Combinatorics, Probability and Computing (1999) 8, 347–361. Printed in the United Kingdom

One, Two and Three Times log n/n for Paths in a Complete Graph with Random Weights

SVANTE JANSON

Department of Mathematics, Uppsala University, PO Box 480, S-751 06 Uppsala, Sweden
(e-mail: svante.janson@math.uu.se)

Received 12 December 1997; revised 19 October 1998

Consider the minimal weights of paths between two points in a complete graph K_n with random weights on the edges, the weights being, for instance, uniformly distributed. It is shown that, asymptotically, this is $\log n/n$ for two given points, that the maximum if one point is fixed and the other varies is $2\log n/n$, and that the maximum over all pairs of points is $3\log n/n$.

Some further related results are given as well, including results on asymptotic distributions and moments, and on the number of edges in the minimal weight paths.

1. Introduction

Let a random weight T_{ij} be assigned to every edge ij of the complete graph K_n. (Thus $T_{ji} = T_{ij}$. We do not define T_{ij} for $i = j$.) We assume that the $\binom{n}{2}$ weights T_{ij}, $1 \leqslant i < j \leqslant n$, are independent and identically distributed; moreover, we assume that they are non-negative and that their distribution function $\mathbb{P}(T_{ij} \leqslant t) = t + o(t)$ as $t \searrow 0$, the main examples being the uniform $U(0,1)$ and the exponential Exp(1) distributions.

For two vertices i and j, let X_{ij} be the minimal total weight of a path between i and j. Our main theorem is a set of three different asymptotic results for X_{ij} (log denotes the natural logarithm).

Theorem 1.1. *Under the assumptions above, as $n \to \infty$:*

(i) *for any fixed i and j,*

$$\frac{X_{ij}}{\log n/n} \xrightarrow{\text{p}} 1;$$

(ii) *for any fixed i,*

$$\frac{\max_{j\leqslant n} X_{ij}}{\log n/n} \xrightarrow{\text{p}} 2;$$

(iii)

$$\frac{\max_{i,j\leqslant n} X_{ij}}{\log n/n} \xrightarrow{\text{p}} 3.$$

Hence, with high probability, X_{ij} is about $\log n/n$ for any fixed (or random) pair of vertices, but there are pairs of vertices for which it is larger: up to $2\log n/n$ if i is fixed and up to $3\log n/n$ globally.

Similarly, defining $Y_i = \max_{j\leqslant n} X_{ij}$, we see from (ii) and (iii) that Y_i typically is about $2\log n/n$, but that it is larger for a few vertices with $\max_i Y_i$ being about $3\log n/n$. A companion result shows that, in contrast, Y_i is not significantly smaller than $2\log n/n$ for any vertex i.

Theorem 1.2. *As $n \to \infty$,*

$$\frac{\min_{i\leqslant n} \max_{j\leqslant n} X_{ij}}{\log n/n} \xrightarrow{\text{p}} 2.$$

In other words, interpreting the weights as distances, most pairs of vertices are at a distance of about $\log n/n$, the radius of the graph is about $2\log n/n$ and the diameter is about $3\log n/n$.

Remark 1. Theorem 1.1(i),(ii) may alternatively be stated in terms of first-passage percolation on the complete graph (the time to reach a given vertex is about $\log n/n$ and the time to reach all is $2\log n/n$).

For completeness and comparison, we also state the corresponding simple (and well-known) results for the *minimal* distance from a vertex. In this case there is less concentration and we obtain convergence (in distribution) to a nondegenerate random variable instead of to a constant.

Theorem 1.3. *Let $Z_i = \min_{j\neq i} X_{ij} = \min_{j\neq i} T_{ij}$. As $n \to \infty$:*

(i) *for any fixed i,*

$$nZ_i \xrightarrow{\text{d}} \mathrm{Exp}(1);$$

(ii)

$$n^2 \min_{i\leqslant n} Z_i = n^2 \min_{i,j\leqslant n} T_{ij} \xrightarrow{\text{d}} \mathrm{Exp}(2);$$

(iii)

$$\frac{\max_{i\leqslant n} Z_i}{\log n/n} \xrightarrow{\text{p}} 1. \qquad \square$$

The proofs of (i) and (ii) are simple exercises, while (iii) is, in disguise, the well-known threshold for existence of isolated vertices in a random graph [1, Exercise III.2]; consider the graph with edges $\{ij : T_{ij} < t\}$. We leave the details to the reader. (Note that if $T_{ij} \in \mathrm{Exp}(1)$, then $(n-1)Y_i \in \mathrm{Exp}(1)$ and $n(n-1)\min_i Y_i \in \mathrm{Exp}(2)$ exactly.)

Using Theorem 1.3(iii), we can give a simple informal explanation of the discrepancy between the three parts of Theorem 1.1 as follows, interpreting the weights as travel times. Most vertices are connected by efficient highways, which take you to almost any other vertex within about $\log n/n$ (but rarely much quicker). Some vertices, however, are remote villages (like Oberwolfach), from which it takes up to $\log n/n$ to get to any other vertex at all. Hence, starting at a typical vertex, most travel times are about $\log n/n$, but it takes an extra $\log n/n$ (just for the final step in the path) to reach a few remote vertices. Similarly, if we start at one of the very remote vertices, it takes about $\log n/n$ to get to any other vertex at all, $2\log n/n$ to get to most other vertices and $3\log n/n$ to get to the other very remote vertices.

Some further results on asymptotic distributions and moments are given in Section 3. The lengths of the minimum weight paths are studied in Section 4.

2. Proofs

We first observe that the distribution of T_{ij} does not affect the results, as long as it satisfies the condition above. This is seen by the following standard coupling argument, which we include for completeness.

Let $F^{-1}: [0,1) \to [0,\infty)$ be the inverse function of the distribution function $F(t) = \mathbb{P}(T_{ij} \leqslant t)$ of T_{ij}. If $U_{ij} \in U(0,1)$ are independent uniform random variables, then $F^{-1}(U_{ij})$ has the same distribution as T_{ij}, so we may without loss of generality assume that $T_{ij} = F^{-1}(U_{ij})$. By assumption, $F(t)/t \to 1$ as $t \searrow 0$, and thus also $F^{-1}(t)/t \to 1$. Let $\varepsilon > 0$. If $X_{ij} < 10\log n/n$, say, for some i and j, then $T_{kl} = F^{-1}(U_{kl}) < 10\log n/n$ for each edge kl in the minimum weight path from i to j, and thus, provided n is large enough, $1-\varepsilon < T_{kl}/U_{kl} < 1+\varepsilon$. Consequently, the sum of the U_{kl} for the same path is at most $(1-\varepsilon)^{-1}X_{ij}$, and thus, using X'_{ij} to denote the minimal path weight defined by $\{U_{ij}\}$, $X'_{ij} \leqslant (1-\varepsilon)^{-1}X_{ij}$. Conversely, by the same argument, if $X'_{ij} < 10\log n/n$ then $X_{ij} < (1+\varepsilon)X'_{ij}$. It follows that, if either $X_{ij} < 9\log n/n$ or $X'_{ij} < 9\log n/n$, and n is large enough, then both $X_{ij} < 10\log n/n$ and $X'_{ij} < 10\log n/n$ hold, and moreover $(1-\varepsilon)X'_{ij} < X_{ij} < (1+\varepsilon)X'_{ij}$. It now follows immediately that, if any part of Theorem 1.1 or 1.2 holds either for T_{ij} or for the uniform U_{ij}, then it holds for both. In particular, a proof of these results for any distribution with $F(t)/t \to 1$ as $t \searrow 0$ implies the same results for $U(0,1)$, and then for any other such distribution.

We may thus choose a convenient distribution of T_{ij}; we use the exponential distribution because of its excellent Markov properties. Hence, in the sequel we assume that $T_{ij} \in \mathrm{Exp}(1)$.

Proof of Theorem 1.1. For parts (i) and (ii), we may assume that $i = 1$. We adopt the first-passage percolation viewpoint (see Remark 1), so we regard vertex 1 as initially infected, and assume that the infection spreads along each edge with an Exp(1)-distributed waiting time. We first study when the other vertices get infected, considering them in order of infection and ignoring their labels.

Since there are $n-1$ neighbours of the initially infected vertex, the time V_1 until the second vertex is infected is exponentially distributed with expectation $1/(n-1)$. More

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

generally, when $k < n$ vertices have been infected, there are $k(n-k)$ edges connecting the infected and non-infected vertices, and thus the time V_k until the next vertex is infected is $\mathrm{Exp}\big(1/(k(n-k))\big)$; moreover, this time is independent of $V_1, \dots, V_{k-1}$. In other words, the time S_m until m vertices have become infected can be written

$$S_m = \sum_1^{m-1} V_k$$

where $V_1, \dots, V_{n-1}$ are independent with $V_k \in \mathrm{Exp}\big(1/(k(n-k))\big)$.

The times $\{S_m\}_{m=2}^n$ are just the minimal path weights $\{X_{1j}\}_{j=2}^n$, arranged in increasing order. In particular,

$$Y_1 = \max_{j \geqslant 2} X_{1j} = S_n = \sum_1^{n-1} V_k. \tag{2.1}$$

Hence

$$\begin{aligned} \mathbb{E} Y_1 &= \sum_1^{n-1} \mathbb{E} V_k = \sum_1^{n-1} \frac{1}{k(n-k)} = \frac{1}{n} \sum_1^{n-1} \left(\frac{1}{k} + \frac{1}{n-k} \right) = \frac{2}{n} \sum_1^{n-1} \frac{1}{k} \\ &= 2 \frac{\log n}{n} + O\left(\frac{1}{n} \right), \end{aligned} \tag{2.2}$$

and similarly

$$\begin{aligned} \operatorname{Var} Y_1 &= \sum_1^{n-1} \operatorname{Var} V_k = \sum_1^{n-1} \left(\frac{1}{k(n-k)} \right)^2 \leqslant 2 \sum_1^{n/2} \frac{1}{k^2(n-k)^2} \\ &\leqslant \frac{8}{n^2} \sum_1^{n/2} \frac{1}{k^2} = O(n^{-2}). \end{aligned} \tag{2.3}$$

Part (ii) now follows by Chebyshev's inequality.

For part (i), fix $j = 2$. Observe that, if N is the number of vertices infected before vertex 2, then

$$X_{12} = S_{N+1} = \sum_1^N V_k, \tag{2.4}$$

where, by symmetry, N is uniformly distributed over $1, \dots, n-1$ and independent of $V_1, \dots, V_{n-1}$. We rewrite this equation as $X_{12} = \sum_1^{n-1} \mathbf{1}[N \geqslant k] V_k$, using indicator functions to eliminate the random summation limit. Hence,

$$\begin{aligned} \mathbb{E} X_{12} &= \sum_1^{n-1} \mathbb{E}(\mathbf{1}[N \geqslant k] V_k) = \sum_1^{n-1} \mathbb{P}(N \geqslant k) \mathbb{E} V_k \\ &= \sum_1^{n-1} \frac{n-k}{n-1} \frac{1}{k(n-k)} = \sum_1^{n-1} \frac{1}{k(n-1)} \\ &= \frac{\log n}{n} + O\left(\frac{1}{n} \right). \end{aligned} \tag{2.5}$$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

In order to estimate the variance, we further rewrite the sum as

$$\begin{aligned} X_{12} &= \sum_1^N (V_k - \mathbb{E}V_k) + \sum_1^N \frac{1}{n}\left(\frac{1}{k} + \frac{1}{n-k}\right) \\ &= \sum_1^N (V_k - \mathbb{E}V_k) + \frac{1}{n}\big(\log N + \log n - \log(n-N)\big) + O\left(\frac{1}{n}\right). \end{aligned} \tag{2.6}$$

We consider the three terms on the right-hand side separately. Since $N, V_1, \ldots, V_{n-1}$ are independent,

$$\begin{aligned} \operatorname{Var}\left(\sum_1^N (V_k - \mathbb{E}V_k)\right) &= \mathbb{E}\left(\sum_1^N (V_k - \mathbb{E}V_k)\right)^2 = \mathbb{E}\left(\sum_1^N \operatorname{Var} V_k\right) \\ &\leqslant \sum_1^{n-1} \operatorname{Var} V_k = \sum_1^{n-1} \frac{1}{k^2(n-k)^2} \\ &\leqslant \sum_1^{n/2} \frac{4}{k^2 n^2} + \sum_{n/2}^{n-1} \frac{4}{n^2(n-k)^2} = O\left(\frac{1}{n^2}\right). \end{aligned}$$

For the second term, we observe that

$$\mathbb{E}\big(\log N - \log(n-1)\big)^2 = \mathbb{E}\left(\log \frac{N}{n-1}\right)^2 \to \int_0^1 (\log x)^2\, dx < \infty$$

as $n \to \infty$. Hence $\operatorname{Var}(\log N) = \operatorname{Var}\big(\log(n-N)\big) = O(1)$, and it follows that the variance of the second term in (2.6) is also $O(n^{-2})$. The same is trivially true for the third term.

Consequently, $\operatorname{Var} X_{12} = O(n^{-2})$, which together with (2.5) yields part (i).

The proof of (iii) is divided into two parts, considering upper and lower bounds separately. First, by (2.1), for $-\infty \leqslant t < 1 - 1/n$,

$$\mathbb{E}e^{tnY_1} = \prod_1^{n-1} \mathbb{E}e^{ntV_k} = \prod_1^{n-1}\left(1 - \frac{nt}{k(n-k)}\right)^{-1}. \tag{2.7}$$

Hence, for every $a > 0$, choosing $t = 1 - 1/\log n$ ($n \geqslant 3$),

$$\begin{aligned} &\mathbb{P}(Y_1 > a \log n/n) \leqslant \mathbb{E}e^{tnY_1 - ta\log n} = e^{-ta\log n} \prod_1^{n-1}\left(1 - \frac{nt}{k(n-k)}\right)^{-1} \\ &= \left(1 - \frac{nt}{n-1}\right)^{-2} \exp\left(-ta\log n + \sum_2^{n-2} -\log\left(1 - \frac{nt}{k(n-k)}\right)\right) \\ &\leqslant \left(1 - \frac{nt}{n-1}\right)^{-2} \exp\left(-ta\log n + \sum_2^{n-2}\left(\frac{nt}{k(n-k)} + \left(\frac{nt}{k(n-k)}\right)^2\right)\right) \\ &= \big(1 - t + O(n^{-1})\big)^{-2} \exp\big(-ta\log n + 2t\log n + O(1)\big) = O\big(n^{2-a}\log^2 n\big). \end{aligned} \tag{2.8}$$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

This evidently implies

$$\mathbb{P}\big(\max_i Y_i > a\log n/n\big) \leqslant n\mathbb{P}\big(Y_1 > a\log n/n\big) = O\big(n^{3-a}\log^2 n\big),$$

which tends to 0 as $n \to \infty$ for every fixed $a > 3$.

For the lower bound, let $\varepsilon > 0$ be small. Partition the vertex set $\{1,\dots,n\}$ of K_n into the sets $A = \{1,\dots,n_A\}$ and $B = \{n_A+1,\dots,n\}$, where $n_A = \lceil n^{1-\varepsilon}\rceil$. Let $n_B = |B| = n - n_A$.

For $i \in A$, let $U_i = \min_{j\in B} T_{ij}$. Then the random variables U_i, $i \in A$, are independent with $U_i \in \mathrm{Exp}(1/n_B)$. In particular,

$$\begin{aligned}\mathbb{P}\big(U_i > (1-2\varepsilon)\log n/n\big) &= \exp\left(-(1-2\varepsilon)\frac{n_B}{n}\log n\right)\\ &\geqslant \exp\big(-(1-2\varepsilon)\log n\big) = n^{2\varepsilon-1}\end{aligned}$$

and thus

$$\mathbb{P}\big(U_i \leqslant (1-2\varepsilon)\log n/n \text{ for every } i \in A\big) \leqslant \big(1-n^{2\varepsilon-1}\big)^{n^{1-\varepsilon}} < e^{-n^{\varepsilon}}. \tag{2.9}$$

For $k \in A$, let $\mathscr{E}_k$ be the event that $U_k > (1-2\varepsilon)\log n/n$ but $U_i \leqslant (1-2\varepsilon)\log n/n$ for $i \leqslant k$. Then the events $\mathscr{E}_k$ are disjoint and, by (2.9),

$$\sum_{k\in A}\mathbb{P}(\mathscr{E}_k) = \mathbb{P}\left(\bigcup_{k\in A}\mathscr{E}_k\right) > 1 - e^{-n^{\varepsilon}}. \tag{2.10}$$

The idea of the proof is to show that, conditioned on $\mathscr{E}_k$, Y_k is with high probability close to $3\log n/n$; in fact, as is shown in detail below, conditioning on $U_k > (1-2\varepsilon)\log n/n$ typically increases Y_k (which is usually about $2\log n/n$) by $(1-2\varepsilon)\log n/n$, while conditioning on $U_i \leqslant (1-2\varepsilon)\log n/n$ for $i \leqslant k$ hardly affects the result.

We will use the following lemma.

Lemma 2.1. *Suppose that μ, $b > 0$ and $X \in \mathrm{Exp}(\mu)$, and define*

$$f(x) = -\mu\log\big(e^{-b/\mu} + (1-e^{b/\mu})e^{-x/\mu}\big).$$

(i) *The distribution of $f(X)$ equals the conditional distribution of X given $X \leqslant b$.*

(ii) *If, further, $0 \leqslant \alpha < 1$ and $b/\mu \geqslant \alpha(1-\log\alpha)/(1-\alpha)$, then $f(x) \geqslant \alpha x$ when $0 \leqslant x \leqslant \alpha^{-1}b - \mu$. Consequently,*

$$\mathbb{P}(f(X) < \alpha X) \leqslant \mathbb{P}(X > \alpha^{-1}b - \mu) = e^{1-\alpha^{-1}b/\mu}.$$

Proof. We may for simplicity, by homogeneity, assume that $\mu = 1$. Then e^{-X} is uniformly distributed on $[0,1]$, and thus for $0 \leqslant t \leqslant b$,

$$\begin{aligned}\mathbb{P}\big(f(X) \leqslant t\big) &= \mathbb{P}\big(e^{-b} + (1-e^{-b})e^{-X} \geqslant e^{-t}\big) = \mathbb{P}\left(e^{-X} \geqslant \frac{e^{-t}-e^{-b}}{1-e^{-b}}\right)\\ &= \frac{1-e^{-t}}{1-e^{-b}} = \mathbb{P}(X \leqslant t \mid X \leqslant b),\end{aligned}$$

which proves part (i).

For part (ii) we observe that (when $\mu = 1$) $f(x) \geqslant \alpha x$ if and only if

$$e^{-b} + (1-e^{-b})e^{-x} \leqslant e^{-\alpha x}. \tag{2.11}$$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

Letting $y = e^{-x}$, the left-hand side of (2.11) is a linear function of y, while the right-hand side y^α is concave; hence, in order to verify (2.11) for the interval $0 \leqslant x \leqslant \alpha^{-1}b - 1$, it suffices to verify it for the endpoints.

For $x = 0$, (2.11) is a trivial identity, while for $x = \alpha^{-1}b - 1$ it is

$$e^{-b} + (1 - e^{-b})e^{-\alpha^{-1}b+1} \leqslant e^{-b+\alpha}. \tag{2.12}$$

Now, by assumption, $\alpha^{-1}b = b + b(1-\alpha)\alpha^{-1} \geqslant b + 1 - \log\alpha$, and thus

$$e^{-b} + e^{-\alpha^{-1}b+1} \leqslant e^{-b} + e^{-b+\log\alpha} = (1+\alpha)e^{-b} \leqslant e^{\alpha}e^{-b};$$

this implies (2.12), which completes the proof of the lemma. □

Continuing with the proof of Theorem 1.1(iii), let $k \in A$ be fixed, let f be as in Lemma 2.1 with $\mu = 1/n_B$ and $b = (1-2\varepsilon)\log n/n$, and define

$$U_i' = \begin{cases} f(U_i), & i < k, \\ U_i + b, & i = k, \\ U_i, & i > k. \end{cases}$$

Then, by Lemma 2.1(i) for $i < k$ and the standard lack-of-memory property of exponential distributions for $i = k$, the distribution of U_i' equals the conditional distribution of U_i given $\mathscr{E}_k$ for every $i \in A$; moreover, by our independence assumptions, this extends to the joint distribution. Furthermore, by the same lack-of-memory property, the family of random variables $\{T_{ij} - U_i\}_{j\in B}$ is independent of U_i, for each $i \in A$ separately and thus for all $i \in A$ jointly too; hence the joint distribution of $\{T_{ij} - U_i\}_{i\in A, j\in B}$ is not affected by conditioning on $\mathscr{E}_k$. It follows that if we define T_{ij}' for $1 \leqslant i < j \leqslant n$ by

$$T_{ij}' = \begin{cases} T_{ij} - U_i + U_i', & i \in A \text{ and } j \in B, \\ T_{ij}, & \text{otherwise}, \end{cases} \tag{2.13}$$

and let $T_{ji}' = T_{ij}'$ for $j > i$, then the family $\{T_{ij}'\}$ has the same distribution as the conditional distribution of $\{T_{ij}\}$ given $\mathscr{E}_k$. Note in particular that $T_{kj}' = T_{kj} + b$ when $j \in B$.

Suppose that $\{T_{ij}\}$ are such that

$$U_i' \geqslant (1-2\varepsilon)U_i, \quad \text{for every } i \in A, \tag{2.14}$$

$$T_{ik} \geqslant 3\frac{\log n}{n}, \quad \text{for every } i \in A, \tag{2.15}$$

and

$$Y_k \geqslant (2-\varepsilon)\frac{\log n}{n}. \tag{2.16}$$

We observe first that, by (2.13) and (2.14), then

$$T_{ij}' \geqslant (1-2\varepsilon)T_{ij}, \quad \text{for every } i \text{ and } j \neq i. \tag{2.17}$$

Now consider the minimal path weights X_{ij}' defined by the edge weights T_{ij}' and the corresponding $Y_i' = \max_j X_{ij}'$. By (2.16), there exists a vertex l such that every path $i_0 = k, i_1, \ldots, i_m = l$ from k to l has weight $W = \sum_1^m T_{i_{s-1}i_s} \geqslant (2-\varepsilon)\log n/n$. Consider such a path and the corresponding weight $W' = \sum_1^m T_{i_{s-1}i_s}'$. Either $i_1 \in A$, and then, by

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

(2.13) and (2.15), $W' \geqslant T'_{ki_1} = T_{ki_1} \geqslant 3\log n/n$, or $i_1 \in B$, and then $T'_{ki_1} = T_{ki_1} + b$, which together with (2.17) yields

$$W' \geqslant b + (1-2\varepsilon)W \geqslant (1-2\varepsilon)\frac{\log n}{n} + (1-2\varepsilon)(2-\varepsilon)\frac{\log n}{n} \geqslant (3-7\varepsilon)\frac{\log n}{n}.$$

Hence $W' \geqslant (3-7\varepsilon)\log n/n$ for every path from k to l, and thus $X'_{kl} \geqslant (3-7\varepsilon)\log n/n$ and finally $Y'_k \geqslant X'_{kl} \geqslant (3-7\varepsilon)\log n/n$.

We have shown that, if (2.14)–(2.16) hold, then $Y'_k \geqslant (3-7\varepsilon)\log n/n$. Consequently,

$$\begin{aligned}\mathbb{P}\big(Y_k \geqslant (3-7\varepsilon)\log n/n \mid \mathscr{E}_k\big) &= \mathbb{P}\big(Y'_k \geqslant (3-7\varepsilon)\log n/n\big)\\ &\geqslant \mathbb{P}((2.14)\text{–}(2.16)\text{ hold}).\end{aligned}$$

Let q denote the probability that (2.14)–(2.16) hold. We have so far kept k fixed, but q is independent of k, and summing over k we obtain

$$\begin{aligned}\mathbb{P}\big(\max_i Y_i \geqslant (3-7\varepsilon)\log n/n\big) &\geqslant \sum_{k\in A}\mathbb{P}\big(Y_k \geqslant (3-7\varepsilon)\log n/n \mid \mathscr{E}_k\big)\mathbb{P}(\mathscr{E}_k)\\ &\geqslant q\sum_{k\in A}\mathbb{P}(\mathscr{E}_k). \qquad (2.18)\end{aligned}$$

Now, by Lemma 2.1(ii) with $\alpha = 1-2\varepsilon$, if n is large enough,

$$\begin{aligned}\mathbb{P}\big((2.14)\text{ fails}\big) &\leqslant \sum_{i\in A}\mathbb{P}\big(U'_i < (1-2\varepsilon)U_i\big) \leqslant n_A e^{1-n_B\log n/n}\\ &= O\big(n^{1-\varepsilon}n^{-1}\big) = o(1).\end{aligned}$$

Similarly,

$$\mathbb{P}\big((2.15)\text{ fails}\big) \leqslant \sum_{i\in A}\mathbb{P}\left(T_{ik} < 3\frac{\log n}{n}\right) \leqslant 3n_A\frac{\log n}{n} = o(1),$$

while $\mathbb{P}\big((2.16)\text{ fails}\big) = o(1)$ by the already proven part (ii) of the theorem.

Consequently, $q = 1 - o(1)$, which by (2.18) and (2.10) yields $\mathbb{P}\big(\max_i Y_i \geqslant (3-7\varepsilon)\log n/n\big) \to 1$ as $n\to\infty$. This completes the proof of (iii). □

Proof of Theorem 1.2. We use (2.7), replacing t by $-t$, and, for every a and $t>0$, obtain

$$\begin{aligned}\mathbb{P}(Y_1 < a\log n/n) &\leqslant \mathbb{E}e^{ta\log n - tnY_1} \leqslant e^{ta\log n}\prod_1^{n-1}\left(1+\frac{nt}{k(n-k)}\right)^{-1}\\ &= \exp\left(ta\log n + \sum_1^{n-1} -\log\left(1+\frac{nt}{k(n-k)}\right)\right)\\ &\leqslant \exp\left(ta\log n + \sum_1^{n-1}\left(-\frac{nt}{k(n-k)} + \frac{1}{2}\left(\frac{nt}{k(n-k)}\right)^2\right)\right)\\ &= \exp\Big(at\log n - 2t\log n + O(t) + O(t^2)\Big).\end{aligned}$$

If $0<a<2$, we thus obtain for any constant t

$$\mathbb{P}(\min_i Y_i < a\log n/n) \leqslant n\mathbb{P}(Y_1 < a\log n/n) = O\big(n^{1+(a-2)t}\big),$$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

which is $o(1)$ provided $t > 1/(2-a)$. On the other hand, Theorem 1.1(ii) implies

$$\mathbb{P}\big(\min_i Y_i > (2+\varepsilon)\log n/n\big) \leqslant \mathbb{P}\big(Y_1 > (2+\varepsilon)\log n/n\big) \to 0$$

for every $\varepsilon > 0$, and the proof is complete. □

3. Asymptotic distributions and moments

The method above also yields the asymptotic distributions of X_{ij} and Y_i: these are not normal. More precisely, we have the following result. (We have to impose a slightly stronger condition on the distribution of the T_{ij}; the condition is satisfied for the exponential and uniform distributions.)

Theorem 3.1. *Suppose that the distribution function* $\mathbb{P}(T_{ij} \leqslant t) = t + o(t/|\log t|)$ *as* $t \searrow 0$. *Then, as* $n \to \infty$,

$$nX_{ij} - \log n - \gamma \xrightarrow{\mathrm{d}} \sum_1^\infty \tfrac{1}{k}(\xi_k - 1) + \zeta \tag{3.1}$$

and

$$nY_i - 2\log n - 2\gamma \xrightarrow{\mathrm{d}} \sum_1^\infty \tfrac{1}{k}(\xi_k - 1) + \sum_1^\infty \tfrac{1}{k}(\xi'_k - 1), \tag{3.2}$$

where γ *is Euler's constant, and the random variables* ξ_k, ξ'_k, $k \geqslant 1$, *and* ζ *are independent with* $\xi_k, \xi'_k \in \mathrm{Exp}(1)$ *while* ζ *has the logistic distribution* $\mathbb{P}(\zeta \leqslant x) = e^x/(1+e^x)$.

Proof. By a slight modification of the coupling argument in the proof of Theorem 1.1, it suffices to consider the case $T_{ij} \in \mathrm{Exp}(1)$; we omit the details.

We write $A_n \approx B_n$ to mean that $\mathbb{E}(A_n - B_n)^2 = o(1)$ as $n \to \infty$. In the exponential case, (2.4) and (2.1) imply that

$$\begin{aligned}
nX_{12} &\overset{\mathrm{d}}{=} \sum_1^N \frac{n}{k(n-k)}\xi_k = \sum_1^N \frac{n}{k(n-k)}(\xi_k - 1) + \sum_1^N \left(\frac{1}{k} + \frac{1}{n-k}\right) \\
&\approx \sum_1^N \frac{1}{k}(\xi_k - 1) + \log N + \gamma + \log n - \log(n-N) \\
&\approx \sum_1^\infty \frac{1}{k}(\xi_k - 1) + \log \frac{N/n}{1 - N/n} + \log n + \gamma,
\end{aligned}$$

and

$$\begin{aligned}
nY_1 &\overset{\mathrm{d}}{=} \sum_1^{n-1} \frac{n}{k(n-k)}\xi_k = \sum_1^{n-1} \frac{n}{k(n-k)}(\xi_k - 1) + 2\sum_1^{n-1} \frac{1}{k} \\
&\approx \sum_1^{\lfloor n/2 \rfloor} \frac{1}{k}(\xi_k - 1) + \sum_{\lfloor n/2 \rfloor + 1}^{n-1} \frac{1}{n-k}(\xi_k - 1) + 2\log n + 2\gamma
\end{aligned}$$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

$$\stackrel{\mathrm{d}}{=} \sum_{1}^{\lfloor n/2 \rfloor} \frac{1}{k}(\xi_k - 1) + \sum_{1}^{\lceil n/2 \rceil - 1} \frac{1}{k}(\xi_k' - 1) + 2\log n + 2\gamma.$$

The result follows, since $N/n \xrightarrow{\mathrm{d}} \eta$, where $\eta \in U(0,1)$, and $\zeta = \log\bigl(\eta/(1-\eta)\bigr)$ has the logistic distribution. $\square$

Since the moment generating function $\mathbb{E}e^{t\xi_k}$ of ξ_k equals $(1-t)^{-1}$, $\mathrm{Re}\,t < 1$, it follows that the moment generating function of $\sum \frac{1}{k}(\xi_k - 1)$ is

$$\begin{aligned}
\prod_{k=1}^{\infty}(1-t/k)^{-1}e^{-t/k} &= \lim_{n\to\infty}\left(\prod_{k=1}^{n}\frac{k}{k-t}\right)e^{-t\sum_{l=1}^{n}\frac{1}{l}} \\
&= \lim_{n\to\infty}\frac{\Gamma(n+1)\Gamma(1-t)}{\Gamma(n+1-t)}e^{-t\log n - t\gamma + o(1)} \\
&= \Gamma(1-t)e^{-t\gamma}, \qquad \mathrm{Re}\,t < 1;
\end{aligned}$$

hence the moment generating function of $W = \sum \frac{1}{k}(\xi_k - 1) + \gamma$ equals $\Gamma(1-t)$, $\mathrm{Re}\,t < 1$. Now, if $T \in \mathrm{Exp}(1)$, then $-\log T$ has the moment generating function $\mathbb{E}e^{-t\log T} = \mathbb{E}T^{-t} = \int_0^\infty x^{-t}e^{-x}\,dx = \Gamma(1-t)$ too. Thus $W \stackrel{\mathrm{d}}{=} -\log T$. (Recall that the restriction of the moment generating function to the imaginary axis yields the characteristic function, which determines the distribution.) Hence,

$$\mathbb{P}(W \leqslant x) = \mathbb{P}(\log T \geqslant -x) = \mathbb{P}(T \geqslant e^{-x}) = e^{-e^{-x}}, \qquad -\infty < x < \infty, \tag{3.3}$$

which is one of the standard extreme value distributions [2].

Consequently, the right-hand side of (3.2) can be written $W + W' - 2\gamma$, where W and W' are independent random variables with the distribution (3.3).

Moreover, the logistic distribution has the moment generating function, for $|\mathrm{Re}\,t| < 1$, with $\eta \in U(0,1)$ as above,

$$\mathbb{E}e^{t\log(\eta/(1-\eta))} = \int_0^1 x^t(1-x)^{-t}\,dx = B(1+t, 1-t) = \Gamma(1+t)\Gamma(1-t),$$

which equals the moment generating function of the symmetrization $W - W'$. Thus $\zeta \stackrel{\mathrm{d}}{=} W - W'$.

We can now restate Theorem 3.1 as follows.

Theorem 3.2. *Suppose that the distribution function $\mathbb{P}(T_{ij} \leqslant t) = t + o(t/|\log t|)$ as $t \searrow 0$. Then, as $n \to \infty$,*

$$nX_{ij} - \log n \xrightarrow{\mathrm{d}} W_1 + W_2 - W_3 \tag{3.4}$$

and

$$nY_i - 2\log n \xrightarrow{\mathrm{d}} W_1 + W_2, \tag{3.5}$$

where W_1, W_2, W_3 are independent random variables with the same extreme value distribution $\mathbb{P}(W_i \leqslant x) = e^{-e^{-x}}$. $\square$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

The variables on the right-hand sides of (3.4) and (3.5) have the moment generating functions $\Gamma(1-t)^2\Gamma(1+t)$, $|\mathrm{Re}t| < 1$, and $\Gamma(1-t)^2$, $\mathrm{Re}t < 1$, respectively, and thus the characteristic functions $\Gamma(1-it)^2\Gamma(1+it)$ and $\Gamma(1-it)^2$. The limit $W_1 + W_2$ in (3.5) has a density function that can be expressed using modified Bessel functions as $2e^{-x}K_0(2e^{-x/2})$ (*cf.*, for instance, [3, (5.10.23)]). We do not know any simple expression for the density function of $W_1 + W_2 - W_3$.

Using the fact that the variance of the logistic distribution is $\pi^2/3$ (which follows from its moment generating function $\Gamma(1+t)\Gamma(1-t) = \pi t/\sin \pi t$, $|\mathrm{Re}t| < 1$, or from the representation $W - W'$ above), it is easily seen that the limiting variables in (3.1) and (3.2) have expectations 0 and variances $\sum_1^\infty k^{-2} + \pi^2/3 = \pi^2/2$ and $2\sum_1^\infty k^{-2} = \pi^2/3$, respectively. Since all approximations and limits in the proof hold in L^2 sense, we obtain that these are the limits of the expectations and variances of the left-hand sides in (3.1) and (3.2) too, provided $T_{ij} \in \mathrm{Exp}(1)$. This carries over to other distributions as well, in particular to the uniform distribution; we have the following theorem.

Theorem 3.3. *Suppose that the distribution function* $\mathbb{P}(T_{ij} \leqslant t) = t + o(t/|\log t|)$ *as* $t \searrow 0$, *and that* $\mathbb{E}T_{ij}^p < \infty$ *for some* $p > 0$. *Then all moments converge in* (3.1), (3.2), (3.4) *and* (3.5)*; in particular,*

$$\begin{aligned}
\mathbb{E}X_{ij} &= \frac{\log n}{n} + \frac{\gamma}{n} + o\left(\frac{1}{n}\right), \\
\mathbb{E}Y_i &= 2\frac{\log n}{n} + \frac{2\gamma}{n} + o\left(\frac{1}{n}\right), \\
\mathrm{Var}X_{ij} &\sim \frac{\pi^2}{2n^2}, \\
\mathrm{Var}Y_i &\sim \frac{\pi^2}{3n^2}.
\end{aligned}$$

Proof. It suffices to prove that $\mathbb{E}(nX_{ij} - \log n)^m = O(1)$ and $\mathbb{E}(nY_i - 2\log n)^m = O(1)$ for every even integer m and n large enough, since this implies convergence of all moments of order $< m$ by a standard result on uniform integrability.

When T_{ij} is exponentially distributed, this can be done as for the case $m = 2$ in the proof of Theorem 1.1; we omit the details.

In general, we let a and b be two constants, to be chosen later, and split the expectation into three parts. (We treat only X_{ij}; the same argument applies to Y_i.)

First, $\mathbb{E}\big((nX_{ij} - \log n)^m \mathbf{1}[X_{ij} \leqslant a\log n/n]\big) = O(1)$ by comparison with the exponential case, using the coupling argument as in earlier proofs.

Secondly, by (2.8), $\mathbb{E}\big((nX_{ij})^m \mathbf{1}[a\log n/n < X_{ij} \leqslant b]\big) \leqslant b^m n^m \mathbb{P}(X_{ij} > a\log n/n) = O(n^{m+2-a}\log^2 n)$; choosing $a = m+3$ this becomes bounded.

Finally, considering only the $n-2$ paths of length 2 between i and j, we see that

$$\begin{aligned}
\mathbb{P}(X_{ij} > x) &\leqslant \mathbb{P}(T_{ik} > x/2 \text{ or } T_{jk} > x/2 \text{ for every } k \neq i, j) \\
&\leqslant \big(2\mathbb{P}(T_{ij} > x/2)\big)^{n-2}.
\end{aligned}$$

Now, if $\mathbb{E}T_{ij}^p < \infty$, then $x^p \mathbb{P}(T_{ij} > x) \to 0$ as $x \to \infty$; it follows that if b is large enough, then $2\mathbb{P}(T_{ij} > x/2) < x^{-p}$ when $x \geqslant b$, and thus

$$\mathbb{P}(X_{ij} > x) \leqslant x^{-(n-2)p}, \qquad x \geqslant b.$$

Consequently,

$$\begin{aligned} \mathbb{E}\big((nX_{ij})^m \mathbf{1}[X_{ij} > b]\big) &= n^m b^m \mathbb{P}(X_{ij} > m) + n^m \int_b^\infty m x^{m-1} \mathbb{P}(X_{ij} > x)\,dx \\ &= O(n^m b^{-np}) = O(1), \end{aligned}$$

provided $n > 2 + m/p$.

Combining these estimates we find $\mathbb{E}(nX_{ij} - \log n)^m = O(1)$ as required. □

Remark 2. The asymptotic variances can also be obtained by refining the estimates used in the proof of Theorem 1.1.

Remark 3. The condition that $\mathbb{E}T_{ij}^p < \infty$ for some $p > 0$ is necessary too; if it fails then X_{ij} has no finite moment for any n. In fact, suppose that, for instance, $\mathbb{E}X_{ij} < \infty$ for some n; then $\mathbb{P}(X_{ij} > t) < 1/t$ for large t. Since $\mathbb{P}(X_{ij} > t) \geqslant \mathbb{P}(T_{ik} > t \text{ for every } k \neq i) = \mathbb{P}(T_{ij} > t)^{n-1}$, this yields $\mathbb{P}(T_{ij} > t) < t^{-1/(n-1)}$ (t large), and thus for example $\mathbb{E}T_{ij}^{1/n} < \infty$.

We do not know any similar results for $\max_{i,j} X_{ij}$.

Problem 1. *What is the asymptotic distribution of* $\max_{i,j} X_{ij}$ *(presuming that some exists)?*

Problem 2. *What is the order of* $\operatorname{Var}(\max_{i,j} X_{ij})$*? Is it* $\sim c/n^2$*? If so, what is the constant* c*?*

4. Lengths of minimal paths

We have so far studied the weights of the minimal paths, but one might also ask how long they are, disregarding their weights, that is, how many edges they contain. Let L_{ij} be the length of the path between i and j that has minimal weight.

For the case of exponentially distributed T_{ij}, these lengths can be studied by observing that the proof of Theorem 1.1 shows that the collection of minimal weight paths from a given vertex, 1 say, form a tree (rooted at 1) which can be constructed as follows. Begin with a single root and add $n-1$ vertices one by one, each time joining the new vertex to a (uniformly) randomly chosen old vertex. This type of random tree is known as a random recursive tree, and it is known that, if D_k is the depth of the kth vertex, then $D_n/\log n \xrightarrow{\text{p}} 1$ [4] and $\max_{k \leqslant n} D_k/\log n \xrightarrow{\text{p}} e$ [5] as $n \to \infty$; see also the survey [6].

This leads to the following result; our condition on the distribution of T_{ij} is presumably much stronger than really required.

Theorem 4.1. *Suppose that* T_{ij} *has a density function* $f(t) = 1 + O(t)$ *for* $t > 0$*. Then, as* $n \to \infty$*:*

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

(i) *for any fixed i and j,*

$$\frac{L_{ij}}{\log n} \xrightarrow{\mathrm{p}} 1;$$

(ii) *for any fixed i,*

$$\frac{\max_{j\leqslant n} L_{ij}}{\log n} \xrightarrow{\mathrm{p}} e.$$

Proof. The case when $T_{ij} \in \mathrm{Exp}(1)$ follows from the discussion before the theorem: we have $L_{ij} = D_N$, where N is uniformly distributed over $2, \dots, n$, and $\max_{j\leqslant n} L_{ij} = \max_{k\leqslant n} D_k$.

In general, we first observe that we may, for a given n, modify the distribution of T_{ij} on the interval $t \geqslant 5\log n/n$ without affecting the result, since, by Theorem 1.1, edges with such large weights are hardly ever used. Hence we may assume that its density function is $1 + O(\log n/n)$ times the density function e^{-t} of the exponential distribution, uniformly for all $t > 0$. It is now easy to see that the minimum weight paths from $i = 1$ form a random tree, obtained by adding vertices one by one as above, with the modification that the probability that the kth vertex (in order of insertion) is joined to the lth, for $l < k$, may depend on the previous history of the tree but is always $(1 + O(\log n/n))/(k-1)$. We may couple this random tree growing process with the one with equal probabilities $1/(k-1)$ in such a way that the probability that a vertex k is joined to different preceding vertices in the two trees is $O(\log n/n)$, even if we condition on the previous history. It follows that, if we fix the end vertex j, the path from $i = 1$ to j is the same in both trees with probability $1 - O(\log^2 n/n)$, which, by the result for the exponential case, implies (i) for a general distribution.

For (ii) we observe that, if D_k is the depth of the kth vertex (in order of insertion) in the tree, and $\bar{D}_k$ is the depth in the random recursive tree with uniformly chosen ancestors, then, by the above, $D_k = \bar{D}_k$ for every $k \leqslant n_1 = n/\log^2 n$ with probability $1 - O(n_1 \log n/n) = 1 - O(1/\log n)$. Since $\max_{k\leqslant n_1} \bar{D}_k/\log n_1 \xrightarrow{\mathrm{p}} e$ by the result quoted above [5], it follows that for every $\varepsilon > 0$, with probability $1 - o(1)$,

$$\max_{k\leqslant n} D_k \geqslant \max_{k\leqslant n_1} D_k = \max_{k\leqslant n_1} \bar{D}_k \geqslant (e-\varepsilon)\log n_1 = \big(e - \varepsilon - o(1)\big)\log n,$$

which by $\max_{j\leqslant n} L_{ij} = \max_{k\leqslant n} D_k$ proves one half of the result.

For the other half, define the generating functions

$$F_m(t) = \mathbb{E}\sum_{k=1}^{m} t^{D_k}$$

and

$$\bar{F}_m(t) = \mathbb{E}\sum_{k=1}^{m} t^{\bar{D}_k}.$$

The recursive definition of the tree yields $\mathbb{E}t^{\bar{D}_{m+1}} = \frac{t}{m}\bar{F}_m(t)$ and thus

$$\bar{F}_{m+1}(t) = \left(1 + \frac{t}{m}\right)\bar{F}_m(t),$$

https://doi.org/10.1017/S0963548399003892 Published online by Cambridge University Press

which together with $\bar{D}_1 = 0$ yields

$$\bar{F}_m(t) = \frac{\Gamma(m+t)}{\Gamma(m)\Gamma(1+t)}.$$

Choosing $t = e$ we obtain, for every $a > e$,

$$\mathbb{P}(\max_{k\leqslant n} \bar{D}_k \geqslant a \log n) \leqslant \mathbb{P}\left(\sum_{k=1}^{n} e^{\bar{D}_k} \geqslant n^a\right) \leqslant n^{-a}\bar{F}_n(e) \sim n^{-a+e}/\Gamma(e+1),$$

which tends to 0 as $n \to \infty$.

For D_k we similarly obtain the inequalities, for some $C < \infty$ and all $t > 0$,

$$\begin{aligned} \mathbb{E}t^{D_{m+1}} &\leqslant \frac{t}{m}\left(1 + C\frac{\log n}{n}\right)F_m(t), \\ F_{m+1}(t) &\leqslant \left(1 + \frac{t}{m}\left(1 + C\frac{\log n}{n}\right)\right)F_m(t), \end{aligned}$$

and thus

$$F_m(t) \leqslant \bar{F}_m\left(t\left(1 + C\frac{\log n}{n}\right)\right).$$

which yields, as above,

$$\mathbb{P}(\max_{k\leqslant n} D_k \geqslant a \log n) \leqslant n^{-a}F_n(e) \leqslant n^{-a}\bar{F}_n(e + Ce\log n/n) \sim n^{-a+e}/\Gamma(e+1),$$

which tends to 0 as $n \to \infty$ for $a > e$. □

Problem 3. *How large is* $\max_{i,j} L_{ij}$*?*

We can show that, if $\alpha \approx 3.591$ is defined by $\alpha \log \alpha - \alpha = 1$, then, for every $\varepsilon > 0$, $\mathbb{P}(e - \varepsilon < \max_{i,j} L_{ij}/\log n < \alpha + \varepsilon) \to 1$. Hence it is natural to conjecture that $\max_{i,j} L_{ij}/\log n$ converges in probability to a constant in $[e, \alpha]$. Which?

Acknowledgements

Most of this work was done during the meeting *Random Graphs and Combinatorial Structures* at Oberwolfach, September–October 1997; I thank several participants, in particular Jim Fill and Johan Håstad, for their helpful comments and questions. The proof of the main theorem was completed a few weeks later, while I tried to get my daughter Sofie back to sleep one night; I thank her for giving me this opportunity.

References

[1] Bollobás, B. (1985) *Random Graphs*, Academic Press, London.
[2] Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.
[3] Lebedev, N. N. (1972) *Special Functions and their Applications*, Dover, New York. (Translated from Russian.)

[4] Moon, J. (1974) The distance between nodes in recursive trees. In *Combinatorics* (*British Combinatorial Conference, Aberystwyth, 1973*), Vol. 13 of *London Math. Soc. Lecture Notes*, Cambridge University Press, pp. 125–132.

[5] Pittel, B. (1994) Note on the heights of random recursive trees and random *m*-ary search trees. *Random Struct. Alg.* **5** 337–347.

[6] Smythe, R. T. and Mahmoud, H. (1995) A survey of recursive trees. *Theory Probab. Math. Statist.* **51** 1–27.