T$. Using this expression, we wish to obtain information about the cumulative generating functions $H_k(z)$ of the powers of root betweenness centrality for the class $\mathcal{T}$:

$$H_k(z) = \sum_{T\in\mathcal{T}} g(T)^k z^{|T|}\,.$$

It is from these cumulative generating functions (CGFs) that moments will be derived.

A key observation that follows from (2.1) will yield the limiting distribution: if one of the branches (without loss of generality, $T_1$) is "large", while the others combined contain a fixed number $k$ of vertices, the root betweenness centrality is dominated by paths between the large branch and the other branches: if the branch sizes are $n_1 = n-k-1, n_2, n_3, \ldots, n_\omega$, so that $n_2+n_3+\cdots+n_\omega = k$ is fixed while the size $n$ of the tree tends to infinity, then we have

$$\begin{aligned} g(T) &= (n-k-1)\sum_{i=2}^{\omega} n_i + \sum_{1<i<j} n_i n_j \\ &= nk + O(k^2)\,. \end{aligned} \tag{2.2}$$

As it turns out, this observation applies to random simply generated trees with high probability.

**2.1 Average root betweenness centrality.** Combinatorially, $H_1(z)$—the CGF of $g(T)$—is equivalent to the generating function of the class of $\mathcal{T}$-trees with two vertices from distinct branches distinguished. This class can be constructed symbolically. Firstly, note that the act of selecting (distinguishing) vertices is encoded by the pointing operator $\Theta\mathcal{T}$; and secondly, that this operation can be restricted to certain substructures (in this case, branches) by making use of a variation of the substitution operator $\circ$.

DEFINITION 2.1. *The partial substitution of $k$ of the $\mathcal{T}$-structures in an admissible construction $\mathcal{A}(\mathcal{T})$ with $\mathcal{P}$-structures is denoted by the operator $\circ_k$, so that if $\mathcal{S}$ is the resulting class, we write*

$$\mathcal{S} = \mathcal{A}(\mathcal{T}) \circ_k \mathcal{P}\,.$$

*The generating function $S(z)$ of $\mathcal{S}$ is given by*

$$\begin{aligned} S(z) &= P(z)^k \sum_{n\geq k} a_n \binom{n}{k} T(z)^{n-k} \\ &= \frac{1}{k!} P(z)^k A^{(k)}(T(z))\,, \end{aligned}$$

*where $P(z)$, $A(z) = \sum a_n z^n$, and $T(z)$ are the generating functions of $\mathcal{P}$, $\mathcal{A}$, and $\mathcal{T}$ respectively.*

The class of trees $\mathcal{H}$ which have two vertices from distinct branches distinguished can now be defined symbolically as

$$\mathcal{H} = \mathcal{Z} \times (\textsc{Seq}_\Omega(\mathcal{T}) \circ_2 \Theta\mathcal{T})\,,$$

which implies that the generating function of $\mathcal{H}$—that is, the CGF $H_1(z)$—satisfies

$$H_1(z) = \frac{z^3}{2}\, y'(z)^2\, \phi''(y(z))\,,$$

since the generating function of the pointed class $\Theta\mathcal{T}$ is $zy'(z)$. For comparison, we remark that the cumulative generating function of path length in simply generated trees is $z^2\, y'(z)^2\, \phi'(y(z))/\phi(y(z))$.

To reduce $H_1(z)$ to its asymptotic form, we make use of the square-root expansion of $y(z)$ at $z = \rho$, and recall that $\phi(u)$ is analytic at $u = \tau$:

$$\begin{aligned} H_1(z) &\sim \frac{\rho}{2}\left(\frac{\gamma}{2}\right)^2 \phi''(\tau)\, \left(1-\frac{z}{\rho}\right)^{-1} \\ &= \frac{\tau}{4}\, \left(1-\frac{z}{\rho}\right)^{-1}\,. \end{aligned}$$

The average root betweenness centrality of a tree of $n$ vertices is then the quotient of the $n$th coefficient of $H_1(z)$ and $y_n$.

THEOREM 2.1. *Let $\mathcal{T}$ be a family of simply generated trees. Then the root vertices of trees $\mathcal{T}_n$ of size $n$ have an average betweenness centrality of order $n^{3/2}$. In particular:*

$$\mathbb{E}_{\mathcal{T}_n}(g(T)) = \frac{[z^n]\, H_1(z)}{[z^n]\, y(z)} \sim \frac{\gamma^{-1}\tau}{2}\sqrt{\pi n^3}\,.$$

Here, for example, are the asymptotic expressions of root betweenness centrality for a few common simply generated trees:

Binary trees ($\phi(u) = (1+u)^2$): $\sqrt{\pi n^3/16}\,.$

Catalan trees ($\phi(u) = (1-u)^{-1}$): $\sqrt{\pi n^3/4}\,.$

Labelled trees ($\phi(u) = e^u$): $\sqrt{\pi n^3/8}\,.$

**2.2 Higher moments of root betweenness centrality.** A similar procedure can be followed to determine the second- and higher-order moments of root betweenness centrality, as long as one takes some care when considering $g(T)^k$ in the CGF $H_k(z)$. For example, the square of $g(T)$ is

$$\begin{aligned} g(T)^2 &= \left(\sum_{a<b} |T_a||T_b|\right)^2 \\ &= \sum_{a<b} |T_a|^2|T_b|^2 + 2\sum_{a}\sum_{\substack{b<c\\ \neq a}} |T_a|^2|T_b||T_c| \\ &\quad + 6\sum_{a<b<c<d} |T_a||T_b||T_c||T_d|\,, \end{aligned}$$

where $T_1, \ldots, T_\omega$ are again the branches of $T$. The sums enumerate the ways in which the exponents (in this case either 0, 1, or 2) can be configured amongst the branches. The second CGF $H_2(z)$ can thus be split into a sum of three simpler generating functions—each of which can be interpreted symbolically. The first counts trees with four distinguished vertices divided equally between two distinct branches; the second, trees with two distinguished vertices in one branch, and two more chosen from two other branches; and the third counts trees with four pointed branches. In general, the cumulative generating function of $g(T)^k$ can be split into smaller CGFs according to the different ways in which $2k$ vertices can be selected from a collection of distinct branches, with the additional restriction that no more than $k$ vertices may be chosen from a single branch (since the exponent of any $|T_i|^\alpha$ cannot exceed $\alpha = k$).

Each of these generating functions is amenable to singularity analysis if one only notes that the selection of $j$ vertices from a single branch can be encoded by the substitution of a normal branch $\mathcal{T}$ with a $j$-pointed branch $\Theta^{(j)}\mathcal{T}$. The effect that substitutions of this form have on the relevant generating function is quantified by the following lemma.

LEMMA 2.1. *Let $\mathcal{T}$ be a tree-like class whose generating function $T(z)$ permits a square-root expansion about its singularity $z = \rho$. Then the partial substitution of $m$ branches of each tree with $m$ pointed branches (each of which may possibly distinguish multiple vertices), which in total contain $d$ distinguished vertices, yields a generating function with a singularity of the form $(1-(z/\rho))^{-d+m/2}$.*

*Proof.* We consider here a family of simply generated trees $\mathcal{T}$ specifically, although the lemma holds for subcritical graph classes (see the following section) as well. If $y(z)$ is the generating function of $\mathcal{T}$, then the altered generating function $\hat{y}(z)$—obtained by the above partial substitution—is given by

$$\hat{y}(z) = \frac{z}{m!}\left[\prod_{i=1}^{m} P_{j_i}(z)\right]\phi^{(m)}(y(z))\,,$$

in which $P_{j_i}(z)$ is the generating function of the $i$th pointed branch $\Theta^{(j_i)}\mathcal{T}$, from which $j_i$ vertices have been selected. It is these branch generating functions that affect the overall asymptotic order, since a $j$-pointed branch has a generating function $P_j(z)$ that satisfies

$$P_j(z) = \sum_{l=1}^{j} \left\{ {j \atop l} \right\} z^l y^{(l)}(z)\,,$$

where $\left\{ {j \atop l} \right\}$ denotes the Stirling numbers of the second kind. Since $y(z)$ permits a square-root expansion, $y^{(l)}(z)$ is of order $(1-(z/\rho))^{-l+1/2}$, and $P_j(z)$ is dominated by a single term:

$$P_j(z) \sim z^j y^{(j)}(z) \sim \frac{\gamma(2j-2)!}{2^{2j-1}(j-1)!}\left(1-\frac{z}{\rho}\right)^{-j+1/2}\,.$$

Lemma 2.1 then follows by noting that $\phi(u)$ is analytic at $u = \tau$ and $\sum_i j_i = d$.

The implication here is that when selecting vertices, the singularity order of the resulting generating function decreases with the number of affected branches. Recalling that $g(T)^k$ is split according to such selections, it follows that the CGF of $g(T)^k$ is dominated by the term which counts the ways in which $2k$ vertices can be chosen from *two* branches:

$$H_k(z) \sim \sum_{T\in\mathcal{T}} \left( \sum_{a<b} |T_a|^k |T_b|^k \right) z^{|T|} .$$

This simplification in turn simplifies the derivation of the asymptotic behaviour of all higher-order moments.

THEOREM 2.2. *The kth moment of the betweenness centrality of the root vertex in a family of simply generated trees $\mathcal{T}_n$ of size n is of order $n^{2k-1/2}$, and satisfies, for $k \geq 1$,*

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

*Proof.* For the $k$th moment, we have the CGF

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

so that $\mathbb{E}_{\mathcal{T}_n}(g(T)^k)$ can be derived in the same way as that in which the mean was calculated. In particular, for the second moment:

$$\mathbb{E}_{\mathcal{T}_n}(g(T)^2) = \frac{[z^n]\, H_2(z)}{[z^n]\, y(z)} \sim \frac{\gamma^{-1}\tau}{16} \sqrt{\pi n^7} .$$

The variance of root betweenness centrality in simply generated trees is thus of the same order as the second moment, namely $n^{7/2}$. The fact that the $k$th moment is of order $n^{2k-1/2}$ can be explained intuitively by the phenomenon that the rather unlikely event (whose probability is only of order $n^{-1/2}$) of having two large branches with a number of vertices linear in $n$ dominates the asymptotic behaviour.

**2.3 Limiting distribution of root betweenness centrality.** To obtain a limiting distribution for $g(T)$, one must work with a scaled version of the function; it turns out to be sufficient to scale $g(T)$ linearly, and regard instead $g(T)/n$. This is a consequence of the unbalanced nature of simply generated trees—although trees with root betweenness centrality of order $n^2$ influence the measure's moments, these balanced trees are increasingly rare in the limit $n \to \infty$, and as such, the limiting distribution can be described fully using only trees of linear root betweenness centrality.

To prove this, we define subclasses $\mathcal{L}_k \subseteq \mathcal{T}$ of trees which contain one large branch, and a few small branches of total size $k$. Formally, $(\mathcal{L}_k)_n$, the class of trees in $\mathcal{L}_k$ of size $n$, simply consists of all trees of $\mathcal{T}_n$ with one distinguished branch of size $n-k-1$. Note that a tree can thus a priori belong to more than one such class. For fixed $k$, each of these subclasses engenders predictable, linear root betweenness centrality, and, in the limit, the classes $(\mathcal{L}_k)_n$ together describe $\mathcal{T}_n$.

THEOREM 2.3. *For simply generated trees $\mathcal{T}_n \subset \mathcal{T}$ of size n, the linearly scaled betweenness centrality of the root vertex, $g(T)/n$, converges weakly as $n \to \infty$ to the discrete distribution defined by*

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

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

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

*Proof.* Firstly, recall our earlier observation that the root betweenness centrality of a tree $T \in (\mathcal{L}_k)_n$ is of linear order for large $n$: if $T$ has a branch of size $n-k-1$, while the other branches contain $k$ vertices for some fixed $k$, then by (2.2) we have

$$g(T) = nk + O(k^2).$$

Secondly, note that for large enough $n$, any two subclasses $\mathcal{L}_k$ and $\mathcal{L}_l$ $(k \neq l)$ are disjoint: we have $(\mathcal{L}_k)_n \cap (\mathcal{L}_l)_n = \emptyset$ 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 a constant probability $p_k$ as $n$ grows, and that the sum of these limiting probability

masses is 1. Begin by considering the generating function $L_k(z)$ of a subclass $\mathcal{L}_k$, which allows for a single growing branch, and counts the $\omega$ possible points of attachment for this branch, as well as the $[z^k]\, y(z)^{\omega-1}$ configurations of the remaining (non-root) vertices:

$$\begin{aligned} L_k(z) &= z^{k+1} y(z) \sum_{\omega\in\Omega} \omega\, \phi_\omega\, [z^k]\, y(z)^{\omega-1} \\ &= z^{k+1} y(z)\, [z^k]\, \phi'(y(z))\,. \end{aligned}$$

Note that the maximum root degree of a tree in $\mathcal{L}_k$ is $\omega = k+1$, accounted for by the fact that $[z^k]\, y(z)^{\omega-1} = 0$ whenever $\omega - 1 > k$. From this generating function, the limiting probability $\lim_{n\to\infty} \mathbb{P}_{\mathcal{T}_n}(T \in (\mathcal{L}_k)_n) = p_k$ is

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

The sum of these constants is indeed 1, so they describe a probability distribution:

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

and the limiting distribution of $g(T)$ can be fully characterised using only the limit behaviour of the subclasses $\mathcal{L}_k$.

Again, we can provide concrete examples (valid for fixed $k \geq 0$):

Binary trees: $p_k = 2^{-(2k+1)} \frac{1}{k+1} \binom{2k}{k}$.

Catalan trees: $p_k = 4^{-(k+1)} \frac{1}{k+2} \binom{2k+2}{k+1}$.

Labelled trees: $p_k = e^{-(k+1)} \frac{(k+1)^{k-1}}{k!}$.

It is also worth pointing out that

$$p_k \sim \frac{\gamma^{-1}\tau}{\sqrt{\pi k^3}}$$

as $k \to \infty$ for any simply generated family of trees.

**2.4 Limiting distribution of the betweenness centrality of a randomly chosen vertex.** The previous sections characterise the moments and typical behaviour of *root* betweenness centrality in simply generated trees, and in addition, for these families of trees, the behaviour of a randomly chosen vertex can be described as well. In the exceptional case of labelled trees (with $\phi(u) = e^u$), all of the preceding results hold for non-root vertices as well, because there is a natural mapping between unrooted and rooted labelled trees: each unrooted tree of size $n$ engenders $n$ rooted trees—one for each label. This implies that by considering all of the root vertices in a class of rooted labelled trees $\mathcal{T}_n$, one is effectively considering each of the individual vertices in the relevant class of unrooted trees. In general, however, this mapping does not hold for other simply generated families. Still, by making use of a construction quite similar to that of the previous section, we can show that a randomly chosen vertex from a simply generated tree usually has betweenness centrality of linear order.

THEOREM 2.4. *The linearly scaled betweenness centrality $g(v)/n$ of a randomly chosen vertex $v$ in a simply generated tree $T \in \mathcal{T}_n$ of size $n$ converges weakly as $n \to \infty$ to the discrete distribution defined by*

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

*Specifically, if $\mathcal{T}'_n$ is the class of trees of size $n$ with a distinguished vertex $v$, and $g(T') \equiv g(v)$ in $T' \in \mathcal{T}'_n$, then for fixed $k$ and every $0 < \varepsilon < 1$:*

$$\mathbb{P}_{\mathcal{T}'_n}(\left|g(T')/n - k\right| < \varepsilon) \xrightarrow[n\to\infty]{} q_k\,.$$

*Proof.* Following a course similar to that of the proof of Theorem 2.3, we construct subclasses of trees of size $n$ which have distinguished vertices of linear-order betweenness centrality, and show that in the limit $n \to \infty$, these classes describe $\mathcal{T}'_n$ fully.

As usual, $\mathcal{T}$ denotes the simply generated family, and has the generating function $y(z)$. Consider a distinguished vertex $v$, which has one large branch of size $n-k-1$ and a collection of smaller branches of total size $k$. Symbolically, we can consider $v$ as a leaf vertex of a rooted tree of size $n-k$, to which a forest of size $k$ has been grafted. If $(\mathcal{L}_k)_n$ is the class of trees built around such distinguished vertices, then the generating function $L_k(z)$ that counts these trees must account for the $[z^k]\, \phi(y(z))$

configurations of the smaller branches, as well as the possible ways to select a leaf from a tree of size $n-k$. The generating function that counts distinguished leaf vertices can be derived from the bivariate generating function $y(z,u)$ which marks leaves with an auxiliary variable $u$, by taking the partial derivative with respect to $u$, and then setting $u=1$. Accordingly, we have for a class $\mathcal{L}_k$:

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

in which $y(z,u) = z\phi(y(z,u)) + (u-1)\phi_0 z$. The appearance of $\phi_0^{-1}$ 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 $\phi(y(z))$.

As in the proof of Theorem 2.3, the distinguished vertex of a tree $T' \in (\mathcal{L}_k)_n$ has betweenness centrality of order $nk + O(k^2)$, and any two such classes $(\mathcal{L}_k)_n$ 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 positive probability $q_k$ of belonging to $(\mathcal{L}_k)_n$, we must express $L_k(z)$ asymptotically. Quickly note that, by differentiating $y(z) = z\phi(y(z))$:

$$\big(1 - z\phi'(y(z))\big)^{-1} = zy'(z)y(z)^{-1}.$$

With this in mind, it follows that

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

with which the limiting probability $q_k$ is found:

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

And finally, we see that $\sum_{k\geq 0} q_k = y(\rho)/\tau = 1$.

**2.5 Maximum betweenness centrality and the centroid of a tree.** We have seen in the previous sections that the average root betweenness centrality of a simply generated tree is of order $n^{3/2}$. On the other hand, if we choose, for example, a random vertex of a random labelled tree (which amounts to picking a random rooted labelled tree), the "typical" betweenness centrality is only of linear order. By contrast, the maximum betweenness centrality is always of quadratic order, as we will show now. In fact, vertices with quadratic betweenness centrality (which are comparatively rare) dominate the moments, as mentioned earlier. A trivial lower bound for the maximum betweenness centrality 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 the branches that stem from it contain more than half of the vertices of the tree.

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 [13], see for instance [12, Chapter 4].

It is easy to see that the betweenness centrality of a vertex decreases when vertices are transferred from one branch to another branch of greater or equal size. Therefore, the smallest possible betweenness centrality 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 betweenness centrality is $\lfloor (n-1)^2/4 \rfloor \geq (n^2-2n)/4$. This provides a lower bound for the maximum betweenness centrality, as mentioned earlier.

Even though a centroid vertex must necessarily have fairly large betweenness centrality, 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 betweenness centrality of about $n^2/4$ in this case, while the centre of the star has about $n^2/3$.

In spite of this counterexample, the centroid will play a major role in our analysis of the 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 betweenness centrality of a random tree, rescaled by a factor $n^{-2}$, has a limiting distribution. This limiting distribution, unlike the distribution of the betweenness centrality of a randomly chosen vertex, is even independent of the specific class of simply generated trees.

Let us first review a connection to random triangulations of the circle that is due to Aldous [4], as well as some results of Meir and Moon [16] on centroid branches.

The limit object of simply generated trees is the celebrated continuum random tree (see the work of Aldous [1–3] and [5, 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 (linear order); the lengths of the three arcs defined by a triangle correspond to the sizes of the branches. It is important in this context that vertices with more than four large branches (order greater than $n^{1-\epsilon}$ for some small $\epsilon > 0$, say) do not occur asymptotically almost surely, so vertices with three large branches capture all large-scale branching. 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 betweenness centrality of the corresponding branching vertex. The maximum betweenness centrality corresponds to the maximum weight of a triangle, and the distribution of this maximum is the limiting distribution of the betweenness centrality. 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 [16] showed, among other things, that the average betweenness centrality of the centroid of a random tree from a simply generated class 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 betweenness centrality, which is actually not far from the truth: 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 betweenness centrality converges to a constant whose numerical value is 0.621. The limiting distribution function of the normalised maximum betweenness centrality is shown in Figure 1. 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. This is done by first generating the centroid triangle according to the known density $(12\pi)^{-1}(x_1x_2x_3)^{-3/2}$ on the set of all triples $(x_1, x_2, x_3)$ such that $0 < x_1, x_2, x_3 < 1/2$ and $x_1 + x_2 + x_3 = 1$ (see [4]). Then one repeats this recursively for all branches. Once it is no longer possible to fit further triangles with a greater weight than the current maximum, one can stop the process.

If all arcs of the centroid triangle have length less than $4/9$, its weight is greater than $8/27$, while the weights of any other triangle is at most $8/27$. Since this happens with positive probability, we have an immediate proof of the fact that the centroid has maximum betweenness centrality with positive limiting probability. A similar argument shows that the probability is strictly less than 1, and one can also show in the same fashion that the density of the distribution function shown in Figure 1 has the interval $[1/4, 1/3]$ as its support.

Let us summarise our findings in the following theorem:

THEOREM 2.5. *The maximum betweenness centrality of a random tree with $n$ vertices from a simply generated family, divided by $n^2$, converges weakly to a limiting distribution that is independent*

*of the specific family of trees. The probability that the maximum betweenness centrality is attained by the centroid tends to a positive constant that is also independent of the specific family.*

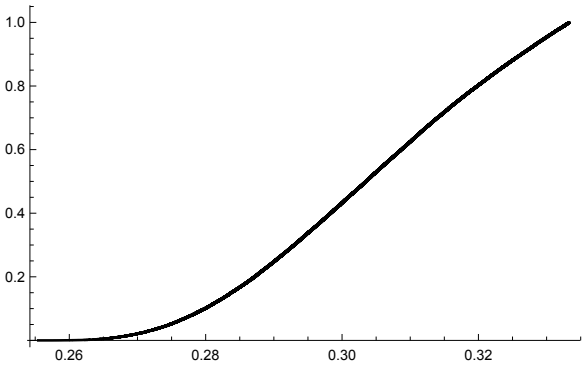


Figure 1: The distribution function of the limiting distribution of maximum betweenness centrality in simply generated trees.

## 3 Subcritical graphs

The same reasoning that yielded moments for root betweenness centrality in simply generated trees can be applied to subcritical graphs, which are also tree-like in nature, and give rise to a square-root singularity [6]. Important special cases of such graphs include cacti, outerplanar graphs and series-parallel graphs. We consider only the case of subcritical graphs with labelled vertices, so that if $\mathcal{C}$ is a class of rooted, connected graphs, and $\mathcal{B}$ the class of 2-connected subgraphs—or "blocks"—obtainable via block decomposition on $\mathcal{C}$, then each graph satisfies the symbolic definition

$$\mathcal{C} = \mathcal{Z} \times \textsc{Set}(\mathcal{B}' \circ \mathcal{C}) \,,$$

in which $\mathcal{B}'$ denotes blocks with a single removed vertex. This leads to the generating function

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

where $B(z)$ is the generating function of $\mathcal{B}$. The subcriticality property asserts that the radii of convergence $\rho$ of $C(z)$ and $\eta$ of $B(y)$ satisfy $C(\rho) < \eta$, so that $B'(y)$ is analytic at $y = \tau = C(\rho)$, and $C(z)$ conforms to the same square-root inverse schema that enabled the investigation of simply generated trees (sufficient conditions for this schema were settled on by Meir and Moon [15]). In particular, we have $\rho^{-1} = \exp(B'(\tau))B''(\tau)$, and expansions of $B'(y)$ and $C(z)$ around $\tau = C(\rho)$ and $\rho$ respectively:

$$B'(y) = B'(\tau) + B''(\tau)(y-\tau) + O\left((y-\tau)^2\right) \,,$$
$$C(z) = \tau - \mu\sqrt{1-\frac{z}{\rho}} + O\left(1-\frac{z}{\rho}\right) \,.$$

In this case, $\mu = \sqrt{2/(B''(\tau)^2 + B^{(3)}(\tau))}$.

In the following, we consider the betweenness centrality of the root once again. Since we consider labelled graphs, this is equivalent to taking the betweenness centrality of a random vertex. The only significant caveat when working with subcritical graphs is that root betweenness centrality is no longer solely determined by paths between vertices in distinct branches (branches, here, take the form of blocks with subgraphs rooted to their vertices, and will be referred to by means of the class $\mathcal{W} = \mathcal{B}' \circ \mathcal{C}$). In addition to between-branch paths, one must consider shortest paths between subgraphs of a branch's root block, as they may also pass through the root vertex.

Since shortest paths within blocks are not necessarily unique, the contribution of paths between a branch's subgraphs to $g(C) \equiv g(r)$—the overall betweenness centrality of the root $r$ of a graph $C \in \mathcal{C}$—is

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

where $v$ and $w$ are vertices in the branch's root block $B \in \mathcal{B}'$; $\sum g_{vw}(B) = g(B)$ is the betweenness centrality of $B$'s removed vertex with respect to paths contained within the block; and $C_v$ and $C_w$ are the subgraphs rooted at $v$ and $w$ respectively.

From this, an expansion for $g(C)$ in terms of two kinds of path is obtained:

$$\begin{aligned} g(C) &= \sum_{a<b} |W_a||W_b| + \sum_B \sum_{v<w} g_{vw}(B)|C_v||C_w| \\ &= g_1(C) + g_2(C) \,. \end{aligned}$$

**3.1 Average root betweenness centrality.** The cumulative generating function of $g(C)$ is the sum of those of $g_1(C)$ and $g_2(C)$; the first of which

can be derived in precisely the same way as the CGF of $g(T)$ was for simply generated trees—that is, by the partial substitution of branches with pointed branches. If $g_1(C)$ and $g_2(C)$ have CGFs $U_1(z)$ and $U_2(z)$ respectively, and the branch generating function is $W(z) = B'(C(z))$, then

$$U_1(z) = \frac{z^3}{2} \exp(W(z))\, W'(z)^2 = \frac{z^2}{2} C(z) W'(z)^2 .$$

From the expansions of $B'(y)$ and $C(z)$, we have

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

so that $U_1(z)$ satisfies

$$U_1(z) \sim \frac{\tau}{2} \left(\frac{\mu}{2} B''(\tau)\right)^2 \left(1 - \frac{z}{\rho}\right)^{-1} .$$

The CGF of $g_2(C)$ can be derived in a similar way, as long as one assigns the weight $g_{vw}(B)$ to each possible pair of substitution points $v, w$ in the block $B$: the generating function of a branch in which two vertices have been selected from distinct subgraphs in this way is

$$\begin{aligned} L(z) &= \sum_B \sum_{v<w} g_{vw}(B)\, C(z)^{|B|-2}\, (zC'(z))^2 \\ &= \sum_B g(B)\, C(z)^{|B|-2}\, (zC'(z))^2 \\ &= M_1(C(z)) \frac{(zC'(z))^2}{C(z)^2} , \end{aligned}$$

where $M_1(y)$ denotes the CGF of $g(B)$ over all blocks:

$$M_1(y) = \sum_B g(B)\, y^{|B|} = m_2 y^2 + m_3 y^3 + \cdots .$$

We remark that $M_1(y)$ has the same (or possibly even greater) radius of convergence as $B(y)$, since $g(B)$ can be bounded trivially by $|B|^2$. Substituting a single branch of a graph $C$ with one on which the above operation has been performed yields $U_2(z)$:

$$U_2(z) = M_1(C(z)) \frac{(zC'(z))^2}{C(z)} .$$

Noting that $C(z)^{-1}$ also permits a square-root expansion around $z = \rho$, beginning $1/\tau + \cdots$, the asymptotic form of $U_2(z)$ is

$$U_2(z) \sim \frac{1}{\tau} \left(\frac{\mu}{2}\right)^2 M_1(\tau) \left(1 - \frac{z}{\rho}\right)^{-1} .$$

Thus both kinds of path contribute equally in order to root betweenness centrality in a subcritical graph, and the CGF (of the first power) of $g(C)$ must be represented as the sum $H_1(z) = U_1(z) + U_2(z)$, from which coefficients can be extracted.

THEOREM 3.1. *Let $\mathcal{C}$ be the family of labelled subcritical graphs constructed from the block class $\mathcal{B}$. Then the root vertices of graphs $\mathcal{C}_n$ of size $n$ have an average betweenness centrality of order $n^{3/2}$, satisfying*

$$\mathbb{E}_{\mathcal{C}_n}(g(C)) = \frac{[z^n]\, H_1(z)}{[z^n]\, C(z)} \sim K_1 \sqrt{\pi n^3} ,$$

*where*

$$K_1 = \frac{\mu}{2} \left(\frac{\tau}{2} B''(\tau)^2 + \frac{1}{\tau} M_1(\tau)\right) .$$

**3.2 Higher moments of root betweenness centrality.** The higher-order moments of $g(C)$ are more interesting, because they involve the function

$$g(C)^k = \sum_{j=0}^{k} \binom{k}{j} g_1(C)^{k-j}\, g_2(C)^j .$$

The behaviours of $g_1(C)^k$ and $g_2(C)^k$, which arise when $j = 0$ and $k$ respectively, are predictably similar to that of $g(T)^k$ in the case of simply generated trees: combinatorially they equate to the selection of $2k$ vertices from between 2 and $2k$ distinct branches/subgraphs. The basic concept of Lemma 2.1 holds once again—even when substituting subgraphs of a root block—and this implies that the fastest growth is awarded by the fewest branchings, so that

$$\begin{aligned} g_1(C)^k &\sim \sum_{a<b} |W_a|^k |W_b|^k , \\ g_2(C)^k &\sim \sum_B \sum_{v<w} g_{vw}(B)^k |C_v|^k |C_w|^k . \end{aligned}$$

Both of these terms lead to cumulative generating functions of order $(1-(z/\rho))^{-2k+1}$.

The question, however, is whether the remaining terms—which involve both $g_1(C)$ and $g_2(C)$, are of lower or equal order. Note that the smallest number of substitutions that can be made when constructing a CGF involving both $g_1(C)$ and $g_2(C)$ is 3: some vertices must be selected from at least two branches, and the rest from at least two subgraphs. At best, one of the pointed branches could be accounted for by selecting extra vertices from the two subgraphs (which would be of the same branch), however in this case 3 substitutions must still be made. Since each replacement of a branch or subgraph with an $\alpha$-pointed structure contributes $(1-(z/\rho))^{-\alpha+1/2}$ to the final order of the constructed class, the remaining terms in $g(C)^k$ engender CGFs of lower order than those of the first and last terms. This simplifies the asymptotic behaviour of $g(C)^k$ greatly:

$$g(C)^k \sim g_1(C)^k + g_2(C)^k\,.$$

It follows that the $k$th moment satisfies an expression 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 $g(C)$ is of this order as well.

THEOREM 3.2. *The kth moment of the betweenness centrality of the root vertex in a class $\mathcal{C}_n$ of subcritical graphs of size $n$ is of order $n^{2k-1/2}$. Specifically, for $k \geq 1$:*

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

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

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

*Proof.* The asymptotic behaviour of $H_k(z)$, the CGF of $g(C)^k$, is

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

in which

$$M_k(y) = \sum_B \sum_{v<w} g_{vw}(B)^k\, y^{|B|}\,.$$

The desired moment is then obtained as a quotient of coefficients from $H_k(z)$ and $C(z)$.

**3.3 Limit behaviour of root betweenness centrality.** Again it can be shown that balanced structures become increasingly rare as $n \to \infty$, and that the majority of subcritical graphs possess linear root betweenness centrality. To do so, we define unbalanced subclasses $\mathcal{L}_{k,m} \subseteq \mathcal{C}$, which not only have $k$ non-root vertices outside their largest branch, but also have a dominant subgraph within that branch, with the remaining $m$ vertices distributed amongst the branch's minor subgraphs.

The class $\mathcal{L}_{k,m}$ has the exponential generating function

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

where $\Lambda_{k,m}$ counts the configurations of the minor structures:

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

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

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

These proportions account for the entire limiting distribution:

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

Finally, the root betweenness centrality of a graph $C \in (\mathcal{L}_{k,m})_n$ is of linear order, since there are linearly many of either kind of path:

$$g(C) \sim (n-k-m-1)\left(\sum_{i=2}^{\omega}k_i + \sum_{j=2}^{\beta}g_{vw_j}(B)\,m_j\right) = nk + n\sum_{j=2}^{\beta}g_{vw_j}(B)\,m_j + O((k+m)^2)\,,$$

where the $k_i$ and $m_j$ are the minor branch and subgraph sizes respectively. Noting that $0 \le g_{vw_j}(B) \le 1$, we have a linear bound on $g(C)$:

$$k \le \lim_{n\to\infty} \frac{g(C)}{n} \le k + m\,.$$

If more information on the blocks of the specific subcritical class—and in particular their betweenness centralities—is available, it is also possible to provide a more precise limit law, as for simply generated trees.

## 4 Recursive trees

Turning now to a class of trees that does not satisfy the square-root inverse schema of simply generated trees and subcritical graphs, we consider the class $\mathcal{R}$ of recursive trees—a form of rooted, labelled, increasing tree; see [5, Section 1.3.1]. These trees can also be obtained from a growth process, where the vertex with label $l$ attaches to one of the previous vertices, each with probability $1/(l-1)$. In contrast with simply generated trees, recursive trees have an average height of order $\log n$, and furthermore, the majority of a recursive tree's vertices lie at a distance of order $\log n$ from the root. With this in mind, it can perhaps be anticipated that the $k$th moment of the root betweenness centrality $g(R)$ of a recursive tree is of order $n^{2k}$.

By definition, the removal of the root vertex from a recursive tree leaves us with a set of branches, so that the generating function $R(z)$ that counts recursive trees is given by the differential equation $y'(z) = \exp(y(z))$, and has the explicit solution $y(z) = -\log(1-z)$.

**4.1 Average root betweenness centrality.** No unfamiliar method is needed when deriving the mean value of root betweenness centrality in recursive trees; one need only apply the partial substitution operation to sets of branches, described by the derivative $y'(z)$, instead of directly dealing with the generating function $y(z)$. The exponential CGF $H_1(z)$ that counts root betweenness centrality by considering sets of branches can be symbolically constructed by replacing two of these branches with pointed branches:

$$H_1(z) = \frac{z^2}{2} y'(z)^2 \exp(y(z)) = \frac{z^2}{2}(1-z)^{-3}\,.$$

THEOREM 4.1. *The expected betweenness centrality of the root vertex in a random recursive tree of size $n$ is*

$$\mathbb{E}_{\mathcal{R}_n}(g(R)) = \frac{[z^{n-1}]\, H_1(z)}{[z^{n-1}]\, y'(z)} = \frac{1}{2}\binom{n-1}{2} \sim \frac{n^2}{4}\,.$$

**4.2 Higher moments and the limiting distribution of root betweenness centrality.** As in the previous sections, higher-order moments can be found by expanding $g(R)^k$ and grouping terms which have the same set of exponents. Unlike the functions in those sections, however, the $k$th moment of $g(R)$ is not dominated by the behaviour of its leading sum

$$g(R)^k = \sum_{a<b} |R_a|^k |R_b|^k + \cdots\,.$$

On the contrary: for recursive trees, each sum contributes a quantity of order $(1-z)^{-2k-1}$ to the CGF of $g(R)^k$ (which will be denoted by $H_k(z)$, and is again computed via $y'(z)$).

In proving this, it will be desirable to know the values of the coefficients involved in the expansion; and for this reason, we consider instead the form of $g(R)^k$ which eschews the double sum: writing $s(R)$ for the sum $\sum_a |R_a|^2$ of the squared branch sizes, and assuming that $\omega$ branches are present, we have

$$\begin{aligned} g(R)^k &= \left(\frac{1}{2}\left(\sum_{a=1}^{\omega} |R_a|\right)^2 - \frac{1}{2}\sum_{a=1}^{\omega} |R_a|^2\right)^k \\ &= \frac{1}{2^k}\left((n-1)^2 - s(R)\right)^k \\ &= \frac{1}{2^k}\sum_{j=0}^{k} \binom{k}{j} (-1)^j s(R)^j (n-1)^{2k-2j}\,. \end{aligned}$$

The $k$th moment can then be deduced from the cumulative generating functions of the $s(R)^j$, after they have been expressed in closed form.

THEOREM 4.2. *The $k$th moment of root betweenness centrality in a recursive tree of size $n$ is of*

*order $n^{2k}$, and satisfies, for $k \geq 1$*

$$\mathbb{E}_{\mathcal{R}_n}(g(R)^k) \sim \frac{n^{2k}}{2^k} \sum_{j=0}^{k} \binom{k}{j} \frac{(-1)^j}{(2j)!} D_j \,,$$

*in which the constant $D_j$ is given by the coefficient of $u^j$ in the exponential generating function* $\exp\left(\sum_{h\geq 1}(2h-1)!u^h/h!\right)$.

*Proof.* We begin by expanding $s(R)^j$, grouping as usual according to the symbolic selection of $2j$ vertices (now two at a time) from a number of branches, so that each group corresponds to an integer partitioning of $j$. Take, for example, the partition $\sigma$ of $j$: $a_1 + 2a_2 + \cdots + ja_j = j$, and let $\sum a_i = d$ be the number of affected branches. The coefficient which precedes this group in the expansion of $s(R)^j$ is

$$B_\sigma = \frac{j!}{1!^{a_1} 2!^{a_2} \cdots j!^{a_j}} \,,$$

obtained by choosing, from $s(R)^j$, the $i$ factors $s(R)$ which each branch term $|R_a|^{2i}$ should appear in. The exponential CGF of this group is built by substituting $a_1$ 1-pointed branches, $a_2$ 2-pointed branches, etc., and satisfies

$$M_\sigma(z) = B_\sigma \prod_{i=1}^{j} P_{2i}(z)^{a_i} \times \sum_{\omega \geq 0} \binom{\omega}{a_1} \cdots \binom{\omega - d + a_j}{a_j} \frac{y(z)^{\omega - d}}{\omega!} \,,$$

in which $P_i(z)$ is the exponential generating function of the $i$-pointed branch class $\mathcal{P}_i = \Theta^{(i)}\mathcal{R}$. Once again the pointed generating functions are dominated by a single term, since $y'(z) = (1-z)^{-1}$:

$$P_i(z) \sim z^i y^{(i)}(z) \sim (i-1)!\,(1-z)^{-i} \,.$$

The asymptotic form of the CGF of group $\sigma$ is thus

$$M_\sigma(z) \sim C_\sigma\,(1-z)^{-2j-1} \,,$$

where the constant $C_\sigma$ is the product of $B_\sigma$, the coefficients of the derivatives $P_{2i}(z)$, and the denominators of the binomial coefficients in $M_\sigma(z)$:

$$C_\sigma = \frac{3!^{a_2} 5!^{a_3} \cdots (2j-1)!^{a_j}}{2!^{a_2} 3!^{a_3} \cdots j!^{a_j}} \frac{j!}{a_1! a_2! \cdots a_j!} \,.$$

Finally, we have an asymptotic expression for the CGF of $s(R)^j$ in terms of the possible partitions $\sigma$ of $j$: with $D_j = \sum_\sigma C_\sigma$,

$$U_j(z) \sim D_j (1-z)^{-2j-1} \,.$$

Specifically, the first few such generating functions are

$$U_0(z) \sim (1-z)^{-1} \,, \qquad U_3(z) \sim 139(1-z)^{-7} \,,$$
$$U_1(z) \sim (1-z)^{-3} \,, \qquad U_4(z) \sim 5665(1-z)^{-9} \,.$$
$$U_2(z) \sim 7(1-z)^{-5} \,,$$

Now that $U_j(z)$ has been derived, it can be combined with the binomial expression of $g(R)^k$ to obtain an expression for the $k$th moment of $g(R)$:

$$\frac{[z^{n-1}]\,H_k(z)}{[z^{n-1}]\,y'(z)} \sim \frac{1}{2^k} \sum_{j=0}^{k} \binom{k}{j} (-1)^j n^{2k-2j} [z^{n-1}]\,U_j(z)$$

which is of order $n^{2k}$. The variance of root betweenness centrality in recursive trees, in particular, is

$$\mathbb{V}_{\mathcal{R}_n}(g(R)) \sim \frac{n^4}{96} \,,$$

and the first three moments begin the progression

$$\frac{1}{4}n^2, \frac{7}{96}n^4, \frac{131}{5760}n^6, \ldots$$

Following on from this, we see that all moments of the scaled random variable $g(R)/n^2$ converge to a limit:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{R}_n}\left(\frac{g(R)^k}{n^{2k}}\right) = c_k \,.$$

Since the betweenness centrality is trivally bounded above by $(n-1)(n-2)/2$, we automatically obtain $c_k \leq 2^{-k}$, which means that the generating function of the constants $c_k$ converges in a neighbourhood of 0 and represents a moment generating function. Thus, in view of [7, Theorem C.2], $g(R)/n^2$ converges weakly to a distribution that is characterised by the moments $c_k$. Figure 2 shows the distribution function of the limiting distribution (which has the interval $[0, 1/2]$ as support). Since we only have an expression for its moment generating function, it was obtained by means of Monte Carlo sampling. The picture suggests that

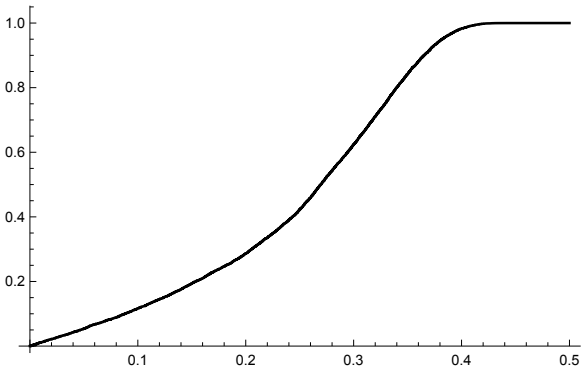


Figure 2: The distribution function of the limiting distribution of root betweenness centrality in the case of random recursive trees.

the distribution has a density, but in order to obtain an analytic expression for this density, some further insights will be required.

An alternative form of this characterisation is given in terms of $1 - (g(R)/(n^2/2))$ instead: the limiting distribution of $1 - (g(R)/(n^2/2))$ has support $[0, 1]$, and the exponential moment generating function

$$M(t) = \sum_{k \geq 0} \frac{C_k}{(2k)!} \frac{t^k}{k!} .$$

We finally remark that the same methods can be used to show that for any fixed label $l$, the betweenness centrality of vertex $l$, divided by $n^2$, converges to a limiting distribution (that depends on $l$).

**Acknowledgements**

The authors are particularly grateful to the reviewers for their insights, comments, and suggestions.

**References**

[1] David Aldous, *The continuum random tree. I*, Ann. Probab. **19** (1991), no. 1, 1–28.

[2] ———, *The continuum random tree. II. An overview*, Stochastic analysis (Durham, 1990), London Math. Soc. Lecture Note Ser., vol. 167, Cambridge Univ. Press, Cambridge, 1991, pp. 23–70.

[3] ———, *The continuum random tree. III*, Ann. Probab. **21** (1993), no. 1, 248–289.

[4] ———, *Recursive self-similarity for random trees, random triangulations and Brownian excursion*, Ann. Probab. **22** (1994), no. 2, 527–545.

[5] Michael Drmota, *Random trees: An interplay between combinatorics and probability*, 1st ed., Springer, 2009.

[6] Michael Drmota, Eric Fusy, Mihyun Kang, Veronika Kraus, and Juanjo Rué, *Asymptotic study of subcritical graph classes*, SIAM Journal on Discrete Mathematics **25** (2011), no. 4, 1615–1651.

[7] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, 1st ed., Cambridge University Press, New York, NY, USA, 2009.

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

[9] Silvia Gago, Jana Coroničová Hurajová, and Tomáš Madaras, *Betweenness centrality in graphs*, Quantitative graph theory, Discrete Math. Appl. (Boca Raton), CRC Press, Boca Raton, FL, 2015, pp. 233–257.

[10] Michelle Girvan and Mark E. J. Newman, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. USA **99** (2002), no. 12, 7821–7826.

[11] Kwang-Il Goh, Eulsik Oh, Hawoong Jeong, Byungnam Kahng, and Doochul Kim, *Classification of scale-free networks*, Proc. Natl. Acad. Sci. USA **99** (2002), 12583–12588.

[12] Frank Harary, *Graph theory*, Addison-Wesley Publishing Co., Reading, Mass.-Menlo Park, Calif.-London, 1969.

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

[14] Amram Meir and John W. Moon, *On the altitude of nodes in random trees*, Canadian Journal of Mathematics **30** (1978), 997–1015.

[15] ———, *On an asymptotic method in enumeration*, Journal of Combinatorial Theory, Series A **51** (1989), no. 1, 77–89.

[16] ———, *On centroid branches of trees from certain families*, Discrete Math. **250** (2002), no. 1-3, 153–170.

[17] Mark E. J. Newman, *Networks. An introduction*, Oxford University Press, Oxford, 2010.