First passage percolation on the random graph

Remco van der Hofstad Gerard Hooghiemstra Piet Van Mieghem

Abstract

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, we show that 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. We compute the distribution of the number hops of the shortest path between two arbitrary nodes.

The results can be generalized 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 $\frac{Np_N}{(\log N)^3} \to \infty$. The result can be used to model shortest paths in the Internet (cf. [7]).

Key words and phrases: first passage percolation, random graphs

AMS 1991 *Subject classifications: Primary 60K25 Secondary 60J25*

1 Introduction

Consider the complete graph K_N with N nodes and $N(N-1)/2$ (undirected) edges. To each edge (i,j) we attach an exponential random variable E_{ij} with mean equal to 1. We take these random variables mutually independent. For two arbitrary nodes, which we label 1 and N, we are interested in the (random) number of edges H_N of the shortest path that connects these two nodes. More precisely, if Q denotes a path connecting node 1 with node N and is given by $1 \to i_1 \to \ldots \to i_{h-1} \to N$, then we define the weight V_Q of that path by

$$V_Q = E_{1i_1} + E_{i_1i_2} + \ldots + E_{i_{h-1}N}.$$

The shortest path from 1 to N is the path that connects 1 and N and that has minimal weight. Let W_N be the weight of this shortest path, i.e., $W_N = \min_Q V_Q$, where we minimize over all possible paths Q from 1 to N. We are interested in the (asymptotic) distribution for large N of the number of edges (hops) H_N of this shortest path. We refer to H_N as the *hopcount*.

We start with the result. We prove that for the complete graph K_N, the probability generating function of the hopcount is equal to

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

where $\varphi_N(z)$ is the generating function

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

and where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, $x > -1$, denotes the gamma function. This identity implies that with γ being Euler's constant

$$E(H_N) \sim \log N + \gamma - 1, \tag{3}$$
$$\mathrm{Var}(H_N) \sim \log N + \gamma - \pi^2/6. \tag{4}$$

By Stirling's approximation we have

$$E(z^{H_N}) = \frac{N^{z-1}}{\Gamma(z+1)}(1 + O(N^{-1})). \tag{5}$$

This implies that

$$P(H_N = k) = \sum_{m=0}^{k} c_{m+1} \frac{\log^{k-m} N}{N(k-m)!}(1 + O(N^{-1})). \tag{6}$$

Here c_m are the Taylor coefficients of $1/\Gamma(z)$ (cf. [1], 6.1.34). The sequence $\{c_m\}$ tends to zero rapidly, so that the law of H_N is close to the law of the Poisson distribution with parameter $\log N$. Intuitively, 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 $\frac{L^k}{k!}$. Multiplying out, we find that $P(H_N = k, W_N \leq L) \approx \frac{(LN)^k}{Nk!}$. 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 $\frac{\log N}{N}$. Substitution of this value gives $P(H_N = k) \approx \frac{(\log N)^k}{Nk!}$, in accordance to (5).

Our result is surprisingly simple. Moreover it is quite robust in the sense that it is also valid 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. In fact, we can even take $p = p_N \to 0$ as long as $Np_N \to \infty$ at a certain rate. As before the edges are equipped with i.i.d. exponential weights with mean 1. In fact, note that the weight W_N now has to be of the order $\frac{\log N}{Np_N}$, 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, the minimum over Np exponential random variables has the *same* distribution as $1/p$ times the minimum over N exponential random variables. This explains that the value of p is only a scale factor. The limiting distribution of the hopcount remains unchanged. The insensitivity with respect to the value of p of the law of the hopcount can be understood by adapting the above heuristic to the case where $W_N \approx \frac{\log N}{Np_N}$ and 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 graph $G_p(N)$ with $p_N \to 0$ but $Np_N \to \infty$ can serve as a first order model for the Internet. In particular, the distribution of the hopcount in the Internet and the corresponding number of routers can be obtained from the present theory. Also dynamic routing effects such as the flooding time after a topology update in one router can be estimated. We refer to the paper [7] for further details.

2 The complete graph

For the complete graph K_N the proof of (1) - (4) is as follows. We consider a continuous time Markov chain $\{X(t)\}_{t\geq 0}$, 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 $\{X(t)\}_{t\geq 0}$ 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 there is a perfect probabilistic coupling between the state $X(t)$ and 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. More precisely, the two *processes* are identical in distribution. 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. This explains the rate $\lambda_n = n(N-n)$. When $X(t) = n$, the (not previously used) edges between the n nodes can be left out. These edges cannot be used by the shortest path otherwise they would have been chosen 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 randomly to one of the existing nodes in the associated tree. The hopcount is then proven to be equal to the *height* L_N of particle N in the recursive tree. It is well known (cf. [6]) that the height of an arbitrary point (including the root) has generating function (2). In our situation, N cannot be the root so that the result (1) for the complete graph follows.

To see that the tree described above is indeed uniform over all trees with N nodes, we argue as follows. Start with the root labeled 1 which corresponds to the starting value $X(0) = 1$ of the chain. If the root gives birth to a node we connect that new node to the root by a unit edge. We next repeat this construction inductively. Suppose that $X(t) = n$ and that the associated tree has n nodes. After an exponential time with rate $n(N-n)$ the Markov chain gives birth to a new node, which is born with equal probability out of any of the n nodes (parents). In the tree we connect the new node to one of the n existing nodes by a unit edge with equal probability. By the induction hypothesis that the tree of size n is uniform, it follows that the resulting tree of size $n+1$ is a uniform tree. Completing the induction shows that the final tree with N nodes is uniform.

Using the above description, we can also compute the generating function of the weight 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 $X_1 + \ldots + X_k$, where X_i is exponentially distributed with parameter $i(N-i)$, equals:

$$Ee^{t(X_1+\ldots+X_k)} = \prod_{i=1}^{k} \frac{i(N-i)}{i(N-i)-t}.$$

Hence

$$Ee^{tW_N} = \frac{1}{N-1}\sum_{k=1}^{N-1}\prod_{i=1}^{k} \frac{i(N-i)}{i(N-i)-t}. \tag{7}$$

3 The random graph and heuristics

In Section 4 we extend the results (3) and (4) to the class $G_p(N)$, where $p = p_N$ is chosen such that

$$\frac{Np_N}{(\log N)^3} \to \infty. \tag{8}$$

This is a technical condition. Form the famous connectivity theorem[1] of Erdös and Rényi and their observation that many important properties of graphs appear quite suddenly [2, Preface

[1] If $p_N = \frac{\log N + x + o(1)}{N}$, the $P\left(G_p(N) \text{ is connected}\right) \to e^{-e^{-x}}$ as $N \to \infty$ (see chapter VII and in particular pp. 150, Theorem 3 in [2]).

ix] at a threshold (which is here $p_N = \frac{\log N}{N}$), we expect that the results remain true as long as $\frac{Np_N}{\log N} \to \infty$ for $N \to \infty$. This means that the expected number of links per node (i.e. the degree) tends to infinity faster than $\log N$, which is the (asymptotic) condition to have a connected graph (such as the Internet).

For the random graph $G_p(N)$, each node has a random number of edges incident to this node. 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, 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

$$\lceil (N-n)p_N \pm \sqrt{A(N-n)p_N(1-p_N)\log N} \rceil, \tag{9}$$

which expression is defined as zero when (9) becomes negative and where A is a positive number to be determined later. To each of this constant number of outgoing links, there belongs a continuous time Markov chain $X^{\pm}(t)$, which is a pure birth process with state space $\{1, 2, \ldots, N^{\pm}\}$ where

$$N^{\pm} = \lceil N\left(1 \pm A(1-p_N)\log N/(Np_N)\right) \rceil, \tag{10}$$

and with birth rate

$$\lambda_n^{\pm} = \lceil (N-n)p_N \pm \sqrt{A(N-n)p_N(1-p_N)\log N} \rceil. \tag{11}$$

Observe that the size $N^{\pm}$ equals the smallest value of n for which $\lambda_n^{\pm} \leq 0$. We next show that with high probability the shortest path of the *uniform* trees belonging to the Markov chains $X^{\pm}(t)$ are the same. This immediately implies that the shortest path for the random graph $G_p(N)$ is also the same, and that (3) and (4) hold when $\log N^{\pm} = \log N + o(1)$, which implies that $\frac{Np_N}{\log N} \to \infty$. In fact, in the technical part of the proof, we need that $\frac{Np_N}{(\log N)^{\beta}} \to \infty$, where the value of β depends on whether we wish to couple the respective random variables, prove convergence of the mean or convergence of the variance.

We close this section with some comments on the choice of the exponential weights E_{ij}. It is well known that the minimum of n independent exponential random variables with mean 1 converges after norming (the minimum must be multiplied by n) to (again) the exponential distribution. The same is true for all distributions F with

$$F(x) \sim xL(x), \quad x \to 0, \tag{12}$$

with L a slowly varying function (cf. [5]). Therefore we expect (although we have no proof for the general case) that the result on the hopcount also holds when the exponential weights are replaced by i.i.d. weights with distribution function satisfying (12). This includes of course uniform weights. Below we give an heuristic argument to explain what we expect to happen if we replace the exponential or uniform weights by i.i.d. weights on the edges with distribution function

$$F_\alpha(x) = x^\alpha 1_{[0,1]}(x) + 1_{[1,\infty)}(x) \tag{13}$$

where $\alpha > 0$ is arbitrary. For this case we conjecture that the hopcount H_N satisfies:

$$E(H_N) \sim \log N/\alpha \tag{14}$$

$$\operatorname{Var}(H_N) \sim \log N/\alpha^2, \tag{15}$$

and that H_N is, asymptotically, normally distributed. Furthermore the asymptotic expected weight $L = EW_N$ of the shortest path will be of order

$$\frac{\log N}{((Np)^{1/\alpha}(\alpha\Gamma(\alpha))^{1/\alpha})}. \tag{16}$$

We start from

$$P_k(L) = P(H_N = k, W_N \leq L) \approx p^k N^{k-1} L^{\alpha k} \frac{(\alpha\Gamma(\alpha))^k}{\Gamma(\alpha k + 1)}, \tag{17}$$

which follows from the fact that the probability that the sum of k independent random variables with distribution F_α in (13) is less than L is given by $L^{\alpha k}\frac{(\alpha\Gamma(\alpha))^k}{\Gamma(\alpha k+1)}$. For typical values of L the norming equation $\sum_k P_k(L) = 1$ should be satisfied. Therefore

$$\begin{aligned} 1 &= \frac{1}{N}\sum_{k=1}^{N-1} \frac{(\alpha Np\,\Gamma(\alpha)L^\alpha)^k}{\Gamma(\alpha k + 1)} \\ &\approx \frac{1}{N}\int_0^\infty \frac{(\alpha Np\,\Gamma(\alpha)L^\alpha)^x}{\Gamma(\alpha x + 1)}\,dx = \frac{1}{\alpha N}\int_0^\infty \frac{(\alpha Np\,\Gamma(\alpha)L^\alpha)^{u/\alpha}}{\Gamma(u+1)}\,du \\ &\approx \frac{1}{\alpha N}\sum_{k=0}^{\infty} \frac{(\alpha Np\,\Gamma(\alpha)L^\alpha)^{k/\alpha}}{\Gamma(k+1)} = \frac{1}{\alpha N}\exp\{(\alpha Np\,\Gamma(\alpha))^{1/\alpha}L\}. \end{aligned}$$

We find

$$L \approx \frac{\log N}{(Np)^{1/\alpha}(\alpha\Gamma(\alpha))^{1/\alpha}},$$

in accordance with (16). Substitution of this result in $P_k(L)$ then yields that

$$P(H_N = k) \approx \frac{1}{\alpha N}\frac{(\log N)^{\alpha k}}{\Gamma(\alpha k + 1)}. \tag{18}$$

The first order approximation for the gamma function is $\Gamma(\alpha k) \sim (\alpha k)^{\alpha k}$ which suggests that $\alpha k \sim \log N$. To confirm this and to calculate the asymptotic variance of the hopcount, we substitute

$$k = \frac{1}{\alpha}\log N + v, \tag{19}$$

in (18). If we take the correct value of v the substitution of (19) should give the normal density with the correct variance. Using Stirling's formula:

$$\begin{aligned} P(H_N = k) &\approx \frac{1}{\alpha}\left(\frac{\log N}{\log N + \alpha v}\right)^{\log N + \alpha v} \frac{e^{\alpha v}}{\sqrt{2\pi(\log N + \alpha v)}} \\ &\approx \frac{e^{\alpha v}}{\alpha\sqrt{2\pi \log N}}\exp\left\{-\log\left(1 + \frac{\alpha v}{\log N}\right)\cdot(\log N + \alpha v)\right\}. \end{aligned}$$

Now use $\log(1+x) = x - x^2/2$, up to second order to obtain

$$P(H_N = k) \approx \frac{e^{-\alpha^2 v^2/(2\log N)}}{\alpha\sqrt{2\pi\log N}}.$$

This shows that H_N is roughly normal with mean $\log N/\alpha$ and variance $\log N/\alpha^2$, as conjectured in (14-15).

Our choice of $\alpha = 1$ and subsequently for exponential weights on the edges is motivated by two reasons:

(i) Recent Internet measurements indicate that for the empirical hopcount the ratio of the expectation and variance equals 1. Moreover statistical tests do support the hypothesis that the distribution (6) gives a good fit. (cf. [7] and papers cited there for further details). This indicates that we should take $\alpha = 1$, because it is the only choice of weight distribution of the form (12) satisfying the quoted ratio.

(ii) A special case for $\alpha = 1$ is the exponential distribution for which distribution the calculations are tractable through the use of continuous time Markov chains.

4 On the class $G_p(N)$

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 such that $\limsup_N p_N < 1$, so that the random graph is truly random. The main result is the following theorem:

Theorem 4.1 *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 (10), such that*

(i) If $Np_N/(\log N)^3 \to \infty$, then $P(H_N \neq H_N^-) = o(1)$.

(ii) If $Np_N/(\log N)^6 \to \infty$, then $E(H_N) = \log N + \gamma - 1 + o(1)$.

(iii) If $Np_N/(\log N)^9 \to \infty$, then $Var(H_N) = \log N + \gamma - \pi^2/6 + o(1)$.

The proof is divided in a number of different steps. We first sketch these steps and then formulate and prove them in a series of lemmas. Finally, we prove Theorem 4.1.

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 of extreme importance for the proof of our theorem, because it gives estimates how to compare the hopcount of the shortest path of $G_p(N)$ with the hopcount in a uniform tree associated with the Markov chain $\{X^-(t)\}$.

   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 4.2 we will show that $P(Np_N W_N > B \log N) \leq N^{-\delta B}$, for some $\delta > 0$. We prove this lemma with the help of Cramérs theorem (cf. [3] p. 26).

2. Using Lemma 4.2, we prove that the bound $H_N \leq B^2 \log N$ holds with overwhelming probability. This will be shown in Lemma 4.3.

3. For a binomial random variable X with parameters k_N and $p = p_N$ such that $\log N/(k_N p_N(1 - p_N)) \to 0$,

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

   This will be proven in Lemma 4.4.

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 $N^- = \lceil N\left(1 - A(1-p_N)\log N/(Np_N)\right)\rceil$. Let
$$A_N = \{H_N = H_N^-\}.$$
The main ingredient to the proof is that $P(A_N^c) \to 0$ at a certain rate that depends on how $Np_N \to \infty$. The random variable H_N^- has generating function
$$E(z^{H_N^-}) = \frac{N^-}{N^- - 1}\left(\varphi_{N^-}(z) - \frac{1}{N^-}\right), \tag{20}$$
where $\varphi_N(z)$ is the generating function in (2). Hence, the ratio of the generating functions $Ez^{H_N^-}$ and $\varphi_N(z)$ tends to one as long as $\frac{Np_N}{\log N} \to \infty$.

5. The asymptotic equalities of $P(H_N \neq H_N^-)$, EH_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 4.2 *There exists constants $\delta > 0$ and B such that for Np_N large,*
$$P(Np_NW_N > B\log N) \leq N^{-\delta B}. \tag{21}$$

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 which are 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 $\sqrt{N}$ 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 $\sqrt{Np_N}\cdot\sqrt{Np_N} = Np_N$ and hence the minimal weight of the connecting edges is of the same order $(Np_N)^{-1}$.

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,
$$E_{(1)}^i < E_{(2)}^i < \ldots$$
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 endpoints 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:
$$k = \left\lceil \frac{\log\sqrt{N}}{\log 2}\right\rceil, \tag{22}$$
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,
$$E_{(1)}^i = \min_j E_j^i \overset{d}{=} \frac{E_1}{X_i}, \qquad E_{(2)}^i \overset{d}{=} \frac{E_1}{X_i} + \frac{E_2}{X_i - 1},$$

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

$$E^i_{(1)} \leq \frac{2E_1}{Np_N}, \qquad E^i_{(2)} \leq \frac{2E_1 + 2E_2}{Np_N}, \tag{23}$$

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

From (23) and the fact that the minimal weight of the connecting edges can also be bounded by $\frac{2E_1+2E_2}{Np_N}$ 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

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

Now apply Cramér's theorem to S_{4k+1} with k given in (22). □

As a Corollary to Lemma 4.2 we have

Lemma 4.3 *There exists constants $\delta > 0$ and B such that for Np_N sufficiently large,*

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

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.

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

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

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

To see that the same bound also holds for R_N, use that

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

where we use (5) for N large enough. Pick $t = \log B$ to get

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

This bound is in fact sharper than the upper bound for $P(H_N > B^2 \log N)$.
□

Lemma 4.4 *For a binomial random variable X_N with parameters k_N and p_N satisfying $(\log N)/(k_N p_N(1 - p_N)) \to 0$, we have uniformly in k_N and p_N for large N,*

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

PROOF: Put for $A > 0$,

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

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

$$P(X_N > kp_N + C_N) \leq \inf_{t>0} P(e^{tX_N} > e^{tkp_N + C_N}) \leq \inf_{t>0} \left\{ e^{-t(k_N p_N + C_N)} \left(\phi(t)\right)^{k_N} \right\},$$

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:

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

From this we obtain

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

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

$$P(X_N > kp_N + C_N) \leq 2 \exp(-C_N^2/(\sigma_N^2 - C_N p_N)) \leq 2N^{-A}.$$

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

$$P(X_N < k_N p_N - C_N) = P(k_N - Y_N < k_N p_N - C_N) = P(Y_N > k_N(1 - p_N) + C_N).$$

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

Lemma 4.5 *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 (10), such that for $Np_N \to \infty$ and* $\limsup p_N < 1$,

$$P(H_N \neq H_N^-) = O\left(\frac{\log N}{[Np_N]^{\frac{1}{3}}}\right). \tag{24}$$

Moreover,

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

as long as $\frac{Np_N}{\log N} \to \infty$.

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

$$(\log N)/((k_N p_N(1 - p_N)) \to 0,$$

as $N \to \infty$. This is the condition of Lemma 4.4 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 $n \geq 2$ 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. 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

$$\lfloor (N-n)p_N - \sqrt{A(N-n)p_N(1-p_N)\log N} \rfloor \tag{25}$$

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 (25). 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 generating function given by (20).

We now introduce three events that will be used to bound the probability $P(H_N \neq H_N^-)$. 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

$$P(D_N^c) = O(k_N/N). \tag{26}$$

Now consider the tree of $G_p(N)$, when its size is equal to $N - k_N$. Let X_{ij}, $1 \leq i \leq N-k_N, j \leq i$ be the number of outgoing links from node j when the cluster contains precisely $i \leq N - k_N$ 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

$$E_N = \cap_{i=1}^{N-k_N} \cap_{j \leq i} \{X_{ij} \in I_{N,i}\}, \tag{27}$$

where

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

According to Lemma 4.4 and Boole's inequality,

$$P(E_N^c) \leq \sum_{i=1}^{N-k_N} 4i \cdot N^{-A} \leq 2N^{2-A}. \tag{28}$$

Finally set

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

then

$$P(F_N^c) \leq 2N^{-\delta B}, \tag{29}$$

because the number of edges in the shortest paths is at most $B^2 \log N$ with probability not exceeding $2N^{-\delta B}$, for some positive δ, according to Lemma 4.3. This estimate holds in the random graph $G_p(N)$. From (26), (28) and (29),

$$\begin{aligned} & P(H_N \neq H_N^-) \qquad (30) \\ & \quad = P\left(H_N \neq H_N^-, D_N \cap E_N \cap F_N\right) + P\left(H_N \neq H_N^-, (D_N \cap E_N \cap F_N)^c\right) \\ & \quad \leq P\left(H_N \neq H_N^-, D_N \cap E_N \cap F_N\right) + P(D_N^c) + P(E_N^c) + P(F_N^c) \\ & \quad \leq (2B^2 \log N) \frac{\sqrt{Ak_N p_N \log N}}{k_N p_N} + O\left(\frac{k_N}{N}\right) + 2N^{2-A} + 2N^{-\delta B} \\ & \quad = O\left(\frac{k_N}{N}\right) + O\left(\frac{(\log N)^{3/2}}{\sqrt{k_N p_N}}\right), \end{aligned}$$

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 have been erased for H_N^- 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 $2\frac{\sqrt{Ak_N p_N \log N}}{k_N p_N}$, when all the binomial random variables are in between the bounds given in (27). The choice $k_N = O((N \log N)/(Np_N)^{1/3})$ follows from optimizing the right hand side of (30) over k_N.

□

PROOF OF THEOREM 4.1: The proof of (i) is immediate from the previous lemma. We only prove statement (ii), the proof of (iii) being similar. As before $A_N = \{H_N = H_N^-\}$. Then

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

We have that

$$E(H_N^- 1_{A_N}) - E(H_N^-) = E(H_N^- 1_{A_N^c}) \to 0, \quad \text{and} \quad E(H_N 1_{A_N^c}) \to 0. \qquad (31)$$

Indeed, let $F = \{\max(H_N, H_N^-) \leq B^2 \log N\}$

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

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

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

This can be obtained from Lemma 4.5 by taking $\frac{Np_N}{(\log N)^6} \to \infty$ which is the condition in part (ii) of the theorem. Moreover, it is easy to check from the explicit formula in (1) that the expectation of H_N^- is asymptotically equal to the r.h.s. of (3) as long as $\frac{Np_N}{\log N} \to \infty$.

□

Acknowledgement: We thank Yuval Perez for pointing out the connection with birth processes.

References

[1] M. Abramowitz and I.A. Stegun *Handbook of Mathematical Functions*, Dover, 1968.

[2] B. Bollobas *Random Graphs*, Academic Press, 1985.

[3] A.Dembo and O.Zeitouni, *Large deviations Techniques and Applications*, Jones and Barlett Publishers, England, 1992.

[4] W. Feller *An Introduction to Probability Theory and Its Applications*, Volume II, 2nd edition, Wiley, New York, 1971.

[5] M.R. Leadbetter, G. Lindgren and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York, 1983.

[6] R.T. Smythe and H.M. Mahmoud *A survey of recursive trees*, Theor. Probability and Math. Statist. **51**,1-27, 1995.

[7] P. Van Mieghem, G. Hooghiemstra and R. van der Hofstad *A scaling law for the hopcount*, submitted to ACM Sigcom 2000, http://ssor.twi.tudelft.nl/ gerardh/

ITS
Department of Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft, the Netherlands
E-mail: R.W.vanderHofstad@its.tudelft.nl
G.Hooghiemstra@its.tudelft.nl

ITS
Department of Electrical Engineering
Delft University of Technology
Mekelweg 4
2628 CD Delft, the Netherlands
E-mail: p.vanmieghem@its.tudelft.nl