Probability in the Engineering and Informational Sciences, **35**, 2021, 276–296.
doi:10.1017/S0269964819000342

SMALL-WORLD EFFECT IN GEOGRAPHICAL ATTACHMENT NETWORKS

QUNQIANG FENG, YONGKANG WANG, AND ZHISHUI HU

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, 230026, China

E-mails: fengqq@ustc.edu.cn; tjwyk@mail.ustc.edu.cn; huzs@ustc.edu.cn

In this work, we use rigorous probabilistic methods to study the asymptotic degree distribution, clustering coefficient, and diameter of geographical attachment networks. As a type of small-world network model, these networks were first proposed in the physical literature, where they were analyzed only with heuristic arguments and computational simulations.

Keywords: degree distribution, diameter, geographical attachment, random networks, small-world

Jel Classification: Primary: 05C80, 90B15, Secondary: 60J85

1. INTRODUCTION

The small-world effect can be readily found in many real-world networks such as theWorld Wide Web (WWW) [8], networks of scientific collaboration [20], protein networks [26], social networks [27], metabolic networks [6], communication networks [29], and so on. An earlier famous example is the small-world experiment of social psychologist Stanley Milgram, who was one of the first to point out the existence of small-world effects in social networks [18]. He performed experiments in 1960s, which suggested that there are only about six intermediate acquaintances separating any two people on the planet. The discovery of small-world effect has enabled researchers to understand, improve, and manipulate these networks for useful applications [30]. For instance, the spread of diseases or epidemics has been well studied in social networks and small-world effect has been used to devise strategies to isolate and quarantine infected communities [19].

A large number of network models, deterministic or random, have been proposed to generate artificial networks that exhibit the small-world effect. The first and the most widely-studied model is the simple and attractive small-world network model of Watts and Strogatz (WS model) [28], which triggered a sharp interest in the studies of the different properties of small-world networks. The WS model is constructed by randomly rewiring each edge in a ring lattice. Then it is a *static* network model, i.e., the initial state has a fixed number of nodes and the network size does not grow. Several variations of the WS model and other small-world network models have been also proposed in the literature (see, e.g., [15,22]).

© Cambridge University Press 2019 0269-9648/19 $25.00

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

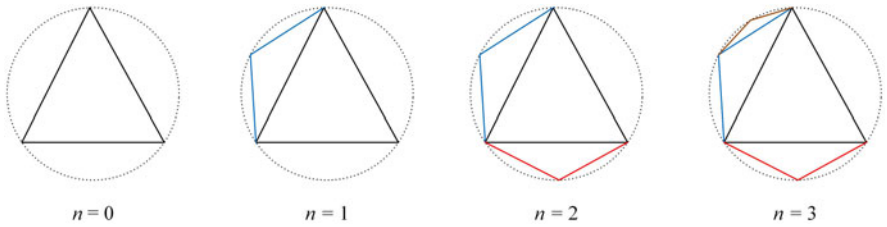


FIGURE **1.** Illustration of a geographical attachment network until time $n = 3$.

Many real-world networks have their topology influenced by geographical constraints. The nodes are separated by some physical distance and thus their ability to know the complete state of all the network nodes at a given time is restricted. In particular, we are interested in *dynamic* network models with the restriction on geographical distance. That is, the networks could grow in time from small size to large size by the successive addition of a new node, which is only connected to those preexisting nodes that are geographically close to it.

Also with the help of a ring, the *geographical attachment network* (GAN) model was first proposed in the physics literature [23], where the small-world effect of this model was analyzed with heuristic arguments and computational simulations.

We now precisely describe the GAN model, which can be generated in an iterative manner. Let $m \geq 2$ be an even integer, and GAN(n) denote the network at time $n \geq 0$. Then the network can be generated in the following way. We begin with an initial state (time $n = 0$) of $m + 1$ nodes distributed on a ring all of which are connected to one another. At time $n \geq 1$, the network GAN(n) is obtained from GAN($n - 1$) as follows: A new node is placed in an internode interval chosen uniformly at random from the $n + m$ existing ones along the ring, and connected to its m nearest neighbors ($\frac{m}{2}$ on either side). Distance, in this case, refers to the *geographical distance*, i.e., the number of internode intervals along the ring. Figure 1 shows an illustration of this network for a special case of $m = 2$ at time $n = 0,1,2$, and 3. Another randomized version of the GAN model is proposed in [31], whereas a deterministic GAN model is also studied in [32]. For other geographical networks, we refer the reader to [13] for an extensive survey.

To describe several elementary properties on the topology of GAN model, we need some graph-theory terminology, which can be found in standard texts. Let G be a simple graph (or network) with the node set V and the edge set E. A pair of nodes $u, v \in V$ are said to be *adjacent* in G, if there exists an edge between them. The *degree* of node v is the number of its adjacent nodes. A *path* in G is a sequence of edges which connect a sequence of nodes which are all distinct from one another(except, possibly, the first and last node). The *length* of a path is the number of edges in it. A *cycle* is a closed path, i.e., the first node in this path is also the last one. A graph is *connected* if, for each pair of distinct nodes, there is a path from one to the other; a graph which is not connected is *disconnected*. Graph G is said to be *planar* if it can be embedded in the plane, i.e., it can be drawn on the plane in such a way that its edges intersect only at their nodes. Then, one can observe a few of simple facts on the GAN model as follows. There are exactly $n + m + 1$ nodes and $(n + 1)m + 1$ edges at time n in the GAN model, since at each increment time one node and m edges are added into the network. Also for any $n \geq 0$ and even number $m \geq 2$, the network GAN(n) is a *Hamiltonian* graph, i.e., there exists a cycle that visits each node exactly once (note that the ring helps us to see this fact). However, we argue that the GAN model with large m is much more complicated to handle than the simplest case $m = 2$. For instance, only the case

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

$m = 2$ is a planar model, and in this case if n is not too small, by deletion of the three initial nodes the networks are always disconnected. Moreover, we remark that for the GAN model in the special case $m = 2$, the generating rule is similar to random Apollonian networks [33], but the latter model has the scale-free property, i.e., its asymptotic degree distribution is a power law [16], compared with the comments below the proof of Theorem 1.

The main contribution of this work is to analyze the small-world effect in the GAN model with rigorous probabilistic methods, which are significantly different from those in [23]. It is not surprising that our results stated here in theorems and propositions agree with those in [23]. More precisely, we establish several limit laws for the asymptotic degree distribution, clustering coefficient, and diameter (see their definitions in Section 2 below) of the network GAN(n), as time n grows to infinity. These results show that the GAN model meets the criteria on the small-world effect in the literature (see e.g., [28]), when its network size is large enough.

The rest of the paper is organized as follows. In the next section, we present briefly several basic definitions and technical preliminaries. In Section 3, we obtain that the asymptotic degree distribution of the GANs is related to a geometric distribution, and briefly discuss the clustering coefficient of the GAN network in the simplest case $m = 2$. Finally, in Section 4, we give a sharp estimate for the diameter of a typical GAN also in this special case.

2. PRELIMINARIES

First of all, we introduce several useful definitions that are usually used to measure the network topology. The *degree distribution* in a network (graph) is defined to be the fraction of nodes with degree $k \geq 0$. In other words, for any $k \geq 0$, the degree distribution shows the probability that the degree of a node picked uniformly at random in the network is k. The *clustering coefficient* of a node is the proportion of the number of existing edges between its neighbors, compared to the number of all possible edges between them. The clustering coefficient of the whole network is defined as the average of clustering coefficients over all the nodes. For any two given nodes n_1 and n_2 in a network, their *graph distance* (or *geodesic distance*), denoted by $d(n_1, n_2)$, is the minimal length of the paths between them. In a network, the *average path length* (or *characteristic path length*) is the average of graph distances between all possible pairs of nodes, and the *diameter* is the largest distances for all pairs of nodes.

The small-world networks always have a couple of basic structural properties [21]: small average path length and high clustering coefficient. With an additional property of node degrees, Ozik, Hunt and Ott [23] identified three criteria on the small-world effect for dynamic network models as the network size $N \to \infty$: (a) the average node degree is $O(1)$; (b) the average path length is proportional to $\ln N$; and (c) the average clustering coefficient does not decay with N, and thus is $O(1)$. As mentioned above, experimentally and heuristically with lack of rigor they have shown that the GAN model meets these criteria.

Throughout this work, we shall use the notation listed in Table 1, where a summarizing phrase is provided for each item.

To study the asymptotic degree distribution of the network GAN(n), we shall use the Azuma–Hoeffding inequality (see [2] or [14]).

LEMMA 1: *If the sequence* $\{M_k, k \geq 0\}$ *is a martingale such that* $|M_k - M_{k-1}| \leq c$ *holds almost surely for some constants* $c, k = 1, 2, \ldots, n$. *Then for any* $\varepsilon > 0$,

$$\mathbb{P}\left(|M_n - \mathbb{E}[M_n]| \geq \varepsilon\right) \leq 2 \exp\left\{-\frac{\varepsilon^2}{2nc^2}\right\}, \quad n \geq 1.$$

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

TABLE 1. The notation used throughout this work.

Notation	Definition
m	even; the model parameter
log	natural logarithm in base the constant e
$\|G\|$	size of network/graph G; the number of nodes in G
$\mathbf{1}_{\mathcal{E}}$	the indicator of event $\mathcal{E}$
w.h.p.	with high probability; events $\{\mathcal{E}_n, n \geq 1\}$ occurs w.h.p., if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$
Exp(λ)	the exponential distribution with mean $1/\lambda > 0$
Gamma(k)	the gamma distribution with parameter integer $k \geq 1$; the distribution of the sum of k independent Exp(1) random variables
Beta(α, β)	the beta distribution with parameters α and β
$X \preceq Y$	random variable X is stochastically smaller than Y; $\mathbb{P}(X > x) \leq \mathbb{P}(Y > x)$ for all x
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
GAN(n)	geographical attachment network at time n
$\mathscr{F}_n$	the σ-field generated by the structure of GANs up to time n
$d(\cdot,\cdot)$	graph distance of a pair of nodes in a network/graph
$X_{k,n}$	number of nodes with degree k in GAN(n)
$P_{k,n}$	empirical degree distribution in GAN(n)
C_n	clustering coefficient of GAN(n) with parameter $m = 2$
diam(n)	diameter of GAN(n) with parameter $m = 2$

The next lemma, which is a special case of [9, Thm. 1], plays an important role in our analysis of the diameter of the GANs in the special case $m = 2$. Before stating this result, we introduce some necessary notation. Let $b \geq 1$ be an integer, and T_∞ an infinite rooted b-ary tree. Namely, T_∞ has a distinguished node as its root, so that the number of children of each node in T_∞ is exactly b. We also let E be a non-negative, non-degenerate random variable that satisfies $\mathbb{P}(E = 0) = 0$, $\inf\{x : \mathbb{P}(E > x) > 0\} = 0$, and its moment generating function $M(\lambda) = \mathbb{E}[\exp\{\lambda E\}]$ exists for some $\lambda > 0$. Let $\mathbf{B}$ denote a b-dimensional random vector with each element distributed as E (but not necessarily independent). For each node u in T_∞, we label the edges to its children using an independent copy of $\mathbf{B}$. Then one can obtain the following result (see also Theorem 5.4 in [10]).

LEMMA 2: *For any $t \geq 0$, let H_t be the height of the subtree of T_∞ induced by the nodes for which the sum of the labels on their path to the root is at most t. Then as $t \to \infty$, the height H_t is w.h.p. asymptotic to ct, where the constant c is the unique solution to*

$$\sup\left\{\frac{\lambda}{c} - \log\left(M(\lambda)\right) : \lambda \leq 0\right\} = \log b.$$

3. ASYMPTOTIC DEGREE DISTRIBUTION

In this section, we consider the asymptotic degree distribution of the geographical attachment networks. For the sake of convenience, we label the nodes in the network as follows. The $m+1$ initial nodes are respectively labeled $0, 1, \ldots, m$ (e.g., in clockwise order); the node added in the network at time $n \geq 1$ is labeled with $n + m$. Note that the degree of any node is at least m at any time in this network. For any $k \geq m$, let $X_{k,n}$ be the number of

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

nodes with degree k in the network GAN(n). Then this quantity can be expressed as

$$X_{k,n} = \sum_{i=0}^{n+m} \mathbf{1}_{\{D_i(n)=k\}},$$

where $D_i(n)$ denotes the degree of node i in the network GAN(n), and the empirical degree distribution at time $n \geq 0$ is

$$P_{k,n} = \frac{X_{k,n}}{n+m+1}, \quad k = m, m+1, \ldots. \tag{1}$$

To compute the mean of $X_{k,n}$, we first consider the case $k = m$. At any time $n \geq 0$, as mentioned in Section 1, there are exactly $n+m+1$ nodes on the ring. Accordingly, the same number of internode intervals are induced at that time. Conditioning on $\mathscr{F}_{n-1}$, the σ-field generated by the structure of GANs up to time $n-1$, the degree of a node is unchanged in GAN(n), if and only if none of its m nearest internode intervals are chosen at time n. It follows that for any $n \geq 1$,

$$\mathbb{E}[\mathbf{1}_{\{D_i(n)=m\}}|\mathscr{F}_{n-1}] = \frac{n}{n+m}\mathbf{1}_{\{D_i(n-1)=m\}}, \quad 0 \leq i \leq n+m-1. \tag{2}$$

Note the simple fact that the new node added at time n also has degree m. We thus have

$$\mathbb{E}[X_{m,n}] = \sum_{i=0}^{n+m} \mathbb{E}[\mathbf{1}_{\{D_i(n)=m\}}] = 1 + \frac{n}{n+m}\mathbb{E}[X_{m,n-1}].$$

With the initial condition $X_{m,0} = m+1$, this recurrence gives that

$$\mathbb{E}[X_{m,n}] = \frac{n+m+1}{m+1} + \frac{m}{\binom{n+m}{m}}, \quad n \geq 0. \tag{3}$$

Analogously to (2), for the case $k \geq m+1$ we have the simple relation

$$\mathbb{E}[\mathbf{1}_{\{D_i(n)=k\}}|\mathscr{F}_{n-1}] = \frac{n}{n+m}\mathbf{1}_{\{D_i(n-1)=k\}} + \frac{m}{n+m}\mathbf{1}_{\{D_i(n-1)=k-1\}},$$

from which one can get that for $k \geq m+1$,

$$\mathbb{E}[X_{k,n}] = \frac{n}{n+m}\mathbb{E}[X_{k,n-1}] + \frac{m}{n+m}\mathbb{E}[X_{k-1,n-1}], \quad n \geq 1. \tag{4}$$

Using the recurrence (4) with the initial condition (3), we now can obtain the explicit formula of $\mathbb{E}[X_{k,n}]$ for general $k \geq m$. For any integers $0 \leq j \leq n$, let $H_n^{(0)} \equiv 1$,

$$H_n^{(j)} = \sum_{i=j}^{n} \frac{H_{i-1}^{(j-1)}}{i}, \quad j = 1, 2, \ldots, n.$$

Then the result for the mean of $X_{k,n}$ is the following.

PROPOSITION 1: *Let $X_{k,n}$ denote the number of nodes with degree k in the network GAN(n). Then for any $k \geq m$ and $n \geq k-m$, we have*

$$\mathbb{E}[X_{k,n}] = \frac{m^{k-m}(n+m+1)}{(m+1)^{k-m+1}} + \frac{m^{k-m}}{\binom{n+m}{m}}\left(mH_n^{(k-m)} - \frac{\sum_{i=0}^{k-m-1}(m+1)^i H_n^{(i)}}{(m+1)^{k-m}}\right), \tag{5}$$

where $\binom{n}{m}$ is the usual binomial coefficient for integers n and m.

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

PROOF: We prove (5) by induction for all $k \geq m$ on $n \geq 0$. When $n=0$, all $m+1$ nodes in the initial network GAN(0) have exactly degree m. That is, we have that $X_{m,0} = m+1$ and $X_{k,0} = 0$ for all $k \neq m$. This initializes the induction.

To advance the induction, we take $n \geq 1$, and note that (5) holds for the case $k=m$ by (3). Thus, we may assume that $k \geq m+1$. We also note that, by the definition of $H_n^{(j)}$,

$$H_{n-1}^{(j)} + \frac{H_{n-1}^{(j-1)}}{n} = H_n^{(j)}, \quad j \geq 0, \tag{6}$$

where, by convention, we define that $H_n^{(j)} = 0$ for $j=-1$ and all $j \geq n+1$. By (4), we apply the induction hypothesis both to $\mathbb{E}[X_{k,n-1}]$ and $\mathbb{E}[X_{k-1,n-1}]$, to obtain that for $k \geq m+1$,

$$\begin{aligned}
\mathbb{E}[X_{k,n}] &= \frac{n}{n+m}\mathbb{E}[X_{k,n-1}] + \frac{m}{n+m}\mathbb{E}[X_{k-1,n-1}] \\
&= \frac{n}{n+m}\left[\frac{m^{k-m}(n+m)}{(m+1)^{k-m+1}} + \frac{m^{k-m}}{\binom{n+m-1}{m}}\left(mH_{n-1}^{(k-m)} - \frac{\sum_{i=0}^{k-m-1}(m+1)^i H_{n-1}^{(i)}}{(m+1)^{k-m}}\right)\right] \\
&\quad + \frac{m}{n+m}\left[\frac{m^{k-m-1}(n+m)}{(m+1)^{k-m}} + \frac{m^{k-m-1}}{\binom{n+m-1}{m}}\right. \\
&\quad \left.\times\left(mH_{n-1}^{(k-m-1)} - \frac{\sum_{i=0}^{k-m-2}(m+1)^i H_{n-1}^{(i)}}{(m+1)^{k-m-1}}\right)\right] \\
&= \frac{m^{k-m}(n+m+1)}{(m+1)^{k-m+1}} + \frac{m^{k-m}}{\binom{n+m}{m}}\left(mH_{n-1}^{(k-m)} + \frac{m}{n}H_{n-1}^{(k-m-1)}\right. \\
&\quad \left. - \frac{\sum_{i=0}^{k-m-1}(m+1)^i H_{n-1}^{(i)}}{(m+1)^{k-m}} - \frac{\sum_{i=0}^{k-m-2}(m+1)^i H_{n-1}^{(i)}}{n(m+1)^{k-m-1}}\right).
\end{aligned}$$

Since the last term in the above formula can be rewritten as

$$\frac{\sum_{i=0}^{k-m-2}(m+1)^i H_{n-1}^{(i)}}{n(m+1)^{k-m-1}} = \frac{\sum_{i=1}^{k-m-1}(m+1)^{i-1} H_{n-1}^{(i-1)}}{n(m+1)^{k-m-1}} = \frac{\sum_{i=0}^{k-m-1}(m+1)^i H_{n-1}^{(i-1)}}{n(m+1)^{k-m}},$$

we have

$$\begin{aligned}
\mathbb{E}[X_{k,n}] = \frac{m^{k-m}(n+m+1)}{(m+1)^{k-m+1}} + \frac{m^{k-m}}{\binom{n+m}{m}}\left[m\left(H_{n-1}^{(k-m)} + \frac{H_{n-1}^{(k-m-1)}}{n}\right)\right. \\
\left. - \frac{\sum_{i=0}^{k-m-1}(m+1)^i (H_{n-1}^{(i)} + H_{n-1}^{(i-1)}/n)}{(m+1)^{k-m}}\right],
\end{aligned}$$

from which, together with (6), one can see (5) holds for $n \geq 1$. This advances the induction, and completes the proof of Proposition 1. ■

From Proposition 1, one can see that

$$\lim_{n\to\infty} \frac{\mathbb{E}[X_{k,n}]}{n+m+1} = \frac{m^{k-m}}{(m+1)^{k-m+1}}, \quad k = m, m+1, \ldots.$$

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

This result coincides with that in [23]. More precisely, we have that for any fixed $k \geq m$, as $n \to \infty$,

$$\left| \frac{\mathbb{E}[X_{k,n}]}{n+m+1} - \frac{m^{k-m}}{(m+1)^{k-m+1}} \right| = O\left(\frac{(\log n)^{k-m}}{n^{m+1}} \right). \tag{7}$$

In fact, by (5), it is sufficient to show that $H_n^{(j)} = O((\log n)^j)$ as $n \to \infty$, for any fixed integer $j \geq 1$. If $j = 1$, it follows that $H_n^{(1)}$ is the n-th harmonic number, i.e, $H_n^{(1)} = \sum_{i=1}^{n} \frac{1}{i} = O(\log n)$. Then, we have that for $j = 2$,

$$H_n^{(2)} = \sum_{i=2}^{n} \frac{H_i^{(1)}}{i} = O\left(\int_1^n \frac{\log x}{x} \mathrm{d}x \right) = O\left((\log n)^2 \right).$$

Similarly, we thus can get the desired result on $H_n^{(j)}$ for any $j \geq 3$.

The result characterizes the asymptotic degree distribution of the network GAN(n) in the following.

THEOREM 1: *Let $P_{k,n}$ denote the empirical proportion of nodes with degree k in the network GAN(n). Then for any constant $c > 4\sqrt{2}(m-1)$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k \geq m} \left| P_{k,n} - \frac{m^{k-m}}{(m+1)^{k-m+1}} \right| \geq c\sqrt{\frac{\log n}{n}} \right) = 0.$$

The method we use in the proof of Theorem 1 is an adaption of the standard martingale method and is similar to that in [7,12].

PROOF: For any fixed $k \geq m$, we denote

$$M_i := \mathbb{E}[X_{k,n} | \mathscr{F}_i], \quad i = 0, 1, 2 \ldots, n.$$

Then the sequence $\{M_i\}_{i=0}^n$ is a Doob-type martingale with respect to the filtration $\{\mathscr{F}_i\}_{i=0}^n$. In particular, we have that $M_0 = \mathbb{E}[X_{k,n}]$ and $M_n = X_{k,n}$. For the martingale difference M_i and M_{i-1}, we shall show that for any $1 \leq i \leq n$,

$$|M_i - M_{i-1}| \leq 4(m-1). \tag{8}$$

To prove (8), here we use the method of coupling. Let us consider the two different realizations G and G' of generating GAN(n) as follows. At time $i-1$ the networks are identical, but at time i two distinct internode intervals are chosen for them, respectively. As a result, the two networks G and G' are different from time i on. Suppose that at time i, the node labeled $i+2$ (i.e., the new node added to G and G' at time i) is connected to nodes $n_1, n_2, \ldots, n_m$ in G, and to nodes $n_1', n_2', \ldots, n_m'$ in G'. Denote the collection $\mathcal{N}_i = \{n_1, n_2, \ldots, n_m, n_1', n_2', \ldots, n_m'\}$, which may contain the same elements. Without loss of generality, we let the nodes $n_1, n_2, \ldots, n_{\frac{m}{2}}, i+2, n_{\frac{m}{2}+1}, \ldots, n_m$ be in clockwise order in G, and denote by $\overline{\mathcal{N}}_i$ the collection of the $\frac{m}{2}-1$ nearest neighbors of n_1 in the anticlockwise direction, together with the $\frac{m}{2}-1$ nearest neighbors of n_m in the clockwise direction in G. We also do the same assumption and definition in G'. At time $i+1$, if an internode interval, of which both endpoints are not in $\mathcal{N}_i$, is chosen for G, we do the same choice for G'; and if an internode interval with at least one endpoint in $\mathcal{N}_i$ is chosen for G, we can make an isomorphic for G' (for instance, if the internode interval between n_1 and n_2 is chosen for G, we chose the internode interval between n_1' and n_2' for G'). Hence, the networks G and G'

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

at time $i+1$ can be coupled in such a way that each node j $(0 \le j \le i+3)$ has the same degree except for the nodes in the collection $\mathcal{N}_i^* = \mathcal{N}_i \cup \overline{\mathcal{N}}_i \cup \overline{\mathcal{N}}_i'$. Since the degrees of nodes added later than i are not affected by what happened before time i, we can extend this coupling up to time n such that only the nodes in $\mathcal{N}_i^*$ have different degrees. Note that the number of nodes in such a collection is at most $2m + 4(\frac{m}{2} - 1) = 4(m-1)$. Then taking expectation with respect to $\mathscr{F}_{i-1}$ shows that (8) holds.

Applying the Azuma–Hoeffding inequality, we now have that for any $\varepsilon > 0$,

$$\mathbb{P}\left(|X_{k,n} - \mathbb{E}[X_{k,n}]| > \varepsilon\right) = \mathbb{P}(|M_n - M_0| > \varepsilon) \le 2\exp\left\{-\frac{\varepsilon^2}{32(m-1)^2 n}\right\}.$$

Note that the above inequality holds for any $k \ge m$. Then, setting $\varepsilon = c\sqrt{n\log n}$ for some constant $c > 4\sqrt{2}(m-1)$, we have as $n \to \infty$,

$$\begin{aligned}\mathbb{P}\left(\max_{k\ge 2}|X_{k,n} - \mathbb{E}[X_{k,n}]| > c\sqrt{n\log n}\right) &\le \sum_{k=2}^{n+m}\mathbb{P}\left(|X_{k,n} - \mathbb{E}[X_{k,n}]| > c\sqrt{n\log n}\right)\\ &\le 2\exp\left\{\left(\frac{\log(n+m)}{\log n} - \frac{c^2}{32(m-1)^2}\right)\log n\right\}\\ &\to 0,\end{aligned}$$

which, by (1) and (7), implies that Theorem 1 holds. ■

An immediate consequence of the above theorem is that for any fixed $k \ge m$,

$$P_{k,n} \xrightarrow{p} \frac{m^{k-m}}{(m+1)^{k-m+1}}, \tag{9}$$

as $n \to \infty$. That is, the asymptotic degree distribution of GANs is a geometric distribution with parameter $1/(m+1)$, supported on the set $\{m, m+1, \ldots\}$. Obviously, the geometric distribution is not a power law tail, but an exponential tail. However, many small-world networks including the WS model belong to this type [21]. Furthermore, the expectation of the asymptotic degree distribution given in (9) is $2m$, which is also shown in [23]. This implies that for large n, the network GAN(n) is a sparse graph whose nodes have many fewer connections than possible.

By using the asymptotic degree distribution, we can now investigate rigorously the clustering coefficient of GANs for the special case $m = 2$. It is an elementary observation in this case that the clustering coefficient of any node with degree k is a constant, regardless of time n. In fact, when the degree of any node increases by 1 at an increment time, then the number of edges between its neighbors increases exactly by 1. This implies that c_k, the clustering coefficient of a node with degree k in GAN(n), is given by

$$c_k = \frac{(k-2)+1}{\binom{k}{2}} = \frac{2}{k}.$$

Proposition 2: *Let C_n be the clustering coefficient of the network GAN(n) in the special case $m = 2$. Then C_n converges in probability to $\frac{3}{2}\log 3 - 1 = 0.6479184\cdots$, as $n \to \infty$.*

Proof: By the definition of clustering coefficient, we have

$$C_n = \frac{1}{n+3}\sum_{k=2}^{n+2} c_k X_{k,n} = 2\sum_{k=2}^{n+2}\frac{P_{k,n}}{k}.$$

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

Note that the identity

$$\sum_{k=2}^{\infty} \frac{2}{k} \cdot \frac{2^{k-2}}{3^{k-1}} = \frac{3}{2} \log 3 - 1.$$

For any fixed integer $l > 2$, we thus have

$$\begin{aligned} \left| C_n - \frac{3}{2} \log 3 + 1 \right| &\le 2 \sum_{k=2}^{l} \frac{1}{k} \left| P_{k,n} - \frac{2^{k-2}}{3^{k-1}} \right| + 2 \sum_{k=l+1}^{n+2} \frac{P_{k,n}}{k} + 2 \sum_{k=l+1}^{\infty} \frac{2^{k-2}}{k \cdot 3^{k-1}} \\ &< 2 \sum_{k=2}^{l} \frac{1}{k} \left| P_{k,n} - \frac{2^{k-2}}{3^{k-1}} \right| + \frac{2}{l} + 2 \sum_{k=l+1}^{\infty} \frac{2^{k-2}}{k \cdot 3^{k-1}}. \end{aligned}$$

Letting $n \to \infty$, by (9) with $m = 2$ it follows that the first term on the right-hand side converges in probability to 0. And then letting $l \to \infty$ gives that both the second and third terms converge to 0. The proof of Proposition 2 is complete. ■

Since the clustering coefficient of any network only takes value in $[0, 1]$, the convergence result in Proposition 2 can be strengthened to convergence in L, by the dominated convergence theorem. Using a heuristic argument together with a simulation result, it is shown in [23] that the limit of $\mathbb{E}[C_n]$ is equal to $\frac{3}{2} \log 3 - 1$ as n grows to infinity, which, of course, agrees with Proposition 2. But here we give a rigorous proof of this result.

4. DIAMETER

Only using computational simulations for the case $m = 2$, even without any heuristic arguments, it was argued in [23] that the average path length of the network GAN(n) is proportional to the logarithm of its size. Instead of the average path length, we shall study the diameter of GANs in the special case $m = 2$. By definition, one can see that the diameter of a network is not smaller than its average path length. Indeed, the average path length and diameter of small-world networks are usually of the same order, i.e., both proportional to the logarithm of the network size (see, e.g., [25, Thm. 7.1 and 7.2]). For simplicity, for the GANs we only refer to this special case in this section.

In the description of the GAN model, the ring can help us to know that it is a planar graph. However, for the sake of convenience we can redefine this simplest case without the ring as follows: Start with an initial graph of a triangle, at each subsequent discrete time, we add a new node, and connect it to both the endpoints of an edge chosen uniformly at random from the edges that have not been chosen before. At any time n, we say an edge is *external* if it can be chosen to add a new node, and *internal* otherwise. Then, by this generating rule, one can see that there are only three external edges at the initial time $n = 0$, both of the numbers of the internal and external edges increase by 1 at each time $n \ge 1$, and each node has exactly two external edges adjacent to it. Then in the network GAN(n), there are exactly n internal edges and $n + 3$ external edges.

For a sequence of random variables $\{X_n\}$ and a function $f(n)$, we say X_n is w.h.p. asymptotic to $f(n)$ if for any fixed $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(1 - \varepsilon \le \frac{X_n}{f(n)} \le 1 + \varepsilon\right) = 1,$$

and we say X_n is w.h.p. $o(f(n))$ if $\mathbb{P}(X_n \le \varepsilon f(n)) = 1$ for any $\varepsilon > 0$, as $n \to \infty$. Our main result on the diameter of the network GAN(n) is first stated in the following.

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

THEOREM 2: *As $n \to \infty$, the diameter of the network GAN(n) in the special case $m = 2$ is w.h.p. asymptotic to $2c \log n$ with the constant*

$$c = \frac{x_0 - 1}{x_0} \left(\log \frac{2x_0^2}{1 - x_0} \right)^{-1} = 1.6738050\ldots,$$

where $x_0 = 0.2550568\ldots$ is the unique solution in the interval $(0, 1)$ to the equation

$$x - \log \frac{2x^2}{1 - x} = 2.$$

Our proof of this main theorem relies on the results in Broutin and Devroye [9], see the proofs of Lemmas 4 and 6 below. Also using those results in [9], the diameter of another network model, the so-called random Apollonian networks, is analyzed in [10,12]. Our method used here is an adaptation of theirs.

Precisely, we shall divide the proof of Theorem 2 into four steps. The details of each step will be provided in a subsection. In Subsection 4.1, we start by constructing three subnetworks so that the diameter of GAN(n) is w.h.p. asymptotic to twice of the radius of each subnetwork, as n grows to infinity. To obtain the asymptotics of the radius, in Subsection 4.2, we generate an auxiliary random tree according to the topology of the subnetwork, and give a correspondence relation between the radius of the subnetwork and the height of its auxiliary tree. For the height of auxiliary tree, in Subsection 4.3 we use coupling to show that it is stochastically bounded between the heights of two other related random trees, both of which are properly embedded into the evolutions of continuous-time branching processes. Finally, in Subsection 4.4, we give the asymptotics of the heights of the these two random trees in explicit forms as time grows to infinity, and show that both of them (divided by time) tend to a desired constant in some manner.

4.1. Subnetworks

Consider the initial network GAN(0), where the ring consists of three initial internode intervals, say $\mathbf{I}_1$, $\mathbf{I}_2$, and $\mathbf{I}_3$ (see Figure 2). We treat them as open intervals. In the network GAN(n), we denote by GAN$_i(n)$ the subgraph induced by the nodes in $\mathbf{I}_i$, including both the endpoints of $\mathbf{I}_i$. Consider the graph distance between any two nodes in distinct subnetworks. By the generating rule of our simplest GAN model, any path between them must contain one of the three original nodes (with label 0, 1, and 2). As a result, the diameter of the network GAN(n) is related to the maximum graph distance of subsequent nodes in each GAN$_i(n)$ to the endpoints 0, 1, and 2. In this subsection, we shall describe the relation between the diameter of GAN(n) and such a maximum graph distance.

Let $Y_{i,n}$ be the number of nodes in $\mathbf{I}_i$ at time n. Then, for any time $n \geq 0$,

$$Y_{1,n} + Y_{2,n} + Y_{3,n} = n,$$

and $|\text{GAN}_i(n)|$, the size of the subnetwork GAN$_i(n)$, is equal to $Y_{i,n} + 2$, where 2 refers to the two endpoints of $\mathbf{I}_i$.

To derive the asymptotic distribution of $Y_{i,n}$ ($i = 1, 2$ and 3), we can employ the well-studied Eggenberger–Pólya urn (see [11]). For a given i, let us color internode intervals inside $\mathbf{I}_i$ with white and all other internode intervals with blue. Consider the network GAN(0) to be the initial condition of an urn having 1 white and 2 blue balls. When a ball (white or blue) is picked, its corresponding interval is no longer an internode one. However, new two internode intervals are born to replace it. That is, if a ball is picked at any time, it will be

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

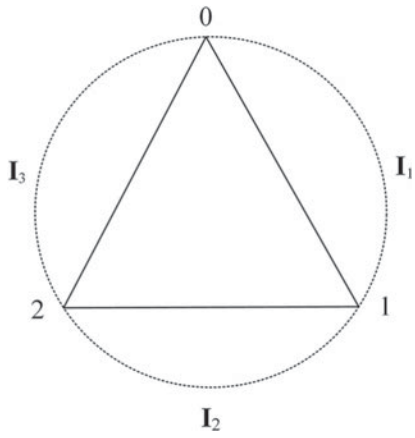


FIGURE **2.** Initial internode intervals in GAN(0).

replaced by two balls with the same color. Then, the corresponding Pólya urn scheme has the replacement matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This is the simplest case of the Eggenberger–Pólya urn. Note that after n draws, the number of white balls corresponds to the number of internode intervals in $\mathbf{I}_i$ in the network GAN(n) plus 1. Following the known result (see, e.g., [17, Thm. 3.2]), we thus have that

$$\frac{Y_{i,n}}{n} \xrightarrow{d} \text{Beta}(1,2), \quad i = 1,2,3,$$

as $n \to \infty$, where Beta(α, β) denotes the beta distribution with parameters α and β. Then for any constant $0 < \varepsilon < 1$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{Y_{i,n}}{n} \le \frac{\varepsilon}{2}\right) = 2\int_0^{\varepsilon/2} (1-x)dx = \varepsilon - \frac{\varepsilon^2}{4} < \varepsilon, \quad i = 1,2,3.$$

Fix a small constant $\delta > 0$. Hence, there exists a sufficiently small number $\varepsilon > 0$, such that for any given $i = 1,2$, and 3,

$$\mathbb{P}\left(\frac{Y_{i,n}}{n} < \varepsilon\right) < \frac{\delta}{3},$$

if n is sufficiently large. Let the event

$$E_n := \left\{\frac{\min\{Y_{1,n}, Y_{2,n}, Y_{3,n}\}}{n} \ge \varepsilon\right\}.$$

Then, by the union bound, it follows that for sufficiently large n,

$$\mathbb{P}(E_n) \ge 1 - \delta. \tag{10}$$

That is, w.h.p. the random variables $Y_{1,n}, Y_{2,n}$, and $Y_{3,n}$ have the same order of magnitude. As a result, for the limiting behavior of the graph distances between nodes in the subnetwork GAN$_i(n)$, we can only look at one of them, say GAN$_1(n)$, and assume that its size $|\text{GAN}_1(n)|$ is w.h.p. of order n.

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

In the subnetwork $\mathrm{GAN}_1(n)$, without loss of generality we assume that the nodes in the initial internode interval $\mathbf{I}_1$ are labeled with $3 \le n_1 < n_2 < \cdots < n_{Y_{1,n}} \le n+2$. Let us define the *radius* of $\mathrm{GAN}_1(n)$ by $\max_{1\le i\le Y_{1,n}} d_{\min}(n_i)$, where $d_{\min}(n_i) = \min\{d(n_i,0), d(n_i,1)\}$ denotes the minimum graph distance of n_i to the endpoints 0 and 1. The result on the radius of $\mathrm{GAN}_1(n)$ is given in the following.

LEMMA 3: *The radius of the subnetwork $GAN_1(n)$ is w.h.p. asymptotic to $c\log n$, where c is a constant defined in Theorem 2.*

We first show that this lemma implies Theorem 2.

Proof of Theorem 2: Let $\mathrm{diam}(n)$ denote the diameter of the network $\mathrm{GAN}(n)$. Fix any two small constants $\varepsilon, \delta > 0$. Analogously to (10), it follows by Lemma 3 that with probability at least $1-\delta$, the radius of the subnetwork $\mathrm{GAN}_i(n)$ is at least $(1-\varepsilon)c\log n$ for $i=1,2$ and 3. Therefore, with probability at least $1-\delta$ there exists a node $u_1 \in \{n_1, n_2, \ldots, n_{Y_{1,n}}\}$ such that both $d(u_1,0)$ and $d(u_1,1)$ are at least $(1-\varepsilon)c\log n$, and also there exists a node u_2 in $\mathbf{I}_2$ such that both $d(u_2,1)$ and $d(u_2,2)$ are at least $(1-\varepsilon)c\log n$. Note that any path between u_1 and u_2 must contain one of the three initial nodes 0, 1, and 2. Then, we have

$$\mathrm{diam}(n) \ge d(u_1,u_2) \ge 2(1-\varepsilon)c\log n,$$

which implies that

$$\mathbb{P}\left(\mathrm{diam}(n) \ge 2(1-\varepsilon)c\log n\right) \ge 1-\delta.$$

Consider the radii of $\mathrm{GAN}_2(n)$ and $\mathrm{GAN}_3(n)$, and note that Lemma 3 also holds for them. By the fact that the largest distances for all pairs of nodes in the whole network $\mathrm{GAN}(n)$ is at most twice of the maximum of the radii of the three subnetworks plus 1, then the desired statement for the upper bound follows by Lemma 3. ■

To complete the proof of Theorem 2, it is now sufficient to prove Lemma 3. As mentioned above, the proof of this lemma is divided into several parts, in each of which we also summarize the key point in a lemma (specifically, Lemma 4, 5, and 6).

4.2. Auxiliary Tree

As shown in Lemma 3, we can only focus on the typical subnetwork $\mathrm{GAN}_1(n)$. At the first step to prove Lemma 3, based on the evolving process of $\mathrm{GAN}_1(n)$ we shall generate a random binary tree T_n, which we call the *auxiliary tree*. We shall give the corresponding relation between the radius of $\mathrm{GAN}_1(n)$ and the height of T_n in this subsection.

To avoid confusion, in the rest of this work we use the term "vertices" for the nodes in trees, so that "nodes" refer to the nodes in our network model. There is a bijection between the edges in $\mathrm{GAN}_1(n)$ and the vertices in T_n. For any edge e in $\mathrm{GAN}_1(n)$, we denote by v^e its corresponding vertex in T_n. For convenience, the second definition of GAN model, given in the beginning of this section, is used here. Then according to the evolving process of the subnetwork $\mathrm{GAN}_1(n)$, the auxiliary tree T_n is generated as follows.

We start with a single vertex as its root, which corresponds to the initial external edge with endpoints 0 and 1 in $\mathrm{GAN}_1(0)$. Whenever an external edge e is chosen and then two external edges e_1 and e_2 with a common endpoint are added, the vertex v^e gives birth to two children v^{e_1} and v^{e_2}. One can extend the correspondence in the natural manner. Therefore, it follows that each vertex in T_n has either 0 or two children, and the leaves (or internal

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

vertices) in T_n correspond to the external (or internal) edges in $\mathrm{GAN}_1(n)$. Then there are exactly $Y_{1,n}$ internal vertices and $Y_{1,n}+1$ leaves in T_n, if $Y_{1,n}$ is given.

Let e be an edge in $\mathrm{GAN}_1(n)$ with endpoints u_1 and u_2, where $u_1, u_2 \in \{0, 1, n_1, n_2, \ldots, n_{Y_{1,n}}\}$. Recall the definition of $d_{\min}(u)$, which is equal to $\min\{d(u,0), d(u,1)\}$. It follows that $|d_{\min}(u_1) - d_{\min}(u_2)| \le 1$, since there is an edge between the two nodes. We say edge e is *balanced* if $d_{\min}(u_1) = d_{\min}(u_2)$, and *unbalanced* otherwise. The terms "balanced" and "unbalanced" are also used for the corresponding vertices in T_n.

At time $n=0$, there is only one edge in $\mathrm{GAN}_1(0)$ with endpoints 0 and 1 satisfying that $d_{\min}(0) = d_{\min}(1) = 0$, then this single edge is balanced. Consider that an external edge e with endpoints u_1 and u_2 in the subnetwork $\mathrm{GAN}_1(n)$ is chosen to add two new external edges e_1 (with endpoints u_1 and u_3) and e_2 (with endpoints u_2 and u_3). Without loss of generality, we assume that the nodes $0, u_1, u_3, u_2, 1$ are placed sequentially, in a clockwise direction, on a ring. When edge e is balanced, one can see that $d_{\min}(u_1) = d(u_1, 0) = d_{\min}(u_2) = d(u_2, 1)$. Since u_3 lies between u_1 and u_2, we have $d_{\min}(u_3) = d(u_3, 0) = d(u_3, 1) = d_{\min}(u_1) + 1 = d_{\min}(u_2) + 1$, which implies that both e_1 and e_2 are unbalanced. When e is unbalanced, on the other hand, we have that $|d_{\min}(u_1) - d_{\min}(u_2)| = 1$. If $d_{\min}(u_1) = d_{\min}(u_2) + 1$, it follows that $d_{\min}(u_3) = d(u_3, 1) = d(u_2, 1) + 1 = d_{\min}(u_1) = d_{\min}(u_2) + 1$. Then, at this time, edge e_2 is balanced while e_3 is unbalanced. Similarly, if $d_{\min}(u_2) = d_{\min}(u_1) + 1$, one can see that edge e_3 is balanced while e_2 is unbalanced. Equivalently, collecting these results above, we can observe several simple facts on the vertices in the auxiliary tree T_n as follows.

Claim 1: *In T_n, the following assertions hold.*

(a) The root is balanced;

(b) A balanced vertex has two unbalanced children;

(c) An unbalanced vertex has one balanced child and one unbalanced child.

For any edge e in $\mathrm{GAN}_1(n)$ with endpoints u_1 and u_2, we define

$$d_{\min}(e) = \min\{d_{\min}(u_1), d_{\min}(u_2)\}.$$

Consider two balanced edges e_1 and e_2 in $\mathrm{GAN}_1(n)$ such that v^{e_1} is the grandparent of v^{e_2} in T_n. Then Claim 1(b) indicates that $d_{\min}(e_2) = d_{\min}(e_1) + 1$, whereas Claim 1(c) indicates that for an unbalanced edge e_3, we have $d_{\min}(e_3) = d_{\min}(e_3^*)$, where $v^{e_3^*}$ is the parent of v^{e_3} in T_n. Thus, for any edge e in $\mathrm{GAN}_1(n)$, the distance $d_{\min}(e)$ is one less than the number of balanced vertices in the path from the root to the corresponding vertex v^e in T_n. From now on, we call $d_{\min}(e)$ the *auxiliary depth* of vertex v^e, and define the *auxiliary height* of T_n as the maximal auxiliary depth of its vertices, i.e.,

$$\mathrm{ah}(T_n) = \max_{v^e \in T_n} d_{\min}(e).$$

If an external edge e in the network is chosen to add a new node u, the simple relation $d_{\min}(u) = d_{\min}(e) + 1$ holds. Hence, we reach the following result.

Claim 2: *The radius of the subnetwork $GAN_1(n)$ is equal to either* $\mathrm{ah}(T_n)$ *or* $\mathrm{ah}(T_n) + 1$.

Instead of constructing T_n from the subnetwork $\mathrm{GAN}_1(n)$, we can also generate the random tree T_n in an alternative way. Start with a single root vertex, at each subsequent

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

discrete time, we choose a leaf uniformly at random and add two leaves as its children (then the chosen leaf turns into an internal vertex). The generating process is repeated until the tree reaches the desired size. Finally, the type of each vertex is assigned according to Claim 1(a–c). More precisely, let the root be a balanced vertex, and determine the types of other vertices in a top-down manner. For a balanced vertex, let its children be both unbalanced. And for an unbalanced vertex, choose one of its children independently and uniformly at random, let that child be balanced and the other unbalanced.

To obtain the limiting behavior of $\mathrm{ah}(T_n)$, the auxiliary height of the random tree T_n, it is now natural to embed the evolution of the T_n into the evolution of continuous-time branching process [1], or a Bellman–Harris process [4]. Embedding in such stochastic processes is a powerful tool to analyze the height of random trees (see, e.g., [5,24]). At initial time $t=0$, a single vertex (the root), which is balanced, is born. Whenever a vertex is born, after an Exp(1) time it gives birth to two children and dies, independently of all the other vertices. The types of vertices are assigned similarly to those in the alternative definition of T_n. Since the exponential distribution is memoryless, at any moment the next leaf to die is chosen uniformly at random. For any time $t \geq 0$, we denote by $T^{(t)}$ the tree obtained by the random process until time t. Then for any $t>0$, conditioning on the event that the sizes of trees $|T^{(t)}|$ and $|T_n|$ are equal, the two random trees $T^{(t)}$ and T_n have the same distribution.

The following lemma describes the relation between $\mathrm{ah}(T^{(t)})$ and the radius of the subnetwork $\mathrm{GAN}_1(n)$.

LEMMA 4: *Assume that there exists a constant $c>0$ such that $\mathrm{ah}(T^{(t)})$, the auxiliary height of $T^{(t)}$, is w.h.p. asymptotic to ct, as $t \to \infty$. Then the radius of the subnetwork $GAN_1(n)$ is w.h.p. asymptotic to $c \log n$, as $n \to \infty$.*

PROOF: Given $Y_{1,n} = n_0$ for a sufficiently large integer n_0. It follows that the tree T_n has exactly n_0 internal vertices and n_0+1 leaves, and thus of size $2n_0+1$. For any fixed $\varepsilon > 0$, we define

$$t_1 = (1-\varepsilon)\log(2n_0+1), \quad t_2 = (1+\varepsilon)\log(2n_0+1).$$

We also let random variable τ be a stopping time such that $T^{(\tau)}$ has exactly $2n_0+1$ vertices. Then a direct application of [9, Prop. 2] yields to that

$$\frac{1}{t}\log\left|T^{(t)}\right| \to 1$$

almost surely, as $t \to \infty$. This implies that w.h.p. as $n_0 \to \infty$,

$$\log\left|T^{(t_1)}\right| \sim t_1 = (1-\varepsilon)\log(2n_0+1)$$

holds, and thus $|T^{(t_1)}| < 2n_0+1$, or equivalently, the stopping time $\tau > t_1$. By symmetry, one can show that w.h.p. as $n_0 \to \infty$, we have $\tau < t_2$. Hence, it follows that w.h.p. as $n_0 \to \infty$,

$$\mathrm{ah}\left(T^{(t_1)}\right) \leq \mathrm{ah}\left(T^{(\tau)}\right) \leq \mathrm{ah}\left(T^{(t_2)}\right). \tag{11}$$

By the assumption, we have that w.h.p. as $n_0 \to \infty$,

$$\mathrm{ah}\left(T^{(t_1)}\right) \sim (1-\varepsilon)c\log(2n_0+1), \quad \mathrm{ah}\left(T^{(t_2)}\right) \sim (1+\varepsilon)c\log(2n_0+1).$$

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

Note that, as mentioned above, the random tree $T^{(\tau)}$ has the same distribution as T_n. By (11), we thus have that w.h.p. as $n_0 \to \infty$,

$$1 - \varepsilon \le \frac{\mathrm{ah}(T_n)}{c \log(2n_0 + 1)} \le 1 + \varepsilon.$$

Then, by Claim 2 and the fact that $\log Y_{1,n} = (1 + o(1)) \log n$ holds w.h.p., the arbitrariness of ϵ implies Lemma 4. ■

4.3. Stochastic Domination

It is difficult to directly verify the assumption in Lemma 4. For our purpose, we shall give a lower and an upper bound on the auxiliary height of $T^{(t)}$ in this subsection, and prove that the bounds are asymptotically equal eventually in the last step.

We now consider the random process defined above in the following equivalent way. Let T be an infinite binary tree, in which each vertex has a type assigned according to Claim 1(a–c), and is associated with an independent Exp(1) random variable as its lifetime. For later convenience, for each edge in T, from a parent to its child, we also assign a number that is equal to the lifetime of the parent. Then for any vertex $v \in T$, its *birth time* can be defined as the total sum of the labels of edges in the path from the root to v, and the birth time of the root is defined to be 0. Then, for any given $t \ge 0$, the tree $T^{(t)}$ is the subtree induced by the vertices with birth time at most t, and is finite with probability 1.

Let $k \ge 2$ be a fixed positive integer. We now define two random infinite trees $\overline{T}_k$ and $\widetilde{T}_k$ as follows. The tree $\overline{T}_k$ is obtained using the same generating rules as T except that each unbalanced vertex, whose distance to its closest balanced ancestor is exactly k, does not give birth to any children when it dies. The tree $\widetilde{T}_k$ is also generated as same as T, except that each unbalanced vertex, whose distance to its closest balanced ancestor is exactly k, has lifetime 0 instead of Exp(1), and immediately gives birth to three balanced children and dies. We say that such a vertex in $\widetilde{T}_k$ is *annoying*. For any given $t \ge 0$, the random trees $\overline{T}_k^{(t)}$ and $\widetilde{T}_k^{(t)}$ are defined similarly to $T^{(t)}$. The relation among the auxiliary heights of random trees $\overline{T}_k^{(t)}$, $T^{(t)}$, and $\widetilde{T}_k^{(t)}$ is the following.

LEMMA 5: *For any $t \ge 0$ and fixed $k \ge 2$, we have*

$$\mathrm{ah}\left(\overline{T}_k^{(t)}\right) \preceq \mathrm{ah}\left(T^{(t)}\right) \preceq \mathrm{ah}\left(\widetilde{T}_k^{(t)}\right).$$

PROOF: A useful tool to analyze stochastic domination is coupling. First, the labels of edges (or the lifetimes of vertices) in $\overline{T}_k^{(t)}$ and $T^{(t)}$ can easily be coupled using a common sequence of independent Exp(1) random variables in such a way that for any time $t \ge 0$, the tree $\overline{T}_k^{(t)}$ is a subtree of $T^{(t)}$. Therefore, we have $\mathrm{ah}(\overline{T}_k^{(t)}) \preceq \mathrm{ah}(T^{(t)})$.

For the coupling between the labels of edges in trees T and $\widetilde{T}_k$, it is enough to choose them using a common sequence of independent Exp(1) random variables, and define an injective mapping ϕ from vertices in T to those in $\widetilde{T}_k$ as follows. For each vertex u in T, the auxiliary depth of $\phi(u)$ is not less than the auxiliary depth of u, and the birth time of u is not less than the birth time of $\phi(u)$. For an unbalanced vertex, whose distance to its closest balanced ancestor is exactly k, the coupling and the mapping ϕ are shown down to their grandchildren in Figure 3, where the annoying vertex with children and grandchildren in T is shown on the left and the corresponding structure in $\widetilde{T}_k$ is shown on the right. The type of each vertex is written inside a circle ("B" for balanced and "U" for unbalanced).

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

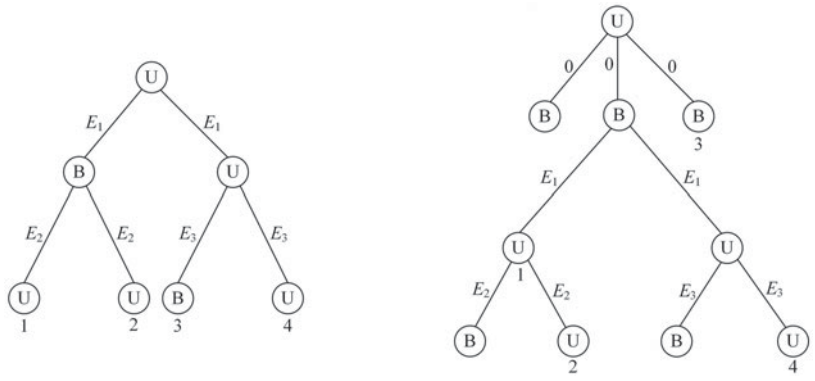


FIGURE **3.** Illustration of the coupling in Lemma 5 for an annoying vertex in T.

The coupling of labels of edges is labeled with $E_1, E_2, E_3 \cdots$, whereas the label 0 is used in the case of $\widetilde{T}_k$. The mapping ϕ is defined by the labels below the vertices (positive integers $1, 2, \ldots$). That is, each vertex with a label in T is mapped to the vertex with the same label in $\widetilde{T}_k$. This can be easily extended in a natural way to all other vertices in the tree. Then the relation $\mathrm{ah}(T^{(t)}) \preceq \mathrm{ah}(\widetilde{T}_k^{(t)})$ holds, which completes the proof of Lemma 5. ■

4.4. Asymptotics of Auxiliary Heights

Next we turn to study the auxiliary heights of these random trees. As part of our plan to prove Lemma 3, we would like to prove the following result.

LEMMA 6: *There exist constants $\overline{c}_k$ and $\widetilde{c}_k$, such that for any sufficiently large k, the auxiliary heights $\mathrm{ah}(\overline{T}_k^{(t)})$ and $\mathrm{ah}(\widetilde{T}_k^{(t)})$ are w.h.p. asymptotic to $\overline{c}_k t$ and $\widetilde{c}_k t$ as $t \to \infty$, respectively.*

For any $i \geq 1$, the ith level in a rooted tree, we refer to the set of all vertices that have the same depth i. The fact in the following will be used below.

CLAIM 3: *At any level $i \geq 2$ in T, there are exactly two balanced vertices and two unbalanced vertices such that the root is the only balanced vertex among their ancestors.*

PROOF: We prove this claim by induction on $i \geq 2$. By Claim 1(a and b), there are exactly two vertices, which are both unbalanced, at the first level in T. Then, Claim 3 holds for $i = 2$ by Claim 1(c). This initializes the induction. To advance the induction, it follows by the induction hypothesis that, at level $i-1$, there are exactly two unbalanced vertices such that the root is the only balanced vertex among their ancestors. Therefore, only four children of such two unbalanced vertices satisfy the desired condition, which implies that Claim 3 holds for i, and advances the induction. ■

By using only the balanced vertices in $\overline{T}_k$, we may generate a random infinite tree $\overline{T}_k'$ as follows. For any two vertices v_1 and v_2 in $\overline{T}_k'$ such that v_1 is the closest ancestor of v_2 in $\overline{T}_k$, there is an edge between v_1 and v_2 in $\overline{T}_k'$, whose label is equal to the sum of the labels of the edges in the unique path from v_1 to v_2 in $\overline{T}_k$. Then it follows by Claim 3 that

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

$\overline{T}'_k$ is an infinite $2(k-1)$-ary tree, and for any vertex u in $\overline{T}'_k$ and any $2 \le i \le k$, there are exactly two children whose birth times are equal to the birth time of u plus a Gamma(i) random variable. Moreover, by the generating rule of $\overline{T}'_k$, we have that for any $t \ge 0$, the height of the subtree of $\overline{T}'_k$ induced by vertices with birth time not greater than t is equal to $\mathrm{ah}(\overline{T}_k^{(t)})$, the auxiliary height of $\overline{T}_k^{(t)}$.

Recall that Gamma(j) denotes the gamma distribution with shape parameter j and rate parameter 1. Let $\overline{E}_k$ be a random variable distributed as Gamma(U_k), where U_k is a random variable, uniformly distributed on the set $\{2, 3, \ldots, k\}$. To apply Lemma 2, we need the label of each edge to have the same distribution. For this, starting from the root and in a top-down manner, we can randomly rearrange the ordering of children with the same parent. This rearrangement results in an infinite $2(k-1)$-ary tree, in which each edge has a random label distributed as $\overline{E}_k$. The labels of edges from a vertex to its children are dependent, but the $2(k-1)$-dimensional random vector of labels of edges from a vertex to its children is independent of all others, as required for Lemma 2. It is straightforward to get that for any real number $\lambda < 1$,

$$\mathbb{E}\left[\exp\{\lambda \overline{E}_k\}\right] = \frac{1}{k-1}\sum_{i=2}^{k} \frac{1}{(1-\lambda)^i}, \tag{12}$$

since the moment generating function of the distribution Gamma(i) is $1/(1-\lambda)^i$ for any integer $i \ge 1$ and $\lambda < 1$. Let $\overline{c}_k$ be the solution to

$$\sup\left\{\frac{\lambda}{c} - \log\left(\mathbb{E}\left[\exp\{\lambda \overline{E}_k\}\right]\right) : \lambda \le 0\right\} = \log(2k-2). \tag{13}$$

Since $\mathrm{ah}(\overline{T}_k^{(t)})$ is equal to the height of the subtree of $\overline{T}'_k$ induced by vertices with birth time not greater than t, by Lemma 2 we have that the auxiliary height of $\overline{T}_k^{(t)}$ is w.h.p. asymptotic to $\overline{c}_k t$, as $t \to \infty$.

Analogously, one can generate a random infinite $2(k+2)$-ary tree $\widetilde{T}'_k$ with the labels of edges distributed as $\widetilde{E}_k$, of which the distribution is Gamma(V_k) with

$$\mathbb{P}(V_k = i) = \frac{1}{k+2}, \quad i = 2, 3, \ldots, k-1; \quad \mathbb{P}(V_k = k) = \frac{4}{k+2}.$$

As we have done for (12), one can get that

$$\mathbb{E}\left[\exp\{\lambda \widetilde{E}_k\}\right] = \frac{4}{(k+2)(1-\lambda)^k} + \frac{1}{k+2}\sum_{i=2}^{k-1} \frac{1}{(1-\lambda)^i}. \tag{14}$$

Then by a similar argument, the auxiliary height of $\widetilde{T}_k^{(t)}$ is w.h.p. asymptotic to $\widetilde{c}_k t$, where the constant $\widetilde{c}_k$ is the solution to

$$\sup\left\{\frac{\lambda}{c} - \log\left(\mathbb{E}\left[\exp\{\lambda \widetilde{E}_k\}\right]\right) : \lambda \le 0\right\} = \log(2k+4).$$

To formulate the solutions $\overline{c}_k$ and $\widetilde{c}_k$, a series of functions on the interval $(0,1)$ are defined in the following. First, we define the functions f and h on the interval $(0,1)$ as

$$f(x) = 2\sum_{i=2}^{\infty} x^i = \frac{2x^2}{1-x},$$

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

and

$$h(x) = \frac{x(x-1)f'(x)}{f(x)} - \log f(x) = x - 2 - \log \frac{2x^2}{1-x}.$$

Note that $h(x)$ has a unique root x_0. In fact, the function $h(x)$ is strictly decreasing on $(0,1)$, since $h(0+) = \infty, h(1-) = -\infty$, and the derivative

$$h'(x) = -\frac{(x-1)^2+1}{x(1-x)} < 0.$$

Moreover, numerical calculations give $x_0 = 0.2550568\ldots$. For any integer $k \geq 2$, we also define the functions $\overline{f}_k$ and $\widetilde{f}_k$ on the interval $(0,1)$ as

$$\overline{f}_k(x) = 2\sum_{i=2}^{k} x^i, \quad \widetilde{f}_k(x) = 8x^k + 2\sum_{i=2}^{k-1} x^i.$$

Then, by (12) and (14), we have that for any $0 < \lambda < 1$,

$$2(k-1)\mathbb{E}\left[\exp\{\lambda \overline{E}_k\}\right] = \overline{f}_k\left(\frac{1}{1-\lambda}\right), \quad 2(k+2)\mathbb{E}\left[\exp\{\lambda \widetilde{E}_k\}\right] = \widetilde{f}_k\left(\frac{1}{1-\lambda}\right). \tag{15}$$

Finally, we set

$$\overline{h}_k(x) = \frac{x(x-1)\overline{f}_k'(x)}{\overline{f}_k(x)} - \log \overline{f}_k(x), \quad \widetilde{h}_k(x) = \frac{x(x-1)\widetilde{f}_k'(x)}{\widetilde{f}_k(x)} - \log \widetilde{f}_k(x). \tag{16}$$

Since for any $0 < x < 1$,

$$\lim_{k\to\infty} \overline{f}_k(x) = f(x), \quad \lim_{k\to\infty} \overline{f}_k'(x) = f'(x), \quad \lim_{k\to\infty} \widetilde{f}_k(x) = f(x), \quad \lim_{k\to\infty} \widetilde{f}_k'(x) = f'(x),$$

we have that both the sequences $\{\overline{h}_k(x)\}_{k=2}^{\infty}$ and $\{\widetilde{h}_k(x)\}_{k=2}^{\infty}$ converge pointwise to $h(x)$ on the interval $(0,1)$. Furthermore, since both $\overline{f}_k(x)$ and $\widetilde{f}_k(x)$ are monotonically increasing in k for all $0 < x < 1$, by Dini's theorem (see, e.g., [3, Thm. 8.2.6]) we have that both the sequences $\{\overline{f}_k(x)\}_{k=2}^{\infty}$ and $\{\widetilde{f}_k(x)\}_{k=2}^{\infty}$ converge uniformly to $f(x)$ on any closed subinterval of $(0,1)$, say $[0.1, 0.9]$. And for the sequences $\{\overline{h}_k(x)\}_{k=2}^{\infty}$ and $\{\widetilde{h}_k(x)\}_{k=2}^{\infty}$, we have more information in the following.

Claim 4: *There exists a positive integer $k_0 \geq 2$ and sequences $\{\overline{x}_k\}_{k=k_0}^{\infty}$ and $\{\widetilde{x}_k\}_{k=k_0}^{\infty}$, such that for any $k \geq k_0$, $\overline{x}_k$ and $\widetilde{x}_k$ are respectively the roots of $\overline{h}_k(x)$ and $\widetilde{h}_k(x)$, and*

$$\lim_{k\to\infty} \overline{x}_k = \lim_{k\to\infty} \widetilde{x}_k = x_0,$$

where x_0 is the unique root of $h(x)$ in $(0,1)$.

Proof: Fix a small $\varepsilon > 0$, then we have that $h(x_0 - \varepsilon) > 0$ and $h(x_0 + \varepsilon) < 0$. By the pointwise convergence of $\overline{h}_k(x)$ to $h(x)$, there exists an integer $\overline{k}_0$ such that for any $k > \overline{k}_0$, both $\overline{h}_k(x_0 - \varepsilon) > 0$ and $\overline{h}_k(x_0 + \varepsilon) < 0$ hold. Note that the continuity of $\overline{h}_k(x)$ implies the existence of its roots in the interval $(x_0 - \varepsilon, x_0 + \varepsilon)$, and that the set $\{x : \overline{h}_k(x) = 0\}$ is closed. Then for any $k > \overline{k}_0$, we can choose a root $\overline{x}_k$ of $\overline{h}_k(x)$, which is the closest to x_0. Hence, by the arbitrariness of ϵ, the limit of the sequence $\{\overline{x}_k\}_{k=k_0}^{\infty}$ is x_0. In an analogous

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

way, one can show the existences of $\widetilde{k}_0$ and the sequence $\{\widetilde{x}_k\}_{k=k_0}^{\infty}$ with $\widetilde{x}_k \to x_0$ as $k \to \infty$. Finally, the desired result follows by letting $k_0 = \max\{\overline{k}_0, \widetilde{k}_0\}$. ■

To prove Lemma 6, we now define the sequences $\{\overline{c}_k\}_{k=k_0}^{\infty}$ and $\{\widetilde{c}_k\}_{k=k_0}^{\infty}$ as

$$\overline{c}_k = \frac{\overline{x}_k - 1}{\overline{x}_k \log \overline{f}_k(\overline{x}_k)}, \quad \widetilde{c}_k = \frac{\widetilde{x}_k - 1}{\widetilde{x}_k \log \widetilde{f}_k(\widetilde{x}_k)}, \quad k \geq k_0, \tag{17}$$

where $\overline{x}_k, \widetilde{x}_k$, and k_0 are given in Claim 4.

Proof of Lemma 6: We only show that $\mathrm{ah}(\overline{T}_k^{(t)})$ is w.h.p. asymptotic to $\overline{c}_k t$ for any fixed $k \geq k_0$, as $t \to \infty$; the argument for $\mathrm{ah}(\widetilde{T}_k^{(t)})$ is similar. First, we have that $\log(\mathbb{E}[\exp\{\lambda \overline{E}_k\}])$ is a strictly convex function of λ on the interval $(-\infty, 0]$. In fact, for any $\lambda_1 < \lambda_2 \leq 0$ and $0 < \alpha < 1$, by Hölder's inequality,

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\alpha\lambda_1 \overline{E}_k + (1-\alpha)\lambda_2 \overline{E}_k\right\}\right] &= \mathbb{E}\left[\left(\exp\{\lambda_1 \overline{E}_k\}\right)^{\alpha}\left(\exp\{\lambda_2 \overline{E}_k\}\right)^{1-\alpha}\right] \\ &< \left(\mathbb{E}\left[\exp\{\lambda_1 \overline{E}_k\}\right]\right)^{\alpha}\left(\mathbb{E}\left[\exp\{\lambda_2 \overline{E}_k\}\right]\right)^{1-\alpha},\end{aligned}$$

where the strictness in the inequality comes from that $\overline{E}_k$ is not a degenerate random variable. Then taking logarithms on both sides of this inequality gives the convexity of the desired function. It follows that for any constant c, the function of λ given by

$$\frac{\lambda}{c} - \log\left(\mathbb{E}\left[\exp\{\lambda \overline{E}_k\}\right]\right)$$

is strictly concave on $(-\infty, 0]$, and thus attain its supremum at a unique point $\lambda \leq 0$.

On the other hand, it follows by (15) and (17) that

$$\frac{\overline{x}_k - 1}{\overline{c}_k \overline{x}_k} - \log\left(\mathbb{E}\left[\exp\left\{\frac{(\overline{x}_k - 1)\overline{E}_k}{\overline{x}_k}\right\}\right]\right) = \log(2k-2), \tag{18}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda} \log\left(\mathbb{E}\left[\exp\{\lambda \overline{E}_k\}\right]\right)\Big|_{\lambda=(\overline{x}_k-1)/\overline{x}_k} &= \frac{\mathrm{d}}{\mathrm{d}\lambda} \log \overline{f}_k\left(\frac{1}{1-\lambda}\right)\Big|_{\lambda=(\overline{x}_k-1)/\overline{x}_k} \\ &= \overline{x}_k^2 \frac{\overline{f}_k'(\overline{x}_k)}{\overline{f}_k(\overline{x}_k)}.\end{aligned} \tag{19}$$

By Claim 4, we have $\overline{h}_k(\overline{x}_k) = 0$, which implies that

$$\overline{x}_k^2 \frac{\overline{f}_k'(\overline{x}_k)}{\overline{f}_k(\overline{x}_k)} = \overline{x}_k \frac{\log \overline{f}_k(\overline{x}_k)}{\overline{x}_k - 1} = \frac{1}{\overline{c}_k},$$

by the definition of $\overline{h}_k(x)$ given by (16). By (19), we thus have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{\lambda}{\overline{c}_k} - \log\left(\mathbb{E}\left[\exp\{\lambda \overline{E}_k\}\right]\right)\right]\Big|_{\lambda=(\overline{x}_k-1)/\overline{x}_k} = 0,$$

which, together with (18), implies that $\overline{c}_k$ is the unique solution to Eq. (13). Hence, the desired result for $\mathrm{ah}(\overline{T}_k^{(t)})$ holds. ■

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

The last step to prove Lemma 3 is to show the sequences $\{\overline{c}_k\}$ and $\{\widetilde{c}_k\}$ having the same desired limit c as $k \to \infty$.

Proof of Lemma 3: By Lemma 4, it is sufficient to show that $\mathrm{ah}(T^{(t)})$, the auxiliary height of the random tree $T^{(t)}$ is w.h.p. asymptotic to ct as $t \to \infty$, where the constant

$$c = \frac{x_0 - 1}{x_0} \log^{-1} \frac{2x_0^2}{1 - x_0} = \frac{x_0 - 1}{x_0 \log f(x_0)}.$$

Note that $0.1 < x_0 < 0.9$ and both the sequences $\{\overline{f}_k(x)\}_{k=2}^{\infty}$ and $\{\widetilde{f}_k(x)\}_{k=2}^{\infty}$ converge uniformly to $f(x)$ on $[0.1, 0.9]$. Recall the sequences $\{\overline{x}_k\}_{k=k_0}^{\infty}$ and $\{\widetilde{x}_k\}_{k=k_0}^{\infty}$ given in Claim 4. Then it follows by (17) that

$$\lim_{k\to\infty} \overline{c}_k = \lim_{k\to\infty} \frac{\overline{x}_k - 1}{\overline{x}_k \log \overline{f}_k(\overline{x}_k)} = \frac{x_0 - 1}{x_0 \log f(x_0)} = c,$$

and

$$\lim_{k\to\infty} \widetilde{c}_k = \lim_{k\to\infty} \frac{\widetilde{x}_k - 1}{\widetilde{x}_k \log \widetilde{f}_k(\widetilde{x}_k)} = \frac{x_0 - 1}{x_0 \log f(x_0)} = c.$$

By Lemma 6, we thus have that both the auxiliary heights $\mathrm{ah}(\overline{T}_k^{(t)})$ and $\mathrm{ah}(\widetilde{T}_k^{(t)})$ are w.h.p. asymptotic to ct as $t \to \infty$, if k is sufficiently large. Then the desired result follows by Lemma 5. The proof of Lemma 3 is complete. ■

5. DISCUSSION

In this work, we show the small-world effect of geographical attachment networks through rigorous probabilistic methods. There is still room to close the gap between the physics and probability literatures on this subject. Although we provide explicit results for asymptotic degree distribution of general geographical attachment networks, the weak convergence of the clustering coefficient and diameter is established only for the simplest case of this network model (when the model parameter m is equal to 2). Additionally, the asymptotic distributions of such random variables are as yet unknown even for the case $m = 2$. We are further interested in other important topics in geographical attachment networks, such as the average path length, in our future studies.

Acknowledgments

We thank a member of editorial board and the anonymous referee for the useful suggestions that greatly improved the presentation of the paper. This work was supported by NSFC (grant number 11671373) and NSFC (grant number 11771418).

References

1. Athreya, K.B. & Ney, P.E. (1972). *Branching processes*. Berlin: Springer-Verlag.
2. Azuma, K. (1967). Weighted sums of certain dependent variables. *Tohoku Mathematical Journal* 3: 357–367.
3. Bartle, R.G. & Sherbert, D.R. (2011). *Introduction to Real analysis* (4th ed.). USA: John Wiley & Sons.
4. Bellman, R. & Harris, T.E. (1948). On the theory of age-dependent stochastic branching processes. *Proceedings of the National Academy of Sciences of the United States of America* 34: 601–604.
5. Bhamidi, S., Steele, J.M., & Zaman, T. (2015). Twitter event networks and the superstar model. *The Annals of Applied Probability* 25: 2462–2502.

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press

6. Bilke, S. & Peterson, C. (2001). Topological properties of citation and metabolic networks. *Physical Review E* 64: 036106.
7. Bollobás, B., Riordan, O., Spencer, J., & Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures and Algorithms* 18: 279–290.
8. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the web. *Computer Networks* 33: 309–320.
9. Broutin, N. & Devroye, L. (2006). Large deviations for the weighted height of an extended class of trees. *Algorithmica* 46: 271–297.
10. Ebrahimzadeh, E., Farczadi, L., Gao, P., Mehrabian, A., Sato, C.M., Wormald, N., & Zung, J. (2014). On the longest paths and diameter in random Apollonian networks. *Random Structures and Algorithms* 45: 703–725.
11. Eggenberger, F. & Pólya, G. (1923). Über die statistik verketteter vorgäge. *Zeitschrift für Angewandte Mathematik und Mechanik* 3: 279–289.
12. Frieze, A. & Tsourakakis, C.E. (2012). On certain properties of random Apollonian networks. In *Algorithms and Models for the Web Graph*. Springer, Berlin, pp. 93–112.
13. Hayashi, Y. (2006). A review of recent studies of geographical scale-free networks. *IPSJ Digital Courier* 2: 155–164.
14. Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58: 13–30.
15. Kleinberg, J.M. (2000). The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*. Association of Computing Machinery, New York, pp. 163–170.
16. Kolossváry, I., Komjáthy, J., & Vágó, L. (2016). Degrees and distances in random and evolving apollonian networks. *Advances in Applied Probability* 48: 865–902.
17. Mahmoud, H. (2008). *Pólya urn models*. Boca Raton: Chapman & Hall/CRC.
18. Milgram, S. (1967). The small world problem. *Psychology Today* 1: 60–67.
19. Moore, C. & Newman, M.E.J. (1999). Epidemics and percolation in small-world networks. *Physical Review E* 61: 5678–5682.
20. Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the USA* 98: 404–409.
21. Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review* 45: 167–256.
22. Newman, M.E.J. & Watts, D.J. (1999). Renormalization group analysis of the small-world network model. *Physics Letters A* 263: 341–346.
23. Ozik, J., Hunt, B.R., & Ott, E. (2004). Growing networks with geographical attachment preference: Emergence of small worlds. *Physical Review E* 69: 026108.
24. Pittel, B. (1994). Note on the heights of random recursive trees and random m-ary search trees. *Random Structures and Algorithms* 5: 337–347.
25. Van der Hofstad, R. (2018). *Random graphs and complex networks: Volume II*. Unpublished manuscript. http://www.win.tue.nl/rhofstad/NotesRGCNII.pdf.
26. Van Noort, V., Snel, B., & Huynen, M.A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports* 5: 280–284.
27. Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
28. Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393: 440–442.
29. Wu, J., Tse, C.K., Lau, F.C.M., & Ho, I.W.H. (2013). Analysis of communication network performance from a complex network perspective. *IEEE Transactions on Circuits and Systems* 60: 3303–3316.
30. Zaidi, F. (2013). Small world networks and clustered small world networks with random connectivity. *Social Network Analysis and Mining* 3: 51–63.
31. Zhang, Z., Rong, L., & Comellas, F. (2006). Evolving small-world networks with geographical attachment preference. *Journal of Physics A: Mathematical and General* 39: 3253–3261.
32. Zhang, Z., Rong, L., & Guo, C. (2006). A deterministic small-world network created by edge iterations. *Physica A* 363: 567–572.
33. Zhou, T., Yan, G., & Wang, B.-H. (2005). Maximal planar networks with large clustering coefficient and power-law degree distribution. *Physical Review E* 71: 046141.

https://doi.org/10.1017/S0269964819000342 Published online by Cambridge University Press