

Distances in a geographical attachment network model

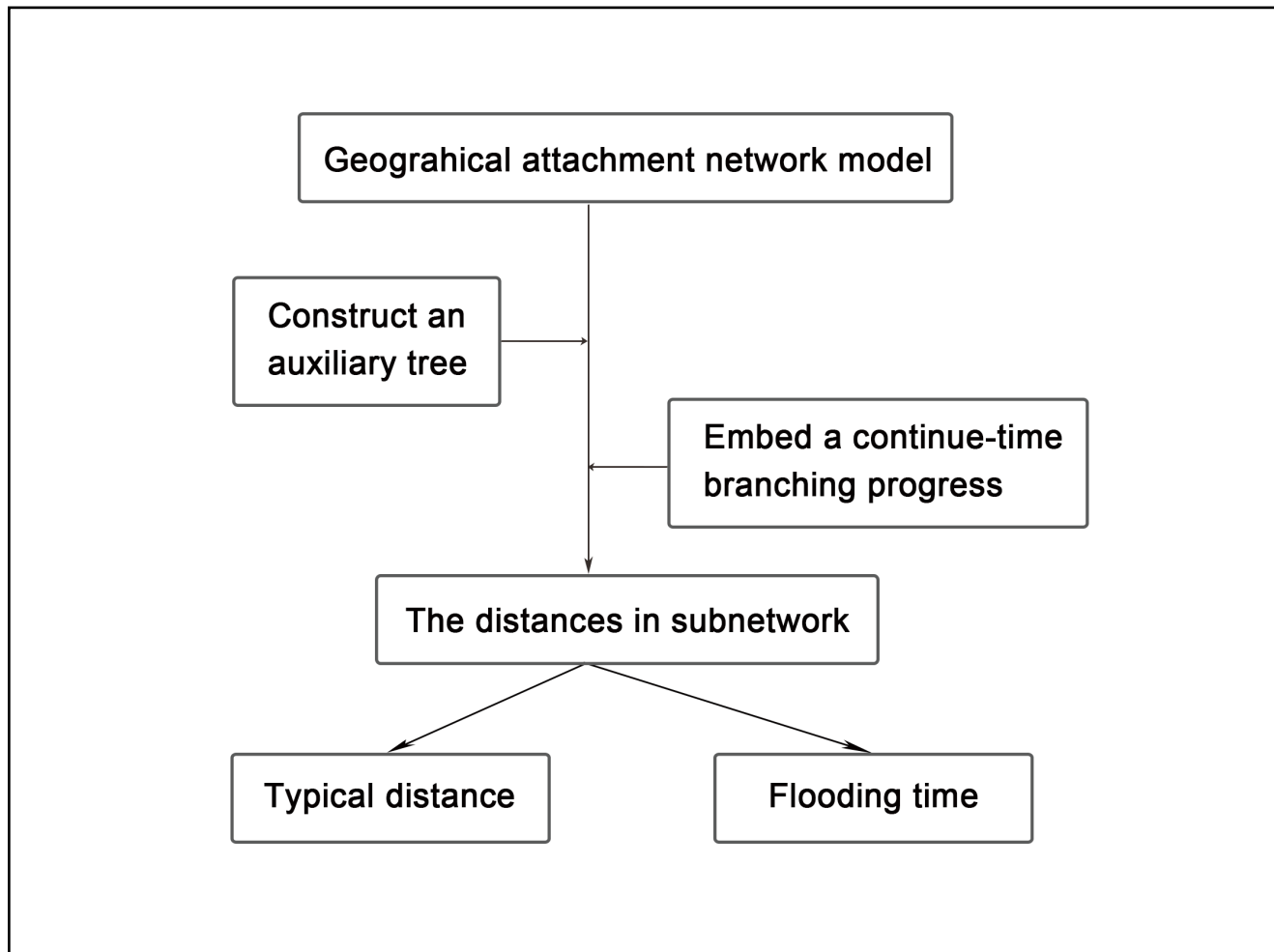
Ziling Xu, and Qunqiang Feng 

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

 Correspondence: Qunqiang Feng, E-mail: fengqq@ustc.edu.cn

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Graphical abstract



Distances in a geographical attachment network model.

Public summary

- The asymptotic properties of the typical distance and the flooding time in a geographic attachment network (GAN) model are studied.
- The typical distance of GAN is asymptotically normal.
- The flooding time of GAN converges to a given constant in probability.

Distances in a geographical attachment network model

Ziling Xu, and Qunqiang Feng 

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

 Correspondence: Qunqiang Feng, E-mail: fengqq@ustc.edu.cn

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Cite This: JUSTC, 2023, 53(11): 1104 (8pp)



Read Online

Abstract: Distances between nodes are one of the most essential subjects in the study of complex networks. In this paper, we investigate the asymptotic behaviors of two types of distances in a model of geographic attachment networks (GANs): the typical distance and the flooding time. By generating an auxiliary tree and using a continuous-time branching process, we demonstrate that in this model the typical distance is asymptotically normal, and the flooding time converges to a given constant in probability as well.

Keywords: typical distance; flooding time; geographic attachment network; branching process

CLC number: O211.4

Document code: A

2020 Mathematics Subject Classification: 05C80; 60C05; 60F05

1 Introduction

A significant amount of research has focused on networks as a result of the recent increase in interest in social networks^[1], communication networks^[2], scientific collaboration networks^[3], biological networks^[4], and many other types of networks. To analyze these networks, researchers have developed a number of models, many of which are widely used, such as the Watts and Strogatz model and the preference attachment model. Due to various constraints, different models have different topologies. In this paper, we concentrate on the model affected by geographical restrictions.

It is a given in social networks that people who relocate will most likely develop acquaintances with people in the area, such as their neighbors. Motivated by this idea, a geographical attachment network (GAN) model was first proposed in Ref. [5]. In this network model, its size (the number of nodes in it) increases over time, and the newly added nodes are only connected to the nodes that are closest to them.

The following guidelines can be used to generate the GAN model presented in this work. We start with an initial state (at time $n = 0$) of three nodes distributed on a ring, all of which are connected to one another. That is, the initial graph $\text{GAN}(0)$ is a triangle; see Fig. 1a. At time $n \geq 1$, the network $\text{GAN}(n)$ is obtained from $\text{GAN}(n-1)$ in the following manner: A new node is placed in an internode interval chosen uniformly at random from the $n+2$ existing nodes along the ring and connected to its two nearest neighbors (one on either side). This GAN model is the simplest case in Ref. [5]; see also Ref. [6]. For convenience in below, here we may define potential nodes and active intervals. If two endpoints of an interval are adjacent on the ring, the interval is said to be active. Each active interval corresponds to a potential node, which is a node that may be chosen as a new node in the future. Instead, we refer to nodes that are added to the network as actual nodes. There are two potential edges that could connect each

potential node to its neighbors. Fig. 1 shows an illustration of this network at times $n = 0, 1$, and 2. For the variants of GAN models, we refer to Refs. [7–9].

A path in network G is a subnetwork with a sequence of successive edges that joins a sequence of distinct nodes (except, possibly, the first and last node). The length of a path is the number of edges in it. Despite the fact that the geographical distance along the ring is used in the generating process of the GAN model, in this paper, we consider the graphic distance, i.e., the distance between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes. The diameter of a network is the largest distance among all pairs of nodes.

Several properties of the GAN model are obtained in Ref. [5] with heuristic arguments and computational simulations, and Ref. [6] using the rigorous probabilistic method. Let $P_{k,n}$ be the proportion of nodes with degree $k \geq 2$ in the network $\text{GAN}(n)$. It is shown in Ref. [6] that as $n \rightarrow \infty$,

$$P_{k,n} \xrightarrow{p} \frac{1}{3} \left(\frac{2}{3} \right)^{k-2}, \quad k = 2, 3, \dots, n.$$

Then the GAN model is not a scale-free network^[10], in which the limiting degree distribution is a power law. This indicates that the network model considered here is essentially different from the random Apollonian network model (see, for example, Refs. [11–14]), which looks quite similar to the GAN model. For the diameter of the GAN model, the following result is also derived in Ref. [6]. We say that a sequence of events $\{\mathcal{E}_n, n \geq 1\}$ occurs with high probability (w.h.p.) when $\mathbb{P}(\mathcal{E}_n) \rightarrow 1$ as $n \rightarrow \infty$.

Theorem 1. As $n \rightarrow \infty$, the diameter of the network $\text{GAN}(n)$ is with high probability asymptotic to $2c \log n$ with the constant

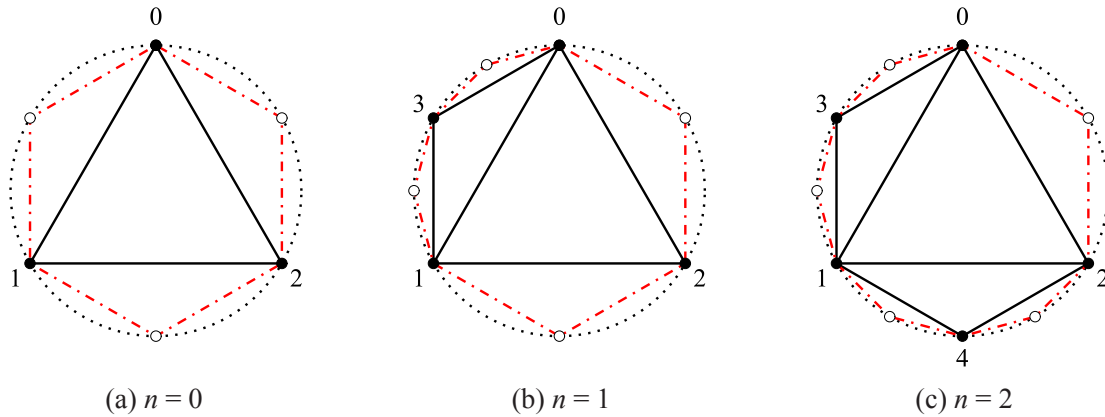


Fig. 1. Illustration of the growing GAN model with potential nodes for time $n = 0, 1$, and 2 , where points $\bullet$ represent nodes in the network, points $\circ$ are potential nodes, and red dashed lines are potential edges.

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

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

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

In this paper, we further investigate another two types of distances in the GAN model: the typical distance and flooding time. The typical distance of a network is the distance between a pair of nodes picked in it uniformly at random (u.a.r.). For the typical distance of other random graph models, we refer to Refs. [15–18]. The flooding time of a network is the greatest distance from a randomly chosen node to other nodes. For more backgrounds and results of flooding time, see Refs. [19–22].

The rest of the paper is organized as follows. In Section 2, we analyze the structure of the subnetworks and obtain some results by building an auxiliary tree and using a continuous-time branching process. Based on these, we derive the results of the typical distance and flooding time of GAN models in Section 3.

2 Subnetworks and auxiliary tree

When the GAN model is in its initial state, i.e., $\text{GAN}(0)$, we can see that the ring is divided into three initial intervals, I_1, I_2 , and I_3 (see Fig. 2). We refer to the subnetwork made up of the nodes and edges in interval I_i , including the endpoints of the interval, as GAN_i , where $i = 1, 2, 3$. It goes without saying that any pair of nodes' paths involves a maximum of two subnetworks. There are only two scenarios to take into account for the typical distance of $\text{GAN}(n)$: (A) Two nodes originate from the same initial internal; (B) two nodes originate from distinct initial internals. Indeed, we only need to consider the typical distance of one of the subnetworks and the distance between any initial node and the node selected u.a.r. in the subnetwork. Therefore, we need to concentrate on the subnetworks first.

Assuming that there are $Y_{1,n}, Y_{2,n}$, and $Y_{3,n}$ nodes in each of the three open intervals at time n , then $Y_{1,n} + Y_{2,n} + Y_{3,n} = n$. We obtain that

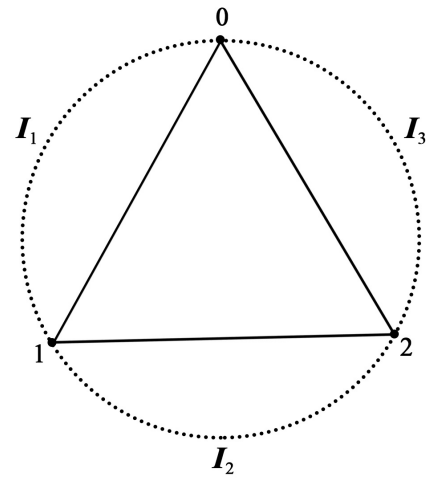


Fig. 2. Initial internode intervals in $\text{GAN}(0)$.

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

as $n \rightarrow \infty$, where $\text{Beta}(\alpha, \beta)$ denotes the beta distribution with parameters α and β . With high probability, the random variables $Y_{1,n}, Y_{2,n}$, and $Y_{3,n}$ have the same order of order n , as proven in Ref. [6]. Then, we can use the fact that

$$\log Y_{i,n} = \log n + O_p(1). \quad (1)$$

The identical distribution of the three random variables is clear to notice. Without loss of generality, we restrict our attention to the subnetwork $\text{GAN}_1(n)$. We suppose that $Y_{1,n} = n_0$, where $n_0 \geq 0$ is an integer, and that a new node is added at each step in this subnetwork. As a result, the size of the subnetwork $\text{GAN}_1(n)$ can be set to $n_0 + 2$, where n_0 varies with n . $\text{GAN}_1(n)$ has two initial nodes that are linked together, hence the distances between any given node in the network and the two initial nodes can differ by up to one.

If we consider the first node added to the subnetwork to be the root of a binary tree and each time a new node is added, the two potential nodes produced are the two children of that node, we can create a binary tree according to this relationship. Afterwards, we may utilize a binary tree to obtain some of this subnetwork's characteristics. Moreover, in each step,

only one potential node is transformed into an actual node, and two new potential nodes are generated. This network's creation resembles a unique continuous-time branching process in which each individual splits just once to produce two offspring.

2.1 Auxiliary tree structure

In this section, we construct an auxiliary binary tree and embed a branching process in the subnetwork $\text{GAN}_1(n)$. There is only one initial interval and one potential node in the initial state of subnetwork GAN_1 . An interval becomes two new active intervals when a new node u is added to it, and these two new intervals also correspond to two new potential nodes. The potential node that lies to the left (or right) of node u is called the left (or right) child of u . Instead, u is the parent of these two nodes. The ancestral line of node u is a path that leads from the first nodes added in $\text{GAN}_1(n)$ to u , where the previous node is the parent of the next node. The network in Fig. 3a can be redrawn using the shape of the binary tree to obtain Fig. 3b. The auxiliary tree in this paper is constructed in a manner similar to that in Ref. [6], but the nodes in the auxiliary tree correspond to the edges of the network in the latter. As shown in Fig. 3b, we can embed a continuous-time branching process (CTBP) (see, for example Refs. [23, 24]) into $\text{GAN}_1(n)$.

Consider a CTBP as follows: At the beginning, there is a single individual who serves as the process's root. This initial individual then splits into two individuals (producing two offspring) before becoming inactive. Each individual has an i.i.d. exponential lifespan with a mean of 1. As a result, after birth, each individual is active for the entirety of its lifespan before splitting, going inactive, and giving birth to two offspring who then become active for their own i.i.d. $\text{Exp}(1)$ lifespans. There are $n_0 + 1$ active individuals if n_0 individuals split during the process. Let t_i represent the time when the i th individual splits.

There is a mapping relation between the subnetwork

$\text{GAN}_1(n)$ and the CTBP with n_0 splitters as follows. The first node in $\text{GAN}_1(n)$ (i.e., the first node added to the subnetwork $\text{GAN}_1(n)$) corresponds to the root of the CTBP. When the n_0 th individual in the CTBP splits, the already split individuals are the actual nodes in $\text{GAN}_1(n)$, while the individuals who are still active in the CTBP are the potential nodes in $\text{GAN}_1(n)$. This is true because, in each step, two new potential nodes are created in place of the actual node when a new node is added to $\text{GAN}_1(n)$. Furthermore, a potential node is randomly chosen in each step, which is equivalent to the CTBP's next individual splitting being a randomly chosen active individual due to the memoryless nature of exponential variables. In fact, there is a maximum difference of one between any node's distance from the first node and its distance from either of the initial nodes.

The generation of an individual in the CTBP is equivalent to the length of the ancestral line leading to the corresponding node in $\text{GAN}_1(n)$. For active individuals, we can directly apply the conclusion of Corollary 1.1 in Ref. [24]. Let L_u be the generation of the individual in the CTBP corresponding to potential node u selected u.a.r. in $\text{GAN}_1(n)$. Then as $n_0 \rightarrow \infty$,

$$\frac{L_u - 2 \log n_0}{\sqrt{2 \log n_0}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (2)$$

2.2 Shortcuts

Obviously, for each node, there is a shortcut that connects it to other ancestors in addition to the connection to its parent. To calculate the distance between each pair of potential nodes, it is necessary to identify the shortcuts between the first node and each node.

The ancestral line shows one of the paths from the node to the first node in $\text{GAN}_1(n)$, and we can use the ancestral line to identify the shortcuts. We could also define a sequence for each node's ancestral line to discover the rule governing the existence of shortcuts. First, the left and right child nodes are denoted by symbols l and r , respectively. The ancestral line

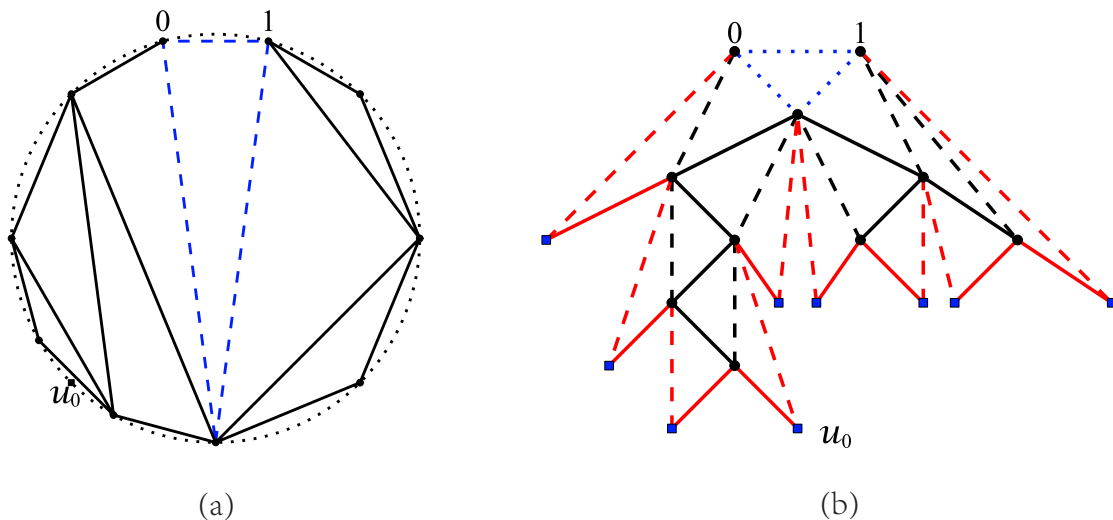


Fig. 3. (a) is the subnetwork GAN_1 after adding several nodes without marking potential nodes where the nodes labeled 0 and 1 are the initial nodes. By redrawing the network, we can obtain (b). The nodes marked as black circles and the nodes marked as blue squares are actual nodes and potential nodes, respectively. Except for the blue dotted line, the solid lines and the dotted lines represent the ancestral line of each node and shortcuts, respectively. Furthermore, the black lines and red lines represent existing edges and potential edges, respectively. u_0 is one of the potential nodes of this subnetwork.

of each node provides a sequence of l and r , where the i th symbol of the sequence denotes whether the $(i+1)$ th ancestor is the left or right child of the i th ancestor. For instance, the sequence of the node labeled u_0 in Fig. 3b is $lrllrr$. The sequence of the first node is represented by $\emptyset$. Each node's sequence is therefore unique, with the exception of the initial nodes. The sequence is sometimes used to represent the node. Furthermore, instead of trying to distinguish between the two initial nodes, we use the symbol Λ to represent all initial nodes.

Assuming there are two nodes u and v , we can represent their sequences using the formulas $\mathbf{u} = u_1 \dots u_p$ and $\mathbf{v} = v_1 \dots v_q$, respectively, where $u_i, v_j \in \{l, r\}$, $i = 1, \dots, p, j = 1, \dots, q$, and p and q are the lengths of the sequence, i.e., $|\mathbf{u}| = p$ and $|\mathbf{v}| = q$. We use η_i to represent the last position of $i \in \{l, r\}$ in a sequence and define a truncation operator T ,

$$T_i \mathbf{u} := \begin{cases} u_1 \dots u_{\eta_i-1}, & \text{if symbol } i \text{ appears in} \\ & \text{the sequence of the node;} \\ \Lambda, & \text{if symbol } i \text{ does not appear in} \\ & \text{the sequence of the node.} \end{cases}$$

The parent of $\mathbf{u}$ is obviously either $T_l \mathbf{u}$ or $T_r \mathbf{u}$. In particular, the first node has two parents, both of which are initial nodes. Along the ancestral line of $\mathbf{u}$, the next node of $T_l \mathbf{u}$ is $(T_l \mathbf{u})l$, so u is the left offspring of $T_l \mathbf{u}$, i.e., $T_l \mathbf{u}$ is to the right of u . Similarly, $T_r \mathbf{u}$ is to the left of u .

Based on the definition of the sequence of nodes and the above assumptions, we obtain the following facts:

(a) The prefixes of the sequence of node u refer to the ancestors of u , i.e., the nodes with the sequence $u_1 \dots u_k$, $k = 1, \dots, p-1$, are ancestors of u .

(b) The sequence $\mathbf{u} \wedge \mathbf{v} := \{u_1 \dots u_k : u_i = v_i, i = 1, \dots, k; u_{k+1} \neq v_{k+1}; k \leq \min\{p, q\}\}$ represents the latest common ancestor of u and v , which can be written as $u \wedge v$.

(c) The shortest path between u and v must pass through some of their common ancestors. We must ascend to their common ancestors to determine the shortest path between them.

Claim 1. $T_l \mathbf{u}$ and $T_r \mathbf{u}$ are two ancestors of $\mathbf{u}$ connected to $\mathbf{u}$, i.e., the neighbors of $\mathbf{u}$ when node $\mathbf{u}$ is added to the network as a new node.

Proof. We prove this claim by induction. The sequence of the first node is $\emptyset$, and its children are l and r . Then the result is valid for the first node. Assume that the claim is valid for the most recently added node $\mathbf{u}$, that is, $\mathbf{u}$ links to $T_l \mathbf{u}$ and $T_r \mathbf{u}$. The endpoints of the interval to the left of $\mathbf{u}$ are u and $T_l \mathbf{u}$. When the left child of $\mathbf{u}$, ul , is added to the network, $T_l(ul) = \mathbf{u}$ and $T_r(ul) = T_l \mathbf{u}$. These two results are exactly the endpoints of the interval of the left child of $\mathbf{u}$. Therefore, the claim holds for ul . Similarly, the claim also holds for ur .

As a result, the shortcut connects any given node u to either $T_l \mathbf{u}$ or $T_r \mathbf{u}$. For any node u , observe the difference between the sequence of nodes connected by the shortcut and the sequence of u . From back to front, this difference collects exactly two symbols. For example, the shortcut of node u_0 with the sequence $lrllrr$ in Fig. 2 connects to the node with the sequence lr , and their difference is lrr . Consequently, we can partition the sequence. From the back to the front, when

exactly two symbols have been collected completely, i.e. the occurrence of the sequence lr or rl , we divide them into a block and obtain a shortcut. The division is then repeated in the same way to obtain the number of blocks. The last block may not have a complete collection of two symbols, but this block represents a node that is 1 away from one of the initial nodes.

Reverse the sequence and define an event $\mathcal{E}$: the occurrence of the sequence lr or rl . Let Y be the number of symbols needed for event $\mathcal{E}$ to occur for the first time in the reversed sequence. It is easy to calculate

$$\mathbb{E}[Y] = 3 \quad \text{and} \quad \text{Var}(Y) = 1.$$

When event $\mathcal{E}$ occurs for the first time, we remove the symbols in the reverse sequence that precede it and those that represent it, and then we need to find the location of event $\mathcal{E}$ in the remaining sequence. This operation is repeated until no event $\mathcal{E}$ occurs in the remaining sequence. The number of times an event $\mathcal{E}$ is repeated in the inverse sequence of a node is equal to the distance between this node and the initial node minus 1. Suppose there is a sequence of length m and let S_m be the number of occurrences of event $\mathcal{E}$ in that sequence. By (6.8) in Chapter XIII of Ref. [25],

$$\mathbb{E}[S_m] \sim \frac{m}{3}, \quad \text{Var}(S_m) \sim \frac{m}{27},$$

where $\sim$ indicates that the ratio of the two sides tends to 1, and S_m has an asymptotically normal distribution, i.e., as $m \rightarrow \infty$,

$$\frac{S_m - m/3}{\sqrt{m}} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{27}\right). \quad (3)$$

Furthermore, we use L_u to represent the length of the ancestral line of node u . If node u is the ancestor of node v , then the shortest distance between u and v only needs to consider the difference in the two nodes' sequences, and the length of the difference is $L_u - L_v$.

Lemma 1. Suppose that u is a potential node picked u.a.r. in $\text{GAN}_1(n)$ with size $n_0 + 2$ and let S_{L_u} be the shortest distance from u to the first node. If $n_0 \rightarrow \infty$, S_{L_u} satisfies

$$\frac{S_{L_u} - \frac{2}{3} \log n_0}{\sqrt{\log n_0}} \xrightarrow{d} \mathcal{N}\left(0, \frac{8}{27}\right). \quad (4)$$

Proof. It is easy to obtain from (2), that

$$\mathbb{E}[L_u] \sim 2 \log n_0, \quad \text{Var}(L_u) \sim 2 \log n_0.$$

We can calculate

$$\mathbb{E}[S_{L_u}] = \mathbb{E}[\mathbb{E}[S_{L_u}|L_u]] \sim \frac{1}{3} \mathbb{E}[L_u] \sim \frac{2}{3} \log n_0, \quad (5)$$

$$\begin{aligned} \text{Var}(S_{L_u}) &= \text{Var}(\mathbb{E}[S_{L_u}|L_u]) + \mathbb{E}[\text{Var}(S_{L_u}|L_u)] \sim \\ &\frac{1}{3^2} \text{Var}(L_u) + \frac{1}{3^3} \mathbb{E}[L_u] \sim \frac{8}{27} \log n_0. \end{aligned} \quad (6)$$

To show the asymptotic normality of S_{L_u} , observe that

$$\frac{S_{L_u} - \frac{2}{3} \log n_0}{\sqrt{\frac{8}{27} \log n_0}} = \frac{S_{L_u} - \frac{1}{3} L_u}{\sqrt{L_u}} \sqrt{\frac{L_u}{\frac{8}{27} \log n_0}} + \frac{\frac{1}{3} L_u - \frac{2}{3} \log n_0}{\sqrt{\frac{8}{27} \log n_0}}.$$

According to formula (2), $\frac{L_u}{2 \log n_0} \xrightarrow{p} 1$, as $n_0 \rightarrow \infty$ and the second term converges to $\mathcal{N}(0, 3/4)$ in distribution. Condition on L_u with $L_u \rightarrow \infty$, $\frac{S_{L_u} - (1/3)L_u}{\sqrt{L_u}}$ converges to $\mathcal{N}(0, 1/4)$ in distribution by (3). Since $S_{L_u}|L_u$ is independent of L_u , by Slutsky's lemma, we can obtain conclusion (4).

2.3 Common ancestor

According to fact (c) in Section 2.2, the shortest path between any two potential nodes must pass through some of their common ancestors. To find the distance between any two potential nodes, we need to find the relationship between the distance of this pair of nodes and their latest common ancestor. First, we want to divide the shortest path into two parts.

Claim 2. For any pair of nodes u and v with sequences $\mathbf{u}$ and $\mathbf{v}$, respectively,

$$\text{dist}(u, v) = \text{dist}(u, u \wedge v) + \text{dist}(v, u \wedge v),$$

or

$$\text{dist}(u, v) = \text{dist}(u, u \wedge v) + \text{dist}(v, u \wedge v) - 1.$$

Proof. For convenience, we use the sequence of the node to represent the node in the following.

If $\mathbf{u}$ is the ancestor of $\mathbf{v}$ or vice versa, then this conclusion is easy to reach. If not, based on the previous facts in Section 2.2, we can assume that the shortest path between them is $\mathbf{u} \rightarrow \dots \rightarrow \mathbf{x} \rightarrow \dots \rightarrow \mathbf{v}$ where $\mathbf{x}$ is either $\mathbf{u} \wedge \mathbf{v}$ or one of its ancestors. For path $\mathbf{u} \rightarrow \dots \rightarrow \mathbf{x}$, each node is the parent of the previous node; for path $\mathbf{x} \rightarrow \dots \rightarrow \mathbf{v}$, each node is the parent of the next node. If $\mathbf{x}$ is $\mathbf{u} \wedge \mathbf{v}$, the conclusion is obvious. If $\mathbf{x}$ is one of the ancestors of $\mathbf{u} \wedge \mathbf{v}$, assume that the two nodes of the child of $\mathbf{u} \wedge \mathbf{v}$ closest to $\mathbf{x}$ are $(\mathbf{u} \wedge \mathbf{v})\mathbf{w}_1$ and $(\mathbf{u} \wedge \mathbf{v})\mathbf{w}_2$, where $\mathbf{w}_1$ and $\mathbf{w}_2$ represent the sequences where all the symbols are the same. Because of symmetry, we can assume that $\mathbf{w}_1 = l\dots l$, $\mathbf{w}_2 = r\dots r$, $|\mathbf{w}_1| \geq 0$, and $|\mathbf{w}_2| \geq 0$.

Since it is impossible for the left offspring of $\mathbf{u} \wedge \mathbf{v}$ and the right offspring of $\mathbf{u} \wedge \mathbf{v}$ to have a shortcut to the same common ancestor at the same time, except for $\mathbf{u} \wedge \mathbf{v}$, there must be another common ancestor (defined as $\mathbf{y}$) in this path. Without loss of generality, we can write the path as

$$\mathbf{u} \rightarrow \dots \rightarrow (\mathbf{u} \wedge \mathbf{v})\mathbf{w}_1 \rightarrow \mathbf{x} \rightarrow \dots \rightarrow \mathbf{y} \rightarrow (\mathbf{u} \wedge \mathbf{v})\mathbf{w}_2 \rightarrow \dots \rightarrow \mathbf{v}. \quad (7)$$

Obviously, there is a shortcut between $(\mathbf{u} \wedge \mathbf{v})\mathbf{w}_1$ and $\mathbf{x}$ and between $(\mathbf{u} \wedge \mathbf{v})\mathbf{w}_2$ and $\mathbf{y}$. $T_r((\mathbf{u} \wedge \mathbf{v})\mathbf{w}_1) = T_r(\mathbf{u} \wedge \mathbf{v}) = \mathbf{x}$ and $T_l((\mathbf{u} \wedge \mathbf{v})\mathbf{w}_2) = T_l(\mathbf{u} \wedge \mathbf{v}) = \mathbf{y}$ means that there is a shortcut between $\mathbf{u} \wedge \mathbf{v}$ and $\mathbf{x}$ and there is at most one shortcut between $\mathbf{u} \wedge \mathbf{v}$ and $\mathbf{y}$. There is a path through $\mathbf{u} \wedge \mathbf{v}$,

$$\mathbf{u} \rightarrow \dots \rightarrow (\mathbf{u} \wedge \mathbf{v})\mathbf{w}_1 \rightarrow \mathbf{x} \rightarrow \mathbf{u} \wedge \mathbf{v} \rightarrow \mathbf{y} \rightarrow (\mathbf{u} \wedge \mathbf{v})\mathbf{w}_2 \rightarrow \dots \rightarrow \mathbf{v},$$

where $\mathbf{u} \rightarrow \dots \rightarrow (\mathbf{u} \wedge \mathbf{v})\mathbf{w}_1$ and $(\mathbf{u} \wedge \mathbf{v})\mathbf{w}_2 \rightarrow \dots \rightarrow \mathbf{v}$ are the same as those in path (7).

Since path (7) is the shortest path and it is possible that node $\mathbf{y}$ is node $\mathbf{u} \wedge \mathbf{v}$, the distance between $\mathbf{x}$ and $\mathbf{y}$ is at least 1. Therefore, if we look for the shortest path through $\mathbf{u} \wedge \mathbf{v}$, the length of this path differs from the length of the path (7) by at most one. The proof of Claim 2 is complete.

Next, we focus on the length of the sequence difference between two nodes and their latest common ancestor. For any pair of nodes u and v picked u.a.r. in $\text{GAN}_1(n)$ with sequences $\mathbf{u}$ and $\mathbf{v}$, respectively, assume that the individual in CTBP corresponding to $\mathbf{u} \wedge \mathbf{v}$ was born at the $\tau_{\mathbf{u} \wedge \mathbf{v}}$ th split. According to Section 2 in Ref. [24], we can write $L_u = \sum_{i=1}^{n_0} I_i$, where I_i are conditionally independent and equal 0 or 1 depending on whether the individual in the ancestral line is new-born at time t_i . Since we know that the number of active individuals (potential nodes) generated at each split is 2 and the total number of active individuals at the i th split is $i+1$, we can see that the indicators I_i are independent and $\mathbb{P}(I_i = 1) = \frac{2}{i+1}$. In the same way, we define $L_v = \sum_{i=1}^{n_0} I'_i$. Note that the event $(I_i, I'_i) = (1, 1)$ means that the ancestral lines of two nodes merge at the i th split. The joint conditional distribution of I_i and I'_i can be written as

$$\mathbb{P}((I_i, I'_i) = (1, 0) | \tau_{\mathbf{u} \wedge \mathbf{v}} < i) = \frac{2}{i+1} \frac{i+1-2}{i+1-1} = \frac{2(i-1)}{i(i+1)}, \quad (8)$$

$$\mathbb{P}((I_i, I'_i) = (0, 1) | \tau_{\mathbf{u} \wedge \mathbf{v}} < i) = \frac{2}{i+1} \frac{i+1-2}{i+1-1} = \frac{2(i-1)}{i(i+1)}, \quad (9)$$

$$\mathbb{P}((I_i, I'_i) = (1, 1), \tau_{\mathbf{u} \wedge \mathbf{v}} = i | \tau_{\mathbf{u} \wedge \mathbf{v}} \leq i) = \frac{2(2-1)}{i(i+1)} = \frac{2}{i(i+1)}. \quad (10)$$

By Eqs. (8)–(10),

$$\sum_{i=1}^{\infty} \mathbb{P}(\tau_{\mathbf{u} \wedge \mathbf{v}} = i | \tau_{\mathbf{u} \wedge \mathbf{v}} \leq i) = \sum_{i=1}^{\infty} \frac{2}{i(i+1)} < \infty.$$

It follows from

$$\mathbb{P}(\tau_{\mathbf{u} \wedge \mathbf{v}} \leq k) = \prod_{i=k+1}^{n_0} [1 - \mathbb{P}(\tau_{\mathbf{u} \wedge \mathbf{v}} = i | \tau_{\mathbf{u} \wedge \mathbf{v}} \leq i)],$$

that $\tau_{\mathbf{u} \wedge \mathbf{v}}$ has a limiting distribution, and $(\log n_0 - \log \tau_{\mathbf{u} \wedge \mathbf{v}}) / \log n_0$ converges to 1 in probability. This also means that $L_{\mathbf{u} \wedge \mathbf{v}}$ has a limiting distribution, independent of n_0 . Considering the length of the sequence difference between a potential node and its ancestor, we can obtain the following result.

Proposition 1. Suppose u and v are two potential nodes picked u.a.r. in $\text{GAN}_1(n)$ with size $n_0 + 2$, then,

$$\left(\frac{L_u - L_{\mathbf{u} \wedge \mathbf{v}} - 2 \log n_0}{\sqrt{2 \log n_0}}, \frac{L_v - L_{\mathbf{u} \wedge \mathbf{v}} - 2 \log n_0}{\sqrt{2 \log n_0}} \right) \xrightarrow{d} (Z, Z'), \quad (11)$$

where Z and Z' are independent standard normally distributed random variables.

Proof. The joint distribution of $L_u - L_{\mathbf{u} \wedge \mathbf{v}}$ and $L_v - L_{\mathbf{u} \wedge \mathbf{v}}$ can be written as

$$(L_u - L_{u \wedge v}, L_v - L_{u \wedge v}) = \left(\sum_{i=\tau_{u \wedge v}}^{n_0} I_i, \sum_{i=\tau_{u \wedge v}}^{n_0} I'_i \right).$$

Using Lindeberg central limit theorem and Eqs. (8)–(10) for linear combinations of $\sum_{i=\tau_{u \wedge v}}^{n_0} (aI_i + bI'_i)$, where a and b are two arbitrary constants, the two variables in (11) converge jointly to a two-dimensional standard normal variable in distribution.

3 Typical distance and flooding time

As prepared in the previous section, we now consider the typical distance and flooding time in the GAN model. We first state our main results in the following.

Theorem 2. (Typical distance) Let D_n be the typical distance of $\text{GAN}(n)$. Then as $n \rightarrow \infty$,

$$\frac{D_n - \frac{4}{3} \log n}{\sqrt{\log n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{16}{27}\right). \quad (12)$$

Theorem 3. (Flooding time) Let F_n be the flooding time of $\text{GAN}(n)$. Then as $n \rightarrow \infty$,

$$\frac{F_n}{\log n} \xrightarrow{p} \frac{2}{3} + c,$$

where c is a constant defined in Theorem 1.

To prove these two theorems, we need to use several lemmas as follows. In fact, we can define the radius of the network GAN as the length of the longest path from one of the initial nodes since the longest path of the network GAN must pass through one of the initial nodes.

Lemma 2^[6]. The radius of the subnetwork $\text{GAN}_i(n)$ is w.h.p. asymptotic to $c \log n$ with the constant c is defined in Theorem 1.

Lemma 3. Let R_{n_0} be the distance between the node picked u.a.r. in $\text{GAN}_i(n)$ and either of the initial nodes. If $n_0 \rightarrow \infty$, then

$$\frac{R_{n_0} - \frac{2}{3} \log n_0}{\sqrt{\log n_0}} \xrightarrow{d} \mathcal{N}\left(0, \frac{8}{27}\right). \quad (13)$$

Proof. For any existing node picked u.a.r. in $\text{GAN}_i(n)$, there is a potential node that makes the distance between these two nodes equal to 1. By Lemma 1, the conclusion (13) is easy to obtain.

Lemma 4. Let D_n^1 be the typical distance of $\text{GAN}_i(n)$. If $n_0 \rightarrow \infty$, then

$$\frac{D_n^1 - \frac{4}{3} \log n_0}{\sqrt{\log n_0}} \xrightarrow{d} \mathcal{N}\left(0, \frac{16}{27}\right).$$

Proof. Pick a pair of potential nodes u and v u.a.r. from $\text{GAN}_i(n)$ whose sequences are $\mathbf{u}$ and $\mathbf{v}$, respectively, and $u \wedge v$ is their latest common ancestor with the sequence $\mathbf{u} \wedge \mathbf{v}$. In the following descriptions, we will directly represent nodes as the sequence of nodes for convenience. Define the distinct postfixes $\tilde{\mathbf{u}}, \tilde{\mathbf{v}}$ after $\mathbf{u} \wedge \mathbf{v}$ by

$$\mathbf{u} = (\mathbf{u} \wedge \mathbf{v})\tilde{\mathbf{u}}, \quad \mathbf{v} = (\mathbf{u} \wedge \mathbf{v})\tilde{\mathbf{v}}.$$

By Lemma 1 and Claim 2 (constant 1 can be ignored), the length of the shortest path between $\mathbf{u}$ and $\mathbf{v}$ satisfies

$$\text{dist}(\mathbf{u}, \mathbf{v}) \stackrel{d}{=} S_{L_u - L_{u \wedge v}} + S_{L_v - L_{u \wedge v}}. \quad (14)$$

Similar to the proof of Lemma 1 and using the conclusion that $L_{u \wedge v}$ has a limiting distribution, we calculate

$$\mathbb{E}[S_{L_u - L_{u \wedge v}}] = \mathbb{E}[\mathbb{E}[S_{L_u - L_{u \wedge v}} | L_u - L_{u \wedge v}]] \sim \frac{2}{3} \log n_0,$$

and

$$\text{Var}(S_{L_u - L_{u \wedge v}}) \sim \frac{8}{27} \log n_0.$$

Observe that

$$\frac{S_{L_u - L_{u \wedge v}} - \frac{2}{3} \log n_0}{\sqrt{\frac{8}{27} \log n_0}} = \frac{S_{L_u - L_{u \wedge v}} - \frac{1}{3}(L_u - L_{u \wedge v})}{\sqrt{L_u - L_{u \wedge v}}} \sqrt{\frac{L_u - L_{u \wedge v}}{\frac{8}{27} \log n_0}} + \frac{\frac{1}{3}(L_u - L_{u \wedge v}) - \frac{2}{3} \log n_0}{\sqrt{\frac{8}{27} \log n_0}}.$$

Since $S_{L_u - L_{u \wedge v}} | L_u - L_{u \wedge v}$ is independent of $L_u - L_{u \wedge v}$ and the lengths of the sequences $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are independent, we can obtain the following conclusion in the same way as the proof of Lemma 1.

$$\frac{S_{L_u - L_{u \wedge v}} - \frac{2}{3} \log n_0}{\sqrt{\frac{8}{27} \log n_0}} \xrightarrow{d} \mathcal{N}(0, 1).$$

By conditioning first on $L_{u \wedge v}$ and using the fact that the symbols in the sequence $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are i.i.d., it can be shown that $(S_{L_u - L_{u \wedge v}} - (2/3) \log n_0, S_{L_v - L_{u \wedge v}} - (2/3) \log n_0) / \sqrt{\log n_0}$ converges jointly to two independent copies of $\mathcal{N}(0, 8/27)$ in distribution.

By Eq. (14), the distance between two potential nodes picked u.a.r. in $\text{GAN}_i(n)$ has the following property.

$$\frac{\text{dist}(\mathbf{u}, \mathbf{v}) - \frac{4}{3} \log n_0}{\sqrt{\log n_0}} \xrightarrow{d} \mathcal{N}\left(0, \frac{16}{27}\right). \quad (15)$$

As shown in Fig. 2, each node in $\text{GAN}_i(n)$ is connected to two potential nodes (except the initial nodes). For randomly selected nodes u' and v' in $\text{GAN}_i(n)$ with sequences $\mathbf{u}'$ and $\mathbf{v}'$, respectively, suppose that the nearest potential nodes are $\mathbf{u}$ and $\mathbf{v}$, respectively. Obviously, node $u'(v')$ is one of the ancestors of potential node $\mathbf{u}(\mathbf{v})$. Then,

$$\mathbf{u} = (\mathbf{u} \wedge \mathbf{v})\tilde{\mathbf{u}} = \mathbf{u}'w_1, \quad \mathbf{v} = (\mathbf{u} \wedge \mathbf{v})\tilde{\mathbf{v}} = \mathbf{v}'w_2, \quad (16)$$

where w_1 and w_2 represent the difference between the two sequences. Define the distinct postfixes $\tilde{\mathbf{u}}', \tilde{\mathbf{v}}'$ after $\mathbf{u}' \wedge \mathbf{v}'$ by

$$\mathbf{u}' = (\mathbf{u}' \wedge \mathbf{v}')\tilde{\mathbf{u}}', \quad \mathbf{v}' = (\mathbf{u}' \wedge \mathbf{v}')\tilde{\mathbf{v}}'. \quad (17)$$

By Eqs. (16) and (17),

$$u \wedge v = u' \wedge v'.$$

Then, the distance relationship among these nodes can be stated as follows.

$$\text{dist}(u', v') = \text{dist}(u', u' \wedge v') + \text{dist}(v', u' \wedge v'),$$

$$|\text{dist}(u', u' \wedge v') - \text{dist}(u, u \wedge v)| \leq 1,$$

$$|\text{dist}(v', u' \wedge v') - \text{dist}(v, u \wedge v)| \leq 1.$$

Therefore, the distance between any pair of nodes selected u.a.r. from $\text{GAN}_1(n)$ satisfies the asymptotic normality (15). The proof of this lemma is now complete.

Proof of Theorem 1. For scenario (A), as $n \rightarrow \infty$, since D_n^1 is the typical distance of the subnetwork $\text{GAN}_1(n)$, we can use fact (a) and Lemma 4 to obtain

$$\frac{D_n^1 - \frac{4}{3} \log n}{\sqrt{\log n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{16}{27}\right).$$

This means that the distance between a randomly chosen pair of nodes in a randomly chosen initial interval has the above property due to the symmetry of $\text{GAN}(n)$.

For scenario (B), the path between any two nodes that originate from two different initial intervals must pass through one of the initial nodes. Consider these two nodes to have originated from I_1 and I_2 , respectively. Because of the independence of each node in GAN_1 and GAN_2 , we can see that the distance between the two nodes has the same distribution as $2R_{n_0}$, which is mentioned in Lemma 3. By (13) and (1), as $n \rightarrow \infty$, $2R_{n_0}$ satisfies

$$\frac{2R_{n_0} - \frac{4}{3} \log n}{\sqrt{\log n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{16}{27}\right).$$

Therefore, the distances between any two randomly selected nodes from two of any randomly selected initial intervals satisfy the above property.

After combining the conclusions of these two scenarios, it is not difficult to conclude that the distance between pairs of nodes in $\text{GAN}(n)$ satisfies (12).

Proof of Theorem 2. The flooding time in $\text{GAN}(n)$ can be expressed as

$$\max_v \text{dist}(u, v),$$

where u is a node picked u.a.r. in $\text{GAN}(n)$. With high probability, the node most distant from u is in another initial interval. Therefore,

$$F_n \stackrel{d}{=} R_{n_0} + \text{radius}(\text{GAN}_i). \quad (18)$$

Without loss of generality, we consider the properties of the distance in GAN_1 . By Lemma 3, $\mathbb{E}[R_{n_0}] \sim \frac{2}{3} \log n_0$ and $\text{Var}(R_{n_0}) \sim \frac{8}{27} \log n_0$. If $n_0 \rightarrow \infty$, by Chebyshev's inequality,

$$P\left(\left|\frac{R_{n_0}}{\log n_0} - \frac{2}{3}\right| \geq \varepsilon\right) \leq \frac{\text{Var}(R_{n_0})}{(\log n_0)^2 \varepsilon^2} \rightarrow 0,$$

i.e.,

$$\frac{R_{n_0}}{\log n_0} \xrightarrow{p} \frac{2}{3}.$$

Using fact (a), as $n \rightarrow \infty$, we obtain

$$\frac{R_{n_0}}{\log n} \xrightarrow{p} \frac{2}{3}. \quad (19)$$

The radius of the subnetwork $\text{GAN}_i(n)$ is given in Lemma 2. Therefore, by (18), (19), and Lemma 2, as $n \rightarrow \infty$,

$$\frac{F_n}{\log n} \xrightarrow{p} \frac{2}{3} + c,$$

where the constant c is defined in Theorem 1.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (11771418).

Conflict of interest

The authors declare that they have no conflicts of interest.

Biographies

Ziling Xu is a postgraduate student of the University of Science and Technology of China. Her research mainly focuses on random network.

Qunqiang Feng is currently an Associate Professor at the University of Science and Technology of China (USTC). He received his Ph.D. degree from USTC in 2006. His research mainly focuses on applied probability, random network models, and network data analysis.

References

- [1] Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press, **1994**.
- [2] Wu J, Tse C K, Lau F C M, et al. Analysis of communication network performance from a complex network perspective. *IEEE Transactions on Circuits and Systems*, **2013**, 60: 3303–3316.
- [3] Newman M E J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, **2001**, 98: 404–409.
- [4] Alm E, Arkin A P. Biological networks. *Current Opinion in Structural Biology*, **2003**, 13: 193–202.
- [5] Ozik J, Hunt B R, Ott E. Growing networks with geographical attachment preference: Emergence of small worlds. *Physical Review E*, **2004**, 69: 026108.
- [6] Feng Q, Wang Y, Hu Z. Small-world effect in geographical attachment networks. *Probability in the Engineering and Informational Sciences*, **2021**, 35: 276–296.
- [7] Hayashi Y. A review of recent studies of geographical scale-free networks. *Information and Media Technologies*, **2006**, 1: 1136–1145.
- [8] Zhang Z, Rong L, Comellas F. Evolving small-world networks with geographical attachment preference. *Journal of Physics A: Mathematical and General*, **2006**, 39: 3253.
- [9] Zhang Z, Rong L, Guo C. A deterministic small-world network created by edge iterations. *Physica A: Statistical Mechanics and its*

- Applications*, **2006**, 363: 567–572.
- [10] Barabási A L, Albert R. Emergence of scaling in random networks. *Science*, **1999**, 286: 509–512.
 - [11] Kolossváry I, Komjáthy J, Vágó L. Degrees and distances in random and evolving Apollonian networks. *Advances in Applied Probability*, **2016**, 48: 865–902.
 - [12] Zhou T, Yan G, Wang B. Maximal planar networks with large clustering coefficient and power-law degree distribution. *Physical Review E*, **2005**, 71: 046141.
 - [13] Andrade Jr J S, Herrmann H J, Andrade R F, et al. Apollonian networks: simultaneously scale-free, small world, Euclidean, space filling, and with matching graphs. *Physical Review Letters*, **2005**, 94: 018702.
 - [14] Zhang Z, Rong L, Zhou S. Evolving Apollonian networks with small-world scale-free topologies. *Physical Review E*, **2006**, 74: 046105.
 - [15] Abdullah M A, Bode M, Fountoulakis N. Typical distances in a geometric model for complex networks. arXiv: 1506.07811, **2015**.
 - [16] Dereich S, Mönch C, Mörters P. Typical distances in ultrasmall random networks. *Advances in Applied Probability*, **2012**, 44: 583–601.
 - [17] Bhamidi S, van der Hofstad R, Hooghiemstra G. First passage percolation on the Erdős–Rényi random graph. *Combinatorics, Probability and Computing*, **2011**, 20: 683–707.
 - [18] Bhamidi S, van der Hofstad R. Weak disorder asymptotics in the stochastic mean-field model of distance. *Advances in Applied Probability*, **2012**, 22: 29–69.
 - [19] van der Hofstad R, Hooghiemstra G, van Mieghem P. The flooding time in random graphs. *Extremes*, **2002**, 5: 111–129.
 - [20] Camargo D, Popov S. Total flooding time and rumor propagation on graphs. *Journal of Statistical Physics*, **2017**, 166: 1558–1571.
 - [21] Amini H, Draief M, Lelarge M. Flooding in weighted sparse random graphs. *SIAM Journal on Discrete Mathematics*, **2013**, 27: 1–26.
 - [22] Mountford T, Saliba J. Flooding and diameter in general weighted random graphs. *Journal of Applied Probability*, **2020**, 57: 956–980.
 - [23] Athreya K B, Ney P E. *Branching Processes*. Berlin: Springer, **1972**.
 - [24] Bühler W J. Generations and degree of relationship in supercritical Markov branching processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **1971**, 18: 141–152.
 - [25] Feller W. *An Introduction to Probability Theory and Its Applications*, Vol. 1. 3rd Edition. New York: Wiley, **2008**.