Discrete Applied Mathematics 10 (1985) 47–56
North-Holland

ON THE VALUE OF A RANDOM MINIMUM SPANNING TREE PROBLEM

A.M. FRIEZE

Dept. of Computer Science and Statistics, Queen Mary College, Mile End Road, London E1 4NS, England

Received 23 February 1983
Revised 13 April 1984

Suppose we are given a complete graph on n vertices in which the lengths of the edges are independent identically distributed non-negative random variables. Suppose that their common distribution function F is differentiable at zero and $D=F'(0)>0$ and each edge length has a finite mean and variance. Let L_n be the random variable whose value is the length of the minimum spanning tree in such a graph. Then we will prove the following: $\lim_{n\to\infty} E(L_n)=\zeta(3)/D$ where $\zeta(3)=\sum_{k=1}^{\infty} 1/k^3=1.202\ldots$, and for any $\varepsilon>0$ $\lim_{n\to\infty}\Pr(|L_n-\zeta(3)/D|)>\varepsilon)=0$.

Introduction

Suppose we are given a complete graph on n vertices in which the lengths of the edges are independent identically distributed non-negative random variables. Suppose that their common distribution function F is differentiable at zero and that $D=F'(0)>0$. Let X denote a random variable with this distribution.

Let L_n be the random variable whose value is the length of the minimum spanning tree in such a graph. Then using an overbar to denote expectations, as we will do where convenient throughout the paper, we will prove the following:

Theorem. *If X has finite mean, then*

$$\lim_{n\to\infty} \bar{L}_n=\zeta(3)/D \quad \textit{where } \zeta(3)=\sum_{k=1}^{\infty} 1/k^3=1.202\ldots. \tag{1a}$$

If X has finite variance, then

$$\lim_{n\to\infty} \Pr(|L_n-\zeta(3)/D|>\varepsilon)=0. \quad \Box \tag{1b}$$

The work in this paper was stimulated by Walkup's result [6] that the expected value of a random assignment problem with independent uniform [0,1] lengths is bounded above by 3. An earlier result, based on Walkup's method, that $L_n \le 2(1+\log n/n)$ when the distribution in question is uniform [0,1], was obtained by Fenner and Frieze [2].

See also Steele [5] for the case where n points are scattered in a Euclidean space and Lueker [4] for similar results on problems with normal distributions (in the main).

The uniform case

We first prove the result for the case where X is a uniform [0,1] variable and then extend the result to the general case.

Let $N=\binom{n}{2}$, $V_n=\{1,2,\dots,n\}$ and suppose that the edges $E_n=\{u_1,u_2,\dots,u_N\}$ of our complete graph are numbered so that $l(u_i)\le l(u_{i+1})$, $i=1,2,\dots$ where $l(u)$ is the length of edge u. It follows that

$$E(l(u_i))=i/(N+1),\quad i=1,2,\dots,N. \tag{2}$$

For any positive integer $M\le N$ let G_M denote the graph defined by $u_1,\dots,u_M$. Clearly G_M is a random graph on n vertices and M edges in the sense of Erdös and Rényi [1]. If M is positive but non-integral, then G_M denotes $G_{\lceil M\rceil}$.

Suppose that the minimum length tree is constructed using the Greedy Algorithm of Kruskal [3]. Let $F_0=\emptyset$, $F_1=\{u_1\}$, $F_2,\dots,F_{n-1}$ be the sequence of edge sets of the successive forests produced. Here $|F_i|=i$ and F_{n-1} is the set of edges in the minimum spanning tree.

Next define $T_i=\max(j:u_j\in F_i)$. It follows from (2) that

$$\bar{L}_n=\sum_{i=1}^{n-1}\bar{T}_i/(N+1). \tag{3}$$

We now introduce the function

$$f(a)=\frac{1}{2a}\sum_{t=1}^{\infty}t^{t-2}(2a\mathrm{e}^{-2a})^t/t!,\quad a>0$$

and let $f(0)=0$.

We summarize some of its salient properties: it follows from Erdös and Rényi [1, eq. 6.6] that for $a>0$

$$f(a)=(x-x^2/2)/2a \quad \text{where } x=x(a) \text{ is the unique value satisfying} \quad \text{(i) } 0<x<1,\quad \text{(ii) } x\mathrm{e}^{-x}=2a\mathrm{e}^{-2a}. \tag{4}$$

Thus $x=2a$ and $f(a)=1-a$ for $a<1/2$. Note also that f is strictly monotonic decreasing from 1 down to 0 as a increases from 0 to ∞. This function is needed because of the following lemma (proved later in outline) on random graphs. Throughout the proof $c_1,c_2,\dots$ denote positive constants.

Lemma 1. *If* $1 \le M \le 2n \log n$, *then*

$$\Pr(G_M \textit{ has more than } nf(M/n) + 3n^{4/5} \textit{ components}) \le c_1 n^{-1/6}, \tag{5a}$$

$$\Pr(G_M \textit{ has fewer than } nf(M/n) - n^{4/5} \textit{ components}) \le c_1 n^{-1/6}. \tag{5b}$$

We shall also prove later

Lemma 2. $\Pr(G_{2n \log n}$ *is not connected*$) \le c_2 n^{-3}$. □

We can obtain some bounds on $\bar{T}_k$. For $0<z<1$ we define $a(z) = f^{-1}(1-z)$. We shall now be able to prove that for $1 \le k \le m = \lceil n - 3n^{4/5} \rceil$ that $|T_k\text{-}na(k/n)|$ is 'small enough'.

So let $b_k = a(k/n + 3n^{-1/5})$, which is well defined for $k \le m$. Now clearly

$$\bar{T}_k \le nb_k + 2n \log n \Pr(nb_k < T_k \le 2n \log n) + N \Pr(T_k > 2n \log n). \tag{6}$$

But for any $M \le N$

$$T_k > M \quad \text{if and only if} \quad G_M \text{ has more than } n-k \text{ components}. \tag{7}$$

Thus using (5a) and (7) we obtain

$$\Pr(T_k > nb_k) \le c_1 n^{-1/6} \tag{8}$$

on noting that $n - k = nf(b_k) + 3n^{4/5}$.

Now Lemma 2 implies

$$\Pr(T_k > 2n \log n) \le c_2 n^{-3}. \tag{9}$$

Thus from (6), (8) and (9) we obtain

$$\bar{T}_k \le nb_k + 2c_1 n^{5/6} \log n + c_2/2n \quad \text{for } 1 \le k \le m. \tag{10}$$

Now for $k > m$, we have (crudely)

$$\begin{aligned} \bar{T}_k &\le 2n \log n + N \Pr(T_k > 2n \log n) \\ &\le 2n \log n + c_2/n. \end{aligned} \tag{11}$$

Now from (3), (10) and (11) we obtain

$$\bar{L}_n \le \left(\left(n \sum_{k=1}^{m} a(k/n + 3n^{-1/5}) \right) / (N+1) \right) + u_n \tag{12}$$

where

$$\begin{aligned} u_n &= (n(2c_1 n^{5/6} \log n + c_2/2n) + 3n^{4/5}(2n \log n + c_2/n))/(N+1) \\ &= \mathrm{O}(n^{-1/6} \log n). \end{aligned}$$

Now as $a(z)$ is monotonic increasing we have

$$\sum_{k=1}^{m} a(k/n+3n^{-1/5}) \le nI \quad \text{where } I=\int_0^1 a(z)\,dz.$$

It follows immediately from (12) that

$$\lim_{n\to\infty} \sup \bar{L}_n \le 2I. \tag{13}$$

To get a lower bound for $\bar{T}_k$ we define $b'_k = a(k/n - n^{-1/5})$ for $k \ge n/2$ and note that

$$\begin{aligned} \bar{T}_k &\ge nb'_k \Pr(T_k \ge nb'_k) \qquad (14)\\ &\ge nb'_k(1-c_1 n^{-1/6}) \qquad \text{using (5b) and (7).} \end{aligned}$$

Now clearly $\bar{T}_k \ge k = na(k/n)$ for $k \le n/2$ and hence from (3) and (14) we have

$$\bar{L}_n \ge (n/(N+1))\left(\sum_{k=1}^{\lfloor n/2 \rfloor} a(k/n) + \sum_{k>n/2}^{m} a(k/n - n^{-1/5})\right)(1-c_1 n^{-1/6})$$

from which we deduce $\lim_{n\to\infty} \inf \bar{L}_n \ge 2I$ and in conjunction with (13) we have

$$\lim_{n\to\infty} \bar{L}_n = 2\int_0^1 a(z)\,dz = -2\int_0^\infty af'(a)\,da = 2\int_0^\infty f(a)\,da$$

(on integrating by parts and using $af(a) = (x - x^2/2)/2$ where x is as in (4))

$$= 2\sum_{k=1}^{\infty} (k^{k-2}/k!) \int_0^\infty (2a)^{k-1} e^{-2ak}\,da = \sum_{k=1}^{\infty} 1/k^3$$

which proves (1a) for the case in which the edge weights are uniform on [0,1].

We next prove (1b) by showing that $\mathrm{Var}(L_n) \to 0$ and $n \to \infty$ and deducing our result from the Chebycheff inequality.

We first state a result that can be readily verified by simple integration: let $X_{(p)}$ denote the pth smallest out of N independent uniform [0,1] random variables. Then

$$E(X_{(p)}X_{(q)}) = p(q+1)/((N+1)(N+2)), \quad 1 \le p \le q \le N. \tag{15}$$

Next let s_k, $k=1,\dots,n-1$ denote the length of the kth edge chosen by the Greedy Algorithm. Thus s_k is the T_kth smallest out of N independent uniform [0,1] random variables.

Therefore if $1 \le k \le l \le n-1$

$$\begin{aligned} E(s_k s_l) &= \sum_{p=k}^{N} \sum_{q=p}^{N} E(s_k s_l \mid T_k = p, T_l = q) \Pr(T_k = p, T_l = q)\\ &= \sum_{p=k}^{N} \sum_{q=p}^{N} (p(q+1)/((N+1)(N+2))) \Pr(T_k = p, T_l = q)\\ &= (E(T_k T_l) + E(T_k))/(N+1)(N+2)). \end{aligned}$$

To show that $\mathrm{Var}(L_n) \to 0$, all we have to prove is that

$$\sum_{k=1}^{n-1}\sum_{l=1}^{n-1} E(T_kT_l) \le (1+o(1))\sum_{k=1}^{n-1}\sum_{l=1}^{n-1} \bar{T}_k\bar{T}_l. \tag{16}$$

This is straightforward. For example we have that for $1 \le k \le l \le m$

$$\begin{aligned} E(T_kT_l) \le\ & 2b_k'n^2 \log n \Pr(T_k < b_k'n \text{ and } T_l \le 2n\log n) \\ & + 2b_l'n^2 \log n \Pr(T_k \le 2n\log n \text{ and } T_l < b_l'n) \\ & + 4n^2(\log n)^2 \Pr(b_kn < T_k \le 2n\log n \text{ and } b_ln < T_l \le 2n\log n) \\ & + N^2 \Pr(T_k < 2n\log n) + b_kb_ln^2. \\ \le\ & \bar{T}_k\bar{T}_l + c_3 n^{11/6}(\log n)^2 \end{aligned} \tag{17}$$

after some simple approximations. The contributions when k or $l > m$ to the left hand side of (16) can be shown to be small and (16) follows easily.

We obtain (1b) immediately from the Chebycheff inequality.

Extension to the general case

We now extend our results to the case where the edge weights are independently and identically distributed as a non-negative random variable X with probability functions F, i.e., $\Pr(X \le x) = F(x)$ for $x \ge 0$. Suppose also that

$$\mu = E(X) < \infty \quad \text{and} \quad \nu = E(X^2) < \infty.$$

Suppose now that F is differentiable at $x=0$ and $D = F'(0) > 0$. For a given small $\varepsilon > 0$ there exists $h = h(\varepsilon) > 0$ such that

$$F(x) \ge (D-\varepsilon)x \quad \text{for } 0 \le x \le h$$

Suppose now that we define a new random variable X_ε with probability function F_ε where

$$\begin{aligned} F_\varepsilon(x) &= (D-\varepsilon)x \quad \text{if } 0 \le x \le h, \\ &= F(x) \qquad\quad \text{if } h < x. \end{aligned} \tag{18}$$

Assuming for the present that the edge lengths are now independent random variables distributed like X_ε, then were $L_{n,\varepsilon}$ denotes the random variable which is the length of the minimum spanning tree in the graph produced,

$$\bar{L}_n \le \bar{L}_{n,\varepsilon}. \tag{19}$$

Let $T_{n,\varepsilon}$ denote the minimum spanning tree in the graph and let $E_{n,\varepsilon} = \{e \in E_n: l(e) \le h\}$. For $S \subseteq E_n$,

$$l(S) = \sum_{e \in S} l(e).$$

Now clearly

$$L_{n,\varepsilon}=l(T_{n,\varepsilon}\cap E_{n,\varepsilon})+l(T_{n,\varepsilon}\cap(E_n-E_{n,\varepsilon})). \tag{20}$$

To deal with $l(T_{n,\varepsilon}\cap E_{n,\varepsilon})$ we consider the problem in which edge weights for $e\notin E_{n,\varepsilon}$ are uniformly randomly generated between h and $1/(D-\varepsilon)$. Let $\tilde{T}_{n,\varepsilon}$ be the minimum spanning tree in this graph. Clearly

$$l(\tilde{T}_{n,\varepsilon})\geq l(T_{n,\varepsilon}\cap E_{n,\varepsilon}) \tag{21}$$

and in this problem the edge weights are uniformly distributed in $[0,1/(D-\varepsilon)]$. Scaling the uniform $[0,1]$ case leads to

$$\lim_{n\to\infty}E(l(\tilde{T}_{n,\varepsilon}))=\zeta(3)/(D-\varepsilon), \tag{22a}$$

$$\lim_{n\to\infty}\operatorname{Var}(l(\tilde{T}_{n,\varepsilon}))=0. \tag{22b}$$

To deal with $L=l(T_{n,\varepsilon}\cap(E_n-E_{n,\varepsilon}))$ define the events

A: $-|E_{n,\varepsilon}|\geq(D-\varepsilon)Nh/2$,

B: the graph $(V_n,E_{n,\varepsilon})$ is connected.

Now

$$E(L)=E(L\mid A\cap B)\Pr(A\cap B)+\sum_{Z\in\Omega}E(L\mid E_{n,\varepsilon}=Z)\Pr(E_{n,\varepsilon}=Z) \tag{23}$$

where

$$\Omega=\{Z\subseteq E_n:|Z|<(D-\varepsilon)Nh/2 \text{ or } (V_n,Z) \text{ is not connected}\}.$$

Now if B occurs, then $T_{n,\varepsilon}\subseteq E_{n,\varepsilon}$ and so $E(L\mid A\cap B)=0$. Also, for large n

$$\Pr(\bar{A}\cup\bar{B})\leq\Pr(\bar{A})+\Pr(\bar{B})\leq c_4n\mathrm{e}^{-(D-\varepsilon)nh}. \tag{24}$$

Here $\Pr(A)\leq\mathrm{e}^{-(D-\varepsilon)Nh/8}$ follows from the Chernoff Inequalities for the Binomial Series and $\Pr(B)=O(n\mathrm{e}^{-(D-\varepsilon)nh})$ can be proved in the same way as Lemma 2.

Now let S_n denote the tree $\{\{1,k\}:k=2,\ldots,n\}$. Clearly for $Z\in\Omega$

$$\begin{aligned}E(L\mid E_{n,\varepsilon}=Z)&\leq E(l(S_n)\mid E_{n,\varepsilon}=Z)\\&=(n-1)E(l(1,2)\mid E_{n,\varepsilon}=Z)\\&\leq(n-1)E(l(1,2)\mid l(1,2)\geq h)\\&\leq(n-1)\mu/\Pr(X\geq h).\end{aligned}$$

Combining this with (23) and (24) gives

$$E(L)\leq c_4n^2\mu\mathrm{e}^{-(D-\varepsilon)nh}/\Pr(X\geq h). \tag{25a}$$

A similar argument yields

$$E(L^2) \leq c_4(n^3\mu^2 + n^2\nu)e^{-(D-\varepsilon)nh}/\Pr(X \geq h). \tag{25b}$$

It now follows from (19), (20), (21), (22a), and (25a) that

$$\limsup_{n\to\infty} E(L_n) \leq \zeta(3)/(D-\varepsilon) \quad \text{for all (small) } \varepsilon > 0. \tag{26}$$

Now from (20) and (21)

$$E(L_{n,\varepsilon}^2) \leq E(l(\tilde{T}_{n,\varepsilon})^2) + 2(n-1)hE(L) + E(L^2).$$

Thus from $E(L_n^2) \leq E(L_{n,\varepsilon}^2)$ and (22) and (25) we have

$$\limsup_{n\to\infty} E(L_n^2) \leq (\zeta(3)/(D-\varepsilon))^2 \quad \text{for all (small) } \varepsilon > 0. \tag{27}$$

On the other hand there exists $0 \leq \hat{h} = \hat{h}(\varepsilon) \leq (D+\varepsilon)^{-1}$ such that

$$F(x) \leq (D+\varepsilon)x \quad \text{for } 0 \leq x \leq \hat{h}.$$

We now define a new random variable $\hat{X}_\varepsilon$ with probability function $\hat{F}_\varepsilon$ where

$$\begin{aligned} \hat{F}_\varepsilon(x) &= (D+\varepsilon)x & &\text{if } 0 \leq x \leq \hat{h}, \\ &= \max((D+\varepsilon)h, F(x)) & &\text{if } \hat{h} \leq x. \end{aligned}$$

If edge lengths are now independent random variables distributed like $\hat{X}_\varepsilon$, and $\hat{L}_{n,\varepsilon}$ denotes the length of the minimum spanning tree, then clearly

$$E(L_n) \geq E(\hat{L}_{n,\varepsilon}), \quad \text{etc.}$$

A similar analysis to that for (26) and (27), then yields

$$\liminf_{n\to\infty} E(L_n) \geq \zeta(3)/(D+\varepsilon) \qquad \text{for all } \varepsilon > 0, \tag{28}$$

$$\liminf_{n\to\infty} E(L_n^2) \geq (\zeta(3)/(D+\varepsilon))^2 \quad \text{for all } \varepsilon > 0. \tag{29}$$

Combining (26), (27), (28) and (29) yields the main result of the paper.

Proof sketches for Lemmas 1 and 2

It remains to prove Lemmas 1 and 2. To do this we have to look at the number of components in the random graph G_M. This question was analysed in detail in the classic paper of Erdös and Rényi [1] and it was this that made us suspect that an asymptotically accurate value for $\bar{L}_n$ could be obtained. In their paper they compute the expected number of components in the graph G_{cn} for c fixed and n tending to infinity. We however need to estimate the probabilities that the number of components differs from the expected number by a given amount where c may depend on n. Rather than try the reader's patience by repeating the calculations of Erdös and Rényi we just indicate the main steps of the argument.

Proof of Lemma 1. For $M \le n/4$ the expected number of cycles in G_M is bounded by a constant and as the number of components lies between $n-M$ and $n-M+C$, where C is the number of cycles, using the Markov inequality on C suffices in this case.

The following lemma is useful in calculations for $M > n/4$:

Lemma 3. *Suppose $n/4 < M \le 2n\log n$ and $1 \le p \le 2n^{1/5}$ and $1 \le q \le \binom{p}{2}$. If*

$$u = u(p,q) = \binom{\binom{n-p}{2}}{M-q} \Big/ \binom{\binom{n}{2}}{M}$$

then

$$(1 - c_5((p+q)\log n)^2/n)\, v \le u \le (1 + c_5((p+q)\log n)^2/n)\, v$$

where $v = \mathrm{e}^{-2Mp/n}(2M/n^2)^q$. □

The proof of this lemma is omitted.

Most of the components in G_M are isolated trees with fewer than $n^{1/5}$ vertices and so the following results are useful and can be proved easily with the aid of Lemma 3.

Lemma 4. (a) *Let t_k be the number of components of G_M which are trees with k vertices. Then if $k \le n^{1/5}$*

$$\bar{t}_k = (n/2a)(k^{k-2}/k!)(2a\mathrm{e}^{-2a})^k(1 + \theta(k\log n)^2/n)$$

where $a = M/n$ and $|\theta| \le c_6$.

(b) *If $\sum_{k=1}^{\lfloor n^{1/5} \rfloor} t_k$, then*

$$\mathrm{Var}(T) \le \bar{T} + c_6(\bar{T}\log n)^2 n^{-3/5}. \quad □$$

The number of small components which are not trees is likely to be small:

Lemma 6. *Let P be the number of components of G_M which are not trees and have no more than $n^{1/5}$ vertices. Then $\bar{P} \le c_7 n^{1/5}$.* □

(The proof of this lemma can be based on the following crude estimate: the number of connected labeled graphs with k vertices and $l \ge k$ edges is no more than $k^{k-2}\binom{k}{2}^{l-k+1}$.)

To prove Lemma 1 we note next: If G_M has more than $nf(M/n) + 3n^{4/5}$ components, then

$$G_M \text{ has more than } nf(M/n) + n^{4/5} \text{ components which are trees and have no more than } n^{1/5} \text{ vertices} \tag{31a}$$

or

G_M has more than $n^{4/5}$ components which are not trees, but have no more than $n^{1/5}$ vertices (31b)

or

G_M has more than $n^{4/5}$ components which have at least $n^{1/5}$ vertices – which is clearly impossible. (31c)

Lemma 6 and the Markov inequality deal with (31b), Lemmas 4 and 5 and the Chebycheff inequality deal with (31a).

Similarly we can use Lemmas 4 and 5 and the Chebycheff inequality to show that G_M usually has at least $nf(M/n) - n^{4/5}$ isolated trees. □

Proof of Lemma 2. If now $M = 2n \log n$ and $S \subset \{1, 2, \ldots, n\}$ and $|S| = k \le n/2$, then

$$\Pr(\text{there are no edges in } G_M \text{ joining } S \text{ and } \bar{S})$$
$$= p_k = \binom{\binom{n}{2} - k(n-k)}{M} \Big/ \binom{\binom{n}{2}}{M}.$$

Thus

$$\varrho = \Pr(G_M \text{ is not connected}) \le \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} p_k.$$

But

$$\binom{n}{k+1} p_{k+1} \Big/ \binom{n}{k} p_k \le ((n-k)/(k+1)) \mathrm{e}^{-4(n-2k-1)\log n/n}$$

from which $\varrho = \mathrm{O}(np_1) = \mathrm{O}(n^{-3})$. □

Acknowledgement

I would like to thank Colin McDiarmid for some very useful conversations and a referee for a careful report.

References

[1] P. Erdös and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hungarian Acad. Sci. 5A (1960) 17–60.

[2] T.I. Fenner and A.M. Frieze, On the connectivity of random *m*-orientable graphs and digraphs, Combinatorica 2 (1982).

[3] J.B. Kruskal, On the shortest spanning subtree of a graph and the travelling salesman problem, Proc. Amer. Math. Soc. 7 (1956) 48–50.

[4] G.S. Lueker, Optimisation problems on graphs with independent random edge weights, SIAM J. Comput. 10 (1981) 338–351.

[5] J.M. Steele, Growth rates of minimal spanning trees of multivariate sample, Stanford Univ., Dept. of Statistics, Research Report (1981).
[6] D.W. Walkup, On the expected value of a random assignment problem, SIAM J. Comput. 8 (1979) 440–442.