

ON THE VALUE OF A RANDOM MINIMUM WEIGHT  
STEINER TREE

BÉLA BOLLOBÁS\*, DAVID GAMARNIK, OLIVER RIORDAN,  
BENNY SUDAKOV†

Received April 12, 2001

Consider a complete graph on  $n$  vertices with edge weights chosen randomly and independently from an exponential distribution with parameter 1. Fix  $k$  vertices and consider the minimum weight Steiner tree which contains these vertices. We prove that with high probability the weight of this tree is  $(1 + o(1))(k - 1)(\log n - \log k)/n$  when  $k = o(n)$  and  $n \rightarrow \infty$ .

**1. Introduction**

Given an arbitrary weighted graph with a fixed set of vertices, the *Steiner tree problem* is the task of finding a minimum weight subtree containing all these vertices, where the weight of a tree is the sum of the weights of the edges it contains. Steiner trees are very well studied objects in combinatorial optimization; the interest is motivated by several practical problems such as network design and VLSI design. The Steiner tree problem is well known to be NP-complete; this separates it from the superficially similar minimum spanning tree problem, for which there is a simple polynomial time algorithm. Most of the theoretical work on the Steiner tree problem concerns obtaining approximation algorithms. Currently, the best approximation factor is 1.55, obtained by Robins and Zelikovsky [17]. Arora [2] showed that

---

*Mathematics Subject Classification (2000)*: 05C80, 60C05, 68R10

\* Research supported in part by NSF grant DSM 9971788

† Research supported in part by NSF grants DMS-0106589, CCR-9987845 and by the State of New Jersey. Part of this research was done while visiting IBM T. J. Watson Research Center.

an  $\epsilon$ -approximation can be achieved for every  $\epsilon > 0$  when the underlying weighted graph is Euclidean. On the other hand, unless  $P=NP$ , the Steiner tree problem in general (weighted) graphs can not be approximated within a factor of  $1+\epsilon$  for some  $\epsilon > 0$ , see [4], [7].

In this paper we focus on random instances of the Steiner tree problem. The study of random instances of combinatorial problems has a very rich history, starting with the random Euclidean traveling salesman problem [3] and including studies of the random minimal spanning tree [11], assignment [1], and shortest path problems [8],[13], as well as many other examples. The main motivation is to complement the classical worst case analysis of algorithms with the analysis of the performance of algorithms on typical instances. On the other hand, very little is known about the Steiner tree problem in a random setting. To the best of our knowledge, the only existing results were obtained by Kučera et al. in [15], where the authors studied algorithms for finding minimum Steiner trees in random graphs with all edge weights equal to 1. We consider the problem of finding the minimum weight of a Steiner tree in a complete graph  $K_n$  on  $n$  vertices with edge weights chosen independently from some distribution  $X$ . Essentially, this distribution can be any non-negative distribution with positive density at zero. Rescaling, we shall assume that

$$(1) \quad X \geq 0 \text{ and } \mathbb{P}(X \leq x) = x + o(x),$$

as  $x \rightarrow 0$ . We also assume that  $\mathbb{E}(X)$  is finite; some assumption along these lines is needed for results on expectation. The two distributions of most interest are the exponential distribution  $\text{EXP}(1)$  with mean 1 and the uniform distribution on  $[0,1]$ . Throughout the paper we write  $G$  for the weighted graph obtained from  $K_n$  by taking the edge weights as independent copies of  $X$ . We fix  $2 \leq k \leq n$  vertices  $v_1, \dots, v_k$ , and consider  $W(k)$ , the minimum weight of a tree containing these vertices. Our main result is the following, giving the asymptotic value of  $W(k)$  for all functions  $2 \leq k(n) = o(n)$ . Note that throughout the paper all logarithms have base  $e$ .

**Theorem 1.** *Let  $\epsilon > 0$  be fixed, and let  $k = k(n)$  satisfy  $2 \leq k = o(n)$ . Let*

$$w = \frac{k-1}{n}(\log n - \log k).$$

*Then  $(1-\epsilon)w \leq W(k) \leq (1+\epsilon)w$  holds with probability tending to 1 as  $n \rightarrow \infty$ .*

Note that since the graph is complete and the weight distribution is the same for all the edges, the actual choice of the  $k$  vertices is irrelevant. This justifies the notation  $W(k)$ . When  $\log k = o(\log n)$ , the value  $(k-1)\log n/n$

that we obtain has a simple intuitive explanation. Davis and Prieditis [8] and Janson [13] showed that the minimum weight of a path between a fixed pair of vertices of  $G$  is asymptotically  $\log n/n$ . This immediately implies an upper bound  $(1+o(1))(k-1)\log n/n$  on the value of  $W(k)$  when  $k$  is bounded; we just take shortest paths from  $v_i$  to  $v_{i+1}$  for  $1 \leq i \leq k-1$  and delete any redundant edges. Using Dijkstra’s algorithm these paths can be found in  $O(n^2)$  computation steps. As  $k$  becomes larger, the  $\log k$  ‘correction term’ becomes important. Intuitively, this arises from the many ways of connecting  $v_1, \dots, v_k$  with paths (perhaps using additional branching vertices).  $W(k)$  can be written as the minimum of  $\Theta(k)^k$  quantities each of which has expectation  $(k-1)\log n/n$ ; the smallest of these turns out to be typically  $(k-1)\log k/n$  smaller than its expectation. For the upper bound it turns out to be sufficient to consider the shortest path  $P_r$  from  $v_{r+1}$  to  $\{v_1, \dots, v_r\}$  for  $1 \leq r \leq k-1$ , as this corresponds to minimizing over  $(k-1)! = \Theta(k)^k$  trees.

Note that for  $k$  constant the minimum weight Steiner tree can be found in polynomial time  $O(n^k)$ , by a brute force search. An interesting corollary of our result is that for any  $k = o(n)$ , using only  $O(kn^2)$  computation steps ( $O(n^2)$  to find each path  $P_r$  defined above) we can construct an asymptotically minimum weight Steiner tree, when  $n \rightarrow \infty$ .

Our work is motivated by similar research for other random combinatorial structures. One of the earliest such results was obtained by Frieze [11] who showed that with high probability the weight of the minimum spanning tree in  $G$  is asymptotically  $\zeta(3) = \sum_{k=1}^{\infty} k^{-3}$ , when  $n \rightarrow \infty$ . This result was extended to regular graphs with certain expansion properties by Beveridge, Frieze and McDiarmid [5]. The analysis explicitly uses the greedy algorithm for finding a minimum spanning tree. Note that the minimum spanning tree problem is the special case of our problem where  $k = n$ , yet the analysis in [11],[5] does not extend to the general case since the problem is NP-complete and the greedy search fails.

Before turning to the proof of [Theorem 1](#) we note that the distribution  $X$  chosen for the edge weights is irrelevant, as long as it satisfies (1). This will allow us to use whichever distribution is convenient in the proof. Suppose that  $X$  is a distribution satisfying (1), and let  $\delta = n^{-1/2}\log n = o(1)$ . Let  $x$  and  $y$  be fixed vertices of  $G$ . Writing  $w(e)$  for the weight of an edge  $e$ , for each other vertex  $z$  we have

$$\mathbb{P}(w(xz) + w(yz) \leq \delta) \geq \mathbb{P}(X \leq \delta/2)^2 \sim \delta^2/4,$$

so this probability is at least  $\delta^2/5$  if  $n$  is large enough. Thus the probability that there is no  $xy$  path of weight at most  $\delta$  is at most

$$\left(1 - \delta^2/5\right)^{n-2} \leq e^{-(n-2)\delta^2/5},$$

which tends to zero faster than any constant power of  $n$ . Thus with very high probability every pair  $\{x, y\}$  of vertices of  $G$  is connected by a path of weight at most  $\delta$ , and an edge of weight larger than  $\delta$  cannot form part of the shortest path between two vertices, or of the minimum weight Steiner tree for any set of vertices. Suppose  $Y$  is another distribution satisfying (1). Then there is a strictly increasing function  $f$  such that  $f(X)$  has the distribution of  $Y$ . (We are ignoring possible discontinuities in the distribution, but this is not a real problem). Condition (1) for  $X$  and  $Y$  implies that  $f(x)$  and  $f^{-1}(x)$  are both  $x + o(x)$  when  $x$  is small. Thus we can convert between edge weights distributed as copies of  $X$  and edge weights distributed as copies of  $Y$  changing all weights less than  $\delta$  by a factor of  $1 + o(1)$ . As with high probability any minimum weight Steiner tree uses only such edges, and the conversion goes both ways, the minimum weight of a Steiner tree also changes by a factor of  $1 + o(1)$ , and a result such as [Theorem 1](#) for edge weights with distribution  $X$  implies the corresponding result for the distribution  $Y$ .

The rest of this paper is organized as follows. In the next two sections we prove lower and upper bounds on  $W(k)$  for  $k = o(n)$ . By the above argument, we may and shall assume in our proof that the edge weights are independent  $\text{EXP}(1)$  random variables. In [Section 4](#) we study the minimum weight of a Steiner tree on  $k$  vertices when  $k$  is linear in  $n$ . The last section contains some concluding remarks. Throughout the paper we shall write  $w(e)$  for the weight of an edge  $e$ . Also, when  $T$  is a subgraph of  $G$  (usually a tree) we shall write  $w(T)$  for its weight, i.e., the sum of the weights of its edges. All logarithms in this paper are natural.

## 2. Lower bound

In this section we obtain a lower bound on the random minimum weight Steiner tree on  $k = o(n)$  vertices. The idea of the proof is as follows: suppose  $v_1, \dots, v_k$  are given, and consider a minimum weight Steiner tree  $T$  for  $v_1, \dots, v_k$  in  $G$ . Note that  $T$  has at most  $k$  leaves (vertices of degree 1), since only the vertices  $v_i$ ,  $1 \leq i \leq k$ , can be leaves of  $T$ . As usual, let us say that a vertex  $v$  of  $T$  is a *branching vertex* if it has degree at least 3 in  $T$ . In  $T$ , the sum of the degrees of the vertices is  $2(|T| - 1)$ . Thus, although  $T$  may have many vertices of degree 2, it is easy to see that it will have at most  $k - 2$  branching vertices. Let us say a vertex of  $T$  is *critical* if it is either a leaf or a branching vertex, so  $T$  has at most  $2k - 2$  critical vertices. Any tree with  $c$  critical vertices consists of  $c - 1$  paths joining these vertices (contracting vertices of degree 2 leaves a tree with exactly  $c$  vertices), so  $T$  consists of

$l \leq 2k - 3$  paths  $P_1, \dots, P_l$  between certain pairs of vertices. As  $T$  is a minimum weight Steiner tree, each path is the minimum weight path joining its endvertices, and these paths are disjoint. Let us write  $W_1, \dots, W_l$  for the weights of these paths, and  $W$  for the minimum weight of a path between two fixed vertices. The key observation is that the disjointness of the paths  $P_i$  makes it hard for the  $W_i$  to be simultaneously small. We could just use the van den Berg–Kesten inequality [18] or Reimer’s inequality [16] to show that  $\mathbb{P}(\forall i : W_i \leq w_i) \leq \prod \mathbb{P}(W \leq w_i)$ . However we obtain better results by considering  $\mathbb{P}(\sum W_i \leq w)$ , and it is not clear how to bound this probability in this way. Instead we proceed ‘by hand’, using only the Harris–Kleitman lemma. Throughout  $W$  is (the distribution of) the minimum weight of a path between two fixed vertices of  $G$ . We write  $W^{(l)}$  for the sum of  $l$  independent copies of  $W$ .

**Lemma 2.** *Suppose that  $l \geq 1$  and let  $s_1, \dots, s_l, t_1, \dots, t_l$  be fixed vertices of  $G$ , with  $s_i \neq t_i$  for  $1 \leq i \leq l$ . For  $1 \leq i \leq l$  let  $P_i$  be the minimum weight  $s_i t_i$  path in  $G$ . Then for any  $x > 0$  the probability that  $P_1, \dots, P_l$  are disjoint and have total weight at most  $x$  is at most  $\mathbb{P}(W^{(l)} \leq x)$ .*

**Proof.** Let  $P'_1 = P_1$  be the (with probability 1 unique) minimum weight  $s_1 t_1$  path. For  $2 \leq i \leq l$  let  $P'_i$  be the minimum weight  $s_i t_i$  path edge-disjoint from  $P'_1, \dots, P'_{i-1}$ , if there is one. Let  $W_i$  be the weight of  $P'_i$ , or infinity if there is no such path. If the paths  $P_i$  are disjoint, then  $P'_i = P_i$  for every  $i$ . Hence it suffices to show that

$$(2) \quad \mathbb{P}(W_1 + \dots + W_l \leq x) \leq \mathbb{P}(W^{(l)} \leq x).$$

We shall prove this statement by induction on  $l$ . The case  $l = 1$  is clear as  $P'_1 = P_1$  is the shortest  $s_1 t_1$  path, so its weight  $W_1$  has the same distribution as  $W$ , by the definition of  $W$ .

Suppose now that  $l \geq 2$  and that (2) holds with  $l$  replaced by  $l - 1$ . We wish to consider the distribution of  $W_l$  conditional on  $W_1 + \dots + W_{l-1}$ . In fact we shall condition on the very fine event that  $P'_1, \dots, P'_{l-1}$  are certain particular paths, and that the individual edges of these paths have certain weights. For any  $w$  the event  $W_1 + \dots + W_{l-1} = w$  is a disjoint union of events  $\mathcal{A}$  of the form

$$\mathcal{A} = \{w(e_j) = w_j, j = 1, \dots, m\} \cap \mathcal{U}$$

where the edges  $e_1, \dots, e_m$  of  $K_n$  form edge disjoint paths  $p_i$  from  $s_i$  to  $t_i$ ,  $1 \leq i \leq l - 1$ , the  $w_j$  are real numbers adding up to  $w$ , and  $\mathcal{U}$  is the event that certain paths in  $G' = K_n - \{e_1, \dots, e_m\}$  have at least certain weights (so that the paths  $p_i$  satisfy the minimality conditions defining  $P'_i$ ). Consider one

such event  $\mathcal{A}$ , so the  $e_i$  and hence  $E(G')$  are fixed. Let  $\Omega'$  be the product probability space given by the independent  $\text{EXP}(1)$  edge weights on  $G'$ . Then  $\mathcal{U}$  can be considered as an event in  $\Omega'$ , and is then an *up-set*, i.e., an event preserved by increasing the weight of an edge (in  $G'$ ). Let  $W'$  be the minimum weight of an *siti* path in  $G'$ . Then for any  $w'$  the event  $\{W' \leq w'\}$  is a *down-set*, i.e., an event preserved by decreasing the weight of an edge. Now it is well known that up-sets and down-sets are negatively correlated. (This can be seen from the original correlation inequality proved independently by Harris [12] and Kleitman [14]. Alternatively, for discrete random variables, it follows immediately from the FKG inequality of [10]. The continuous case is easy to deduce by ‘discretizing’ the random variables.) Given  $\mathcal{A}$  we have  $W_l = W'$ , so

$$\begin{aligned} \mathbb{P}(W_l \leq w' \mid \mathcal{A}) &= \mathbb{P}_{\Omega'}(W' \leq w' \mid \mathcal{U}) \\ &\leq \mathbb{P}_{\Omega'}(W' \leq w') \\ &\leq \mathbb{P}(W \leq w'), \end{aligned}$$

as  $W'$  is the shortest path between two fixed vertices using edges of  $G'$ , a subgraph of  $G$ . As this holds for each  $\mathcal{A}$  we have

$$\mathbb{P}(W_l \leq w' \mid W_1 + \dots + W_{l-1} = w) \leq \mathbb{P}(W \leq w'),$$

and (2) follows by induction on  $l$ . ■

We now wish to study the sum of independent copies of the random variable  $W$ , the minimum weight of a path joining two fixed vertices of  $G$ . We first use a standard method to describe the distribution of  $W$ . Then we simply estimate its moment generating function.

**Lemma 3.** *Let  $X_1, \dots, X_{n-1}$  be independent  $\text{EXP}(1)$  random variables, and let  $R$  be uniformly distributed on  $\{1, \dots, n-1\}$  and independent of the  $X_i$ . Then  $W$  is distributed as*

$$(3) \quad \sum_{i=1}^R \frac{X_i}{i(n-i)}.$$

For any real  $x, \epsilon > 0$  and any integer  $l \geq 1$  we have

$$\mathbb{P}(W^{(l)} \leq x) \leq (n-1)^{-l} \epsilon^{-l} e^{(1+\epsilon)nx}.$$

**Proof.** The first part is standard; we include the proof for the sake of completeness. Consider fixing a vertex  $x$  of  $G$  and using Dijkstra’s algorithm to find the distance in  $G$  from  $x$  to each other vertex, writing these distances as  $d_1, \dots, d_{n-1}$  in increasing order. Clearly  $d_1$  is the minimum weight of an

edge from  $x$ ; being the minimum of  $(n - 1)$  independent  $\text{EXP}(1)$  random variables, this has an  $\text{EXP}(n - 1)$  distribution. Suppose that this minimum weight edge is  $xy$ . Then  $d_2 - d_1$  is the smaller of  $\min\{w(yz), z \neq x, y\}$  and  $\min\{w(xz) - d_1, z \neq x, y\}$ . By the memoryless property of the exponential distribution, given  $y$  and  $d_1$  the quantities  $w(xz) - d_1$  again have an  $\text{EXP}(1)$  distribution, so  $d_2 - d_1$  is the minimum of  $2(n - 2)$  independent exponentials. Thus  $d_2 - d_1$  has an  $\text{EXP}(2(n - 2))$  distribution and is independent of  $d_1$ . Continuing, we see that  $d_i - d_{i-1}$  is the minimum of  $i(n - i)$  independent exponentials and hence the  $d_i, 1 \leq i < n$ , have the same distribution as

$$\sum_{j \leq i} \frac{X_j}{j(n - j)}, \quad 1 \leq i < n.$$

Since all vertices are equivalent, the distance from  $x$  to another fixed vertex  $x'$  is equally likely to be any of the  $d_i$ , proving the first part of the lemma.

In order to estimate  $\mathbb{P}(W^{(l)} \leq x)$  we consider the moment generating function of  $W$ . If  $X$  has an  $\text{EXP}(1)$  distribution and  $\alpha > 0$ , then

$$\mathbb{E}(e^{-\alpha X}) = \int_0^\infty e^{-\alpha t} e^{-t} dt = \frac{1}{1 + \alpha}.$$

Thus for  $\theta > 0$  we have

$$\mathbb{E}(e^{-\theta X_i/i}) = \frac{1}{1 + \theta/i} = \frac{i}{\theta + i}.$$

Let  $S_m = \sum_{i=1}^m X_i/i$  and let  $Z = S_R$ . Defining  $W$  by the sum (3), which gives the correct distribution by the first part of the lemma, we have  $nW \geq Z$ . Now

$$\mathbb{E}(e^{-\theta S_m}) = \prod_{i=1}^m \mathbb{E}(e^{-\theta X_i/i}) = \frac{m!}{(\theta + 1) \cdots (\theta + m)}.$$

For  $\theta > 1$  it is easy to verify by induction on  $n$  that

$$\sum_{m=1}^{n-1} \frac{m!}{(\theta + 1) \cdots (\theta + m)} = \frac{1}{\theta - 1} \left( 1 - \frac{n!}{(\theta + 1) \cdots (\theta + n - 1)} \right) \leq \frac{1}{\theta - 1}.$$

Hence, as  $nW \geq Z = S_R$ ,

$$\begin{aligned} \mathbb{E}(e^{-\theta nW}) &\leq \mathbb{E}(e^{-\theta S_R}) \\ &= \frac{1}{n - 1} \sum_{m=1}^{n-1} \mathbb{E}(e^{-\theta S_m}) \\ &\leq \frac{1}{(n - 1)(\theta - 1)}. \end{aligned}$$

Taking  $\theta = 1 + \epsilon$  and applying Markov's inequality,

$$\begin{aligned} \mathbb{P}(W^{(l)} \leq x) &= \mathbb{P}\left(\exp(-\theta n W^{(l)}) \geq \exp(-\theta n x)\right) \\ &\leq \mathbb{E}(\exp(-\theta n W^{(l)})) / \exp(-\theta n x) \\ &= \mathbb{E}(\exp(-\theta n W))^l e^{\theta n x} \\ &\leq (n-1)^{-l} \epsilon^{-l} e^{(1+\epsilon) n x}, \end{aligned}$$

as required. ■

We are now ready to prove the lower bound on  $W(k)$  in the following rather unattractive form.

**Lemma 4.** *Let  $v_1, \dots, v_k$  be fixed vertices of  $G$ , where  $2 \leq k = k(n) \leq n/e^2$ , and let  $W(k)$  be the minimum weight of a Steiner tree for  $v_1, \dots, v_k$  in  $G$ . If  $\omega = \omega(n) \rightarrow \infty$  (arbitrarily slowly) then*

$$\mathbb{P}\left(W(k) \leq \frac{(k-1)(\log n - \log k - 2 \log \log(n/k) - 3) - \omega}{n}\right) \rightarrow 0$$

as  $n \rightarrow \infty$ .

Before turning to the proof note that this gives the lower bound stated in [Theorem 1](#), since if  $k = o(n)$  and  $\omega$  tends to infinity sufficiently slowly then the bound on  $W(k)$  above is asymptotically  $(k-1)(\log n - \log k)/n$ , as required.

**Proof of Lemma 4.** Let  $v_1, \dots, v_k$  be fixed vertices of  $G$  and let  $T$  be a minimum weight Steiner tree for  $v_1, \dots, v_k$ . Recall that a vertex  $v$  of  $T$  is a branching vertex if  $v$  has degree at least 3 in  $T$ . Suppose that the branching vertices of  $T$  not among  $\{v_1, \dots, v_k\}$  are  $v_{k+1}, \dots, v_m$ . To make it clear exactly how we are counting later, we shall insist that  $v_{k+1} \prec \dots \prec v_m$  for some arbitrary order on  $V(G)$ . As noted at the beginning of this section, since  $T$  has at most  $k$  leaves it has at most  $k-2$  branching vertices, so  $m \leq 2k-2$ . The tree  $T$  consists of edge-disjoint paths  $P_i$  from  $v_{s_i}$  to  $v_{t_i}$ ,  $1 \leq i \leq m-1$ , for some  $1 \leq s_i, t_i \leq m$ , such that all internal vertices of every path are the vertices of degree 2 in  $T$ . By minimality, each  $P_i$  is the minimum weight path from  $v_{s_i}$  to  $v_{t_i}$ . The structure of  $T$  can be represented by an auxiliary tree  $T'$  on  $[m] = \{1, 2, \dots, m\}$ ; there is one edge  $s_i t_i$  of  $T'$  for each path  $P_i$  in  $T$ . (Again, for the counting we will assume that the  $s_i$  and  $t_i$  are fixed by  $T$ . In other words, given a tree  $T'$  on  $[m]$  we fix the order of the edges and the starting vertex of each edge according to some arbitrary rule.)

For each choice of  $m, v_{k+1}, \dots, v_m$  and  $T'$ , applying Lemma 2 with  $l = m - 1$  shows that the probability of such a minimum weight Steiner tree  $T$  existing with  $w(T) \leq x$  is at most  $\mathbb{P}(W^{(m-1)} \leq x)$ . This in turn is at most

$$(n - 1)^{-(m-1)} \epsilon^{-(m-1)} e^{(1+\epsilon)nx}$$

for any  $\epsilon > 0$  by Lemma 3. Given  $m$ , there are  $\binom{n-k}{m-k} \leq (n-1)^{m-k} / (m-k)!$  choices for  $v_{k+1}, \dots, v_m$ , and at most  $m^{m-2} < m^{m-1}$  choices for the auxiliary tree  $T'$  on  $[m]$ . (The larger bound will be more convenient in the estimates below.) Thus for any  $x, \epsilon > 0$  we have

$$\begin{aligned} \mathbb{P}(W(k) \leq x) &\leq \sum_{m=k}^{2k-2} \frac{(n-1)^{m-k}}{(m-k)!} m^{m-1} (n-1)^{-(m-1)} \epsilon^{-(m-1)} e^{(1+\epsilon)nx} \\ &= \sum_{m=k}^{2k-2} \frac{m^{m-1}}{(m-k)!} (n-1)^{-(k-1)} \epsilon^{-(m-1)} e^{(1+\epsilon)nx}. \end{aligned}$$

Intuitively, for  $k$  small the dominant term should be  $m = 2k - 2$ , as the minimum weight Steiner tree will have  $v_1, \dots, v_k$  as leaves and all branching vertices will have degree 3. In fact, for the whole range of  $k$  the quantity  $a_m = m^{m-1} / (m-k)!$  increases rapidly with  $m$ ; for  $m > k \geq 2$  we have

$$\frac{a_m}{a_{m-1}} = \left(\frac{m}{m-1}\right)^{m-2} \frac{m}{m-k} \geq \left(1 + \frac{1}{m-1}\right)^{m-2} \geq \frac{3}{2}.$$

For  $\epsilon \leq 1$  the  $\epsilon^{-(m-1)}$  term also increases with  $m$ , so the whole sum above is bounded by  $\sum_{i \geq 0} (2/3)^i = 3$  times the value of the term with  $m = 2k - 2$ . Loosening the estimates slightly to simplify the final expression, and applying Stirling's formula in the weak form  $r! \geq (r/e)^r$ , we obtain

$$\begin{aligned} \mathbb{P}(W(k) \leq x) &\leq 3 \frac{(2k-2)^{2k-3}}{(k-2)!} (n-1)^{-(k-1)} \epsilon^{-(2k-3)} e^{(1+\epsilon)nx} \\ &\leq O(1) \frac{(2k-2)^{2k-2}}{(k-1)!} (n-1)^{-(k-1)} \epsilon^{-2(k-1)} e^{(1+\epsilon)nx} \\ &\leq O(1) (4e)^{k-1} ((k-1)/(n-1))^{k-1} \epsilon^{-2(k-1)} e^{(1+\epsilon)nx} \\ &\leq O(1) (4ekn^{-1} \epsilon^{-2})^{k-1} e^{(1-\epsilon)^{-1}nx}. \end{aligned}$$

Taking logs we see that  $\mathbb{P}(W(k) \leq x_0) \rightarrow 0$  if for some  $\omega \rightarrow \infty$  we have

$$\begin{aligned} x_0 &= (1 - \epsilon)(k - 1)(\log n - \log k - 1 - \log 4 - 2 \log(\epsilon^{-1})) / n - \omega / n \\ &\geq (k - 1)(\log n - \log k - 1 - \log 4 - 2 \log(\epsilon^{-1}) - \epsilon \log(n/k)) / n - \omega / n. \end{aligned}$$

The best choice for  $\epsilon$  is  $\epsilon = 2/\log(n/k)$  which is at most 1 by the assumption on  $k$ . For this  $\epsilon$  the bound above simplifies to

$$x_0 \geq (k-1)(\log n - \log k - 2 \log \log(n/k) - 3)/n - \omega/n,$$

which is the quantity appearing in the statement of the lemma. ■

**Remark.** The bound  $m^{m-2}$  on the number of trees in the calculation above turns out to be a key step—it is from this that the  $-\log k$  term in the final bounds comes. One might thus think that we should bound the number of trees more carefully, using the fact that  $v_{k+1}, \dots, v_m$  all have degree at least 3. Although a method for doing this was suggested to us by Cecil Rousseau, it turns out that the only gain is to improve the constant 3 appearing above. This constant is irrelevant for [Theorem 1](#), while for  $k = \Theta(n)$  the correct constant presumably cannot be obtained in this way. Thus for simplicity we have just used the Cayley bound.

### 3. Upper bound

In this section we complete the proof of [Theorem 1](#) by obtaining an upper bound on the minimum weight of a Steiner tree in  $G$ . The idea is to consider a simple method of constructing an upper bound for  $W(k)$  given  $G$  whose behaviour when  $G$  is random we can analyze. The simplest approach, taking a shortest path from  $v_r$  to  $v_{r+1}$  for  $1 \leq r \leq k-1$ , does not suffice - this would give a bound like  $(k-1)\log n/n$ . Instead, for  $1 \leq r \leq k-1$  let  $M_r$  be the minimum weight of a path from  $v_{r+1}$  to  $\{v_1, \dots, v_r\}$ . It is easy to see that the union of such paths is a connected graph containing  $\{v_1, \dots, v_k\}$  and thus that  $M_1 + \dots + M_{k-1}$  is an upper bound on  $W(k)$ . Indeed, by our construction every vertex  $v_i$  is connected by a path to a vertex with index smaller than  $i$ . This implies that there is a path from every vertex to  $v_1$ . Now it is easy to calculate the expectation of  $M_r$ , which turns out to be about  $\log(n/r)/n$ . This gives us an upper bound on  $\mathbb{E}(W(k))$ . As this matches the lower bound of the previous section, we obtain the required high probability bound as an immediate consequence.

**Lemma 5.** *Let  $\alpha < 1$  be constant. Then*

$$\mathbb{E}(M_r) \sim \frac{\log n - \log r}{n - r}$$

*holds uniformly for  $1 \leq r \leq \alpha n$ .*

**Proof.** Consider running Dijkstra’s algorithm starting from the set  $V_r = \{v_1, \dots, v_r\}$ . Arguing as in the proof of [Lemma 3](#), the distances from  $V_r$  to the remaining  $n - r$  vertices are distributed as  $(T'_1, \dots, T'_{n-r})$ , where

$$T'_m = \sum_{i=r}^{r+m-1} \frac{X_i}{i(n-i)},$$

with the  $X_i$  independent  $\text{EXP}(1)$  random variables as before. Now as  $v_{r+1}$  is equally likely to be the closest, 2nd closest etc. vertex to  $V_r$  we have

$$\begin{aligned} \mathbb{E}(M_r) &= \frac{1}{n-r} \sum_{m=1}^{n-r} \mathbb{E}(T'_m) \\ &= \frac{1}{n-r} \sum_{m=1}^{n-r} \sum_{i=r}^{r+m-1} \frac{1}{i(n-i)} \\ &= \frac{1}{n-r} \sum_{i=r}^{n-1} \sum_{m=i+1-r}^{n-r} \frac{1}{i(n-i)} \\ &= \frac{1}{n-r} \sum_{i=r}^{n-1} \frac{1}{i} \sim \frac{\log n - \log r}{n-r}. \end{aligned}$$

In the last step we used the fact that  $\delta_j = \sum_{i=1}^j i^{-1} - \log j$  is bounded to deal with the case  $r = O(1)$ , and the fact that  $\delta_j$  tends to a constant as  $j \rightarrow \infty$ , while  $\log n - \log r$  is bounded away from zero, to deal with the case  $r \rightarrow \infty$ . ■

Let  $M(k) = \sum_{r=1}^{k-1} M_r$ . Using the lower bound from [Theorem 1](#), [Lemma 5](#) has the following corollary, where

$$w = \frac{k-1}{n} (\log n - \log k)$$

is the quantity appearing in [Theorem 1](#).

**Corollary 6.** *Let  $\epsilon > 0$  be fixed, and let  $k = k(n)$  satisfy  $2 \leq k = o(n)$ . Then*

$$(1 - \epsilon)w \leq W(k) \leq M(k) \leq (1 + \epsilon)w$$

*holds with probability tending to 1 as  $n \rightarrow \infty$ .*

**Proof.** The first inequality was proved in the previous section. The second is the observation made at the start of this section. For the third, as  $k = o(n)$ ,

by [Lemma 5](#) we have

$$\begin{aligned} \mathbb{E}(M(k)) &\sim \sum_{r=1}^{k-1} \frac{\log n - \log r}{n - r} \\ &\sim \sum_{r=1}^{k-1} \frac{\log n - \log r}{n} \\ &\sim \frac{k-1}{n} (\log n - \log k + 1) \sim w. \end{aligned}$$

Fix  $\delta > 0$ . Since for  $n$  large enough  $M(k) \geq (1 - \epsilon\delta/4)w$  holds with probability at least  $1 - \epsilon\delta/4$ , it follows that we cannot have  $M(k) \geq (1 + \epsilon)w$  with probability  $\delta > 0$ . (Otherwise we would have  $\mathbb{E}(M(k)) \geq (1 + \epsilon\delta/2)w$  for large  $n$ .) This proves the corollary, and hence completes the proof of [Theorem 1](#). ■

#### 4. Steiner trees for many vertices

So far we have found asymptotically the minimum weight  $W(k)$  of a Steiner tree for  $k$  vertices of  $G$ , for all functions  $k = o(n)$ . In this section we consider the case when  $k$  is linear in  $n$ . For  $k = n$  such a tree is just a spanning tree in  $G$ , so by the result of Frieze [[11](#)] we have  $W(n) = \zeta(3) + o(1)$  with high probability. It is thus natural to ask what happens in between, when  $k = \alpha n$  with  $\alpha = \alpha(n)$  bounded away from 0 and 1, say. Of course the proof of our main result gives some bounds for free: from [Lemma 4](#) it is easy to see that if  $\alpha < e^{-7}$  then for any  $\epsilon > 0$  we have

$$W(k) \geq \alpha(\log(\alpha^{-1}) - 2 \log \log(\alpha^{-1}) - 3) - \epsilon$$

with probability tending to 1 as  $n \rightarrow \infty$ . In the other direction, writing  $V$  for the set of  $k$  vertices to be connected, one might expect that, as there are so many ways to use the vertices not in  $V$  as part of the Steiner tree, almost all other vertices will be used. Thus one might expect  $W(k) = \zeta(3) + o(1)$  whenever  $k = \alpha n$  with  $\alpha$  bounded away from zero. It turns out that this is not the case. Suppose that  $k = k(n)$  is such that  $\alpha = k/n$  is bounded away from 0 and 1. As before let  $M_r$  be the minimum weight of a path from  $v_{r+1}$  to  $\{v_1, \dots, v_r\}$ , so  $W(k) \leq M(k) = \sum_{r=1}^{k-1} M_r$ . Let us write

$$S = \sum_{r=1}^{k-1} \frac{\log n - \log r}{n - r}.$$

From the uniformity of the bound in [Lemma 5](#) we have  $\mathbb{E}(\sum_{r=1}^{k-1} M_r) \sim S$ . Now the quantity  $(\log n - \log r)/(n - r)$  decreases with  $r$ . Thus, writing

$$I(x, y) = \int_x^y \frac{\log n - \log r}{n - r} dr,$$

we have  $I(1, k) \leq S \leq I(0, k)$ . For  $0 \leq x \leq 1$  let

$$\text{dilog}(x) = \int_x^1 \frac{-\log x}{1 - x} dx.$$

Writing  $k = \alpha n$  and changing the variable in the integration to be  $x = r/n$ , one can easily check that

$$(4) \quad I(0, k) = \log(\alpha^{-1}) \log((1 - \alpha)^{-1}) + \text{dilog}(1 - \alpha).$$

As this quantity is bounded away from zero, while  $I(0, 1) = O(\log n/n) = o(1)$ , we have  $S \sim I(0, k)$ , giving the right hand side of (4) as an asymptotic upper bound for  $\mathbb{E}(W(k))$ .

Unsurprisingly, the bounds above, extracted from the proof for  $k = o(n)$ , give little or no information when  $k$  is close to  $n$ . As  $\alpha \rightarrow 0$  both bounds are asymptotically  $\alpha \log(\alpha^{-1})$ , the bound in [Theorem 1](#). For  $\alpha$  near 1, the argument above gives no lower bound at all, while the upper bound tends to  $\text{dilog}(0) = \pi^2/6 = 1.644\dots$ , which is larger than  $\zeta(3) = 1.202\dots$ , Frieze's bound for  $\alpha = 1$ . It would be interesting to know how  $W(k)$  decreases from  $\zeta(3)$  as  $k$  decreases from  $n$ . For example, one might expect that  $W(k) = \zeta(3) - o(\beta)$  when  $k = (1 - \beta)n$  with  $\beta \rightarrow 0$ . Again this turns out not to be the case. We can obtain upper and lower bounds for  $W(k)$  in this range using the proof in [\[11\]](#) as a basis, showing that  $W(k)$  decreases from  $\zeta(3)$  linearly with  $\beta$ .

Let  $u(0) = 1$  and for  $x > 0$  let

$$(5) \quad u(x) = \frac{1}{x} \sum_{k=1}^{\infty} \frac{k^{k-2}}{k!} (xe^{-x})^k.$$

**Theorem 7.** *Let  $k = k(n)$  be such that  $\beta = 1 - k/n$  is bounded away from 0 and 1, and let  $\epsilon > 0$  be fixed. Then with probability tending to 1 as  $n \rightarrow \infty$  we have*

$$(6) \quad W(k) \geq \zeta(3) - \log(1/(1 - \beta)) - \epsilon,$$

$$(7) \quad W(k) \geq \int_0^{u^{-1}(\beta)} (u(x) - \beta) dx - \epsilon,$$

and

$$W(k) \leq \zeta(3) - \beta/17.$$

The first lower bound is in fact an easy consequence of Frieze's result: let  $V = \{v_1, \dots, v_k\}$  be the set of Steiner vertices, and  $V^c = \{v_{k+1}, \dots, v_n\}$  be the remaining vertices. For  $k+1 \leq i \leq n$  let  $m_i$  be the minimum weight of an edge from  $v_i$  to  $\{v_1, \dots, v_{i-1}\}$ . Together with a Steiner tree for  $V$ , the corresponding edges form a connected graph on  $[n]$ . Thus we have

$$W(n) \leq W(k) + \sum_{i=k+1}^n m_i.$$

The  $m_i$  are independent exponential random variables with means  $\mathbb{E}(m_i) = 1/(i-1)$ ,  $k+1 \leq i \leq n$ . It follows from standard Martingale methods (or just a suitable form of the central limit theorem) that  $\sum_{i=k+1}^n m_i$  is concentrated about its mean of  $\sum_{i=k+1}^n 1/(i-1) \sim \log(n/k) = \log(1/(1-\beta))$ . Since by Frieze's result we have  $W(n)$  concentrated about  $\zeta(3)$ , the first lower bound (6) follows immediately.

For the remaining bounds we use the methods of [11]. It will be convenient to take the edge weights to be uniformly distributed on  $[0, 1]$ . We ignore the probability 0 event that two edges have the same weight.

The proof in [11] that  $W(n)$  is concentrated about  $\zeta(3)$  uses the greedy algorithm to construct the minimum weight spanning tree in  $G$ . Here we use the same algorithm to construct the spanning subgraph  $H_{k-1}$  of  $G$  which has minimum weight among all subgraphs of  $G$  with rank  $k-1$ , i.e., with  $n-k+1$  components. As any connected graph containing any  $k$  vertices has rank at least  $k-1$ , the weight of  $H_{k-1}$  is a lower bound for the minimum of  $MWST(V)$  over all sets  $V$  of  $k$  vertices of  $G$ , where  $MWST(V)$  is the minimum weight of a Steiner tree for  $V$ . In particular,  $W(k) \geq w(H_{k-1})$ .

Let  $H'_0$  be the empty graph on  $V(G)$ . For  $1 \leq r \leq n-1$  let  $e_r$  be the minimum weight edge joining two different components of  $H'_{r-1}$ , and let  $H'_r = H'_{r-1} \cup \{e_r\}$ . Then it is easy to see that  $H'_{k-1}$  is equal to  $H_{k-1}$ . Indeed, suppose that this is not the case. The two graphs have the same number of edges, so there is some  $e_r$ ,  $r \leq k-1$ , not in  $H_{k-1}$ . If  $e_r$  joins two components of  $H_{k-1}$  then  $H_{k-1} \cup \{e_r\}$  has fewer components than  $H'_{k-1}$ , and thus contains an edge  $f$  joining distinct components of  $H'_{k-1}$ , and hence of  $H'_{r-1}$ . By definition of  $e_r$  we have  $w(f) > w(e_r)$ . On the other hand, if  $e_r$  falls within a component of  $H_{k-1}$  then  $H_{k-1} \cup \{e_r\}$  contains a cycle which, by definition of  $e_r$ , meets two components of  $H'_{r-1}$ . Some edge  $f \neq e_r$  of this cycle must also join two components of  $H'_{r-1}$ . Again by the definition of  $e_r$  we then have  $w(f) > w(e_r)$ . In either case  $H_{k-1} \cup \{e_r\} - \{f\}$  has smaller weight than  $H_{k-1}$ , but the same rank, contradicting the definition of  $H_{k-1}$ .

Now

$$(8) \quad w(H_{k-1}) = \sum_{r=1}^{k-1} w(e_r) = \sum_{r=1}^{k-1} (k-r)(w(e_r) - w(e_{r-1})),$$

where  $w(e_0)$  is to be interpreted as 0. For  $0 < p < 1$  let  $G_p$  be the spanning subgraph of  $G$  formed by all edges with weight at most  $p$ . As we are taking the edge weights to be uniformly distributed on  $[0, 1]$ ,  $G_p$  has the same distribution as a random graph from the standard model  $\mathcal{G}(n, p)$  where vertices are joined independently with probability  $p$ . As  $p$  increases from 0 to 1 edges are added to  $G_p$  one at a time. Each new edge joins two components of  $G_p$  if and only if it is one of the  $e_r$ . Thus if  $w(e_{r-1}) \leq p < w(e_r)$ , so  $e_1, \dots, e_{r-1} \in E(G_p)$ , then  $G_p$  has exactly  $n + 1 - r$  components. Writing  $c(H)$  for the number of components of a graph  $H$ , we can write (8) as

$$w(H_{k-1}) = \int_0^{w(e_{k-1})} (c(G_p) - (n + 1 - k)) dp.$$

Note that the integrand decreases, and becomes zero precisely at  $p = w(e_{k-1})$ . So far the argument is similar to that in [11]. To obtain (7) we shall use the standard result below. In the statement of this result  $N = \binom{n}{2}$  and  $\tilde{G} = (G_t)_{t=0}^N$  is the standard random graph process. Thus  $G_0$  is the empty graph on  $[n]$ , and  $G_{t+1}$  is formed from  $G_t$  by adding one of the  $N - t$  edges of  $G_t^c$  chosen uniformly at random. There should be no confusion between  $G_p$  and  $G_t$  as  $0 < p < 1$ , while  $t$  is an integer. Let  $u(x)$  be defined by (5). The following result is Corollary 14 of chapter VI of [6]; we have changed the notation slightly to avoid clashes.

**Theorem 8.** *The probability that for a fixed  $\gamma$  the graph process  $\tilde{G} = (G_t)_{t=0}^N$  satisfies*

$$|c(G_t) - u(2t/n)n| \leq 2(\log n)n^\gamma$$

for every  $t \leq 4n \log n$  is  $1 - o(n^{2-3\gamma})$ . ■

Using this result it is straightforward to deduce the lower bound (7) on  $W(k)$ .

**Proof of second lower bound in Theorem 7.** Let  $G_t$  be the subgraph of  $G$  formed by the  $t$  edges with smallest weights. Then  $(G_t)_{t=0}^N$  forms a standard random graph process. Applying Theorem 8 with  $\gamma = 3/4$ , noting that  $c(G_t)$  is decreasing and that  $u(8 \log n) = o(1)$ , we see that with probability  $1 - o(1)$  we have

$$c(G_t) = (u(2t/n) + o(1))n,$$

for all  $0 \leq t \leq N$ , where the implicit bound in the  $o(1)$  term depends on  $n$  only and is uniform in  $t$ . Now  $G_p = G_t$  whenever there are exactly  $t$  edges of  $G$  with weight at most  $p$ . Standard bounds on the binomial distribution imply that with probability  $1 - o(1)$  the graph  $G_p$  has  $pN + o(n)$  edges for all  $p \leq n^{-1/2}$  simultaneously. As  $u$  is continuous, it follows that with probability  $1 - o(1)$  we have

$$c(G_p) = (u(pn) + o(1))n$$

for all  $0 \leq p \leq n^{-1/2}$ . Now suppose that  $k = k(n)$  is such that  $\beta = 1 - k/n$  is bounded away from 0 and 1. Then  $w(e_{k-1})$  is the minimal  $p$  such that  $c(G_p) = n + 1 - k \sim \beta n$ . Thus with high probability  $w(e_{k-1}) \sim u^{-1}(\beta)/n$ , which is much smaller than  $n^{-1/2}$  for  $n$  large enough, and

$$w(H_{k-1}) = \int_0^{(1+o(1))u^{-1}(\beta)/n} (u(pn) - \beta + o(1))n dp.$$

Substituting  $p = x/n$  we can write this as

$$\int_0^{u^{-1}(\beta)+o(1)} (u(x) - \beta + o(1))dx = \int_0^{u^{-1}(\beta)} (u(x) - \beta)dx + o(1),$$

completing the proof. ■

One can check that, as  $\beta \rightarrow 0$ , the bound (7) proved above gives  $W(k) \geq \zeta(3) - \beta(\log(\beta^{-1}) + O(1))$ , which is considerably worse than the bound  $W(k) \geq \zeta(3) - \beta - O(\beta^2)$  given by (6). For larger  $\beta$ , the more complicated bound (7) is better. Also, as noted above, the right hand side of (7) is actually a lower bound on a smaller quantity than  $W(k)$ , namely the minimum weight of any rank  $k - 1$  subgraph of  $G$ .

We now turn to the upper bound, that if  $k = (1 - \beta)n$  then with high probability  $W(k) \leq \zeta(3) - c\beta$ , for some constant  $c > 0$ . Our proof is again based on Frieze's result for  $k = n$ . The main idea is to show that the minimum weight spanning tree has many leaves, some of which can then be omitted.

For  $x > 1$  a constant let

$$t(x) = \frac{1}{x} \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} (xe^{-x})^k.$$

Erdős and Rényi [9] proved that almost every graph on  $n$  vertices with  $\lfloor xn/2 \rfloor$  edges has a giant component with  $(1 - t(x) + o(1))n$  vertices. Also, trivial estimates of the first and second moments imply that almost every such graph has  $(e^{-x} + o(1))n$  isolated vertices. These results and [Theorem 8](#) above immediately transfer to the random graph  $G_p$  when  $p = x/n$ , so

almost every such graph has  $(u(x) + o(1))n$  components, a giant component with  $(1 - t(x) + o(1))n$  vertices, and  $(e^{-x} + o(1))n$  isolated vertices. Taking  $x = 2$ , the sums defining  $u(x)$  and  $t(x)$  converge fairly rapidly, and it is easy to verify that  $u(2) < 1/6$  while  $t(2) < 1/4$  and  $e^{-2} > 1/8$ .

Let  $p_0 = 2/n$ . From the remarks above, with probability  $1 - o(1)$  the graph  $G_{p_0}$  has the following three properties: (i)  $G_{p_0}$  has a giant component consisting of at least  $3n/4$  vertices, (ii)  $G_{p_0}$  has at least  $n/8$  isolated vertices, and (iii)  $G_{p_0}$  has at most  $n/6$  components.

Let  $H_1, \dots, H_{n-1}$  be the minimum weight forests in  $G$  constructed by the greedy algorithm as above, so  $H_t = H_{t-1} \cup \{e_t\}$ , and let  $t_0$  be maximal subject to  $H_{t_0} \subset G_{p_0}$ . Then by construction of the graph  $H_{t_0}$ , this graph has the same components as  $G_{p_0}$ , and thus has properties (i)-(iii) above. (Note that here and later when we talk about a component we mean its vertex set.) From now on we take  $t_0$  and  $H_{t_0}$  to be fixed, with  $H_{t_0}$  having the properties above, and consider the random process  $\tilde{H} = (H_t)_{t=t_0+1}^{n-1}$ . Note that at step  $t > t_0$  in this process, going from  $H_{t-1}$  to  $H_t$ , the edge  $e_t$  added is chosen uniformly at random from among all edges joining two components of  $H_{t-1}$ .

Let  $I$  be a set of  $\lceil n/8 \rceil$  isolated vertices of  $H_{t_0}$ . We say that a vertex  $x \in I$  becomes a candidate at step  $t$  if  $e_t$  is the first edge from  $x$  in the process  $\tilde{H}$ , and  $e_t$  joins  $x$  to the giant component of  $H_{t-1}$ . In other words,  $x \in I$  becomes a candidate at step  $t > t_0$  if  $x$  is a leaf of the giant component of  $H_t$  but is isolated in  $H_{t-1}$ . Let us say that step  $t > t_0$  is *critical* if the edge  $e_t$  has at least one endvertex isolated in  $H_{t-1}$ . Suppose that  $H_{t-1}$  is given and has  $m$  isolated vertices. Then there are at most  $mn$  possible edges  $e_t$  that would make step  $t$  critical. Since the giant component of  $H_{t-1}$  has at least  $3n/4$  vertices, at least  $3mn/4$  of these would create a new candidate at step  $t$ . Thus at each critical step a new candidate is created with probability at least  $3/4$ , no matter what has happened so far. The number of isolated vertices goes down by at most two at each critical step, and is otherwise unchanged. Thus there are at least  $n/16$  critical steps, and the total number of candidates created stochastically dominates a binomial  $\text{Bi}(n/16, 3/4)$  distribution. Thus (using the central limit theorem, for example), with probability  $1 - o(1)$  a total of at least  $n/22$  vertices become candidates at some stage.

From now on we condition on the components of  $H_t$  for every  $t$ , or equivalently, on which components become joined at which steps. Let  $\mathcal{A}$  be an event of the form ‘for each  $t > t_0$  the edge  $e_t$  joins components  $C_t, C'_t$  of  $H_{t-1}$ ’. Given  $\mathcal{A}$  the only remaining randomness is in which of the  $|C_t||C'_t|$  edges to chose for  $e_t$  at each stage. Note that  $\mathcal{A}$  determines which vertices in  $I$  become candidates at which stages. We write  $L_0 = \{x_1, \dots, x_m\}$  for the vertices which become candidates at some stage.

Let us say that a candidate  $x$  is *eliminated* at stage  $t$  if  $x$  is a leaf in the giant component of  $H_{t-1}$  but not in that of  $H_t$ , so  $e_t = xy$  for some  $y$  not in the giant component of  $H_{t-1}$ . We say that a candidate  $x$  *survives* if it is not eliminated, and write  $L_1$  for the set of surviving candidates, noting that these vertices are leaves of the spanning tree  $H_{n-1}$ .

Let  $\mathcal{B}$  be the event that certain of the candidates  $\{x_1, \dots, x_{s-1}\}$  are eliminated at certain stages, and that the rest survive. We wish to bound the probability that  $x_s$  is eliminated given  $\mathcal{A}$  and  $\mathcal{B}$ . Given these events we know that  $x_s$  becomes a leaf of the giant component at a certain step  $t$ . We also know that the giant component merges with other components at certain later steps  $t_1, \dots, t_r$ , with  $r \leq c(H_t) \leq c(H_{t_0}) \leq n/6$ . At some of these steps we know that one of  $x_1, \dots, x_{s-1}$  is eliminated, so  $x_s$  cannot be. At each of the remaining  $r' \leq r$  steps  $t_i$  we know that an edge  $e_{t_i}$  is added joining the giant component  $C$  to some (particular) other component, and that none of  $x_1, \dots, x_{s-1}$  is eliminated. This leaves at least  $|C| - (s-1) \geq |C| - |I| + 1 \geq 3n/4 - n/8 = 5n/8$  possibilities for the vertex  $y$  of  $e_{t_i}$  which lies in  $C$ . As these are all equally likely, the probability that  $x_s$  is eliminated at stage  $t_i$  is at most  $8/(5n)$ . As there are  $r' \leq r \leq n/6$  stages when  $x_s$  might be eliminated, we have

$$\mathbb{P}(x_s \text{ survives} \mid \mathcal{A} \cap \mathcal{B}) \geq 1 - \frac{n}{6} \frac{8}{5n} = 11/15.$$

As this holds for all  $s$  and all events  $\mathcal{B}$  of the above form, given  $\mathcal{A}$  the number of surviving candidates stochastically dominates a  $\text{Bi}(m, 11/15)$  distribution. Thus, provided the event  $\mathcal{A}$  is such that  $m \geq n/22$ , we have

$$\mathbb{P}(|L_1| \leq n/31 \mid \mathcal{A}) = o(1).$$

As this holds for all  $\mathcal{A}$  for which  $m \geq n/22$ , and  $m \geq n/22$  holds with probability  $1 - o(1)$ , we have

$$\mathbb{P}(|L_1| \leq n/31) = o(1).$$

We are now ready to prove the upper bound on  $W(k)$ .

**Proof of Theorem 7.** We have already proved the lower bounds. Suppose that  $k = k(n)$  is such that  $\beta = 1 - k/n$  is bounded away from 0 and 1. Let  $H_{n-1}$  be the minimum weight spanning tree in  $G$ , and let  $L_1$  be the set defined above. Then each vertex  $x$  in  $L_1$  is a leaf of  $H_{n-1}$ . Furthermore, for each such  $x$  the unique edge  $xy \in E(H_{n-1})$  has weight at least  $p_0 = 2/n$ . This is because we only considered vertices for inclusion in  $L_1$  if they were isolated in  $G_{p_0}$ , i.e., were adjacent to no edges of weight less than  $p_0$ . Now, as shown above, with probability  $1 - o(1)$  we have  $|L_1| \geq n/31$ . Let us consider the weights on

$G$  and thus  $L_1$  as given, and then the set  $V = \{v_1, \dots, v_k\}$  of Steiner vertices as randomly chosen. Supposing that  $|L_1| \geq n/31$ , as  $|V^c| = \beta n$  and  $\beta$  is bounded away from zero we have  $|V^c \cap L_1| \geq \beta n/32$  with probability  $1 - o(1)$ . We may omit these vertices from  $H_{n-1}$  to obtain a tree  $T$  containing all the vertices of  $V$  with  $w(T) \leq w(H_{n-1}) - p_0|V^c \cap L_1|$ . We know from [11] that with high probability  $w(H_{n-1}) = \zeta(3) + o(1)$ . Thus with probability  $1 - o(1)$  we have

$$W(k) \leq w(T) \leq \zeta(3) + o(1) - \frac{2}{n} \frac{\beta n}{32} < \zeta(3) - \beta/17,$$

provided  $n$  is sufficiently large. This completes the proof. ■

### 5. Concluding remarks

In this paper we considered the weight of a minimal Steiner tree in a complete graph on  $n$  vertices with edge weights chosen randomly and independently from some distribution  $X$  satisfying  $\mathbb{P}(X \leq x) = x + o(x)$  as  $x \rightarrow 0$ . We showed that  $W(k)$ , the minimal weight of a tree that contains a given set of  $k = o(n)$  vertices is with high probability  $(1 + o(1))(k-1)(\log n - \log k)/n$ .

In conclusion, let us draw attention to an interesting problem we could not solve. It is very likely that there is a function  $c(\alpha)$  defined on  $[0, 1]$  with  $c(0) = 0$  and  $c(1) = \zeta(3)$  such that if  $k/n \rightarrow \alpha$  then  $\mathbb{E}(W(k)) \rightarrow c(\alpha)$ . Assuming that this function exists, we have found its asymptotic behaviour as  $\alpha \rightarrow 0$ , and given upper and lower bounds for all  $\alpha$ . We do not believe that any of the latter bounds is close to being best possible, and we do not have a conjecture, or even a guess, for the form of  $c(\alpha)$ . The question is to determine  $c(\alpha)$ .

**Acknowledgments.** The second and fourth authors gratefully acknowledge many stimulating discussions with A. Frieze, L. Thoma and D. Coppersmith.

### References

- [1] D. ALDOUS: The  $\zeta(2)$  limit in the random assignment problem, *Random Structures and Algorithms* **18** (2001), 381–418.
- [2] S. ARORA: Polynomial time approximation schemes for Euclidean TSP and other geometric problems, *Journal of the ACM* **45** (1998), 753–782.

- [3] J. BEARDWOOD, J. HALTON and J. HAMMERSLEY: The shortest path through many points, *Proc. Camb. Philos. Society* **55** (1959), 299–327.
- [4] M. BERN and P. PLASSMANN: The Steiner tree problem with edge lengths 1 and 2, *Information Processing Letters* **32** (1989), 171–176.
- [5] A. BEVERIDGE, A. FRIEZE and C. MCDIARMID: Random minimum length spanning trees in regular graphs, *Combinatorica* **18** (1998), 311–333.
- [6] B. BOLLOBÁS: *Random Graphs*, 2<sup>nd</sup> ed., Cambridge University Press, New York, 2001.
- [7] A. CLEMENTI and L. TREVISAN: Improved non-approximability results for minimum vertex cover with density constraints, *Theoret. Comput. Sci.* **225** (1999), 113–128.
- [8] R. DAVIS and A. PRIEDITIS: The expected length of a shortest path, *Information Processing Letters* **46** (1993), 135–141.
- [9] P. ERDŐS and A. RÉNYI: On the evolution of random graphs, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 17–61.
- [10] C. M. FORTUIN, P. W. KASTELEYN and J. GINIBRE: Correlation inequalities on some partially ordered sets, *Comm. Math. Phys.* **22** (1971), 89–103.
- [11] A. FRIEZE: On the value of a random minimum spanning tree problem, *Discrete Appl. Math.* **10** (1985), 47–56.
- [12] T. E. HARRIS: A lower bound for the critical probability in a certain percolation process, *Proc. Cambridge Philos. Soc.* **56** (1960), 13–20.
- [13] S. JANSON: One, two and three times  $\log n/n$  for paths in a complete graph with random weights, *Combinatorics Probability and Computing* **8** (1999), 347–361.
- [14] D. J. KLEITMAN: Families of non-disjoint subsets, *J. Combinatorial Theory* **1** (1966), 153–155.
- [15] L. KUČERA, A. MARCHETTI-SPACCAMELA, M. PROTASI and M. TALAMO: Near optimal algorithms for finding minimum Steiner trees on random graphs, *Mathematical foundations of computer science*, Lecture Notes in Comput. Sci. 233, Springer, Berlin, 1986, 501–511.
- [16] D. REIMER: Proof of the van den Berg-Kesten conjecture, *Combinatorics Probability and Computing* **9** (2000), 27–32.
- [17] G. ROBINS and A. ZELIKOVSKY: Improved Steiner tree approximation in graphs, *Proc. 11th ACM-SIAM Symposium on Discrete Algorithms* (2000), 770–779.
- [18] J. VAN DEN BERG and H. KESTEN: Inequalities with applications to percolation and reliability, *J. Appl. Probab.* **22** (1985), 556–569.

Béla Bollobás

*Department of Mathematical Sciences*

*University of Memphis*

*Memphis TN 38152*

*USA*

[bollobas@memphis.edu](mailto:bollobas@memphis.edu)

*and*

*Trinity College*

*Cambridge CB2 1TQ*

*UK*

[b.bollobas@dpms.cam.ac.uk](mailto:b.bollobas@dpms.cam.ac.uk)

David Gamarnik

*IBM*

*T.J. Watson Research Center*

*Yorktown Heights*

*NY 10598*

*USA*

[gamarnik@watson.ibm.com](mailto:gamarnik@watson.ibm.com)

Oliver Riordan

*Trinity College*

*Cambridge CB2 1TQ*

*UK*

[o.riordan@dpms.cam.ac.uk](mailto:o.riordan@dpms.cam.ac.uk)

Benny Sudakov

*Department of Mathematics*

*Princeton University*

*Princeton*

*NJ 08540*

*USA*

*and*

*Institute for Advanced Study*

*Princeton*

*NJ 08540*

*USA*

[bsudakov@math.princeton.edu](mailto:bsudakov@math.princeton.edu)