The expected value of random minimal length spanning tree of a complete graph

David Gamarnik *

Abstract

We consider the number $c(n,m)$ of connected labeled graphs on n nodes and m edges and the intimately related object, the expected length of the minimal spanning tree of a complete graphs with random edge lengths. We use a very simple recursive procedure for computing the values of $c(n,m)$ for computing the expected length of the minimal spanning tree exactly, under the uniform and the exponential distributions. Our computations are recursive, scale very well with the size of the problem, and we provide the values of the expected minimal length spanning trees for complete graphs K_n with sizes $n \leq 45$, extending recent results of Steele [Ste02], and Fill and Steele [FS04]. The main proof technique is based on introducing an artificial root to a graph and subsequently using a very simple inductive argument.

1 Introduction

Given positive integers n, m let $c(n,m)$ denote the number of connected labeled graphs on n nodes with m edges. The values of $c(n,m)$ have been a subject of a considerable interest in the area of enumerative combinatorics. Various asymptotical results are available for computing $c(n,m)$, see for example Bender, Canfield and McKay [BCM90], Luczak [Luc90], Coja-Oghlan et al. [COMS04]. The methods for computing $c(n,m)$ recursively can be found in Harary and Palmer [HP73]. In this paper we investigate the connection between $c(n,m)$ and the expected length $\mathbb{E}[\boldsymbol{T}_n]$ of the minimal spanning tree of a complete graph K_n with edges equipped with random lengths, generated according to either the uniform or the exponential distribution.

It was established by Frieze [Fri85] that value of $\mathbb{E}[\boldsymbol{T}_n]$ converges to $\zeta(3) = \sum_{k\geq 1} 1/k^3$ as $n \to \infty$, see also Steele [Ste87]. Two recent papers by Steele [Ste02], and Fill and Steele [FS04] have addressed the issue of computing the values $\mathbb{E}[\boldsymbol{T}_n]$ exactly. They used a connection between $\mathbb{E}[\boldsymbol{T}_n]$ and the Tutte polynomial of a complete graph in which the lengths are generated randomly and independently according to the uniform distribution with support $[0,1]$ (denoted $U(0,1)$ henceforth). The values $\mathbb{E}[\boldsymbol{T}_n]$ were computed exactly for $n = 2, 3, \ldots, 9$. Also a somewhat different recursive method was proposed in [FS04] which also led to exact computation of $\mathbb{E}[\boldsymbol{T}_n]$. Unfortunately, both methods do not scale well as n increases and the overall computational effort is an exponential function of n.

In this paper we start with a very simple method for computing the values $c(n,m)$ exactly for any n, m. The computation is recursive but grows only as a polynomial function of n and m. A similar computation can be found in [HP73] and some variations using generating functions are very well known in the area of enumerative combinatorics. Then we propose a simple exact formula for the values of $\mathbb{E}[\boldsymbol{T}_n]$ expressed in terms of $c(n,m)$. The formula is derived both for the uniform $U(0,1)$ and the exponential distribution with parameter 1 (denoted henceforth by Exp(1)). The computation of $\mathbb{E}[\boldsymbol{T}_n]$ arising from these formulas scale very well (in fact polynomially) as a function of n and we compute the exact values of $\mathbb{E}[\boldsymbol{T}_n]$ for $n = 2, 3, \ldots, 45$. It was observed in [Ste02] and [FS04] that the values $\mathbb{E}[\boldsymbol{T}_n]$ are monotonically increasing as a function of n for the derived cases $n = 2, 3, \ldots, 9$, and it was conjectured that the monotonicity remains valid for all n. Interestingly, our computation show that, while for the case of the uniform distribution the values remain to be monotonically increasing, the values corresponding to the exponential distribution increase for $n = 2, 3, \ldots, 7$ and decrease for $n = 8, \ldots, 44$, and we conjecture that the decrease continues for all the values of $n \geq 8$ (see Table 1 and Figure 5). Our proof method is based on a simple idea of introducing an artificial root to the graph and using an inductive argument.

2 Enumeration of connected graphs

Recall that $c(n,m)$ is the total number of labeled connected graphs on n nodes with m edges. The following proposition is used to obtain a simple recursion on $c(n,m)$.

*IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, gamarnik@watson.ibm.com.

PROPOSITION 2.1. *For every* $m \in [n-1, n(n-1)/2]$

$$\binom{\frac{n(n-1)}{2}}{m} = \sum_{i=1}^{n} \sum_{l=0}^{\frac{n(n-1)}{2}} c(i,l) \binom{n-1}{i-1} \binom{\frac{(n-i)(n-i-1)}{2}}{m-l}, \tag{2.1}$$

where for every $m \notin [n-1, n(n-1)/2]$, $c(n,m) = 0$.

Proof. Fix a root $r, r = 1, 2, \ldots, n$ and consider any (not-necessarily connected) graph on n nodes with m edges with the given root r. The total number of such labeled graphs with the given root r is $\binom{\frac{n(n-1)}{2}}{m}$. On the other hand, let $\mathcal{C}(r)$ denote the component in this graph containing r and let i, l denote the number of nodes and edges in $\mathcal{C}(r)$. For a fixed such a component $\mathcal{C}(r)$ there are altogether $\binom{\frac{(n-i)(n-i-1)}{2}}{m-l}$ labeled graphs spanned by the remaining $n-i$ nodes using the remaining $m-l$ edges. For a fixed choice of $i-1$ nodes there exists by definition $c(i,l)$ labeled connected graphs spanning these nodes and the node r. Since there are $\binom{n-1}{i-1}$ choices for these nodes we obtain the formula after cancelling n on both sides. □

Proposition 2.1 provides the following simple recursive formula for $c(n,m)$ the derivation of which is immediate.

$$c(n,m) = \binom{\frac{n(n-1)}{2}}{m} - \sum_{i=1}^{n} \sum_{l \le m-1} c(i,l) \binom{n-1}{i-1} \binom{\frac{(n-i)(n-i-1)}{2}}{m-l} - \sum_{i \le n-1} c(i,m) \binom{n-1}{i-1}, \tag{2.2}$$

where the first double sum corresponds to the number of graphs with n nodes and m edges such that the component containing r has fewer than m edges, and the second sum corresponds to the number of such graphs where the component containing r has exactly m edges but fewer than n nodes. It is easy to see that the computation time required to compute $c(n,m)$ grows as a polynomial function in n (note $m \le n(n-1)/2$). The computation time growth as a function of n is in fact quite moderate and we can compute the values of $c(n,m)$ for n up to 45 in a matter of minutes.

3 Expected minimal length spanning tree

Consider a complete graph K_n on n nodes with edges having non-negative lengths $w_{i,j}, 1 \le i < j \le n$. Let w^* be any value larger than $\max_{i,j} w_{i,j}$. Denote by $T(K_n)$ the minimal total length of a spanning tree of K_n. When the lengths $w_{ij} = \boldsymbol{W}_{i,j}$ are generated at random independently according to some probability distribution, we denote $T(K_n)$ by $\boldsymbol{T}_n$. Our focus is computing $\mathbb{E}[\boldsymbol{T}_n]$ when $\boldsymbol{W}_{i,j}$ are distributed either according to the uniform distribution over $[0,1]$ (denoted $U(0,1)$) or according to the exponential distribution with parameter 1 (denoted Exp(1)).

Given an arbitrary (non-random) graph G on n-nodes, let $\kappa(G)$ denote the number of connected components of G. When the edges of G are equipped with lengths $w_{i,j}$, $1 \le i < j \le n$ and $x > 0$, let $G(x)$ denote the subgraph obtained from G by including only the edges with length $w_{i,j} \le x$. The following formula which was derived first in Avram and Bertsimas [AB92] relates $T(G)$ to the number of the connected components $\kappa(G(x))$ of the subgraph $G(x)$. Originally the formula was developed for the case $w_{i,j} \le 1$. Its extension applicable to arbitrary lengths is obtained immediately by rescaling $w_{i,j}$ to $w_{i,j}/w^*$.

PROPOSITION 3.1. *For every connected graph* G

$$T(G) = \int_0^{w^*} \kappa(G(x))dx - w^*. \tag{3.3}$$

We now state and prove our main result.

THEOREM 3.1. *For every* $n \ge 2$

$$\mathbb{E}[\boldsymbol{T}_n] = -1 + \sum_{k=1}^{n} \sum_{m=k-1}^{\frac{k(k-1)}{2}} \frac{\binom{n}{k} m! c(k,m)}{(\frac{k(k-1)}{2} + k(n-k) + 1 - m) \cdots (\frac{k(k-1)}{2} + k(n-k) + 1)}, \tag{3.4}$$

when the edge length distribution is $U(0,1)$ *and*

$$\mathbb{E}[\boldsymbol{T}_n] = -\sum_{i=1}^{\frac{n(n-1)}{2}} \frac{1}{i} + \sum_{k,m \in F(n)} \frac{\binom{n}{k} m! c(k,m)}{(\frac{k(k-1)}{2} + k(n-k) - m) \cdots (\frac{k(k-1)}{2} + k(n-k))}, \tag{3.5, 3.6}$$

when the edge length distribution is Exp(1)*, where*

$$F(n) = \Big\{(k,m) : k \le n, m \le k(k-1)/2, \; k + m < n + n(n-1)/2\Big\}. \tag{3.7}$$

In light of the derivation in Section 2 which allows us to compute $c(k,m)$ in time polynomial in k and m, the formula above gives us a polynomial in n algorithm for computing expected minimal spanning tree on a complete n-graph. The main trick which allows us to derive the formula above is again creating an artificial root in the graph and relating the expected number of components to the number $c(n,m)$ of connected graphs on n nodes with m edges.

Proof. For any graph G we denote by $n(G)$ and $e(G)$ the cardinality of the node set and the edge set of G, respectively. The following formula is immediate for every graph G.

$$\kappa(G) = \sum_{1\le i\le n} \frac{1}{n(\mathcal{C}(i))}, \tag{3.8}$$

where $\mathcal{C}(i)$ is the component containing i. This formula, while trivial, provides us with a convenient representation for $\kappa(G)$:

$$\kappa(G) = \sum_{1\le i\le n} \sum_{1\le k\le n} \sum_{m=k-1}^{\frac{k(k-1)}{2}} \frac{1\{n(\mathcal{C}(i))=k, e(\mathcal{C}(i))=m\}}{k}, \tag{3.9}$$

When $G = K_n$ and the lengths are random, using symmetry we obtain

$$\mathbb{E}[\kappa(K_n(x))] = n \sum_{1\le k\le n} \sum_{m=k-1}^{\frac{k(k-1)}{2}} \frac{\Pr(n(\mathcal{C}(1))=k, e(\mathcal{C}(1))=m)}{k}. \tag{3.10}$$

Now we focus on the case of the uniform distribution. The proof for the case of exponential distribution is delayed till the next paragraph. Consider any connected subgraph $G_1 \subset K_n$ containing the node 1 which consists of k nodes and m edges. Since the length probability distribution of the edges is $U(0,1)$ and $0 \le x \le 1$, then each edge (i,j) of G belongs to $G(x)$ with probability x and does not with probability $1-x$ independently for all edges. Then, under $U(0,1)$ distribution, the probability that the random graph $G(x)$ is such that the component $\mathcal{C}(1)$ containing 1 is exactly G_1 is equal to $x^m(1-x)^{\frac{k(k-1)}{2}-m+k(n-k)}$ since we must have that exactly m edges of G_1 to have length at most x, and the remaining edges $\frac{k(k-1)}{2} - m$ between pairs of nodes in G_1 as well as $k(n-k)$ between nodes of G_1 and its complement must all have length bigger than x. There are $\binom{n-1}{k-1}$ choices for the remaining $k-1$ nodes to generate the component $\mathcal{C}(1)$ and there are $c(k,m)$ connected graphs on a given collection of k nodes with m edges. We obtain

$$\Pr\Big(n(\mathcal{C}(1))=k, e(\mathcal{C}(1))=m\Big) = \binom{n-1}{k-1} c(k,m) x^m (1-x)^{\frac{k(k-1)}{2}-m+k(n-k)}. \tag{3.11}$$

We use the formula

$$\int_0^1 x^i (1-x)^j dx = \frac{i!}{(j+1)\cdots(j+i+1)},$$

for every $i,j \ge 0$. Note that $c(k,m) = 0$ unless $k-1 \le m \le \frac{k(k-1)}{2}$. Then applying (3.10) we obtain

$$\int_0^1 \mathbb{E}[\kappa(G(x))]dx = n \sum_{1\le k\le n} \sum_{m=k-1}^{\frac{k(k-1)}{2}} \frac{\binom{n-1}{k-1} c(k,m) m!}{k(\frac{k(k-1)}{2} - m + k(n-k) + 1)\cdots(\frac{k(k-1)}{2} + k(n-k) + 1)}.$$

Applying (3.3) and using $w^* = 1$ we obtain (3.4).

When the edges lengths are exponentially distributed use formula (3.3) to observe that

$$\mathbb{E}[\boldsymbol{T}_n] = \mathbb{E}[\int_0^\infty \kappa(G(x)) 1\{\max w_{i,j} > x\} dx - \max_{i,j} w_{i,j}].$$

Indeed when $x > \max_{i,j} w_{ij}$, we have $\kappa(G(x)) = \kappa(G) = \kappa(K_n) = 1$ (graph is complete). Therefore for every $w^* > \max w_{i,j}$

$$\begin{aligned}
&\int_0^\infty \kappa(G(x)) 1\{\max w_{i,j} > x\} dx - \max_{i,j} w_{i,j} \\
&\quad = \int_0^{\max w_{i,j}} \kappa(G(x)) dx - \max_{i,j} w_{i,j} \\
&\quad = \int_0^{w^*} \kappa(G(x)) dx - (w^* - \max_{i,j} w_{i,j}) - \max_{i,j} w_{i,j} \\
&\quad = \int_0^{w^*} \kappa(G(x)) dx - w^*
\end{aligned}$$

Using (3.9) and interchanging the order of integration we obtain

$$\begin{aligned}
\mathbb{E}[\boldsymbol{T}_n] &= \sum_{1\le i\le n} \sum_{1\le k\le n} \sum_{m=k-1}^{\frac{k(k-1)}{2}} \\
&\frac{1}{k} \int_0^\infty \mathbb{E}[1\{n(\mathcal{C}(i))=k, e(\mathcal{C}(i))=m, \max w_{i,j} > x\}] dx \\
&- \mathbb{E}[\max w_{i,j}] \\
&= \sum_{k,m\in F(n)} \frac{n}{k} \int_0^\infty \Pr(n(\mathcal{C}(i))=k, e(\mathcal{C}(i))=m) dx \\
&- \mathbb{E}[\max w_{i,j}],
\end{aligned}$$

where in the second equality we use the fact $\max w_{i,j} < x$ iff $G(x)$ is a complete graph, implying $k = n, m = n(n-1)/2$. Using the argument similar to the one leading to (3.11) we obtain that for the Exp(1) distribution

$$\begin{aligned}(3.12)\qquad &\Pr(n(\mathcal{C}(1)) = k, e(\mathcal{C}(1)) = m) = \\ &\binom{n-1}{k-1} c(k,m)(1-\exp(-x))^m \\ &\exp\Big(-x(\frac{k(k-1)}{2} - m + k(n-k))\Big),\end{aligned}$$

We use the following formula

$$(3.13)\quad \int_0^\infty (1-\exp(-x))^i \exp(-xj)dx = \frac{i!}{j\cdots(j+i)},$$

and note $\mathbb{E}[\max w_{i,j}] = \sum_{1\le i\le n(n-1)/2} \frac{1}{i}$ to obtain

$$\begin{aligned}\mathbb{E}[\boldsymbol{T}_n] = &\sum_{k,m\in F(n)} \binom{n-1}{k-1}\frac{n}{k} \\ &\frac{c(k,m)m!}{(\frac{k(k-1)}{2} - m + k(n-k))\cdots(\frac{k(k-1)}{2} + k(n-k))} \\ &- \sum_{1\le i\le n(n-1)/2} \frac{1}{i},\end{aligned}$$

which is exactly (3.5). □

4 Computations

We have computed the values of $\mathbb{E}[\boldsymbol{T}_n]$ for the case of $U(0,1)$ and Exp(1) distributions for $n = 2, 3, \ldots, 45$. The answers are presented in Table 1, where in addition for every row n we present the difference $\mathbb{E}[\boldsymbol{T}_n] - \mathbb{E}[\boldsymbol{T}_{n-1}]$. We have also plotted the values on Figure 5, where the horizontal line corresponds to $\zeta(3) \approx 1.202$. For the case of uniform distribution our computations show that the monotonicity conjectured in [Ste02] and [FS04] and confirmed for $n = 2, 3, \ldots, 8$, indeed holds for all $n \le 44$. Yet for the case of the exponential distribution we see that $\mathbb{E}[\boldsymbol{T}_n]$ is growing for $n = 2, 3, \ldots, 7$ but starting with $n = 8$ becomes monotonically decreasing. It is natural then to extend the conjecture stated in [Ste02] and [FS04] as follows.

Conjecture 4.1. *Under the* $U(0,1)$ *distribution* $\mathbb{E}[\boldsymbol{T}_n] < \mathbb{E}[\boldsymbol{T}_{n+1}]$ *for all* $n = 2, 3, \ldots,$. *Under* Exp(1) *distribution* $\mathbb{E}[\boldsymbol{T}_n] > \mathbb{E}[\boldsymbol{T}_{n+1}]$ *for all* $n \ge 8$.

5 Acknowledgements

The author wishes to thank Jim Fill, Michael Steele and Ira Gessel for useful and informative conversations.

n	$U(0,1)$	Difference	Exp (1)	Difference
2	0.5000		1.0000	
3	0.7500	0.2500	1.1667	0.1667
4	0.8857	0.1357	1.2167	0.0500
5	0.9665	0.0807	1.2353	0.0187
6	1.0183	0.0519	1.2427	0.0074
7	1.0537	0.0354	1.2454	0.0027
8	1.0791	0.0253	1.2460	0.0005
9	1.0979	0.0188	1.2455	-0.0005
10	1.1124	0.0144	1.2445	-0.0010
11	1.1237	0.0114	1.2432	-0.0013
12	1.1328	0.0091	1.2418	-0.0013
13	1.1403	0.0075	1.2405	-0.0014
14	1.1465	0.0062	1.2391	-0.0013
15	1.1517	0.0052	1.2379	-0.0013
16	1.1561	0.0044	1.2366	-0.0012
17	1.1599	0.0038	1.2355	-0.0012
18	1.1632	0.0033	1.2344	-0.0011
19	1.1661	0.0029	1.2333	-0.0010
20	1.1686	0.0025	1.2323	-0.0010
21	1.1708	0.0022	1.2314	-0.0009
22	1.1728	0.0020	1.2305	-0.0009
23	1.1746	0.0018	1.2297	-0.0008
24	1.1762	0.0016	1.2289	-0.0008
25	1.1777	0.0015	1.2282	-0.0007
26	1.1790	0.0013	1.2275	-0.0007
27	1.1802	0.0012	1.2268	-0.0007
28	1.1813	0.0011	1.2262	-0.0006
29	1.1823	0.0010	1.2256	-0.0006
30	1.1832	0.0009	1.2250	-0.0006
31	1.1841	0.0009	1.2245	-0.0005
32	1.1849	0.0008	1.2240	-0.0005
33	1.1856	0.0007	1.2235	-0.0005
34	1.1863	0.0007	1.2230	-0.0005
35	1.1869	0.0006	1.2225	-0.0004
36	1.1875	0.0006	1.2221	-0.0004
37	1.1881	0.0006	1.2217	-0.0004
38	1.1886	0.0005	1.2213	-0.0004
39	1.1891	0.0005	1.2209	-0.0004
40	1.1895	0.0005	1.2206	-0.0004
41	1.1899	0.0004	1.2202	-0.0003
42	1.1903	0.0004	1.2199	-0.0003
43	1.1907	0.0004	1.2196	-0.0003
44	1.1911	0.0004	1.2192	-0.0003
45	1.1914	0.0003	1.2189	-0.0003

Table 1: Expected lengths and expected difference of the minimal spanning tree lengths under $U(0,1)$ (left two columns) and Exp (right two columns) in $K_n, n = 2, 3, \ldots, 45$.

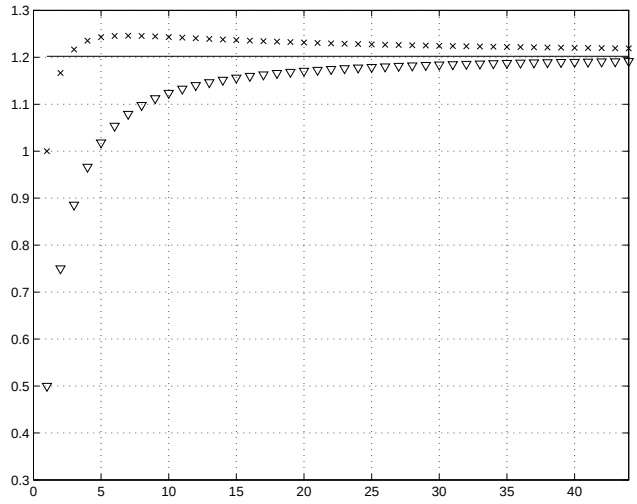

Figure 1: Expected minimal length spanning tree under $\mathrm{Exp}(0,1)$ (top) and $U(0,1)$ (bottom) distributions in $K_n, n = 2, 3, \ldots, 45$.

References

[AB92] F. Avram and D. Bertsimas, *The minimum spanning tree constant in geometric probability and under the independent model: a unified approach*, The Annals of Applied Probability **2** (1992), 113–130.

[BCM90] E. Bender, E. Canfield, and B. McKay, *The asymptotic number of labeled connected graphs with a given number of vertices and edges*, Random structures and algorithms **1** (1990), no. 2, 127–169.

[COMS04] A. Coja-Oghlan, C. Moore, and V. Sanwalani, *Counting connected graphs and hypergraphs via the probabilistic method*, Proceedings of RANDOM, 2004.

[Fri85] A. Frieze, *On the value of a random minimum spanning tree problem*, Discrete Appl. Math. **10** (1985), 47–56.

[FS04] J. Fill and M. Steele, *Exact expectations of minimal spanning trees for graphs with random edge weights.*

[GS96] I. Gessel and B. Sagan, *The Tutte polynomial of a graph, depth-first search, and simplicial complex partitions*, Electronic J. Combinatorics, Foata Festschrift **3** (1996), no. 2, R9.

[HP73] F. Harary and E. M. Palmer, *Graphical enumeration*, Academic Press, 1973.

[Łuc90] T. Łuczak, *On the number of sparse connected graphs*, Random Structures and Algorithms **1** (1990), no. 2, 171–174.

[Ste87] J. M. Steele, *On Frieze's $\zeta(3)$ limit for lengths of minimal spanning trees*, Discrete Applied Mathematics **18** (1987), 99–103.

[Ste02] ———, *Minimal spanning trees for graphs with random edge lenghts*, Mathematics and Computer Science II. Algorithms, Trees, Combinatorics and Probabilities,B. Chauvin, Ph. Flajolet, D. Gardy, and A. Mokkadem (eds.), Birkhäuser, Boston, 2002, pp. 223–245.