Shortest paths and load scaling in scale-free trees

Béla Bollobás[1,2] and Oliver Riordan[2,3]
[1]*Department of Mathematical Sciences, University of Memphis, Memphis, Tennessee 38152, USA*
[2]*Trinity College, Cambridge CB2 1TQ, United Kingdom*
[3]*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, United Kingdom*
(Received 27 October 2003; published 24 March 2004)

Szabó, Alava, and Kertész [Phys. Rev. E **66**, 026101 (2002)] considered two questions about the scale-free random tree given by the $m=1$ case of the Barabási-Albert (BA) model (identical with a random tree model introduced by Szymański in 1987): what is the distribution of the node to node distances, and what is the distribution of node *loads*, where the load on a node is the number of shortest paths passing through it? They gave heuristic answers to these questions using a ''mean-field'' approximation, replacing the random tree by a certain *fixed* tree with carefully chosen branching ratios. By making use of our earlier results on scale-free random graphs, we shall analyze the random tree rigorously, obtaining and proving very precise answers to these questions. We shall show that, after dividing by N (the number of nodes), the load distribution converges to an integer distribution X with $\Pr(X=c)=2/[(2c+1)(2c+3)]$, $c=0,1,2,...$, confirming the asymptotic power law with exponent -2 predicted by Szabó, Alava, and Kertész. For the distribution of node-node distances, we show asymptotic normality, and give a precise form for the (far from normal) large deviation law. We note that the mean-field methods used by Szabó, Alava, and Kertész give very good results for this model.

DOI: 10.1103/PhysRevE.69.036114 PACS number(s): 89.75.Hc, 05.10.−a, 89.70.+c, 89.75.Da

I. INTRODUCTION

Recently there has been much interest in studying ''scale-free'' random graphs as simple models for large-scale real-world networks arising in a wide variety of contexts. We use notation from graph theory: nodes in the network being modeled correspond to *vertices* in the graph. Our graphs will usually have N vertices. Certain pairs of vertices are joined directly by edges; the *degree* of a vertex is the number of edges incident with it. Following widespread, though perhaps unfortunate, terminology, we say a random graph is *scale free* if the (average) degree distribution follows a power law, i.e., if $p_k \sim k^{-\gamma}$ for some $\gamma>0$, where p_k is the limiting fraction of vertices with degree k, as $N\to\infty$ with k fixed. See [1,2] for extensive surveys of this rapidly developing field, and [3] for a survey of the smaller amount of mathematically rigorous work.

A special case of scale-free random graphs is that of scale-free random trees. Graph theoretically, this is a particularly simple case, as well as a very natural one. Since trees are the minimal connected graphs, they are appropriate models for real-world networks in many contexts. For this reason, scale-free trees have been studied recently by several groups, in the form of the $m=1$ case of the Barabási-Albert (BA) scale-free graph model [4]. As will shall note in the next section, scale-free trees in fact predate the BA model by more then a decade; they were introduced in a different context and under a different name by Szymański [5] in 1987.

Since scale-free trees arise in the context of modelling communication networks, for example, shortest paths in these trees are a natural object of study. Szabó, Alava, and Kertész [6] considered the following two questions for the $m=1$ case of the BA model. First, what is the asymptotic distribution of shortest path lengths? Secondly, what is the distribution of the ''load'' on the vertices? Here the *load* on a vertex is the number of shortest paths passing through the vertex. Szabó, Alava, and Kertész [6] study these questions using a ''mean-field'' approximation, i.e., considering a *deterministic* tree, with the same number of vertices in each layer that one expects in the random tree. Using this method, they give a heuristic derivation showing that the load distribution will be asymptotically a power law with exponent -2; the same power law has been obtained by analytic derivation (still a heuristic, not a rigorous proof) in [13]. Szabó, Alava, and Kertész also derive asymptotic normality (and in fact more) for the distribution of the root-node distances. For node-node distances, they give no results (despite appearances), though, as we shall see in the final section, a simple heuristic derivation is possible using the work in [6].

It turns out that the particular model under consideration is simple enough that rigorous analysis is possible. We shall show that, after dividing by N, the number of paths through a random vertex converges in distribution to an integer valued distribution X, with

$$\Pr(X=c)=\frac{2}{(2c+1)(2c+3)}$$

for $c=0,1,2,...$. Also, we shall show that the distribution of node-node distances is asymptotically normal with mean and variance $\log N$, and give bounds on the distribution accurate to within a factor $\Theta(1)$ well into the tails. (See theorem 1.)

In the next section we describe exactly the model, and note that this particular model was introduced in a different context, in a paper not widely known in the physics community. In the subsequent two sections we prove our results on load distribution and node-node distances. At the end of each section, we compare our results with existing heuristic re-

sults, especially those of [6], and extrapolations from them; it turns out that the mean-field heuristic gives very good results in these cases.

II. THE MODEL

Barabási and Albert [4] introduced one of the first models for the growth of a scale-free graph such as the web or internet graph. They fix an integer $m \geqslant 1$. Starting with m_0 vertices and no edges, vertices are added one at a time, and each new vertex is joined to m old vertices, chosen ''with probabilities proportional to their degrees.'' As noted in [7], this is not a complete description of a model: for example, it is not clear how the process should get started, since as written all degrees are initially zero. For $m \geqslant 2$ there is a much more serious problem (see [3,7]), but this need not concern us here. A precise model fitting the rough description of Barabási and Albert is given in [7] (see also [3]). While this precise model has many nice features (such as a simple static description), for the special case $m=1$ it is a little unnatural, in that it produces a forest with loops, rather than a tree.

For $m=1$, the only ambiguity in the Barabási-Albert (BA) model is how to get started. Having started, at each stage a new vertex is added, and joined to an old vertex selected with probability proportional to its degree. Perhaps the most common way to get going is to start with one vertex, the *root*, which has an extra ''virtual'' edge coming in to it from nowhere, so the degree of the root at the start counts as 1. Thus at time t, when there are t vertices, although there are $t-1$ edges in the tree, the sum of the degrees is taken to be $2(t-1)+1=2t-1$.

This precise version of the $m=1$ case of the BA model is not at all new; it is exactly the standard model for random plane-oriented recursive trees. A tree on a labeled vertex set $V=\{1,2,...,t\}$ is *recursive* if each vertex other than 1 is joined to exactly one earlier vertex. In other words, the tree can be grown by adding the vertices in numerical order, joining each new vertex to some old vertex. Two natural ways of constructing random recursive trees have been considered: for the simplest, a ''uniform random recursive tree'' is grown one vertex at a time by joining the new vertex to an old vertex chosen uniformly at random. See, for example, the survey [8]. To introduce the second, we need one more definition.

A *plane-oriented* tree is one with a cyclic order on the edges meeting each vertex, induced, for example, by drawing the tree in the plane. Suppose that a new vertex v is added to a plane-oriented recursive tree T and joined to an existing vertex w, where w has degree d. Then there are d different ways in which the new edge can meet the vertex w, depending on the place in which this edge is inserted into the cyclic order. Thus the number of different plane-oriented recursive trees that may result is d. Hence, a (uniformly selected) random plane-oriented recursive tree may be obtained by adding vertices one at a time, joining each new vertex to an old vertex selected with probability proportional to its degree. In fact, as in the Barabási-Albert model, the standard definition treats the first vertex, the root, differently, effectively imagining an edge from the root going off to infinity. In this way branches of plane-oriented recursive trees are again plane-oriented recursive trees. Random plane-oriented recursive trees were introduced by Szymański [5] in 1987 (although with a slightly different treatment of the root) and have been studied since in several papers, including [5,9–11].

Throughout the paper, by T_N we shall mean the random plane-oriented recursive tree with N vertices, with vertex set $\{1,2,...,N\}$, or, equivalently, the Barabási-Albert scale-free random tree given by (a precise version of) the case $m=1$ of the model introduced in [4]. Formally, T_1 has a single vertex 1 and no edges. For $N \geqslant 2$, given T_{N-1}, the tree T_N is constructed by adding a new vertex N and joining it to an old vertex v, $1 \leqslant v \leqslant N-1$, with

$$\Pr(v=j)=\frac{d_{N-1}(j)}{2N-3}$$

for $j \geqslant 2$, and

$$\Pr(v=1)=\frac{d_{N-1}(1)+1}{2N-3},$$

where $d_{N-1}(j)$ is the degree of the vertex j in the tree T_{N-1}.

We shall say that a vertex v joined to a vertex w is a *child* of w if $v>w$ (so v was added later), and the *parent* of w if $v<w$.

III. LOAD SCALING

The first question we consider here, raised by Szabó, Alava, and Kertész in [6] and considered also by Goh, Kahng, and Kim in [12], and the same group together with Oh and Jeong in [13], concerns the distribution of vertex loads in the tree T_N. Given a general graph with N vertices, for each pair $\{x,y\}$ of distinct vertices in the graph choose a shortest path $S_{x,y}$ between them (uniformly at random if there is more than one). A path *passes through* a vertex v if it contains v in its interior, i.e., not as an end point. The *load* $l(v)$ at v is the number of the paths $S_{x,y}$ passing through v.

There are several minor variations on this definition, for example, counting also shortest paths ending at v, or, for each pair x, y generating more than one shortest path, assigning weights summing to 1 to these paths in one of several ways. In a tree, the situation is simpler, as there is a unique shortest path between each pair of vertices. As we do not count paths ending at v, if the components of T_N-v, the graph formed from T_N by deleting the vertex v, have $s_1,\ldots,s_r$ vertices, then $l(v)=\Sigma_{i<j}s_i s_j$. Now for almost all vertices v in T_N one of these components (that containing the root) is much larger than the others. Suppose that v has $c(v)$ descendants in T_N, where the descendants of a vertex are its children, its children's children, and so on. Then the component of T_N-v containing the root has size $N-1-c(v)$. (We have defined the descendants of v so as to exclude v itself.) Furthermore, the load at v is $c(v)[N-1-c(v)]+O(c(v)^2)$. Here the first term counts paths from descendants of v to other vertices. The second term accounts for the $\binom{c(v)}{2}$ paths between descendants of v; none, some, or all of these may pass through v. In particular, if $c(v)=o(N)$,

which holds for all but a few early vertices, then we have $l(v) \sim c(v)N$.

The quantities $c(v)$ are natural in their own right, and it is not surprising that their distribution has already been given exactly by Mahmoud, Smythe, and Szymański in [11]: the probability that in T_N the vertex k has exactly c descendants is

$$\frac{1\cdot 3\cdots(2c-3)(2c-1)(2k-2)(2k)\cdots(2n-2c-4)}{(2k-1)(2k+1)\cdots(2n-3)} \times \binom{n-k}{c}. \quad (1)$$

This expression can easily be proved by a direct combinatorial argument, or by induction. For the load distribution we wish to know the expected number of vertices with c descendants. This can be obtained by summing Eq. (1) over k, but can also be obtained by the following much simpler direct method.

When a new vertex $N \geqslant 2$ is added to the tree T_{N-1}, what is the probability p_v that N becomes a descendant of a given old vertex $v \geqslant 2$? From the definition of the model, the answer is $\Sigma_{x\in S} d(x)/(2N-3)$, where $d(x)$ is the degree of x in T_{N-1}, and S is the set consisting of v and all its descendants. Now S induces a subtree of T_{N-1} with $|S| = 1 + c(v)$ vertices, and hence $c(v)$ edges, joined to the rest of T_{N-1} by a single edge, that from v to its parent. Thus $\Sigma_{x\in S} d(x) = 2c(v) + 1$, so

$$p_v = \frac{2c(v)+1}{2N-3}.$$

This formula is also valid if $c = 1$.

Let us write $n_t(c)$ for the number of vertices in T_t having c descendants. For $c \geqslant 1$ the number of such vertices in T_N is the same as in T_{N-1} unless either vertex N becomes a descendant of an existing vertex with c descendants, or vertex N becomes a descendant of an existing vertex with $c-1$ descendants. The former event has probability $[(2c+1)/(2N-3)]n_{N-1}(c)$, and the latter probability $([2(c-1)+1]/(2N-3))n_{N-1}(c-1)$. Thus, for $c \geqslant 1$,

$$\mathbb{E}(n_N(c)|T_{N-1}) = n_{N-1}(c) - \frac{2c+1}{2N-3} n_{N-1}(c) + \frac{2c-1}{2N-3} n_{N-1}(c-1).$$

Writing $\lambda_{N,c}$ for $\mathbb{E}(n_N(c))$ and taking the expectation of both sides of the equation above, we find that

$$\lambda_{N,c} = \lambda_{N-1,c} - \frac{2c+1}{2N-3}\lambda_{N-1,c} + \frac{2c-1}{2N-3}\lambda_{N-1,c-1}.$$

For $c = 0$ the only difference is that the new vertex always has 0 descendants, so

$$\lambda_{N,0} = \lambda_{N-1,0} - \frac{1}{2N-3}\lambda_{N-1,0} + 1.$$

Together with the boundary conditions that $\lambda_{1,c} = 0$ for $c \geqslant 1$ and $\lambda_{1,0} = 1$, the above equations easily imply that

$$\lambda_{N,c} = \frac{2N-1}{(2c+1)(2c+3)} \quad (2)$$

for $c \leqslant N-2$, while $\lambda_{N,N-1} = 1$ (at time N the root has $N-1$ descendants) and $\lambda_{N,c} = 0$ for $c \geqslant N$.

Fixing N and varying c, Eq. (2) gives the *load scaling* in the scale-free tree T_N: the expected proportion of vertices having exactly c descendants, and hence load approximately cN, is exactly

$$\frac{2 - 1/N}{(2c+1)(2c+3)}$$

for $0 \leqslant c \leqslant N-2$. This gives a much more precise version of the inverse square power-law described in [6]; the heuristic derivation used there gives only the asymptotic form as $c \to \infty$, since vertices far enough down in the tree that the average branching factor is less than one are omitted. Since such vertices are a constant proportion of all vertices, they must be included to obtain the exact formula given above.

Note that the exponent of 2 is the same as that given with a heuristic derivation in [13], rather than the value ≈ 2.2 originally suggested by some of the same authors in [12]. It is also clear that the $N^{1.8}\log N$ scaling suggested in [12] for the load of the root is not correct for this model, and that the correct answer is $\Theta(N^2)$: with constant probability the vertices 2 and 3 are each joined directly to the root, and on average each has $\Theta(N)$ descendants, giving order N^2 paths passing through the root.

IV. SHORTEST PATHS

We now turn to the main topic of this paper, the distribution of the distances between vertices, i.e., the distribution of the lengths of the $\binom{N}{2}$ paths in T_N. (As T_N is a tree, every path is the shortest path between its end vertices.) As for load scaling, this distribution is studied heuristically by Szabó, Alava, and Kertész in [6] using the mean-field approximation. It is stated in [6] that the distribution of distances in T_N had previously been analyzed precisely; in fact this is not the case. As we have noted in the Introduction, the random tree T_N, i.e., the case $m = 1$ of the Barabási-Albert model, is nothing other than a random plane-oriented recursive tree, or ''nonuniform random recursive tree'' as previously studied in [5,10,11,14], among other papers. This has very different properties from a *uniform* random recursive tree, where there is no preferential attachment. The references given in [6] are indeed to rigorous studies of certain properties of random recursive trees. However, one, [15], deals entirely with the uniform case, while another, [16], gives for the nonuniform case only the distribution of distances from the root. (This distribution was given earlier in [10].) The reference [14] for the diameter (meaning *maximum* length of a shortest path, as is usual in graph theory, as opposed to *average* length) is essentially correct; this paper shows that the height of the tree (largest distance from the root) is almost certainly [1

$+o(1)](2\gamma)^{-1}\log N$ for γ the solution of $\gamma e^{1+\gamma}=1$. It is easy to see that the method gives the diameter as $[1+o(1)]\gamma^{-1}\log N$. Note that this value $\gamma^{-1}=3.591...$ is different from the incorrect value $1+\sqrt{2}/2=1.707...$ given in [6]. The difference is actually not from the mean-field approximation, which turns out to give the right answer, using the method suggested in [6]. The mistake in [6] is to use the normal approximation [their Eq. (7)] to their Eq. (6), which is not valid this far out, and to forget to multiply by two.

Concerning the distribution of distances, despite the statement in [6], to the best of our knowledge the distribution of distances in T_N, the case $m=1$ of the BA model, has not previously been rigorously determined.

Here we shall give a (cumbersome) exact formula for the expected number of paths of length k for any k, and a simple description of the distribution on two scales, one giving the asymptotic form of the central part of the distribution, including almost all paths, and one the rate (and form) of decay of the tails. Let us write $E_k=E_k(N)$ for the expected number of (shortest) paths in T_N of length k, so $\Sigma_{k=1}^{\infty}E_k(N)=\binom{N}{2}$.

Theorem 1. Suppose that $k=k(N)$ is such that $\alpha=\alpha(N)=k/\log N$ is bounded above and below by constants strictly between 0 and e. Then

$$E_k=\Theta(N^{1+\alpha\log(e/\alpha)}/\sqrt{\log N}), \tag{3}$$

as $N\to\infty$. Furthermore, if $k=\log N+x\sqrt{\log N}$ where $x=x(N)=o(\sqrt{\log N})$, then

$$E_k\sim\frac{N^2}{2}\frac{1}{\sqrt{2\pi\log N}}e^{-x^2/2}, \tag{4}$$

as $N\to\infty$.

Note that the second statement says that the distribution of path lengths is asymptotically normal with mean and variance $\log N$. Our main tool will be a result from [3], based on the work in [7], giving the exact probability that a given graph S is present as a subgraph of T_N. Here we mean that the specific edges in S occur in T_N, not that T_N contains a subgraph isomorphic to S. Although the result and its proof are simple, stating the result requires some definitions.

Let S be any graph on $V=\{1,2,...,N\}$ which could possibly occur as a subgraph of T_N. Thus, in S, every vertex j is joined to at most one "earlier" vertex i, where i is *earlier* than j means $i<j$. Considering each edge ij of S with $i<j$ as oriented from j to i, let $V^+(S)$ be the set of vertices sending out at least one edge, and $V^-(S)$ the set of vertices receiving at least one edge. These sets are, of course, in general not disjoint. Furthermore, for $i\in V^-(S)$ let $d_S^{in}(i)$ be the number of edges of S coming in to i. Finally, for $1\leqslant t\leqslant N$ let $C_S(t)$ count the number of edges ij of S with $i<t$ and $j\geqslant t$. Then, corollary 22 of [3] states that the probability that $S\subset T_N$ is given exactly by

$$p_S=\prod_{i\in V^-(S)}d_S^{in}(i)!\prod_{i\in V^+(S)}\frac{1}{2i-3}\prod_{t\notin V^+(S)}\left(1+\frac{C_S(t)}{2t-3}\right). \tag{5}$$

This result is the core of our proof of theorem 1. Most of the rest is a straightforward (but somewhat tedious) estimation of the resulting sums.

Proof of Theorem 1. In the case of paths, which is all we consider here, formula (5) simplifies somewhat: the only paths P which may appear in T_N are of the form $av_1\cdots v_scw_t\cdots w_1b$, where c is the last common ancestor of a and b, and $av_1\cdots v_sc$ and $bw_1\cdots w_tc$ are paths down the tree, so, for example, $a>v_1>\cdots>v_s>c$. (Here we adopt the nonstandard convention that the root of a tree is at the bottom.) We shall assume without loss of generality that $a<b$. Note that $c\leqslant a$, with $c=a$ possible. To apply Eq. (5) it is convenient to regroup the vertices of P, writing $V(P)=\{c\}\cup L\cup\{a\}\cup R\cup\{b\}$, where L is the set of vertices v of P with $c<v<a$, and R the set with $a<v<b$. Every vertex of P has in-degree 0 or 1 except c, which has in-degree 2 (if $a\neq c$). Also, $C_P(t)$, the number of edges of P from vertices before t to t and vertices after, is 2 if $c<t\leqslant a$, 1 if $a<t\leqslant b$, and zero otherwise. Hence, provided $c<a$, the probability that P is present in T_N is

$$\begin{aligned}p_P&=p(a,b,c,L,R)\\&=2\prod_{i\in\{a,b\}\cup L\cup R\setminus\{c\}}\frac{1}{2i-3}\prod_{c<t<a,\,t\notin L}\frac{2t-1}{2t-3}\\&\quad\times\prod_{a<t<b,\,t\notin R}\frac{2t-2}{2t-3}.\end{aligned}$$

This can be rewritten as

$$\begin{aligned}p(a,b,c,L,R)=2\prod_{c<t\leqslant a}\frac{2t-1}{2t-3}\frac{1}{2a-1}\prod_{i\in L}\frac{1}{2i-1}\\\times\prod_{a<t\leqslant b}\frac{2t-2}{2t-3}\frac{1}{2b-2}\prod_{i\in R}\frac{1}{2i-2}.\end{aligned} \tag{6}$$

(The scope of each product is the fraction immediately following it.) If $a=c$ then the initial 2 and the factor $1/(2a-1)$ must be omitted, as c has in degree one, and $a=c\notin V^+(P)$. Note that the path P is not quite specified by a, b, c, L and R: in P each vertex in R must lie on the subpath from c to b, but each vertex in L may be on either this subpath, or the one from c to a. Thus, not that it is very useful, we have the following exact formula for the expected number E_k of paths of length k in T_N:

$$\begin{aligned}E_k=\sum_{1\leqslant c<a<b\leqslant N}\ \sum 2^{|L|}p(a,b,c,L,R)\\+\sum_{1\leqslant a<b}\ \sum_{R\subset\{a+1,...,b-1\},|R|=k-1}p(a,b,a,\varnothing,R),\end{aligned}$$

where the second sum in the first term is over all pairs (L,R) with $L\subset\{c+1,...,a-1\}$, $R\subset\{a+1,...,b-1\}$ and $|L\cup R|=k-2$.

We now aim to find simple descriptions of the distribution of shortest path lengths at various scales. Until the end of the proof we fix a small positive ϵ. We shall be prepared to

ignore multiplicative errors smaller than $1+\epsilon$ in E_k, as well as absolute errors smaller than $N^{1+\epsilon}$. In particular, for any k we shall ignore the at most N paths in T_N of length k for which $c=a$. (There is at most one such path downward in the tree from each upper endvertex b.) Thus we have

$$E_k = \sum_{1\leq c<a<b} \sum_{L,R} 2^{|L|} p(a,b,c,L,R) + O(N),$$

with the same conditions on L and R as before. Now in the formula (6) the first product telescopes, and is just $(2a-1)/(2c-1)$. The third product does not telescope, but is equal to $\sqrt{b/a}(1+O(1/a))$. Thus if $a \geq N^{\epsilon}$, $c<a$ and $|R| = O(\log N)$ then

$$\begin{aligned} p(a,b,c,L,R) &\sim 2\frac{2a-1}{2c-1}\frac{1}{2a-1} \\ &\quad \times \prod_{i\in L}\frac{1}{2i-1}\sqrt{b/a}\,\frac{1}{2b-2}\prod_{i\in R}\frac{1}{2i-2} \\ &\sim \frac{1}{(2c-1)\sqrt{ab}}\prod_{i\in L}\frac{1}{2i-1}\prod_{i\in R}\frac{1}{2i-2}. \end{aligned}$$

As there are only $O(N^{1+\epsilon})$ paths with $a\leq N^{\epsilon}$ or with $b-a\leq N^{\epsilon}$, we can consider only terms with a, $b-a>N^{\epsilon}$. We are only interested in the range where $k=\Theta(\log N)$, since T_N has diameter $O(\log N)$, and there are very few paths of length $o(\log N)$. It follows that we need only consider terms in the sum where $c=O(1)$. In fact, from the form of the sums involved, the contribution to E_k from terms with increasing values of c decreases exponentially as c increases; we shall return to this later.

From now on we only consider terms with $1\leq c<a<b$ where a, $b-a>N^{\epsilon}$ and c is at most some sufficiently large constant C. We write Σ^* for sums over such triples. We also assume that $k=\Theta(\log N)$. As noted above, we then have

$$\begin{aligned} E_k &\sim \sum^{*} \frac{1}{(2c-1)\sqrt{ab}} \sum_{k_1+k_2=k-2} 2^{k_1} \\ &\quad \times \sum_{L\subset\{c+1,\ldots,a-1\},|L|=k_1} \prod_{i\in L}\frac{1}{2i-1} \\ &\quad \times \sum_{R\subset\{a+1,\ldots,b-1\},|R|=k_2} \prod_{i\in R}\frac{1}{2i-2} + O(N^{1+\epsilon}). \end{aligned}$$

It is easy to see that

$$\begin{aligned} &\sum_{R\subset\{a+1,\ldots,b-1\},|R|=k_2} \prod_{i\in R}\frac{1}{2i-2} \\ &\quad \sim \frac{1}{k_2!}\left(\sum_{i=a+1}^{b-1}\frac{1}{2i-2}\right)^{k_2} \sim \frac{[\log(b/a)]^{k_2}}{2^{k_2}k_2!}, \end{aligned}$$

since $k_2=O(\log N)$, so when the power of the sum is expanded the terms which are products of distinct summands dominate.

The corresponding term for L is a little trickier to handle: since $c=O(1)$, a fraction $\Theta(1/\log N)$ of the sum $\Sigma_{i=c+1}^{a-1}[1/(2i-1)]$ comes from each of the first few terms. When we raise this sum to a power $\Theta(\log N)$ this means that a constant fraction of the result comes from repeated terms. In particular, we see that

$$\begin{aligned} &\sum_{L\subset\{c+1,\ldots,a-1\},|L|=k_1} \prod_{i\in L}\frac{1}{2i-1} \\ &\quad = \eta_1\frac{1}{k_1!}\left(\sum_{i=c+1}^{a-1}\frac{1}{2i-1}\right)^{k_1} = \eta_2\frac{[\log(a/c)]^{k_1}}{2^{k_1}k_1!}, \end{aligned}$$

where the "error factors" η_j are functions of c, a and k_1, and using $a>N^{\epsilon}$, $c=O(1)$, and $k_1=O(\log N)$ we have η_1, $\eta_2=\Theta(1)$.

The approximation above is more than enough to give our first result. Indeed, combining the formulas above,

$$\begin{aligned} E_k &\sim \sum^{*} \frac{1}{(2c-1)\sqrt{ab}} \sum_{k_1+k_2=k-2} \eta_2 \\ &\quad \times \frac{[\log(b/a)]^{k_2}[2\log(a/c)]^{k_1}}{2^{k-2}k_1!k_2!} + O(N^{1+\epsilon}). \end{aligned} \tag{7}$$

Thus, recalling that $\eta_2=\Theta(1)$, the binomial theorem implies that

$$E_k = \Theta\left(\sum^{*}\frac{1}{c\sqrt{ab}}\frac{[\log(ab/c^2)]^{k-2}}{2^k(k-2)!}\right) + O(N^{1+\epsilon}).$$

The rest is a matter of straightforward calculation. Fixing a and b, bounding the sum

$$\sum_{1\leq c\leq C}\frac{[\log(ab/c^2)]^{k-2}}{c}$$

above and below by suitable integrals, we see that this sum is $\Theta([\log(ab)]^{k-1}/k)$. Moreover, as $\log(ab)$ and k are both $\Theta(\log n)$, terms with increasing values of c decrease exponentially, provided c is not too large. This justifies our restriction to $c=O(1)$. We thus have

$$\begin{aligned} E_k &= \Theta\left(\frac{1}{2^k(k-1)!}\sum_{1\leq a<b\leq N}\frac{1}{\sqrt{ab}}[\log(ab)]^{k-1}\right) \\ &\quad + O(N^{1+\epsilon}). \end{aligned}$$

Again the sum can be approximated within a constant factor by an integral, namely by

$$\begin{aligned} &N\int_0^1\int_0^1 \frac{[\log(N^2xy)]^{k-1}}{\sqrt{xy}}\,dx\,dy \\ &\quad = \Theta(N[\log(N^2)]^{k-1}) = \Theta(2^kN(\log N)^{k-1}), \end{aligned}$$

using $k=\Theta(\log N)$. This gives

$$E_k = \Theta\left(N \frac{(\log N)^{k-1}}{(k-1)!} \right) + O(N^{1+\epsilon}),$$

or, since $k = \Theta(\log N)$,

$$E_k = \Theta\left(N \frac{(\log N)^{k}}{k!} \right) + O(N^{1+\epsilon}). \tag{8}$$

Using Stirling's formula,

$$E_k = \Theta\left(\frac{N}{\sqrt{\log N}} \left(\frac{e \log N}{k} \right)^k \right) + O(N^{1+\epsilon}).$$

Writing $k = \alpha \log N$, for $0 < \alpha < e$, taking ϵ small enough we obtain

$$E_k = \Theta(N^{1+\alpha \log(e/\alpha)} / \sqrt{\log N}),$$

providing Eq. (3).

Having proved Eq. (3), we have shown that the bulk of the distribution is at $\log N$, i.e., that *all but* $o(N^2)$ *of the* $\binom{N}{2}$ *paths have length* $k \sim \log N$. One can check that for k in this range, the error function $\eta_2 = \eta_2(c,a,k)$ satisfies $\eta_2(c,a,k) \sim \eta_3(c, \log a / \log N)$, where η_3 is continuous in the second argument. Following through the calculation from Eq. (7) to Eq. (8), one can check that the implicit constant in the $\Theta(\cdot)$ notation in Eq. (8) is almost constant and so for $k \sim \log N$ we have

$$E_k \sim \eta N \frac{(\log N)^k}{k!}, \tag{9}$$

for some absolute constant η. Summing over k in the range $k \sim \log N$, we count almost all of the $\binom{N}{2}$ paths, so

$$\binom{N}{2} \sim \sum_{k \sim \log N} E_k \sim \eta N^2.$$

Thus $\eta = 1/2$. The formula given in Eq. (4) now follows from Eq. (9) using Stirling's formula, for example. □

A. Comparison with heuristics

The distribution of shortest path lengths in T_N is studied by Szabó, Alava, and Kertész [6] using the mean-field approximation. It turns out that this heuristic gives very good results; unfortunately, only numerical results are given in [6], despite the impression given in the abstract, introduction and conclusions. In the body of the paper, the distribution of root-node distances is given, in greater generality. Using Eq. (5) in [6], i.e.,

$$b(l) = \frac{1}{2} \frac{\log N}{l}, \tag{10}$$

where $b(l)$ is the mean-field number of children of a vertex at distance l from the root, the approximation

$$n(l) = \Theta(n^{1/2}) \left(\frac{e \log N}{2(l-1)} \right)^{l-1} \tag{11}$$

is obtained for the number of vertices at distance l from the root. This formula is (6) in [6], specialized to the particular model; in particular, we have substituted the value $\Theta(n^{1/2})$ for $b(0) = l(1)$. It is also noted in [6] that the distribution is asymptotically normal, with mean and variance $\log N/2$ (again specializing the more general result to T_N).

No corresponding formula is given in [6] for node-node distances; indeed, it is stated that, in contrast to the root-node distances, ''Eq. (13) and the quantities it is constructed out of turn out to be too complex to handle without numerics.'' However, it is noted later that the main contribution arises from paths passing through the root, leading to a ''convolution-type distribution.'' This heuristic is a good one: nearly all paths pass very close to the root. In particular, asymptotic normality of the node-node distance distribution does follow (heuristically, but this is a very strong heuristic): the convolution of two normal distributions is normal.

Actually, the heuristic goes much further: rather than use the approximation (11), starting from Eq. (10) we obtain

$$n(l) = b(0) \frac{(\log N/2)^{l-1}}{(l-1)!}.$$

Since

$$\sum_{r+s=t} \frac{A^r}{r!} \frac{A^s}{s!} = \frac{(A+A)^t}{t!}$$

by the binomial theorem, this suggests by convolution that the number of paths of length k will be given essentially by

$$\Theta(N) \frac{(\log N)^k}{k!}.$$

[We ignore additive errors of $O(1)$ in the path length.] This is more or less the form given in theorem 1, apart from the normalization. And if, as in [6], one corrects the heuristic by normalizing [the heuristic, together with the real value of $b(0)$, suggests this should not be necessary, but it seems sensible anyway], this agreement is very striking. Note, however, that from the heuristic we have no idea how far the agreement will extend; from our precise work we see that the (heuristically constant) implicit constant in the Θ notation in theorem 1 does actually vary as k varies on the $\log N$ scale; this is due to the fact that paths not passing through the root matter very much when looking at the much smaller numbers of very short paths, for example.

B. General graphs

Let us conclude by very briefly considering the general case $m \geqslant 2$ of the BA model, giving rise to a scale-free random graph that is not a tree. A precise version of this model, the LCD model, was introduced in [7] (see also [3]), where it was shown that for $m \geqslant 2$ the diameter is asymptotically $\log N / \log \log N$ rather than $\log N$. Thus, as far as shortest paths are concerned, the general case behaves very differently from the tree case $m = 1$. The general case is likely to be much harder to analyze precisely.

ACKNOWLEDGMENTS

This research was supported by NSF Grant No. ITR 0225610 and DARPA Grant No. F33615-01-C-1900.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. Dorogovtsev and J. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] B. Bollobás and O. Riordan, in *Handbook of Graphs and Networks*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Weinheim, 2002), pp. 1–34.
[4] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[5] J. Szymański, in *Random Graphs '85 (Poznań, 1985)*, North-Holland Math. Stud. Vol. 144 (North-Holland, Amsterdam, 1987), pp. 297–306.
[6] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E **66**, 026101 (2002).
[7] B. Bollobás and O. Riordan, Combinatorica (to be published).
[8] H. Mahmoud and R. Smythe, Theor. Probab. Math. Statistics **51**, 1 (1995).
[9] F. Bergeron, P. Flajolet, and B. Salvy, in *CAAP '92 (Rennes, 1992)*, Lecture Notes in Comput. Sci. Vol. 581 (Springer, Berlin, 1992), pp. 24–48.
[10] H. Mahmoud, J. Comput. Appl. Math. **41**, 237 (1992).
[11] H. Mahmoud, R. Smythe, and J. Szymański, Random Struct. Algorithms **4**, 151 (1993).
[12] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[13] K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, Proc. Natl. Acad. Sci. USA **99**, 12583 (2002).
[14] B. Pittel, Random Struct. Algorithms **5**, 337 (1994).
[15] R. Dobrow, J. Appl. Probab. **33**, 749 (1996).
[16] R. Dobrow and R. Smythe, Random Struct. Algorithms **9**, 79 (1996).