DIMACS Technical Report 2009-20
September 2009

Statistics for a Random Network Design Problem

by

Fred J. Rispoli[1]
Fred_Rispoli@dowling.edu

Steven Cosares
cosares@aol.com

DIMACS is a collaborative project of Rutgers University, Princeton University, AT&T Labs-Research, Bell LABS, NEC Laboratories and Telcordia Technologies, as well as affiliate members Avaya Labs, HP Labs, IBM Research, Microsoft Research, Stevens Institute of Technology, Georgia Institute of Technology and Rensselaer Polytechnic Institute. DIMACS was founded as an NSF Science and Technology Center.

[1]Department of Mathematics and Computer Science, Dowling College, Oakdale, NY 11769

ABSTRACT

We investigate a random network design problem specified by a complete graph with n nodes whose edges have associated fixed costs that are independent random variables, and variable costs associated that are also independent random variables. The objective is to find a spanning tree whose total fixed cost plus total variable cost is minimum, where the total variable cost is the sum of variable costs along all paths from a source node to every other node. Here we examine the distributions of total fixed cost and total variable cost obtained from random tree generation, and compare the expectations in solving the different components of a random network design problem using simulation.

1 Introduction

Solving network design problems has become a critical optimization problem in today's highly connected world. Finding good networks typically requires a tradeoff between the cost of establishing links and nodes in a network, and usage costs incurred over some time horizon. Design problems are often modeled using the complete graph with n vertices, denoted by K_n, to represent the collection of all possible nodes and links. The edges $\{e_{i,j} = \{i, j\} : i \neq j\}$ of K_n are used to represent potential links, and associated with every edge is both a fixed cost $f_{i,j}$ and a variable cost $v_{i,j}$. Among the most common type of design problems is to find a spanning tree T that minimizes the sum of total fixed costs (TFC) based on the edges used in T, plus total variable cost (TVC) determined by the sum of variable costs along all $n-1$ paths from a root node to every other node in T. When variable costs are negligible, the problem reduces to a minimum spanning tree problem. When fixed costs are negligible, the problem reduces to a shortest paths problem. When both costs are significant, the problem requires a compromise between these two trees, which may be very hard to find. Problems with just a single source are known to be NP-hard. Now suppose that the edges in K_n, have associated fixed costs that are independent random variables, and variable costs that are also independent random variables. Then our design problem is called the *random network design problem*.

Minimum spanning tree and shortest path problems problems with random edge weights have been studied by many (e.g. see [2,3,4,5,6,7]). Randomized approaches have been used to find approximate solutions when deterministic methods are impractical, and are often used when test problems are needed to evaluate new algorithms. But very few studies have been made on random network design problems that require a compromise solution between TFC and TVC. Here, we examine the distribution of total cost over all possible spanning trees of K_n, as well as the two components TFC and TVC. By Cayley's Theorem, we know that the population consists of the costs associated with n^{n-2} spanning trees of K_n. If one were to consider just the distribution of total fixed costs, it would be natural to ask if a random sample of spanning trees yields a mean total fixed cost that is statistically equal to a mean total fixed cost of random sample of subsets of $n-1$ edges? Motivated by the Central Limit Theorem, one can also ask: is the distribution of total fixed costs obtained from spanning trees approximately normal? Futhermore, what is the relative position of the total fixed cost of a minimum spanning tree (MST)? Similar questions can be investigated for the total variable cost structure and the shortest paths tree (SPT). Answers can be used to compare the two problems and determine which component of the problem we should expect to obtain a better solution using random trees. This will also shed some light on how well one should expect simulation to solve the network design problem.

One approach to solving a network design problem is to model the problem as an integer program and use discrete optimization software. But a network design problem with as few as 15 nodes can take a long time to solve (e.g., see [1]). An alternative approach is to consider generating random trees to try to find a good approximate solution. The questions posed above were initially investigated using random trees. After examining many different simulations we shall elaborate on several cases that we beleive to be representaive of our findings. Consider a random network design problem whose fixed costs are uniformly and

independently distributed in the interval $[10, 150]$, from here on denoted by $u[10, 150]$ (or $u[a, b]$ in general), and variable costs from $u[5, 50]$. Using a well known proof of Cayley's Theorem to represent spanning trees as $(n-1)$-tuples included in [8], we generated 1,000 random trees. The TFC vs. TVC were then graphed using a scatter plot and is given in Figure 1. Descriptive statistics for the simulation results are also given in Table 1. Included in the scatterplot is the MST whose costs are represented by the coordinates of the point on the extreme left, and the SPT, whose costs are given by the coordinates of lowest point in the scatterplot. Prim's algorithm was used to obtain the MST which has a TFC of 330, and a TVC of 1,587. The SPT included in the scatterplot of Figure 1 was obtained using Dijkstra's algorithm and has a TVC of 247, and a TFC of 1,322.

	Mean	**Median**	**St.Dev.**	**Min**	**Max**	**MST**	**SPT**	**Edge Weights**
TFC	1,106.4	1,102	130.7	710	1,479	330	1,322	$u[10, 150]$
TVC	1,408.1	1,353	397.7	556	2,786	1587	247	$u[5, 50]$
TC	2,514.5	2,468	411.2	1,635	3,965	1,917	1,569	NA

Table 1. Descriptive Statistics for 1,000 Random Trees, the MST, and the SPT

When considering a minimum spanning tree or shortest path problems with random edge weights, most studies assume that the weights are from $u[0, 1]$. Since in a network design problem with both fixed and variable costs, fixed costs are generally larger than variable costs, we looked at cases that reflect this. Although the case where $f_{i,j} \in u[0, 1]$ and $v_{i,j} \in u[0, 1]$ was examined. The resulting scatterplot of TFC vs. TVC was nearly identical to the scatterplot given in Figure 1 with the man difference being that the total fixed costs were shifted down.

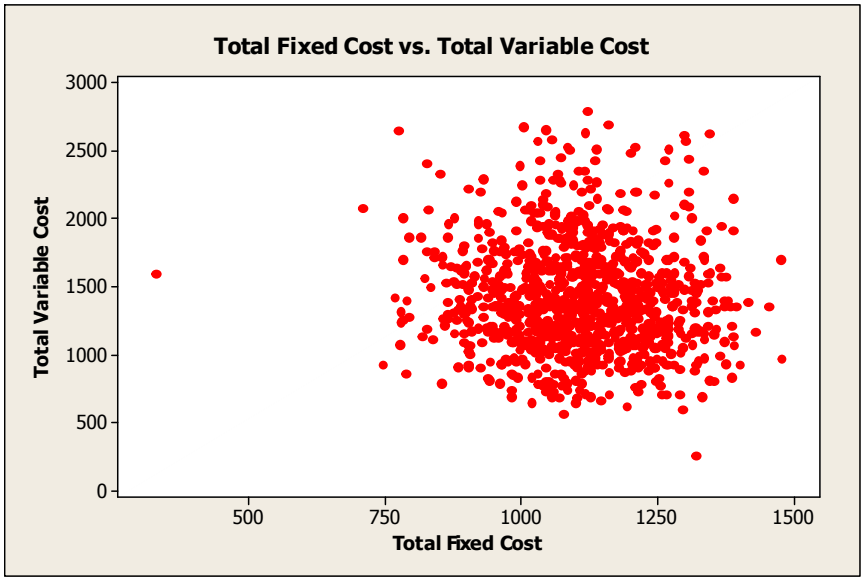

Figure 1. A Scatterplot of TFC vs. TVC for 1,000 Random Trees with $f_{i,j} \in u[10, 150]$ and $v_{i,j} \in u[5, 50]$

Results from the simulations were used to guide subsequent investigation which is organized as follows. In Section 2 we discuss the distribution of TFC and explain why the MST is an extreme outlier. In Section 3 we discuss the distribution of TVC. In the last section we consider the total cost distribution and how our findings relate to the efficiency of random tree generation. Random tree generation was performed using Microsoft's Visual Basic for Applications with the Excel interface. Statistical computations and charts were all obtained using Minitab Release 15.

2 The distribution of total fixed costs

Consider the distribution of the total fixed cost for subsets of edges of size 14 for the random network design problem described above, i.e., $n = 15$ with $f_{i,j} \in u[10, 150]$. It follows from the Central Limit Theorem that this distribution is approximately normal. But what happens when we consider the TFC of spanning trees? In particular, does the tree structure requiring that a subset of edges form a connected subgraph of K_n without any cycles affect the sample mean or the type of distribution? A snapshot of the distribution of TFC obtained from the 1,000 random spanning trees is given by the histogram in Figure 2. To test the hypothesis that there is no significant difference between the two distributions we generated 1,000 random subsets of 14 edges and found the total fixed cost for each subset. This gave a sample mean of total fixed cost $\overline{x} = 1,108$ with a sample standard deviation of $s = 148$. Next, we used a two-tailed z-test for 2 samples and tested the hypothesis that the difference in mean total fixed cost is zero. The result was a p-value of 0.306 implying that there is no statistically significant difference of the means at the 95% confidence level.

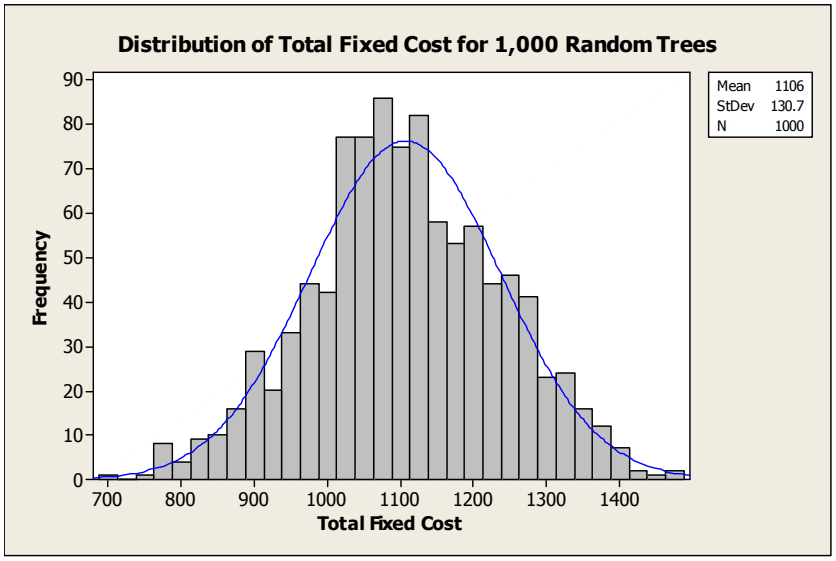

Figure 2. The Distribution of TFC for 1,000 Random Trees with $\mathbf{f}_{i,j} \in u[10, 150]$

To confirm the fact that the distribution of TFC for the spanning trees is normal, we performed an individual distribution identification test in Minitab. This tests the input data

against 14 different possible distributions. Indeed, the normality of the distribution was confirmed at the 95% confidence level with a p-value of 0.082. The results remained true for $f_{i,j} \in u[0,1]$, and when the $f_{i,j}$ are chosen from a normal distribution. However, when fixed costs follow an exponential distribution the TFC distribution failed to be normal.

Property 1 *Given a random network design problem K_n, if the $f_{i,j}$ are uniformly or normally distributed, then the distribution of total fixed cost is approximately normal.*

Observe from Table 1 that the smallest and largest TFC found from the random trees is 710 and 1,479, which yield z-scores of $z = -3.03$ and $z = 2.85$,respectively. Since the distribution is approximately normal, we know that the range captures 99.66% of the data. But, as noted above, th e total fixed cost of the MST is 330 which yields a z-score of $z = -5.94$. If one were to use the best tree obtained from the simulation, the approximate minimum total fixed cost would be more than double the actual minimum. After running many simulations we noticed that the MST always had a z-score below -5.00 which motivated the following analysis.

The random network design problem is a generalized version of the *random minimum spanning tree problem* which seeks to find the minimum spanning tree in K_n whose edge weights are independent random variables. In a remarkable theorem of Frieze [4], he showed that the limit as $n \longrightarrow \infty$ of the expected sum of the MST of K_n whose edges have weights in $u[0,1]$ is $\zeta(3) = \sum_{i=1}^{\infty} i^{-3} = 1.202...$ The function $\zeta(n) = \sum_{i=1}^{\infty} \frac{1}{i^n}$ is the famous Reimann zeta function. Fill and Steele [2] obtained a formula to compute the expected weight of an MST and computed the expected weight for $2 \leq n \leq 9$. This work was continued by Gamarnik [6] who found the expected weight of the MST of K_n for all $n \leq 45$.

It is well known that for the probability distribution $u[a,b]$, the mean is $\mu = \frac{a+b}{2}$ with standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. Since the mean of a sum of random variables is the sum of the means, we know that the mean sum for a spanning tree of K_n whose edges have random weights from $u[0,1]$ is given by $\mu = \frac{n-1}{2}$. Moreover, the variance is the sum of the variances, so the standard deviation of the tree sums is $\sigma = \sqrt{\frac{n-1}{12}}$. For example, with $n = 15$, we expect the mean tree weight to be 7 with standard deviation 1.080.

Let z_n be the z-score of an MST for K_n whose edges have random weights from $u[0,1]$. For sufficiently large n we we know that

$$z_n = \frac{1.202 \ - \ (\frac{n-1}{2})}{\sqrt{\frac{n-1}{12}}} < \frac{1 \ - \ (\frac{n-1}{2})}{\sqrt{\frac{n-1}{12}}} = \frac{2\sqrt{3}}{\sqrt{n-1}} - \sqrt{3}\sqrt{n-1}.$$

Hence, as $n \longrightarrow \infty$ we see that $z_n \longrightarrow -\infty$. For example, if $n = 15$, we see that $z_n \approx -5.37$, and for $n = 101$, we have $z_n \approx -16.90$. In general, we have shown that z_n is $O(\sqrt{n})$.

Property 2 *Given a random network design problem K_n whose fixed costs satisfy $f_{i,j} \in u[0,1]$, let z_n be the z-score for the total fixed cost of an MST. Then $z_n \approx \frac{\sqrt{3}(3.404-n)}{\sqrt{n-1}}$. Moreover, z_n is $O(\sqrt{n})$ and as $n \longrightarrow \infty$, it holds that $z_n \longrightarrow -\infty$.*

3 The distribution of total variable costs

To determine the total variable cost for a spanning tree T with root node r, one must sum the variable costs along the path in T from r to i, for every node i. Thus certain v_{ij} will appear repeatedly in the total variable cost sum. For example in the network given in Figure 3 the shortest paths and their corresponding costs are as given. The TVC is 181. Notice how the variable cost on edge {1,3} appears 3 times in the TVC sum.

From	To	Nodes	Path Cost
1	2	1,2	33
1	3	1,2	23
1	4	1,2,4	33 + 25 = 58
1	5	1,3,5,	23 + 7 = 30
1	6	1,3,5,6,	23 + 7 + 7 = 37

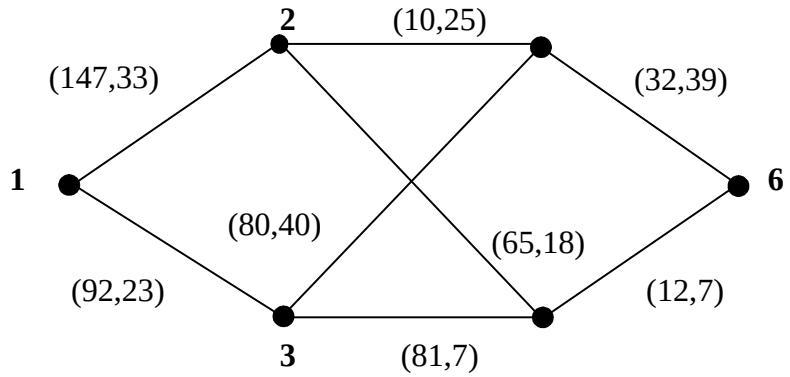

Figure 3 A 6-node network with $f_{i,j} \in u[10, 150]$ and $v_{i,j} \in u[5, 50]$

The distribution of TVC obtained from the 1,000 random trees for the $n = 15$, $v_{i,j} \in u[5, 50]$ problem is illustrated by the histogram given in Figure 4. The smallest TVC found from the random trees was 556, giving $z = -2.14$. Whereas the SPT gives a TVC of 247, implying a z-score of $z = -2.93$. In 50 runs of the simulation we observed that the z-score of the SPT was always greater than -3.00. Moreover, the distribution does not appear to be normal, but rather skewed to the right. An individual distribution identification test indicates that the distribution of TVC is lognormal at the 95% confidence level with a p-value of 0.518.

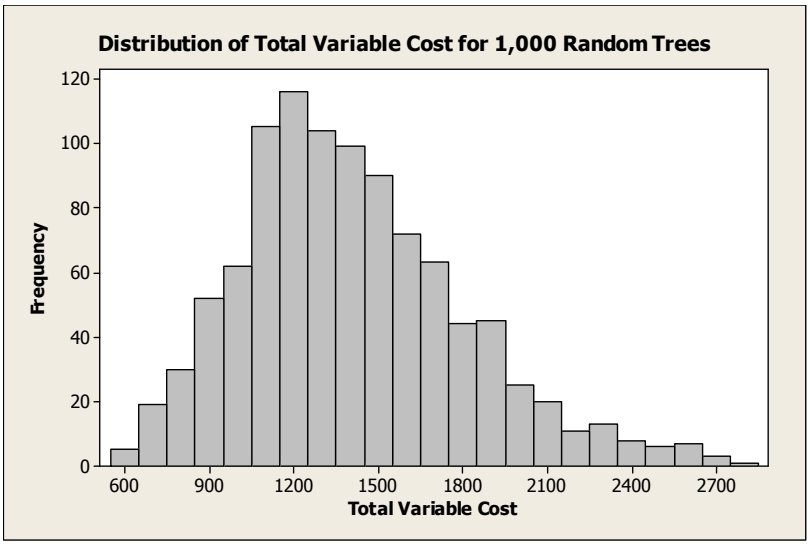

Figure 4 The Distribution of TVC for 1,000 Random Trees with $v_{i,j} \in u[5, 50]$

The lognormal result explains why the z-scores for the SPT is much closer to the minimum obtained from the simulation. However, when the range was changed from $v_{i,j} \in u[5, 50]$ to $v_{i,j} \in u[0, 1]$, the distribution of total variable cost was neither normal nor lognormal. So there is dependence on a and b. This also happened when the $v_{i,j}$ were selected from a normal distribution. One of the reasons why the distribution for total variable cost is no longer normal, but instead skewed to the right is because the number of terms included in the calculation of TVC is no longer constant, but varies from tree to tree. When calculating TFC there are always $n - 1$ terms in the calculation. It is interesting to ask: on average how many of the terms $v_{i,j}$ are there in the total variable cost sum? As noted in Table 1, the average TVC for the 1,000 random trees with v_{ij} chosen from $u[5, 50]$ is 1,408.1. Since the average $v_{i,j}$ is $\frac{5+50}{2} = 27.5$, we obtain an estimate of $\frac{1408.1}{27.5} \approx 51.2$ terms. As a check, we point out that the average number of terms used to compute TFC with f_{ij} chosen from $u[10, 150]$, is $\frac{1106.4}{80} \approx 13.8$, or rounded to the nearest integr is 14, which is what one expects.

The edge-wise standard deviations are 40.4 for $f_{i,j}$ and 13.0 for $v_{i,j}$. However, the relatively large number of terms used in the computation of TVC explains why the standard deviation is much larger for TVC. In Table 1 we see that the standard deviation for TFC is 130.7 and for TVC it is 397.7. This results in smaller z-values for the TVC distribution when comparing it to the TFC distribution, and in particular, a smaller z-value for the SPT compared to the z-value for the MST.

In [5] it was shown that for a random shortest paths problem where edge weights are $u[0, 1]$, the distance between each pair of nodes is bounded by $c(\log n)/n$ almost surely, for some constant c. Moreover, the order of magnitude of this bound cannot be improved. To determine the total variable cost associated with an SPT consider the $n-1$ paths from node 1 to the other $n - 1$ nodes. Since all of these paths have length at most c log n / n, the expected sum of the path lengths of an SPT is at most $c(n-1)(\log n)/n < c' \log n$, or simply $O(\log n)$.

Property 3. *Given a random network design problem K_n whose variable costs satisfy $v_{i,j} \in u[0,1]$, the expected TVC of an SPT is $O(\log n)$.*

Open Problem *Given a random network design problem K_n whose variable costs satisfy $v_{i,j} \in u[0,1]$, let z_n be the z-score for the total variable cost of an SPT. As $n \longrightarrow \infty$ does $z_n \longrightarrow -\infty$ or is z_n bounded by a constant?*

4 Conclusions and the distribution of total cost

We have seen that the minimum fixed cost attained by the MST is an extreme outlier in the distribution of TFC. In contrast, the minimum TVC attained by the SPT is not nearly as extreme when $n = 15$. Hence, if one were to use random trees with a sample size of 1,000 to find an approximate solution, then we can expect to be close to the minimum total TVC, but not very close to the minimum TFC. Random tree generation requires a much larger sample size to get close to a minimum fixed cost. Properties 1 and 2 shed some light on why this happens. Moreover, the standard deviation for the TFC is much smaller than the standard deviation of TVC, because the number of edges used to compute TFC is constant . The problem with fixed costs is that the tails of the simulated distribution fill up very slowly and the MST weight is located beyond -5σ, for n as small as 15. Using the results obtained above we can determine an appropriate sample size for random tree generation. We determined the MST would have a z-score between -5 and -6. Since, in a normal distribution $P(-6.00 < z < -5.00) = 0.00000029$, one must be willing to generate roughly 10,000,000 random trees to expect a z-score less than -5.00.

We have seen that for the TVC distribution, the probability mass is skewed right. So for example when $n = 15$, the z-score of an SPT is between -2.00 and -3.00. For a normal distirbution we know that $P(-3.00 < z < -2.00) = 0.0214$.The probability that z is between -2.00 and -3.00 for the TVC distribution is even larger. So generating 1,000 random trees should produce a tree with a TVC with a z-score below -2.00, and hence close to the TVC obtained by the SPT.

Next we address the distribution of total cost. We have seen the the distributions of total fixed cost is normal, and the distribtion of total variable costs have been, normal, lognormal and sometimes neither. For the problem discussed in the introduction with $n = 15$, $f_{i,j} \in u[10, 150]$ and $v_{i,j} \in u[5, 50]$, the total cost has a lognormal distribution. A simulation similar to the one described above was performed with fixed costs aassociated with edges that follow a normal distibution with mean 80 and standard deviation 20, and edge variable costs also normally distributed with mean 25 and standard deviation 5. Again TFC follows a normal distribution and TVC is lognormal. The z-scores are -2.86 for the SPT and -6.19 for the MST. However, when variable costs and fixed costs were both from $u[0, 1]$, the total variable cost failed to follow a lognormal, nor a normal distribution. But the total fixed costs distribution remained normal.

The above discussion is a good illustration of how simulation may not provide good approximation when trying to obtain an extreme value. This can be a problem when using the random network design problem to test new algorithms.

References

[1] Steve Cosares and Fred J. Rispoli, Selection Criteria for a Network Design Model with Uncertain Demand, *Applications of Management Science* **12**, eds. K.D. Lawrence and R.K. Klimberg, Elsevier, 2006

[2] James A. Fill and J. Michael Steele, Exact Expectations of Minimal Spanning Trees For Graphs with Random Edge Weights, *Steins Method and Applicatins*, (A. Barbour & L. Chen eds.) 169-180, World Publications, Signapore, 2005.

[3] Abraham D. Flaxman, The lower tail of the random minimum spanning tree, *The Electronic Journal of Combinatorics* **14** (2007).

[4] A.M. Frieze, On the value of a random minimum spanning tree problem, *Discrete Applied Math.* **10**, 47-56 (1985).

[5] A. M. Frieze and G. R. Grimmett, The shortest-path problem for graphs with random arc-lengths, *Discrete Applied Math.* **10**, 57-77 (1985).

[6] David Gamarnik, The expected value of random minimal length spanning tree of a complete graph, *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 700-704 (2005).

[7] Refael Hassin and Eitan Zemel, On shortest paths in graphs with random weights, *Mathematics of Operations Research*, **10** (4), 557-564 (1985).

[8] Robin J. Wilson and John J. Watkins, *Graphs: An Introductory Approach*, Wiley, 1990.