Probability in the Engineering and Informational Sciences, **15**, 2001, 225–237. Printed in the U.S.A.

FIRST-PASSAGE PERCOLATION ON THE RANDOM GRAPH

Remco van der Hofstad and
Gerard Hooghiemstra
ITS
Department of Mathematics
Delft University of Technology
2628 CD Delft, The Netherlands
E-mail: R.W.vanderHofstad@its.tudelft.nl
E-mail: G.Hooghiemstra@its.tudelft.nl

Piet Van Mieghem
ITS
Department of Electrical Engineering
Delft University of Technology
2628 CD Delft, The Netherlands
E-mail: p.vanmieghem@its.tudelft.nl

We study first-passage percolation on the random graph $G_p(N)$ with exponentially distributed weights on the links. For the special case of the complete graph, this problem can be described in terms of a continuous-time Markov chain and recursive trees. The Markov chain $X(t)$ describes the number of nodes that can be reached from the initial node in time t. The recursive trees, which are uniform trees of N nodes, describe the structure of the cluster once it contains all the nodes of the complete graph. From these results, the distribution of the number of hops (links) of the shortest path between two arbitrary nodes is derived.

We generalize this result to an asymptotic result, as $N \to \infty$, for the case of the random graph where each link is present independently with a probability p_N as long as $Np_N/(\log N)^3 \to \infty$. The interesting point of this generalization is that (1) the limiting distribution is insensitive to p and (2) the distribution of the number of hops of the shortest path between two arbitrary nodes has a remarkable fit with shortest path data measured in the Internet.

1. INTRODUCTION

The main result in this article is Theorem 2.1. This theorem contains a first-passage result on the random graph $G_p(N)$, where the probability $p = p_N$ of an open edge

 0269-9648/01 $12.50

tends to zero as $N \to \infty$, in such a way that $Np_N \to \infty$. More specifically, to the open edges we attach weights given by independent exponential random variables each with mean 1 and ask for the number of edges (hops) H_N of the shortest path between the nodes 1 and N. The shortest path is the a.s. *unique* path from 1 to N that minimizes the sum of the weights on the edges of the path.

Given a condition on the speed with which $Np_N \to \infty$ (see Thm. 2.1), we prove that the hopcount H_N can be coupled (in an asymptotic sense) to a random variable R_N, with generating function

$$\mathbb{E}(z^{R_N}) = \frac{N}{N-1}\left(\varphi_N(z) - \frac{1}{N}\right), \tag{1}$$

where φ_N is the generating function

$$\varphi_N(z) = \frac{\Gamma(z+N)}{\Gamma(N+1)\Gamma(z+1)} = \frac{N^{z-1}}{\Gamma(z+1)}\,(1 + O(N^{-1})) \tag{2}$$

and where $\Gamma(z)$ denotes the Gamma function (cf. [1, Sect. 6.1.1]). We also show that

$$\mathbb{E}(H_N) \sim \log N + \gamma - 1 \quad \text{if } Np_N/(\log N)^6 \to \infty, \tag{3}$$

$$\mathrm{Var}(H_N) \sim \log N + \gamma - \pi^2/6 \quad \text{if } Np_N/(\log N)^9 \to \infty, \tag{4}$$

where $\gamma \approx 0.5772$, is Euler's constant.

This theorem is a generalization of scattered known results for the complete graph K_N $\left(\text{the graph with } N \text{ nodes and all } \binom{N}{2} \text{ edges present}\right)$ and results for the uniform recursive tree. For K_N, this theorem seems to be part of the folklore of epidemic theory, although we did not find a precise reference to the result that on the complete graph with exponential weights, the hopcount H_N is connected with the height L_N of a uniform recursive tree and

$$\mathbb{E}(z^{L_N}) = \varphi_N(z),$$

with φ_N given by (2) (cf. Smythe and Mahmoud [5]).

The question arises why this generalization to $G_p(N)$ is interesting. Recent measurements, both at Delft University of Technology [7] and the University of Gent [6], of the number of hops that have to be traversed between two arbitrary nodes in the Internet show a remarkable fit between this data and the distribution with generating function (2) (cf. [7]). It is known that the IP packets in the Internet are routed according to a shortest-path algorithm (the so-called Dijkstra algorithm), where the weights between the routing machines are set by the operators.[1] We therefore model the Internet as a graph with weights on the edges. Since many details of the Internet are unknown and the Internet seems rather chaotic in structure, we choose a *random* graph with *random* weights as a model. As far as we know, the result is the first closed-form expression for the hopcount of the Internet, which, according to [8, Sect. 2.2.2], has remained impregnable.

The complete graph K_N is obviously not a good choice to model the Internet, because the number of edges extending from each node (router) equals $N - 1$,

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

whereas in reality, this number is restricted (in most cases by 32). However, every graph with N nodes and, hence, the graph representation of the Internet are subgraphs of K_N. Therefore, we randomly thin the number of edges on the complete graph as far as we could (see Thm. 2.1) by erasing links in an independent and identically distributed (i.i.d.) fashion and prove that the hopcount in the random graph $G_p(N)$ still has the same (asymptotic) distribution as the hopcount of K_N as long as $Np_N/(\log N)^3 \to \infty$. Moreover, simulations of the hopcount of the random graph with $p_N = \lambda/N$, where $\lambda = 10$, and with exponential weights on the edges show that the simulated distribution still resembles a distribution with generating function (2) (see [7]). The number $\lambda = Np_N$, which equals the average number of outgoing links in $G_p(N)$, is close to the genuine number of ports of a router in the Internet.

Our next point is the explanation of the choice of the exponential weights. First, note that exponential weights are comparable with uniform [0,1] weights, because the exponential distribution and the uniform distribution are both in the same *minimal* domain of attraction: For both exponential and uniform (i.i.d.) random variables $X_1, X_2, \ldots$, we have that

$$n \min_{1 \le k \le n} X_k$$

converges to a random variable with an extreme value distribution given by $1 - e^{-x}$, for $x > 0$ (cf. [4]). We only have a statistical motivation to use exponential (or uniform weights); other weights (constant weights or i.i.d. weights with distribution function $F(x) = x^\alpha$, $x \in [0,1]$, with $\alpha \neq 1$) do not fit the data (see [7]). For $p_N = \lambda/N$ and weights constantly equal to 1, we have that the shortest path uses the minimal number of hops. In this case, it can readily be seen that $\mathbb{E}(H_N) \sim \log N/\log \lambda$ and that $\mathrm{Var}(H_N)$ is bounded for $N \to \infty$ (see [7]). Hence, our result can be formulated as the statement that for the Internet there is a great variability for the link weights. In any case, the weights[2] are surely not all equal to 1 as it initially was with RIP (Routing Information Protocol).

To explain the method of proof of our result for the random graph, we return to the complete graph with exponential weights on the edges. For the complete graph K_N, the proof of (1)–(4) is as follows. Consider a continuous-time Markov chain $\{X(t)\}_{t \ge 0}$, which is a pure birth process with state space $\{1,2,\ldots,N\}$ and birth rate $\lambda_n = n(N-n)$. The random variable $X(t)$ represents the number of nodes that can be reached from node 1 in a travel time less than or equal to t. The process $\{X(t)\}_{t \ge 0}$ starts at time 0 with one particle (node) and will eventually be absorbed in state N, when all the nodes can be reached.

Observe that the process that describes the number of distinct nodes (including node 1) that can be reached in K_N over the exponential edges starting from 1 within time t is indeed equal in distribution to the process $\{X(t)\}_{t \ge 0}$. This follows from the memoryless property of the exponential distribution and because when n nodes are reached, each of these n nodes can be connected to the set of $N - n$ remaining nodes over $N - n$ different edges, which explains the rate $\lambda_n = n(N-n)$. When $X(t) = n$, the (not previously used) edges between the n nodes can be omitted. These edges do

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

not belong to the shortest path, otherwise they would have been selected at an earlier time.

Geometrically, the evolution of the above birth process can be visualized by a (random) recursive tree, which is a uniform tree of N nodes. Indeed, each birth in the Markov process corresponds to connecting an edge of unit length to one of the existing nodes in the associated tree. It follows from the Markov property that the new edge is connected *randomly* to one of the existing nodes of the tree, which implies that this tree is a *uniform* tree. The hopcount is hence equal to the *height* L_N of a uniformly chosen particle in the tree. It is well known (cf. [5]) that the height of an arbitrary point (including the root) has generating function (2). In our problem, N cannot be the root, so that the result (1), with $R_N = H_N$, for the complete graph follows.

Using the above description, one can also compute the generating function of the total weight W_N of the shortest path. Since the tree is uniform, each of the $N-1$ possibilities of positions for node N is equally likely; furthermore, the generating function of the (independent) sum $X_1 + \cdots + X_k$, where X_i is exponentially distributed with parameter $i(N-i)$, equals

$$\mathbb{E}(e^{t(X_1+\cdots+X_k)}) = \prod_{i=1}^{k} \frac{i(N-i)}{i(N-i)-t}.$$

Hence,

$$\mathbb{E}(e^{tW_N}) = \frac{1}{N-1} \sum_{k=1}^{N-1} \prod_{i=1}^{k} \frac{i(N-i)}{i(N-i)-t}. \tag{5}$$

In the next section, we extend the results (1)–(4) to the class $G_p(N)$, where $p = p_N$ is chosen such that

$$\frac{Np_N}{(\log N)^3} \to \infty. \tag{6}$$

This is a technical condition. From the famous connectivity theorem of Erdös and Rényi, it follows that the random graph is, with large probability, disconnected when $Np_N/\log N < 1$, whereas it is with large probability connected when $Np_N/\log N > 1$ (see [2]). Therefore, $p_N = (\log N)/N$ is called the *connectivity threshold*. Since the Internet is connected, we can restrict ourselves to the case where $Np_N/\log N > 1$. Moreover, the *percolation threshold* on the complete graph is $p_N = 1/N$. Hence, for $Np_N > 1$, the largest cluster is of the order N, whereas for $Np_N < 1$, the largest cluster is of order $\log N$. Hence, we see that for p_N such that $Np_N \to \infty$, the probability that the source and the destination are in the largest cluster converges to 1. We expect that our limit laws in (1)–(4) remain valid even in this regime, when we condition the source and the destination to be in the largest cluster. Therefore, we believe our results to remain valid below the connectivity threshold. Simulations with $Np_N = 10$ and $N = 210{,}000$ do confirm this.

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

For the random graph $G_p(N)$, each node has a random number of links. The above proof for the complete graph was based on the fact that from each node in a cluster of size n, there are a *constant* number $(N-n)$ of outgoing links (i.e., edges going to nodes outside the present cluster). Now, for the random graph, for each node in the cluster of the root when this cluster has size n, the number of outgoing links is binomial with parameters p and $N-n$. These binomial random variables can be sandwiched in between two *constant* numbers of outgoing links in each node of the cluster of size n equal to

$$\lceil (N-n)p_N \pm \sqrt{A(N-n)p_N(1-p_N)\log N}\,\rceil, \tag{7}$$

which is defined to be zero when (7) becomes negative and where A is a positive number to be determined later. To each of this constant number of outgoing links, there belong continuous-time Markov chains $X^{\pm}(t)$, which is a pure birth process with state space $\{1,2,\ldots,N^{\pm}\}$, where

$$N^{\pm} = \lceil N(1 \pm A(1-p_N)\log N/(Np_N))\rceil, \tag{8}$$

and with birth rates $\lambda_n^{\pm}$ equal to n times the quantity given in (7). Observe that the size $N^{\pm}$ equals the smallest value of n for which $\lambda_n^{\pm} \le 0$. We next show that with high probability, the hopcount of the shortest path of the *uniform* tree belonging to the Markov chains $X^{-}(t)$ and the hopcount of the shortest path of the random graph $G_p(N)$ are the same. Hence, (1)–(4) hold when $\log N^{-} = \log N + o(1)$, which implies that $Np_N/\log N \to \infty$. In fact, in the technical part of the proof, we need that $Np_N/(\log N)^{\beta} \to \infty$, where the value of β depends on whether we wish to couple the respective random variables, prove convergence of the mean, or prove convergence of the variance.

The result for the hopcount of the random graph and the insensitivity with respect to the value of p can also be explained intuitively. From (2), it is seen that the law of R_N is close to the Poisson law with parameter $\log N$. This can be explained as follows. The probability that there is a path of k edges that has a sum of exponentials not exceeding L is approximately equal to the number of such paths times the probability that the sum of k i.i.d. exponential variables with mean 1 is less than L. The number of paths of length k from 1 to N is, for N large, roughly equal to N^{k-1}. The probability that the sum of exponential weights is less than or equal to L is roughly equal to $L^k/k!$. Multiplying out, we find that $\mathbb{P}(H_N = k, W_N \le L) \approx (LN)^k/Nk!$. These probabilities have to sum up to 1 when L is the typical size of the weight of the shortest path, so that L has to be equal to $(\log N)/N$. Substitution of this value gives $\mathbb{P}(H_N = k) \approx (\log N)^k/Nk!$, in accordance to (2). For the random graph $G_p(N)$, where edges of the complete graph are present or absent independently with probability p and $1-p$, respectively, the weight W_N has to be of the order $(\log N)/Np_N$ (i.e., the value of p merely serves as a scale factor). The reason for this is that p only decreases the *number* of links, which means that we take the minimum over less exponential random variables. Now, for integer Np, the minimum over Np exponential random variables has the *same* distribution as $1/p$ times the minimum over N

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

exponential random variables. This explains that p only serves as a scale factor. The limiting distribution of the hopcount remains unchanged. The insensitivity with respect to p of the law of the hopcount can be understood by adapting the above heuristic to the case where $W_N \approx (\log N)/Np_N$ and where the number of paths of lengths k is replaced by the *expected* number of paths of length k which is equal to $p_N^k N^{k-1}$. We see that the factors of p_N cancel out, and we find that the asymptotics of the hopcount is independent of p_N.

The next section, which contains the full proof of our main result, is organized as follows. We start with the statement of the theorem. Then, in Lemmas 2.2 and 2.3, we show with large deviation theory that the hopcount H_N is bounded with overwhelming probability by a large multiple of $\log N$. Lemma 2.4 is used in the proof of our main theorem to show that we can assume that the number of outgoing links is between a fixed upper and lower bound to obtain a uniform tree. We finally couple the height L_N of nodes of this uniform tree to the hopcount of the random graph in Lemma 2.5.

2. THE RANDOM GRAPH

In this section, we investigate the hopcount of the random graph $G_p(N)$ with exponential travel times on the edges. We always assume that we are dealing with sequences p_N satisfying $\limsup_N p_N < 1$, so that the random graph is truly random. The main result is the following theorem.

THEOREM 2.1: *There exists a probability space on which the hopcount H_N of $G_p(N)$ and a random variable H_N^- can be defined simultaneously, and where the marginal distribution of H_N^- has generating function (1) with $N = N^-$ given by (8), such that the following hold:*

(i) *If* $Np_N/(\log N)^3 \to \infty$, *then* $\mathbb{P}(H_N \neq H_N^-) = o(1)$.
(ii) *If* $Np_N/(\log N)^6 \to \infty$, *then* $\mathbb{E}(H_N) = \log N + \gamma - 1 + o(1)$.
(iii) *If* $Np_N/(\log N)^9 \to \infty$, *then* $\mathrm{Var}(H_N) = \log N + \gamma - \pi^2/6 + o(1)$.

The proof is divided into a number of steps. We first sketch these steps and then formulate and prove them in a series of lemmas. Finally, we prove Theorem 2.1.

1. As indicated by the results (3) and (4), we expect that the probability that the hopcount H_N exceeds a large multiple of $\log N$ is small. This result is important for the proof of our theorem, because it gives an upper bound on the number of nodes with which we have to deal.

   If the hopcount H_N is bounded by a multiple of $\log N$, then the exponential weights over the shortest path are likely to be bounded by another multiple of $\log N$ times the typical weight over each edge of the shortest path. These typical weights are of order $(Np_N)^{-1}$. The size of a typical weight of an edge belonging to the shortest path follows, because each node has, on the average, Np_N edges and the minimum of Np_N independent exponentials each with weight 1 has expectation $(Np_N)^{-1}$. In Lemma 2.2, we will show that

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

$\mathbb{P}(Np_N W_N > B \log N) \le N^{-\delta B}$, for some $\delta > 0$. We prove this lemma with the help of Cramér's theorem (cf. [3, p. 26]).

2. Using Lemma 2.2, we prove that the bound $H_N \le B^2 \log N$ holds with overwhelming probability. This will be shown in Lemma 2.3.
3. For a binomial random variable X_N with parameters k_N and $p = p_N$ such that $(\log N)/(k_N p_N(1-p_N)) \to 0$,

$$\mathbb{P}\big(X_N \notin \big[k_N p_N - \sqrt{Ak_N p_N(1-p_N)\log N}, k_N p_N + \sqrt{Ak_N p_N(1-p_N)\log N}\big]\big) \le 4N^{-A}.$$

This will be proven in Lemma 2.4.

4. We couple H_N with a random variable H_N^-, which is the number of hops of a uniformly chosen point in a *uniform* tree of size $N^- < N$, where $N^- = \lceil N(1 - A(1-p_N)\log N/(Np_N))\rceil$. Let

$$A_N = \{H_N = H_N^-\}.$$

The main ingredient of the proof is that $\mathbb{P}(A_N^c) \to 0$ at a certain rate that depends on how $Np_N \to \infty$. The random variable H_N^- has generating function

$$\mathbb{E}(z^{H_N^-}) = \frac{N^-}{N^- - 1}\left(\varphi_{N^-}(z) - \frac{1}{N^-}\right), \tag{9}$$

where φ_N is the generating function in (2). Hence, the ratio of the generating functions $\mathbb{E}(z^{H_N^-})$ and $\varphi_N(z)$ tends to 1 as long as $Np_N/\log N \to \infty$.

5. The asymptotic expressions for $\mathbb{P}(H_N \ne H_N^-)$, $\mathbb{E}(H_N)$, and $\mathrm{Var}(H_N)$ then follow.

We start with Step 1. Let W_N denote the sum of the exponential weights along the shortest path from 1 to N in the graph $G_p(N)$.

LEMMA 2.2: *There exists constants $\delta > 0$ and B such that for Np_N large,*

$$\mathbb{P}(Np_N W_N > B \log N) \le N^{-\delta B}. \tag{10}$$

PROOF: The idea behind this proof is that starting from node 1 we build a *binary* tree by choosing at each node the two shortest edges (shortest with respect to the exponential weights). The size of this tree grows as 2^k, where k is the depth of the tree. Hence, within $k = (\log N)/\log 2$ steps, we have reached all N nodes. However, if $k \approx (\log N)/\log 2$, the number of nodes not yet in the binary tree approaches 0, and, therefore, the weight of the minimal edges has expectation almost 1, which is large compared to $(Np_N)^{-1}$. Therefore, we grow two binary trees: one with root 1 and a second with root N. If we grow both trees until they reach size $\sqrt{N}$, then there are still $N - O(\sqrt{N})$ nodes not in these trees, which implies that all weights in the trees are of order $(Np_N)^{-1}$. Moreover, the number of connections between the two trees is of order $\sqrt{Np_N}\sqrt{Np_N} = Np_N$ and, hence, the minimal weight of the connecting edges is of the same reciprocal order $(Np_N)^{-1}$.

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

Indeed, in $G_p(N)$, we denote the exponentially distributed weights on the edges incident with node i by E_k^i if the edge (i,k) is present. Furthermore,

$$E_{(1)}^i < E_{(2)}^i < \cdots$$

are the ordered weights of the edges incident with i. Define a binary (random) subtree $B_1 \subset G_p(N)$ of depth k in the following way: start at node 1 and take the two edges with weight $E_{(1)}^1$ and $E_{(2)}^1$. Let i and j denote the end points of these two edges. From the collection of edges incident to $i(j)$, we remove the edge $(1,i) = (i,1)((1,j) = (j,1))$ and from the remaining set of edges incident with $i(j)$, we take the two shortest ones. Proceeding this way, we grow a binary tree with depth

$$k = \left\lceil \frac{\log \sqrt{N}}{\log 2} \right\rceil, \tag{11}$$

where $\lceil x \rceil$ is the smallest integer larger than x. If $N \notin B_1$, grow a binary tree of depth k starting from node N, without using any of the nodes in tree B_1.

For $i \in \{1,2,\ldots,N\}$ and with X_i the number of remaining edges incident to i,

$$E_{(1)}^i = \min_j E_j^i \stackrel{d}{=} \frac{E_1}{X_i}, \qquad E_{(2)}^i \stackrel{d}{=} \frac{E_1}{X_i} + \frac{E_2}{X_i - 1},$$

by properties of the exponential distribution. Hence, if $X_i \geq \frac{1}{2}Np_N + 1$, then

$$E_{(1)}^i \leq \frac{2E_1}{Np_N}, \qquad E_{(2)}^i \leq \frac{2E_1 + 2E_2}{Np_N}, \tag{12}$$

where, as earlier, E_1 and E_2 are independent exponential random variables with mean 1.

From (12) and the fact that the minimal weight of the connecting edges can also be bounded by $(2E_1 + 2E_2)/Np_N$, we conclude that $W_N \leq 2S_{4k+1}/Np_N$, where S_n is the sum of n independent exponentials with mean 1. Hence,

$$\mathbb{P}(Np_N W_N \geq B \log N) \leq \mathbb{P}\left(S_{4k+1} \geq \frac{B \log N}{2}\right).$$

Now, apply Cramér's theorem to S_{4k+1} with k given in (11). ■

As a corollary to Lemma 2.2 we have the following lemma.

LEMMA 2.3: *There exists constants $\delta > 0$ and B such that for Np_N sufficiently large,*

$$\mathbb{P}(H_N > B^2 \log N) \leq 2N^{-\delta B}.$$

Moreover, the same bound holds for R_N, which is the number of hops of a uniform chosen point in a uniform tree of size N.

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

PROOF: Intersect the event $\{H_N > B^2 \log N\}$ with the event $\{Np_N W_N > B \log N\}$ and its complement to obtain

$$\begin{aligned}
&\mathbb{P}(H_N > B^2 \log N) \\
&\quad = \mathbb{P}(Np_N W_N > B \log N, H_N > B^2 \log N) \\
&\qquad + \mathbb{P}(Np_N W_N \le B \log N, H_N > B^2 \log N) \\
&\quad \le \mathbb{P}(Np_N W_N > B \log N) + \mathbb{P}(Np_N W_N \le B \log N, H_N > B^2 \log N) \\
&\quad \le N^{-\delta B} + \mathbb{P}(S_{\lceil B^2 \log N \rceil} \le B \log N) \\
&\quad \le 2N^{-\delta B},
\end{aligned}$$

where $\mathbb{P}(S_{\lceil B^2 \log N \rceil} \le \mathrm{B} \log N) \le N^{-\delta B}$ by Cramér's theorem.

To see that the same bound also holds for R_N, the random variable with generating function (2), use

$$\mathbb{P}(R_N > B \log N) \le \min_{t>0} \mathbb{P}(e^{tR_N} > N^{tB}) \le 2 \min_{t>0} N^{-tB} \frac{N^{e^t}}{\Gamma(e^t + 1)},$$

where we use the asymptotic expression in (2) for N large enough. Pick $t = \log B$ to get

$$\mathbb{P}(R_N > B \log N) \le N^{-B(\log B - 1)} \frac{2}{\Gamma(B+1)}.$$

This bound is, in fact, sharper than the upper bound for $\mathbb{P}(H_N > B^2 \log N)$. ■

LEMMA 2.4: *For a binomial random variable X_N with parameters k_N and p_N satisfying $(\log N)/(k_N p_N(1-p_N)) \to 0$, uniformly in k_N and p_N for large N,*

$$\begin{aligned}
\mathbb{P}\big(X_N \notin \big[&k_N p_N - \sqrt{A k_N p_N (1-p_N) \log N}, k_N p_N \\
&+ \sqrt{A k_N p_N (1-p_N) \log N}\big]\big) \le 4N^{-A}.
\end{aligned}$$

PROOF: For $A > 0$, define

$$C_N = \sigma_N \sqrt{A \log N},$$

where $\sigma_N^2 = k_N p_N (1-p_N)$. Then,

$$\mathbb{P}(X_N > k p_N + C_N) \le \inf_{t>0} \mathbb{P}(e^{tX_N} > e^{t k p_N + C_N}) \le \inf_{t>0} \{e^{-t(k_N p_N + C_N)} (\phi(t))^{k_N}\},$$

where $\phi(t) = 1 - p_N + p_N e^t$. For $k_N(1-p_N) > C_N$, we find that the argument t_N of the infimum satisfies

$$e^{t_N} = \frac{\sigma_N^2 + C_N(1-p_N)}{\sigma_N^2 - C_N p_N}.$$

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

From this, we obtain

$$\mathbb{P}(X_N > kp_N + C_N) \le \left(1 + \frac{C_N}{\sigma_N^2 - C_N p_N}\right)^{-(k_N p_N + C_N)} \left(1 + \frac{C_N p_N}{\sigma_N^2 - C_N p_N}\right)^{k_N}.$$

Hence, for $C_N/\sigma_N^2 \to 0$ or, equivalently, $(\log N)/\sigma_N^2 \to 0$, as $N \to \infty$,

$$\mathbb{P}(X_N > kp_N + C_N) \le 2\exp\left(-\frac{C_N^2}{\sigma_N^2 - C_N p_N}\right) \le 2N^{-A}.$$

To treat $\mathbb{P}(X_N < kp_N - C_N)$, define $Y_N = k_N - X_N$; then, Y_N has a binomial distribution with parameters k_N and $1 - p_N$ and

$$\begin{aligned}\mathbb{P}(X_N < k_N p_N - C_N) &= \mathbb{P}(k_N - Y_N < k_N p_N - C_N)\\ &= \mathbb{P}(Y_N > k_N(1 - p_N) + C_N).\end{aligned}$$

The result follows from repeating the above argument with X_N replaced by Y_N and p_N by $1 - p_N$. ■

LEMMA 2.5: *There exists a probability space on which the hopcount H_N of $G_p(N)$ and a random variable H_N^- can be defined simultaneously and where the marginal distribution of H_N^- has generating function (1) with $N = N^-$ given by (8), such that for $Np_N \to \infty$ and* $\limsup p_N < 1$,

$$\mathbb{P}(H_N \ne H_N^-) = O\left(\frac{\log N}{[Np_N]^{1/3}}\right). \tag{13}$$

Moreover,

$$\mathbb{E}(z^{H_N^-}) = \varphi_N(z)(1 + o(1))$$

as long as $Np_N/\log N \to \infty$.

PROOF: The method of proof is described in Step 4 at the beginning of this section.

Define $k_N = O((N \log N)/(Np_N)^{1/3})$ (this choice of k_N will become clear at the end of the proof) and check that $Np_N \to \infty$ together with $\limsup p_N < 1$ imply

$$\frac{\log N}{k_N p_N (1 - p_N)} \to 0,$$

as $N \to \infty$. This is the condition of Lemma 2.4 that guarantees that the binomial random variable X_N with parameters k_N and p_N is with probability larger than $1 - 4N^{-A}$ in between the bounds $k_N p_N \pm C_N$. Take node 1 of $G_p(N)$. The number of edges incident to node 1 is a Bernoulli variable X_1 with parameters $N - 1$ and p_N. We erase edges from node 1 until we reach the nearest integer of $(N-1)p_N - \sqrt{A(N-1)p_N(1-p_N)\log N}$. The edges that we erase are called ghost edges.

Now, take the smallest edge extending from node 1 and form the tree which consists of these two nodes. We now proceed with the induction step. Suppose that the uniform tree contains $n \ge 2$ nodes. In the original graph $G_p(N)$, each of these n nodes has a binomial-distributed number of edges to the $N - n$ remaining nodes.

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

The parameters of these (in total n) marginal distributions are $N-n$ and p_N. Assume that all these binomial random variables are in between $(N-n)p_N \pm \sqrt{A(N-n)p_N(1-p_N)\log N}$. Then, erase edges in graph $G_p(N)$ in a uniform way, until each of the n nodes has precisely

$$\left\lfloor (N-n)p_N - \sqrt{A(N-n)p_N(1-p_N)\log N} \right\rfloor \tag{14}$$

outgoing links. Draw the link to the node which carries the smallest exponential weight. Since this link is connected to any of the nodes of the cluster of size n with equal probability, it gives rise to a uniform tree of size $n+1$. This advances the induction. Furthermore, the above construction also produces a continuous-time Markov chain $X^-(t)$ with birth rate given by n times the quantity in (14). Here, $X^-(t)$ is the number of points in the cluster where the sum of the weights is less than or equal to t. We continue until this Markov chain is in the absorbing state, which is precisely when the cluster contains N^- points. To this Markov chain there is associated a uniform tree of size N^-. Hence, the random variable H_{N^-}, which is the number of hops in this uniform tree, has a generating function given by (9).

We now introduce three events that will be used to bound the probability $\mathbb{P}(H_N \neq H_{N^-})$. Define the event:

$$D_N = \{\text{node } N \text{ is reached when } X^-(t) = N - k_N\}.$$

Since the probability for any order of connections of the $N-1$ nodes other than the root 1 is equally likely, the probability that the node N has not been connected to the tree of $G_p(N)$ when this tree has size $N-k_N$ is $k_N/(N-1)$. Hence, we have

$$\mathbb{P}(D_N^c) = O(k_N/N). \tag{15}$$

Now, consider the tree of $G_p(N)$, when its size is equal to $N-k_N$. Let X_{ij}, $1 \le i \le N-k_N$, $j \le i$, be the number of outgoing links from node j when the cluster contains precisely $i \le N-k_N$ nodes (i.e., the number of links to the $N-i$ nodes not in the tree at that moment). Then, for every j, the marginal distribution of X_{ij} is binomial with parameters $N-i$ and p_N. Let

$$E_N = \bigcap_{i=1}^{N-k_N} \bigcap_{j \le i} \{X_{ij} \in I_{N,i}\}, \tag{16}$$

where

$$I_{N,i} = \left[(N-i)p_N - \sqrt{A(N-i)p_N(1-p_N)\log N}, (N-i)p_N + \sqrt{A(N-i)p_N(1-p_N)\log N}\right].$$

According to Lemma 2.4 and Boole's inequality,

$$\mathbb{P}(E_N^c) \le \sum_{i=1}^{N-k_N} 4i(N^{-A}) \le 2N^{2-A}. \tag{17}$$

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

Finally, we set

$$F_N = \{|H_N| \le B^2 \log N\},$$

so that by Lemma 2.3,

$$\mathbb{P}(F_N^c) \le 2N^{-\delta B}. \tag{18}$$

This estimate holds in the random graph $G_p(N)$. From (15), (17), and (18),

$$\begin{aligned}
&\mathbb{P}(H_N \neq H_N^-) \\
&\quad = \mathbb{P}(H_N \neq H_N^-, D_N \cap E_N \cap F_N) + \mathbb{P}(H_N \neq H_N^-, (D_N \cap E_N \cap F_N)^c) \\
&\quad \le \mathbb{P}(H_N \neq H_N^-, D_N \cap E_N \cap F_N) + \mathbb{P}(D_N^c) + \mathbb{P}(E_N^c) + \mathbb{P}(F_N^c) \\
&\quad \le (2B^2 \log N)\, \frac{\sqrt{Ak_N p_N \log N}}{k_N p_N} + O\left(\frac{k_N}{N}\right) + 2N^{2-A} + 2N^{-\delta B} \\
&\quad = O\left(\frac{k_N}{N}\right) + O\left(\frac{(\log N)^{3/2}}{\sqrt{k_N p_N}}\right),
\end{aligned} \tag{19}$$

where the second inequality follows from Boole's inequality, using that the shortest path has at most $B^2 \log N$ nodes, and from the probability that any given link in the shortest path in $G_N(p)$ is one of the edges that has been erased for H_N^- and is bounded by the number of edges that have been erased divided by the total number of edges extending from the node. This ratio is bounded above by $2\sqrt{Ak_N p_N \log N}/k_N p_N$, when all the binomial random variables are in between the bounds given in (16). The choice $k_N = O((N \log N)/(Np_N)^{1/3})$ follows from optimizing the right-hand side of (19) over k_N. ■

Proof of Theorem 2.1: The proof of (i) is immediate from the previous lemma. We only prove statement (ii); the proof of (iii) is similar. As earlier, $A_N = \{H_N = H_N^-\}$. Then,

$$\mathbb{E}(H_N) = \mathbb{E}(H_N 1_{A_N}) + \mathbb{E}(H_N 1_{A_N^c}) = \mathbb{E}(H_N^- 1_{A_N}) + \mathbb{E}(H_N 1_{A_N^c}).$$

We have that

$$\mathbb{E}(H_N^- 1_{A_N}) - \mathbb{E}(H_N^-) = \mathbb{E}(H_N^- 1_{A_N^c}) \to 0 \quad \text{and} \quad \mathbb{E}(H_N 1_{A_N^c}) \to 0. \tag{20}$$

Indeed, let $F = \{\max(H_N, H_N^-) \le \mathrm{B}^2 \log N\}$; then,

$$\mathbb{E}(H_N 1_{A_N^c}) \le \mathbb{E}(H_N 1_{F^c}) + \mathbb{E}(H_N 1_{A_N^c} 1_F) \le CN^{1-\delta B} + (B^2 \log N)\mathbb{P}(A_N^c)$$

and similarly for $\mathbb{E}(H_N^- 1_{A_N^c})$. From this, we see that it is necessary to have

$$\mathbb{P}(A_N^c) = o\left(\frac{1}{\log N}\right).$$

This can be obtained from Lemma 2.5 by taking $Np_N/(\log N)^6 \to \infty$, which is the condition in part (ii) of the theorem. Moreover, it is easy to check from the explicit

https://doi.org/10.1017/S026996480115206X Published online by Cambridge University Press

formula in (1) that the expectation of H_N^- is asymptotically equal to the right-hand side of (3) as long as $Np_N/\log N \to \infty$. ■

Acknowledgment

We thank Yuval Peres for pointing us to the connection to birth processes.

Notes

1. The actual values are kept confidential by the Internet operators.
2. In Cisco's OSPF implementation, it is suggested to use weights which are inverse proportional to the bandwidth of the link.

References

1. Abramowitz, M. & Stegun, I.A. (1968). *Handbook of mathematical functions.* New York: Dover.
2. Bollobas, B. (1985). *Random graphs.* New York: Academic Press.
3. Dembo, A. & Zeitouni, O. (1992). *Large deviations: techniques and applications.* London: Jones and Barlett.
4. Leadbetter, M.R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes.* New York: Springer-Verlag.
5. Smythe, R.T. & Mahmoud, H.M. (1995). A survey of recursive trees. *Theoretical Probability and Mathematical Statistics* 51: 1–27.
6. Vanhastel, S., Duysburg, B., & Demeester, P. (1999). Performance measurements on the current internet. In 7th IFIP ATM&IP Workshop.
7. Van Mieghem, P., Hooghiemstra, G., & van der Hofstad, R. (2000). A scaling law for the hopcount, Report No. 2000125, Delft University of Technology. (http://wwwtvs.et.tudelft.nl/people/piet/teleconference.html)
8. Watts, D.J. (1999). *Small worlds: The dynamics of networks between order and randomness.* Princeton, NJ: Princeton University Press.