The Probabilistic Minimum Spanning Tree Problem

Dimitris J. Bertsimas
Sloan School of Management, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139

In this paper we consider a natural probabilistic variation of the classical minimum spanning tree problem (MST), which we call the probabilistic minimum spanning tree problem (PMST). In particular, we consider the case where not all the points are deterministically present, but are present with certain probability. We discuss the applications of the PMST and find a closed-form expression for the expected length of a given spanning tree. Based on these expressions, we prove that the problem is *NP-complete*. We further examine some interesting combinatorial properties of the problem, establish the relation of the PMST with the MST and the network design problem, and examine some cases where the problem is solvable in polynomial time. We finally characterize the asymptotic behavior of reoptimization strategies, in which we find the MST or the Steiner tree, respectively, among the points that are present on a particular instance, and the PMST, in the case in which points are randomly distributed in the Euclidean plane and in the case in which the costs of the arcs are randomly distributed. In both cases the PMST is within constant factors from both strategies.

1. INTRODUCTION

The classical minimum spanning tree (MST) problem plays an important role in combinatorial optimization. It possesses the matroidal property that allows the greedy algorithm to solve the problem optimally, and thus it is the prototype for problems solvable in polynomial time. For a summary of its properties and algorithms for its solution, see Papadimitriou and Steiglitz [8]. From a practical point of view, it has important applications in transportation, communications, distribution systems, etc.

In this paper we consider a natural probabilistic variation of this classical problem. In particular, we consider the case where not all the points are deterministically present, but are present with certain probability. Formally, given a weighted graph $G = (V, E)$ and a probability of presence p_i for each vertex i, we want to construct an **a priori** spanning tree of minimum expected length in the following sense: On any given instance of the problem, delete the vertices and their adjacent edges among the set of absent vertices provided that the tree remains connected. The problem of finding an a priori spanning tree

© 1990 John Wiley & Sons, Inc.
CCC 0028-3045/90/030245-031$04.00

of minimum expected length is the probabilistic minimum spanning tree (PMST) problem. In order to clarify the definition of the PMST problem, consider the example in Figure 1. If the a priori tree is T and nodes 2, 7, 9 are the only ones not present, the tree becomes T_1. One can easily observe that if $p_i = 1$ for all $i \in V$, then the problem reduces to the classical MST problem.

This paper is part of a more general investigation of the properties of combinatorial optimization problems when instances are modified probabilistically. Jaillet [6] defined the probabilistic traveling salesman problem (PTSP), examined some of its combinatorial properties, and proved asymptotic limit theorems in the plane. Bertsimas [2] derived further properties of the PTSP and also analyzed the probabilistic vehicle routing problem and probabilistic facility location problems. To our knowledge, the PMST problem has never been defined before in the literature despite its intrinsic interest as well as its applicability.

In the next section, we discuss some applications of the PMST problem, whereas in Section 3 we address the question of finding an explicit expression for the expected length of an a priori tree T. In Section 4, we investigate the complexity of the problem and we prove that even a restricted version of the problem with all weights equal is *NP-complete*, which in view of the simplicity of the MST problem is a quite surprising result. We further examine some special cases that are solvable in polynomial time. In Section 5, we examine some interesting combinatorial properties of the problem. In Section 6, we perform probabilistic analysis under the random Euclidean and the random-length models of the PMST and the two reoptimization strategies, in which we find the MST or the Steiner tree, respectively, among the points that are present at a particular instance. Under the random Euclidean model, we characterize the asymptotic behavior of the two reoptimization strategies and the PMST. In particular, we prove that with probability 1, as the number of points goes to infinity, the expected length of the PMST is within a constant of the MST and the Steiner reoptimization strategies. In the random-length model, using a result of Frieze [4], we prove that in probability the expected length of the PMST is asymptotically smaller than is the expectation of the MST reoptimization strategy. The final section includes some concluding remarks.

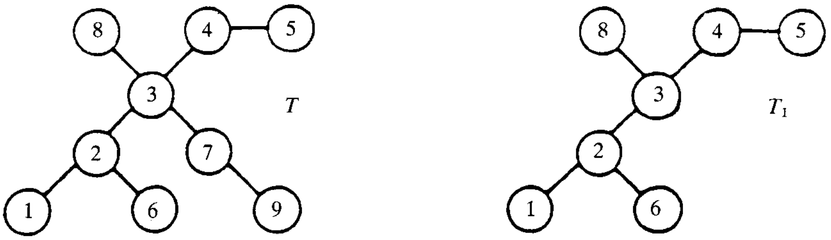

FIG. 1. The PMST methodology.

2. DISCUSSION AND APPLICATIONS OF THE PMST PROBLEM

The PMST problem defines an efficient strategy to update minimum spanning tree solutions when the problem's instances are modified probabilistically because of the absence of certain nodes from the graph. We denote this strategy with Σ_{T_p}, where T_p is the optimal a priori tree. Then, in the instance S, i.e., when only nodes in the set S are present, the strategy produces a tree $T_p(S)$ with length $L_{T_p}(S)$, which is the length of the tree that connects nodes from the set S of present nodes using parts of T_p. In the context of this discussion, the letter Σ denotes the strategy used. Alternatively, we can use either one of the following reoptimization strategies.

1. A reoptimization strategy Σ_{MST}, in which we find the minimum spanning tree (MST) of the set of present nodes in every instance. We denote with $L_{\text{MST}}(S)$ the length of the MST of the nodes in the set S.
2. A reoptimization strategy Σ_{STEINER}, in which we find the minimum Steiner tree of the set of present nodes in every instance. We denote with $L_{\text{STEINER}}(S)$ the length of the Steiner tree of the nodes in the set S, using possibly nodes from the set $V-S$.

Remarks. The above definition of the reoptimization strategy Σ_{STEINER} applies only for the case of a fixed network, as opposed to the case where the points are located in the Euclidean plane. In this case, $L_{\text{STEINER}}(S)$ is the length of the Steiner tree in the plane of the points from the set S. Note also that the PMST strategy uses "Steiner" points.

Why do we not use these reoptimization strategies, Σ_{MST}, Σ_{STEINER}, rather than the strategy Σ_{T_p} that we are proposing?

Concerning the Σ_{STEINER} strategy, it is clear that $L_{\text{STEINER}}(S) \leq L_{T_p}(S)$, because the tree connecting the set S using only parts of the tree T_p is also a solution to the Steiner problem. The disadvantage of the STEINER strategy is that we have to solve an *NP-hard* problem in every instance, something that is feasible only for small problem instances.

With the strategy Σ_{MST}, it is clear that we can compute $L_{\text{MST}}(S)$ in $O(|S|^2)$, using the greedy algorithm, but it is not clear that $L_{\text{MST}}(S) \leq L_{T_p}(S)$. In fact, in Section 5, we construct examples where the probabilistic strategy Σ_{T_p} we are proposing is better than the Σ_{MST}. Furthermore, in Section 6, we prove that asymptotically under reasonable probabilistic assumptions the probabilistic strategy Σ_{T_p} is at least as good as the Σ_{MST}.

What is more important is the fact that in many applications we need a real-time strategy to modify the solution when the instances are modified. Clearly, the PMST strategy satisfies this criterion, since the tree $T(S)$ can be found in $O(n)$ time as follows:

```
1. Start with the a priori tree $T$.
2. Until there are no unmarked leaves in $T$:
   find an unmarked leaf in $T$;
   if $i \in S$ mark it; else delete $i$ from $T$.
3. The resulting tree is the tree $T(S)$.
```

Since we are only looking at every node at most once, this is an O(n) algorithm. Note that the two reoptimization strategies are superlinear. In addition, we may not have the computer resources to reoptimize. An even more important motivation in favor of the Σ_{T_p} strategy is that this strategy does not change the underlying network structure, whereas both the reoptimization strategies can result in a completely different network structure by adding new edges and deleting old ones. In a communication network, for example, it may be very expensive or even impractical to create new communication links for each problem instance.

After this discussion of the various strategies available when problem instances are modified, we will describe some potential application areas of the PMST problem. In a VLSI context, suppose that on a circuit there are n processors subject to failures and processor i becomes inactive with probability p_i. Then we would like to connect the active processors using a spanning tree structure, which minimizes the manufacturing cost. Communication of two active processors through some inctive processors means that the inactive processors allow communication. Since in this example changing the underlying network structure is impractical, the PMST strategy is a good solution to the problem.

In a communication network, nodes may represent communication centers, arcs represent communication links, and link costs are the communication costs among centers. The probability of failure p_i is the probability of blocked communication in center i. If the centers are blocked, they can be used only to establish communication between unblocked centers. Then the problem of finding an a priori network structure of minimum expected cost is the PMST problem.

A more unusual application of the PMST problem is in the area of organizational structures. Suppose the n points that we wish to interconnect represent our agents or spies in a foreign country. They will undertake in the future a series of missions, each mission involving a different subset of agents. A mission, in our context, is an instance of the problem. We are looking for an a priori organizational structure in which, for obvious reasons, each agent will know only the people immediately above or below him/her in the structure; this implies a spanning-tree-like structure. The probability p_i associated with point i is the a priori probability that agent i will have to participate in any random mission undertaken by the network. For any given mission, only that part of the organization that is necessary to interconnect all the agents participating in that particular mission is activated. The distance between points i and j is interpreted as the cost or risk of exposure incurred when agents i and j must communicate or work with each other. Given p_i for $i = 1, 2, \ldots, n$ and the distance matrix for all possible pairs (i, j), the PMST gives the organizational structure that, in the expected value sense, minimizes the risk of exposure of the network on a random mission.

Other applications of the PMST include transportation and strategic planning. One might object that all the examples we have discussed represent some idealization of reality. Nevertheless, the PMST is a generic problem, which in many applications can be a more appropriate model than is the classical MST, in the case where a particular type of randomness is present. It also addresses the question of finding a spanning tree that is optimal on the average, rather than a solution that is optimal on a particular instance. The essential characteristic therefore of the PMST is that it is a more global and more robust problem than is the MST.

Unfortunately, as we prove in Section 4, one pays for these nice properties (robustness, globality) by changing the complexity of the problem radically. Although the MST problem is easily solvable, the PMST problem is *NP-hard*.

3. THE EXPECTED LENGTH OF A GIVEN SPANNING TREE

As we noted in the previous section, the PMST problem defines an efficient strategy for updating spanning tree solutions when problem instances are modified probabilistically in response to the absence of certain nodes from the graph. Given an a priori tree T we define $L_T(S)$ to be the length of the tree that connects nodes from the set S of present nodes using only parts of T. For example, in Figure 1, $S = \{1, 3, 5, 6, 8\}$ and $L_T(S)$ is the length of the tree T_1. Then if the set S of points present has probability $p(S)$, the problem can be defined formally as follows:

Problem definition. Given a graph $G = (V, E)$, not necessarily complete, a cost function $c: E \rightarrow R$, and a probability function $p: 2^V \rightarrow [0, 1]$, we want to find a tree T that minimizes the expected length $E[L_T]$:

$$E[L_T] = \sum_{S \subseteq V} p(S) L_T(S) \tag{1}$$

where the summation is taken over all subsets of V, the instances of the problem.

Note that at this level of generality we can model dependencies among the probabilities of presence of sets of nodes. An additional observation is that with this formulation one would need $O(n2^n)$, $(|V| = n)$ effort to compute the expected length of a given tree T. We would like to be able to compute $E[L_T]$ efficiently. The question we address in this section is for which probability functions $p(S)$ we can compute efficiently $E[L_T]$ for a given tree T.

If we define $h(S) \triangleq Pr\{\text{none of the nodes in S is present}\} = \Sigma_{R \subseteq V-S} p(R)$, then

Theorem 1. Given an a priori tree T its expected length is given by the expression

$$E[L_T] = \sum_{e \in T} c(e)\{1 - h(K_e) - h(V - K_e) + h(V)\} \tag{2}$$

where K_e, $V - K_e$ are the subsets of nodes contained in the two subtrees obtained from T by removing the edge e (see Fig. 2).

Proof. Given a tree T, let us consider how much each edge $e \in T$ contributes

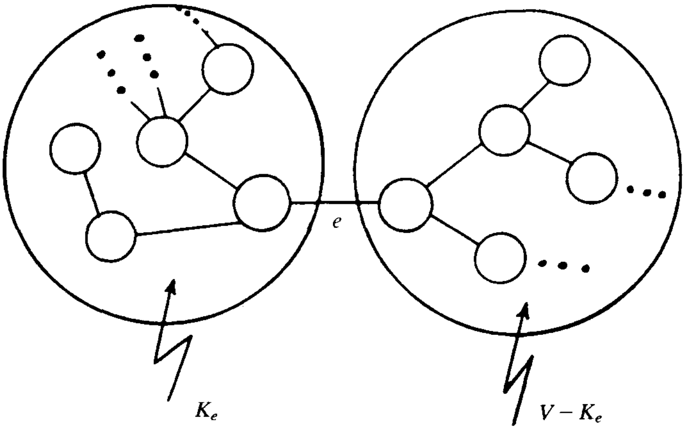

FIG. 2. The sets K_e, $V - K_e$.

to $E[L_T]$. By the definition of the problem, only the edges in T contribute in this expectation. If we define the events: $A(k_e) \triangleq$ at least one node in K_e is present, then the contribution of every edge e is

$$P(e)Pr\{A(K_e) \cap A(V - K_e)\}$$

because the edge e is used if and only if there exists at least one node present in K_e and at least one node present in $V - K_e$. As a result,

$$E[L_T] = \sum_{e \in T} c(e)Pr\{A(K_e) \cap A(V - K_e)\}$$

But

$$\begin{aligned}
&Pr\{A(K_e) \cap A(V - K_e)\} = Pr\{[A^c(K_e) \cup A^c(V - K_e)]^c\} \\
&= 1 - Pr\{A^c(K_e) \cup A^c(V - K_e)\} \\
&= 1 - Pr\{A^c(K_e)\} - Pr\{A^c(V - K_e)\} + Pr\{A^c(K_e) \cup A^c(V - K_e)\}.
\end{aligned}$$

But since $Pr\{A^c(K_e)\} = Pr\{$none of the nodes in K_e is present$\} = h(K_e)$, we easily obtain (2). ■

Thus, if instead of the probability function $p(S)$ we are given the function $h(S)$, we can compute $E[L_T]$ for any given tree T in $O(n)$, assuming we can compute $h(S)$ in $O(1)$, since we can find all the sets K_e for all e in $O(n)$ by starting the computation at the leaves. An interesting case, and important in practice, is when the nodes are present independently. Then we can find an explicit expression for $E[L_T]$.

Theorem 2. If node i is present with probability p_i, then the expectation $E[L_T]$

of a given tree T is given by the expression

$$E[L_T] = \sum_{e \in T} c(e) \left\{ 1 - \prod_{i \in K_e} (1 - p_i) \right\} \left\{ 1 - \prod_{i \in V - K_e} (1 - p_i) \right\} \tag{3}$$

Proof. In this case, because nodes are present independently

$$h(S) = \prod_{i \in S} (1 - p_i)$$

Substituting the above expression in (2), we easily obtain (3). ■

From (3) we can compute $E[L_T]$ in $O(n^2)$, since we can compute $h(S)$ in $O(|S|)$. By organizing the computation carefully, we can compute $E[L_T]$ in $O(n)$ as follows:

```
1. Let $a = \Pi_{i \in V}(1 - p_i)$; let $a_i = 1$; let MARKED = set of leaves.
2. Until node set is empty:
   if $i$ is a leaf, let $a_i = (1 - p_i)\,\Pi_{j \in \text{MARKED}, (i,j) \in T}\, a_j$;
   add $i$ to the set MARKED; delete $i$ from $T$.
3. $E[L_T] = \Sigma_{e=(i,j) \in T}\, c(e)(1 - a_i)(1 - a/a_i)$.
```

An important special case is when $p_i = p$ for all i. Then $E[L_T]$ becomes

$$E[L_T] = \sum_{e \in T} c(e)\{1 - (1 - p)^{|K_e|}\}\{1 - (1 - p)^{n - |K_e|}\}. \tag{4}$$

If we define

$$\phi(k) \triangleq \{1 - (1 - p)^k\}\{1 - (1 - p)^{n-k}\} \tag{5}$$

then

$$E[L_T] = \sum_{e \in T} c(e)\phi(|K_e|) \tag{6}$$

Based on these closed-form expressions, we will prove in the next section that the decision version of the PMST problem, even with $p_i = p$ for all i and $c(e) = 1$, is *NP-complete*. An additional importance of the expressions (6) is that they will assist us in deriving some key combinatorial properties of the optimal solution to the PMST problem.

4. THE COMPLEXITY OF THE PMST PROBLEM

In this section, we prove that the simplest possible case of the PMST problem with equal weights $c(e) = 1$ and $p_i = p$ is *NP-complete*. We first define formally the decision version of this restricted problem.

The Restricted PMST Problem (RPMST)

Instance. Given a graph $G=(V,E)$, costs $c(e)=1$ for all $e \in E$, a rational number $p, 0<p<1$ and a bound B.

Question. Is there a spanning tree T for G with

$$E[L_T]=\sum_{e\in T} c(e)\{1-(1-p)^{|K_e|}\}\{1-(1-p)^{n-|K_e|}\} \le B?$$

In order to prove that the RPMST problem is *NP-complete*, we will need some properties of the function $\phi(k)=(1-x^k)(1-x^{n-k})$, $x=1-p$ defined in (5).

Proposition 3. The function $\phi(k)$ has the following properties:

1. If $k<m<n/2$, then $\phi(k)<\phi(m)$.
2. $\phi(k+m)<\phi(k)+\phi(m)$.
3. $3\phi(3)-2\phi(4)>0$.

Proof. These properties follow easily from elementary algebraic manipulations as follows:

1. $\phi(k)-\phi(m)=(x^m-x^k)(1-x^{n-m-k})<0$ if $k<m$ and $m+k<n/2-n/2=n$.
2. $\phi(k)+\phi(m)-\phi(k+m)=(1-x^k)(1-x^m)(1+x^{n-k-m})>0$.
3. $3\phi(3)-2\phi(4)=3(1-x^3)(1-x^{n-3})-2(1-x^4)(1-x^{n-4})$ $\ge(1-x^{n-4})(1+2x^4-3x^3)=(1-x^{n-4})(1-x)$ $\times[1+x(1-x^2)+x^2(1-x)]>0$. ■

We now have all the required tools to prove that the RPMST problem is *NP-complete*.

Theorem 4. The RPMST problem is *NP-complete*.

Proof. Clearly, RPMST belongs to the class *NP*, since given a tree T we can compute $E[L_T]$ in polynomial time ($O(n)$) and compare it with the given bound B. In order to prove the completeness of the problem, we will reduce the *NP-complete* problem EXACT COVER BY 3-SETS (Garey and Johnson [5]) to it.

Exact Cover by 3-Sets (E-3C)

Instance. A family $S=\{\sigma_1,\ldots,\sigma_s\}$ of 3-element subsets of a set $C=\{c_1,\ldots,c_{3c}\}$

Question. Is there a subfamily $S_1 \subset S$ of pairwise disjoint sets such that $\cup_{\sigma\in S1}\sigma=C$?

Given an instance of the E-3C problem, we define the following instance of the RPMST problem:

$$
\begin{aligned}
&G=(V,E),\\
&V=R\cup S\cup C,\\
&R=\{a_0,\dots,a_r\},\\
&r=s+3c,\\
&E=\{(a_i,a_0),i=1,\dots,r\}\cup\{(a_0,\sigma),\sigma\in S\}\cup\{(\sigma,c),c\in\sigma\}.\\
&p \text{ arbitrary rational with } 0<p<1,\\
&B=(r+3c)\phi(1)+c\phi(4),\\
&\phi(k)=(1-x^k)(1-x^{n-k}),\ x=1-p,\ n=r+1+s+3c.
\end{aligned}
$$

As an example, if $S=\{\{c_1,c_2,c_3\},\{c_2,c_3,c_5\},\{c_2,c_4,c_5\},\{c_4,c_5,c_6\}\}$, $c=2$, $s=4$, the corresponding graph is presented in Figure 3.

Let T be a feasible ($E[L_T]\le B$) spanning tree of G. Clearly, $(a_i,a_0)\in T$. We now show that if $E[L_T]\le B$, then $(a_0,\sigma)\in T$ for all $\sigma\in S$. Suppose first there exists only one $(a_0,\sigma)\notin T$ for some $\sigma\in S$. We will show that $E[L_T]>B$. Since $(a_0,\sigma)\notin T$, there exists $i\in S$ and $j\in C$ such that (a_0,i), (i,j), $(j,\sigma)\in T$ (see Fig. 4a). We define $g_l\triangleq$ the number of nodes in $C-\{j\}$ that are adjacent to l in T. In the example in Figure 4a, $g_i=1$, $g_\sigma=2$.

We also define $s_l\triangleq$ the number of nodes in $S-\{i,\sigma\}$ in T that are adjacent to exactly l vertices from C in T ($l=0,1,2,3$).

From these definitions we get

$$
\begin{aligned}
s_1+2s_2+3s_3=3c-g_i-g_\sigma-1 \Rightarrow s_3\\
=\tfrac{1}{3}(3c-2s_2-s_1-g_i-g_\sigma-1)
\end{aligned}
\tag{7}
$$

We now write an expression for $E[L_T]$:

$$
\begin{aligned}
E[L_T]=r\phi(1)+(3c-g_i-g_\sigma-1)\phi(1)+s_1\phi(2)+s_2\phi(3)+s_3\phi(4)\\
+\phi(g_i+g_\sigma+3)+(g_i+g_\sigma)\phi(1)+\phi(2+g_\sigma)+\phi(1+g_\sigma)
\end{aligned}
$$

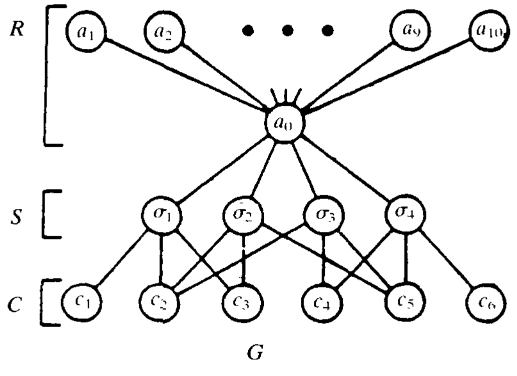

FIG. 3. Equivalent instances of E-3C and RPMST.

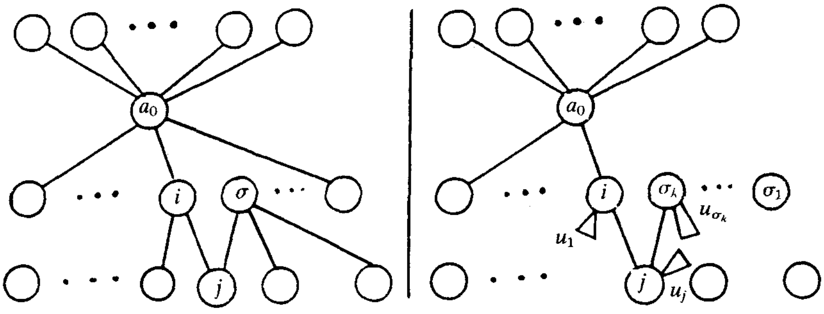

FIG. 4. The cases $(a_0, \sigma) \notin T$ and $(a_0, \sigma_1), \ldots, (a_0, \sigma_k) \notin T$.

where the first term $(r\phi(1))$ is from the contributions of the r edges (a_i, a_0), the second term is from the contributions of the edges connecting the nodes in C except the ones that are connected with i, σ, and the terms $\phi(g_i + g_\sigma + 3)$, $\phi(2 + g_s)$ and $\phi(1 + g_\sigma)$ are from the contributions of the edges (a_0, i), (i, j), and (j, σ), respectively. Then

$$E[L_T] > B = (r + 3c)\phi(1) + c\phi(4) \Leftrightarrow s_1\phi(2) + s_2\phi(3) + s_3\phi(4) \\ + \phi(g_i + g_\sigma + 3) + \phi(2 + g_\sigma) + \phi(1 + g_\sigma) - \phi(1) - c\phi(4) > 0$$

Substituting (7) we get

$$E[L_T] > B \Leftrightarrow \tfrac{1}{3}s_1[3\phi(2) - \phi(4)] + \tfrac{1}{3}s_2[3\phi(3) - 2\phi(4)] \\ + \tfrac{1}{3}[3\phi(g_i + g_\sigma + 3) - (g_i + g_\sigma)\phi(4) + \phi(2 + g_\sigma) \\ + \tfrac{1}{3}[2\phi(2 + g_\sigma) - \phi(4)] + [\phi(1 + g_\sigma) - \phi(1)] > 0.$$

Using Proposition 3 we can easily check that all the terms in [ ] are strictly positive and thus $E[L_T] > B$.

Suppose now that there are $(a_0, \sigma_1), \ldots, (a_0, \sigma_k) \notin T$ (see Fig. 4b). Since T is a tree, there exist $i \in S$ and $j \in C$ such that (a_0, i), (i, j), $(j, \sigma_k) \in T$. Then if we add the edge (a_0, σ_k) and delete the edge (j, σ_k), we get a new tree T_{k-1}, in which there are only $k - 1$ nodes $\sigma_1, \ldots, \sigma_{k-1}$ not connected with a_0. If we denote the tree T with T_k in order to represent the fact that there are k nodes in T not connected to a_0, we claim that

$$E[L_{T_k}] > E[L_{T_{k-1}}].$$

Let u_i, u_j, u_{σ_k} be the number of nodes in the subtrees from nodes i, j, σ_k, respectively (see also Fig. 4b).The contribution of edges in T_k, T_{k-1} that are not involved in the cycle created by adding the edge (a_0, σ_k) is the same. Then

$$E[L_{T_k}] - E[L_{T_{k-1}}] = \phi(u_i + 1 + u_j + 1 + u_{\sigma_k} + 1) + \phi(u_j + 1 + u_{\sigma_k} + 1) + \phi(u_{\sigma_k} + 1) - \phi(u_i + 1 + u_j + 1) - \phi(u_j + 1) - \phi(u_{\sigma_k} + 1)$$

where $\phi(u_i + 1 + u_j + 1 + u_{\sigma_k} + 1)$, $\phi(u_j + 1 + u_{\sigma_k} + 1)$, and $\phi(u_{\sigma_k} + 1)$ are the contributions in T_k of $(a_0, i),(i, j)$, and (j, σ_k), respectively. Similarly in T_{k-1}, the terms $\phi(u_i + 1 + u_j + 1)$, $\phi(u_j + 1)$ and $\phi(u_{\sigma_k} + 1)$ are from (a_0, i), (i, j), and (a_0, σ_k), respectively. By Proposition 3, we have that $\phi(u_i + 1 + u_j + 1 + u_{\sigma_k} + 1) > \phi(u_i + 1 + u_j + 1)$ and $\phi(u_j + 1 + u_{\sigma_k} + 1) > \phi(u_j + 1)$. As a result, $E[L_{T_k}] > E[L_{T_{k-1}}]$. Note that we have used the fact $r = s + 3c$, since in order for Proposition 3 to hold, we need $u_i + 1 + u_j + 1 + u_{\sigma_k} + 1 < s + 3c < n/2 = (r + 1 + s + 3c)/2$.

As a result, the expected cost of T decreases by adding one missing arc (a_0, σ_k). Making this transformation inductively, we find

$$E[L_T] = E[L_{T_k}] > E[L_{T_{k-1}}] > \cdots > E[L_{T_1}]$$

But since the tree T_1 has only one missing arc (a_0, σ_1), we have already proved that in this case $E[L_{T_1}] > B$.

Therefore, it follows that for the tree T to be feasible, all edges $(a_0, \sigma_i) \in T$. We will now show that

$$E[L_T] \leq B \Leftrightarrow \text{E-3C has a solution}$$

But using the quantities s_l $(l = 0, 1, 2, 3)$ defined above, we have

$$s_1 + 2s_2 + 3s_3 = 3c, \qquad s_0 + s_1 + s_2 + s_3 = s$$

The expected cost of T is then given by

$$E[L_T] = (r + 3c)\phi(1) + s_1\phi(2) + s_2\phi(3) + s_3\phi(4)$$

Thus

$$E[L_T] \leq B \Leftrightarrow s_1\phi(2) + s_2\phi(3) + (s_3 - c)\phi(4) \leq 0$$
$$\Leftrightarrow \tfrac{1}{3}s_1[3\phi(2) - \phi(4)] + \tfrac{1}{3}s_2[3\phi(3) - 2\phi(4)] \leq 0. \qquad (8)$$

From Proposition 3, $3\phi(2) - \phi(4) > 0$ and $3\phi(3) - 2\phi(4) > 0$. As a result, inequality (8) holds if and only if $s_1 = s_2 = 0$ and, hence, $s_3 = c$, which is equivalent to E-3C having a solution. Thus, $E[L_T] \leq B \Leftrightarrow$ E-3C has a solution, and hence, the RPMST problem is *NP-complete*. ■

We can add some insight to why the problem is hard by noticing the following remarkable fact. As $p \to 1$, the PMST approaches the MST, which is easily solvable. What is the limit as $p \to 0$? In this case

$$\phi(k) = (1-(1-p)^k)(1-(1-p)^{n-k}) \rightarrow p^2 k(n-k)$$

As a result

$$E[L_T] \rightarrow p^2 \sum_{e \in T} c(e)|K_e|(n - |K_e|)$$

The expression $\Sigma_{e \in T} c(e)|K_e|(n - |K_e|)$ is the objective function of another famous problem, the NETWORK DESIGN PROBLEM on a tree, which is defined as follows:

Network Design Problem

Instance. A graph $G=(V, E)$, a weight $c(e)$ for each $e \in E$, and a bound B.

Question. Is there a spanning tree T for G such that, if $W(\{u, v)\})$ denotes the sum of the weights of the edges on the path joining u and v in T, then

$$f(T) = \sum_{u,v \in V} W(\{u, v\}) \leq B?$$

It is easily seen by considering the contribution of every edge e that $f(T) = \Sigma_{w \in T} c(e)|K_e|(n - |K_e|)$. The network design problem on a tree was proved *NP-complete* in Johnson et al. [7]. Thus, the PMST problem approaches as $p \rightarrow 0$ an *NP-complete* problem, which gives some intuition as to why the problem is hard. In fact, it is this observation that originally led us to suspect that the PMST problem is hard.

We have proved that the restricted version of the PMST with equal costs on a noncomplete graph is *NP-complete*. We now prove that even if the graph is complete, but the costs are either small or large, the problem is still hard.

We have proved that the restricted version of the PMST with equal costs on a noncomplete graph is *NP-complete*. We now prove that even if the graph is complete, but the costs are either small or large, the problem is still hard.

The PMST problem on a complete graph.

Instance. A complete graph K_n, a cost $c(e) \in \{1, M\}$, a bound B, and a probability $p, 0 < p < 1$.

Question. Is there a spanning tree T with $E[L_T] \leq B$?

Theorem 5. The PMST problem on a complete graph is *NP-complete*.

Proof. Clearly, the problem is in *NP* because of the closed-form expressions we have found. To prove that the problem is complete, we use the same reduction as in the proof of Theorem 4. In order to make the graph complete, we add the remaining edges but with very high cost, i.e., $c(e) = [(r+3c)\phi(1) + c\phi(4) + 1]/[\phi(1)]$. Then if we include any edge of this type, its contribution would be $c(e)\phi(|K_e|) \geq c(e)\phi(1) = B+1$, i.e., it cannot be included in the tree. Therefore, the proof remains unchanged since edges with large costs never appear in a tree with $E[L_T] \leq B$. ■

The previous theorems indicate that the problem is hard if either the graph is complete and the costs are 1 or M or the graph is noncomplete but the costs are equal.

The next question concerns the complexity of the problem when we combine the above restrictions, i.e., when we have a complete graph with all costs $c(e) = 1$. We prove a more general theorem, which includes this case and characterizes the optimal solution.

Theorem 6. In the case where $p_i = p$ for all $i \in V$, whenever the optimum solution of the MST problem is a star tree T_*, then T_* is also the solution to the PMST problem.

Proof. For all trees T

$$E[L_T] = \sum_{e \in T} c(e)\phi(|K_e|) \geq \phi(1)L_T$$

from Proposition 3. But

$$E[L_{T*}] = \sum_{e \in T} c(e)\phi(1) = \phi(1)L_{T*}$$

Since T_* is by assumption the MST $L_{T*} \leq L_T$ for all trees. Combining the above inequalities

$$E[L_{T*}] \leq E[L_T]$$

Therefore, the star tree T_* solves the PMST problem. ■

Theorem 6 characterizes the optimal solution whenever the MST is a star tree T_*. Examples in which the MST is a star tree T_* and thus, by Theorem 6, it is also the PMST include a complete graph, with $c(i,j) = c_i + c_j$, $c(i,j) = c_i c_j$, $c_i \geq 0$, $c(i,j) = c_i + c_j + d_i d_j$, with $c_1 = \min c_i$ and $d_1 = \min d_i$ or $c(i,j) = \min(c_i, c_j)$.

Clearly in a complete graph with $c(e) = 1$, the MST is a star tree T_* and thus T_* is also the PMST. Hence, in this case, the optimal solution can be found in $O(n)$ time. Finally, using similar techniques, we can prove that even in the case in which the nodes have different probabilities of presence there are cases in which the star tree is the PMST.

Theorem 7. If the probability of presence of node i is p_i, $p_1 = \min_i p_i$ and the MST is a star tree T_* rooted at node 1, then T_* is the PMST.

5. PROPERTIES OF THE PMST

In this section, we examine the case with equal probabilities $p_i = p$. In this case, we are trying to find a spanning tree that minimizes the expression

$$f(p) \triangleq \min_T f_T(p) = \min_T \left\{ \sum_{e \in T} c(e)\phi(|K_e|) \right\} \tag{9}$$

5.1. Functional Properties of the PMST

Expression (9) is clearly a function of the coverage probability p. For different values of p, the corresponding optimal probabilistic trees that minimize (9) are different. We first address the question of specifying the properties of the function $f(p)$. From the results of Section 4, we have seen that it would be difficult to find $f(p)$ for a particular value of p, but can we find some global properties of this function that will give some insight into the problem? Some initial observations are stated in the following proposition.

Proposition 8. The function $f(p)$ is continuous, increasing, and piecewise differentiable. For $np\rangle 2$, it is also concave if the costs are positive.

Proof. We examine the properties of the function

$$\phi_k(p) \triangleq (1-(1-p)^k)(1-(1-p)^{n-k})$$

We can easily check that $(d/dp)\phi_k(p) > 0$, and $(d^2/dp^2)\phi_k(p) < 0$ for all $k \geq 2$ and $(d^2/dp^2)\phi_1(p) < 0$ if $np > 2$. Thus, the function $f_T(p)$ is continuous and differentiable since it is a polynomial and furthermore it is increasing and

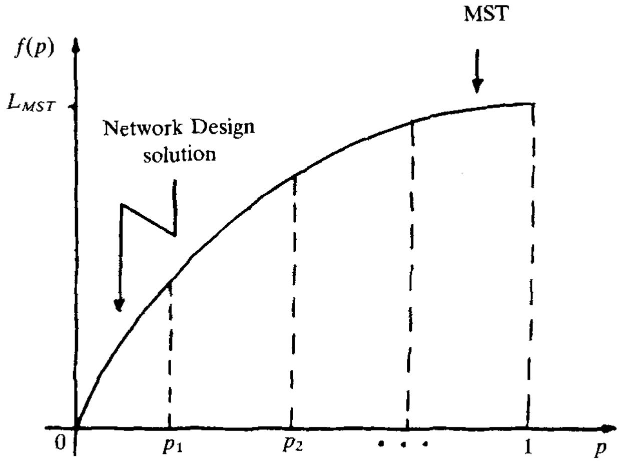

FIG. 5. The PMST problem as a function of the coverage probability p.

concave for $np>2$, since it is a weighted sum with positive weights ($c(e)\geq 0$). Therefore, the function $f(p)$ is concave for $np>2$ and continuous, since it is the minimum of a finite number of concave and continuous functions. Furthermore, $f(p)$ is increasing because for $p_1<p_2$ if $f(p_i)=f_{T_i}(p_i)$, $i=1,2$, then $f(p_1)=f_{T_i}(p_1)\leq f_{T_2}(p_1)<f_{T_2}(p_2)=f(p_2)$. Finally, there is a finite number of trees, which can possibly minimize $f(p)$. Thus, the function $f(p)$ has a finite number of breakpoints. Between successive break points p_i, p_{i+1}, $f(p)=f_{T_i}(p)$, $p_i\leq p\leq p_{i+1}$ for some T_i. Hence, $f(p)$ is piecewise differentiable. ■

We can now combine Proposition 8 and our previous observations that as $p\rightarrow 1$ the PMST tends to the MST, i.e., the optimal tree for p close to 1 is the MST, and as $p\rightarrow 0$, the optimal PMST is the solution to the network design problem, to sketch a possible graph of the function $f(p)$ in Figure 5.

5.2. Bounds for the PMST

Based on the above functional properties of $f(p)$ and some properties of $\phi(k)$ from Proposition 3, we can prove the following proposition.

Proposition 9. If T_p is the optimal PMST and L_T is the length of the tree T, then

$$\begin{aligned}&\max[pL_{\mathrm{MST}}, p(1-(1-p)^{n-1})L_{T_p}]\leq E[L_{T_p}]\\&\leq(1-(1-p)^{[n/2]})^2 L_{\mathrm{MST}}\end{aligned}\tag{10}$$

Proof. From the concavity of the function $f(p)$, we get that

$$f(p)\geq pf(1)+(1-p)f(0)=pL_{\mathrm{MST}}$$

where clearly $L_{\mathrm{MST}}\triangleq$ the length of the minimum spanning tree, which is the solution of PMST for $p=1$.

From Proposition 3 we get

$$\phi(1)\leq\phi(|K_e|)\leq\phi\left(\left[\frac{n}{2}\right]\right)$$

From the closed-form expression (6) for $E[L_T]$, we find

$$\phi(1)L_T=\phi(1)\sum_{r\in T}c(e)\leq E[L_T]=\sum_{e\in T}c(e)\phi(|K_e|)\leq\phi\left(\left[\frac{n}{2}\right]\right)L_T$$

Since $E[L_{T_p}]\leq E[L_{\mathrm{MST}}]$, we easily derive (10). ■

Exploiting these bounds, we address the question of how good is the MST as a solution to the PMST problem. The following is an obvious corollary of the bounds (10).

Corollary 10.

$$\frac{E[L_{\mathrm{MST}}]-E[L_{T_p}]}{E[L_{T_p}]}\leq\frac{(1-p)(1-(1-p)^{[n/2]-1})}{p} \tag{11}$$

Proof. Since $E[L_{T_p}]\leq E[L_{\mathrm{MST}}]\leq\phi([n/2])L_{\mathrm{MST}}\leq(1-(1-p)^{[n/2]})L_{\mathrm{MST}}$, and $E[L_{T_p}]\geq pL_{\mathrm{MST}}$, we can easily derive (11). Note that as $p\to 0$ the bound becomes $O(n)$.

These bounds indicate that for p large enough (say $p>1/2$) the MST solution is a good approximation for the solution of the PMST problem, which is consistent with our intuition. However, as $p\to 0$ and $n\to\infty$, this bound is not informative. In fact, the following example confirms our intuition that the MST can be a very poor solution to the PMST problem.

Consider a complete graph K_{n+1} with cost function: $c(i,i+1)=1$, $i=1,\ldots,n$ and $c(e)=2$ for all $e\neq(i,i+1)$. Note that the cost function in this example satisfies the triangle inequality. If the tree T_1 is the path $1,2,\ldots,n+1$ and T_2 is the star tree rooted at node $n+1$ (see Fig. 6), then clearly T_1 is the MST. Then $E[L_{T2}]=(2n-1)\phi(1)$ and $E[L_{T_1}]=2\sum_{i=1}^{n/2}\phi(i)=n(1+(1-p)^n)-2[(1-p)+p(1-p)^{n/2}-(1-p)^n]/p$ (assuming n is an even number). Then if T_p is the minimal PMST, we obtain

$$\frac{E[L_{\mathrm{MST}}]}{E[L_{T_p}]}\geq\frac{E[L_{T_1}]}{E[L_{T_2}]}=\frac{n(1+(1-p)^n)-2\dfrac{(1-p)+p(1-p)^{n/2}-(1-p)^n}{p}}{(2n-1)p(1-(1-p)^n)}$$

If $P=a/n$ for some constant $a>2$, we easily obtain as $n\to\infty$ that

$$\frac{E[L_{\mathrm{MST}}]}{E[L_{Tp}]}\geq\Omega(n)$$

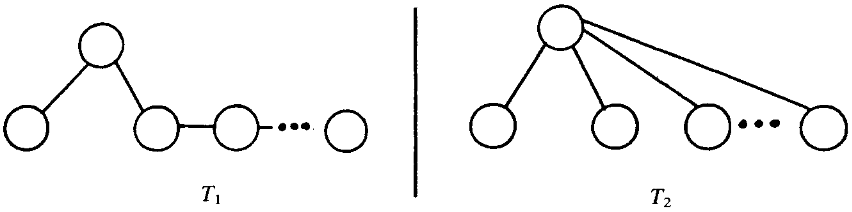

FIG. 6. The trees T_1, T_2.

Thus, from (11) we always have

$$\frac{E[L_{\mathrm{MST}}]}{E[L_{TP}]} = O(n)$$

and we have found an example for which

$$\frac{E[L_{\mathrm{MST}}]}{E[L_{T_P}]} = \Theta(n)$$

As a result, we conclude that the bound (11) is the best possible.

Furthermore, we can address the opposite question. How good is the PMST solution to the MST problem? Similarly we can show

Proposition 11.

$$\frac{L_{T_p} - L_{\mathrm{MST}}}{L_{\mathrm{MST}}} \leq \frac{1-p}{p(1-(1-p)^{n-1})} \tag{12}$$

Proof. Inequality (12) follows from the inequality (10) as follows:

$$p(1-(1-p)^{n-1})L_{T_p} \leq E[L_{TP}] \leq L_{\mathrm{MST}} \qquad \blacksquare$$

5.3. Relation of the PMST Problem and Reoptimization Strategies

As mentioned above, the PMST problem defines an efficient strategy to update the solution to minimum spanning tree problems, when problem instances are modified probabilistically because of the absence of certain nodes from the graph. We have also defined the two alternative reoptimization strategies Σ_{MST} and $\Sigma_{\mathrm{STEINER}}$. If $L_{\mathrm{MST}}(S)$, $L_{\mathrm{STEINER}}(S)$ is the length of the MST (Steiner tree) of the nodes in the set S, we define the expectation of these reoptimization strategies as follows:

$$E[\Sigma_{\mathrm{MST}}] \triangleq \sum_{S \subseteq V} p(S) L_{\mathrm{MST}}(S) \tag{13}$$

$$E[\Sigma_{\mathrm{STEINER}}] \triangleq \sum_{S \subseteq V} p(S) L_{\mathrm{STEINER}}(S) \tag{14}$$

where $p(S)$ was defined earlier to be the probability that only nodes in S are present. In this section, we address the question of comparing the expectation of the reoptimization strategies with the expectation of the PMST strategy. We should emphasize that the PMST strategy uses Steiner points and therefore it is directly comparable with the Steiner reoptimization strategy.

In general, it is difficult to find a closed-form expression for $E[\Sigma_{\mathrm{MST}}]$, since we have to compute a sum of $O(2^n)$ terms. Instead, we will find a bound on the $E[\Sigma_{\mathrm{MST}}]$.

Proposition 12. If every node is independently present with probability p, then

$$E[\Sigma_{\mathrm{MST}}] \geq \frac{np + (1-p)^n - 1}{n-1} L_{\mathrm{MST}} \tag{15}$$

where L_{MST} is the length of the MST.

Proof.

$$E[\Sigma_{\mathrm{MST}}] = \sum_{k=2}^{n} p^k(1-p)^{n-k} \sum_{S \subseteq V, |S| = k} L_{\mathrm{MST}}(S)$$

We define $D_k \triangleq \Sigma_{S \subseteq V, |S| = k} L_{\mathrm{MST}}(S)$ and thus

$$E[\Sigma_{\mathrm{MST}}] = \sum_{k=2}^{n} p^k(1-p)^{n-k} D_k \tag{16}$$

We claim that

$$D_k \geq \frac{k-1}{n-1}\binom{n}{k} L_{\mathrm{MST}} \tag{17}$$

We will prove the above claim by backward induction. Consider the n sets $S_i = V - \{i\}$. Then

$$L_{\mathrm{MST}}(S_i) + c(i,j) \geq L_{\mathrm{MST}}(V) = L_{\mathrm{MST}} \qquad \forall (i,j) \in \mathrm{MST} \tag{18}$$

because the tree created by adding the edge (i,j) to the MST on S_i is a feasible solution to the instance V. We apply (18) for $i = 1 \ldots n$, and since it holds for any edge in the MST, we choose for every i the corresponding edge (i,j) from the MST, such that the $n-1$ edges (i,j) are distinct, and one edge is the one with the minimum cost among all edges in the MST. Summing over all i we get

$$\sum_{i=1}^{n} L_{\mathrm{MST}}(S_i) = \sum_{i=1}^{n} c(i,j) \geq nL_{\mathrm{MST}}$$

In order to choose $n-1$ edges (i,j) to be distinct and the one remaining the least in cost, we perform the following algorithm:

1. Find the edge e^* with smallest cost $c(e^*)$.

2. Until the node set is nonempty,
   if i is a leaf in the MST, then let (i, j_i) be the unique edge in MST.
   If $(i, j_i) \neq e^*$ delete i.
3. For the two remaining nodes, let e^* be their corresponding edge.

Since there are $n-1$ edges (i, j) that are distinct, then

$$\sum_{i=1}^{n} c(i, j) = L_{\mathrm{MST}} + c(e^*) \leq \left(1 + \frac{1}{n-1}\right) L_{\mathrm{MST}}$$

As a result

$$D_{n-1} = \sum_{i=1}^{n} L_{\mathrm{MST}}(S_i) \geq \left(n - 1 - \frac{1}{n-1}\right) L_{\mathrm{MST}} = \frac{n(n-2)}{n-1} L_{\mathrm{MST}}$$

Consider now the $t = \binom{k}{k}$ subsets of V of cardinality $k, A_1, A_2, \ldots, A_t$. For all A_i, let $A_{i,j} \triangleq A_i - \{j\}$. Arguing as before

$$\sum_{j} L_{\mathrm{MST}}(A_{i,j}) \geq \frac{k(k-2)}{k-1} L_{\mathrm{MST}}(A_i)$$

Adding with respect to i, we get

$$\sum_{i,j} L_{\mathrm{MST}}(A_{i,j}) \geq \frac{k(k-2)}{k-1} D_k \tag{19}$$

But

$$\sum_{i,j} L_{\mathrm{MST}}(A_{i,j}) = (n - k + 1) D_{k-1} \tag{20}$$

since in the summation in (20), we count each distinct subset of the $\binom{n}{k-1}$ subsets of V cardinality $k-1$, $n-k+1$ times. Combining (19), (20) we find

$$D_{k-1} \geq \frac{k(k-2)}{(k-1)(n-k+1)} D_k \tag{21}$$

Applying (21) inductively, we easily obtain (17). Then from (16) and (17), we find

$$E[\Sigma_{\mathrm{MST}}] \geq \sum_{k=1}^{n} p^k (1-p)^{n-k} \frac{k-1}{n-1} \binom{n}{k} L_{\mathrm{MST}}$$

Therefore

$$E[\Sigma_{\mathrm{MST}}] \geq \frac{np + (1-p)^n - 1}{n-1} L_{\mathrm{MST}}$$

Note that as $n \to \infty$ the bound becomes

$$E[\Sigma_{\mathrm{MST}}] \geq pL_{\mathrm{MST}} \qquad \blacksquare$$

It is not clear that $E[\Sigma_{\mathrm{MST}}] \leq E[L_{T_p}]$. In fact, we give an example where $E[\Sigma_{\mathrm{MST}}] > \mathrm{E}[\mathrm{L}_{T_p}]$. Let $G = (V, E)$ be a complee graphs K_n with $c(i,j) = c_i + c_j$, $c_1 \leq c_2 \leq \cdots \leq c_n$. Then, the MST is the star tree rooted at node 1, and thus from Theorem 6, the optimal PMST is the same star tree. As a result

$$E[L_{T_p}] = p(1 - (1-p)^{n-1}) \left[(n-1)c_1 + \sum_{k-2}^{n} c_k \right]$$

In this example, we will be able to find a closed-form expression for $E[\Sigma_{\mathrm{MST}}]$ by exploiting the special structure of the cost function. If the ith node is present and the 1st, . . . , i − 1th nodes are not present, then the optimal tree is the star tree rooted at node i. From this observation, we can write a closed-form expression for $E[\Sigma_{\mathrm{MST}}]$

$$E[\Sigma_{\mathrm{MST}}] = \sum_{i=1}^{n-1} p(1-p)^{i-1} E[L_{Ti} \mid \text{node } i \text{ is present}]$$

where $E[L_{T_i}]$ means the expected length in the PMST sense of the star tree rooted at node i with leaves $i+1, \ldots, n$. Since $E[L_{T_i} \mid i \text{ is present}] = pL_{T_i} = p[(n-1-i)c_i + \Sigma_{k=i+1}^{n} c_k]$, after some algebraic manipulations, we easily find that

$$E[\Sigma_{\mathrm{MST}}] = p \sum_{i=1}^{n} c_i [p(n-i)(1-p)^{i-1} + 1 - (1-p)^{i-1}]$$

Choosing $c_i = i$, we find

$$E[L_{T_p}] = \frac{n(n+1)-1}{2} p + n - \frac{3}{p} + \mathrm{O}((1-p)^n)$$

$$E[\Sigma_{\mathrm{MST}}] = \frac{n^2 + 3n - 4}{2} p + \mathrm{O}((1-p)^n)$$

Letting $np = c$ and $n \to \infty$, we see

$$E[L_{T_p}] \to (c/2 + 1 - 3/c)n, \; E[\Sigma_{\mathrm{MST}}] \to cn/2$$

Then, as $n \to \infty$, $E[L_{T_p}] > E[\Sigma_{\mathrm{MST}}]$ for $c > 3$, but $E[L_{T_p}] < E[\Sigma_{\mathrm{MST}}]$ for $c < 3$.

6. PROBABILISTIC ANALYSIS

In this section, we address the important issue of comparing in terms of performance the PMST with the MST and the Steiner reoptimization strategies asymptotically as the number of points tends to infinity. We show that the PMST behaves comparably with the two re-optimization strategies we have defined under the two models that have been widely used in the literature: the random Euclidean model and the random-length model. Under the random Euclidean model introduced in Beardwood et al. [1], the points are uniformly and independently distributed in $[0, 1]^d$. Beardwood et al. [1] analyzed the TSP and Steele [10] develops the theory of subadditive Euclidean functionals to obtain very sharp limit theorems for a broad class of combinatorial optimization problems. For a very nice survey of the area, see Steele [13]. Under the random-length model, we are given a complete network with the costs $c(i, j)$ being uniform random variables in $(0, 1)$. Since we want to find the performance of these strategies as the number of points goes to infinity, we assume that every point has the same probability of presence p and therefore the probability of a set S is $p(S) = p^{|S|}(1-p)^{n-|S|}$, where $n = |V|$.

We first establish the asymptotic behavior of the MST and the Steiner reoptimization strategies in the random Euclidean model. We then characterize the asymptotic behavior of the PMST problem by using the theory of subadditive Euclidean functionals (Steele [10]) and finally we prove that under the random-length model the expectation of the MST reoptimization strategy is asymptotically larger than the expectation of the PMST strategy.

6.1. Reoptimization Strategies in the Random Euclidean Model

Let $X^{(n)} \triangleq (X_1, \ldots, X_n)$ be n points, which are uniformly and independently distributed in $[0, 1]^d$ according to a distribution with bounded support and absolutely continuous part $f(x)$. It is well known (Steele [10]) that if

$$L^n_{\mathrm{MST}}(X^{(n)}) \triangleq \text{the length of the MST on } X^{(n)}$$

and similarly

$$L^n_{\mathrm{STEINER}}(X^{(n)}) \triangleq \text{the length of the Steiner tree on } X^{(n)}$$

then with probability 1 as $n \to \infty$, there are constants $\beta_{\mathrm{MST}}(d)$ and $\beta_{\mathrm{STEINER}}(d)$ such that

$$\lim_{n\to\infty} \frac{\mathrm{L}^n_{\mathrm{MST}}(X^{(n)})}{n^{(d-1)/d}} = \beta_{\mathrm{MST}}(d) \int_{R^d} f(x)^{(d-1)/d}\, dx$$

$$\lim_{n\to\infty} \frac{L^n_{\mathrm{STEINER}}(X^{(n)})}{n^{(d-1)/d}} = \beta_{\mathrm{STEINER}}(d) \int_{R^d} f(x)^{(d-1)/d}\, dx \tag{22}$$

Let the probability of presence of any point be p. The expectation of the MST reoptimization strategy on $X^{(n)}$ is then

$$E[\Sigma^n_{\mathrm{MST}}(X^{(n)})] = \sum_{k=0}^{n} p^k(1-p)^{n-k} \sum_{s:|S|=k} L_{\mathrm{MST}}(X^{(n)}; S)$$

where $L_{\mathrm{MST}}(X^{(n)}; S)$ denotes the MST if the points are $X^{(n)}$ and the set of present points is S. Similarly

$$E[\Sigma^n_{\mathrm{STEINER}}(X^{(n)})] = \sum_{k=0}^{n} p^k(1-p)^{n-k} \sum_{S:|S|=k} L_{\mathrm{STEINER}}(X^{(n)}; S)$$

We now prove the following result.

Theorem 14. With probability 1

$$\lim_{n\to\infty} \frac{E[\Sigma^n_{\mathrm{MST}}(X^{(n)})]}{n^{(d-1)/d}} = \beta_{\mathrm{MST}}(d)p^{(d-1)/d} \int_{R^d} f(x)^{(d-1)/d}\,dx$$

$$\lim_{n\to\infty} \frac{E[\Sigma^n_{\mathrm{STEINER}}(X^{(n)})]}{n^{(d-1)/d}} = \beta_{\mathrm{STEINER}}(d)p^{(d-1)/d} \int_{R^d} f(x)^{(d-1)/d}\,dx$$

Proof. Let W be the number of nodes being present and

$$h_k \triangleq \sum_{S:|S|=k} L_{\mathrm{MST}}(X^{(n)}; S) \Big/ \binom{n}{k}$$

Then

$$E[\Sigma^n_{\mathrm{MST}}(XC^{(n)})] = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} h_k = \sum_{k=0}^{n} \Pr\{W=k\}h_k$$

Fix $\epsilon > 0$. Then

$$\begin{aligned} E[\Sigma^n_{\mathrm{MST}}(X^{(n)})] = &\sum_{k=0}^{\lceil np(1-\epsilon)\rceil-1} \Pr\{W=k\}h_k + \sum_{k=\lceil np(1+\epsilon)\rceil+1}^{n} \Pr\{W=k\}h_k \\ &+ \sum_{k=\lceil np(1-\epsilon)\rceil}^{\lceil np(1+\epsilon)\rceil} \Pr\{W=k\}h_k \end{aligned}$$

Since $L_{\mathrm{MST}}(X^{(n)}; S) < (|S|-1)\sqrt{d}$, since every edge in the tree is less than the largest edge,then $h_k \le n\sqrt{d}$. As a result

$$\sum_{k=0}^{\lceil np(1-\epsilon)\rceil-1} \Pr\{W=k\}h_k + \sum_{k=\lceil np(1+\epsilon)\rceil+1}^{n} \Pr\{W=k\}h_k \le n\sqrt{d}\Pr\{|W-np| > np\epsilon\}$$

From the Chernoff bound (Raghavan [9]), we have

$$\Pr\{|W-np|>np\epsilon\}<2\left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{np}=2\delta^{n},\quad 0<\delta<1$$

The contribution of the first two terms is then

$$\sum_{k=0}^{\lfloor np(1-\epsilon)\rfloor-1}\Pr\{W=k\}h_k+\sum_{k=\lfloor np(1+\epsilon)\rfloor+1}^{n}\Pr\{W=k\}h_k<2n\sqrt{d}\delta^{n},\quad \delta<1$$

For $\lceil np(1-\epsilon)\rceil\leq k\leq\lfloor np(1+\epsilon)\rfloor$, we apply (22) and get that with probability 1 $\forall\epsilon>0,\ \exists k_\epsilon:\forall S$, with $|S|=k\geq k_\epsilon$

$$-\epsilon\leq\frac{L_{\mathrm{MST}}(X^{(n)};S)}{K^{(d-1)/d}}-\beta_{\mathrm{MST}}(d)\int_{R^d}f(x)^{(d-1)/d}\,dx\leq\epsilon$$

and thus

$$-\epsilon\leq\frac{h_k}{k^{(d-1)/d}}-\beta_{\mathrm{MST}}(d)\int_{R^d}f(x)^{(d-1)/d}\,dx\leq\epsilon$$

In addition

$$\sum_{k=\lceil np(1-\epsilon)\rceil}^{\lfloor np(1+\epsilon)\rfloor}\Pr\{W=k\}=\Pr\{|W-np|\leq np\epsilon\}>1-2\delta^{n}$$

Therefore

$$\left(\beta_{\mathrm{MST}}(d)\int_{R^d}f(x)^{(d-1)/d}\,dx-\epsilon\right)(1-2\delta^{n})<\sum_{k=\lceil np(1-\epsilon)\rceil}^{\lfloor np(1+\epsilon)\rfloor}\Pr\{W=k\}\frac{h_k}{k^{(d-1)/d}}$$

$$<\left(\beta_{\mathrm{MST}}(d)\int_{R^d}f(x)^{(d-1)/d}\,dx+\epsilon\right)$$

from which

$$\left(\beta_{\mathrm{MST}}(d)\int_{R^d}f(x)^{(d-1)/d}\,dx-\epsilon\right)(1-2\delta^{n})(p(1-\epsilon))^{(d-1)/d}$$

$$<\sum_{k=\lceil np(1-\epsilon)\rceil}^{\lfloor np(1+\epsilon)\rfloor}\Pr\{W=k\}h_k/n^{(d-1)/d}$$

$$<\left(\beta_{\mathrm{MST}}(d)\int_{R^d}f(x)^{(d-1)/d}\,dx+\epsilon\right)(p(1+\epsilon))^{(d-1)/d}$$

Combining the above bounds, we find that almost surely $\forall \epsilon > 0$, $\forall n \geq k_\epsilon / [p(1-\epsilon)]$

$$\left(\beta_{\mathrm{MST}}(d) \int_{R^d} f(x)^{(d-1)/d}\, dx - \epsilon\right)(1 - 2\delta^n)(p(1-\epsilon))^{(d-1)/d} < \frac{E[\Sigma^n_{\mathrm{MST}}(X^{(n)})]}{n^{(d-1)/d}}$$

$$< \left(\beta_{\mathrm{MST}}(d) \int_{R^d} f(x)^{(d-1)/d}\, dx + \epsilon\right)(p(1+\epsilon))^{(d-1)/d} + 2n^{1/d}\sqrt{d}\delta^n$$

Since ϵ can be arbitrarily small, we let $\epsilon \rightarrow 0$ and thus we prove the theorem. The same argument proves the result for the Steiner reoptimization strategy. ■

The next interesting and natural question to address is what is asymptotically the expectation of the PMST strategy.

6.2. The PMST Strategy in the Random Euclidean Model

We prove in this subsection that we can characterize asymptotically the expected length of the optimal PMST. We define $L_T(X^{(n)}; S)$ to be the length of the tree under the PMT strategy if the points are $X^{(n)}$, the set of present points in S, and the a priori tree is T. If $E[L^n_T(X^{(n)})]$ denotes the expected length of an a priori tree T, then the following theorem holds.

Theorem 15. Let $X^{(n)}$ be a sequence of points distributed independently and uniformly in $[0, 1]^d$ and p the coverage probability of each point. With probability 1, there exists a constant $c(p, d)$

$$\lim_{n \rightarrow \infty} \frac{E[L^n_{T_p}(X^{(n)})]}{n^{(d-1)/d}} = c(p,d)$$

where T_p is the optimal PMST and

$$\max\{\beta_{\mathrm{MST}}(d)p, \beta_{\mathrm{STEINER}}(d)p^{(d-1)/d}\} \leq c(p, d) \leq \beta_{\mathrm{MST}}(d)$$

Proof. We will prove that with probability 1

$$\lim_{n \rightarrow \infty} \frac{E[L^n_{T_p}(X^{(n)})]}{n^{(d-1)/d}}$$

exists. In order to do so, we use subadditivity techniques developed by Steele [10, 12]. Clearly, the functional

$$f(X^{(n)}) \triangleq E[L^n_{T_p}(X^{(n)})]$$

is Euclidean and linear and has finite variance. As we prove in the following claim, it is also subadditive, but, unfortunately, it is not monotone (the MST is not also monotone). Thus, we cannot directly apply Theorem 1 of [10]. Instead, we use some approximate monotonicity property to prove its convergence. First we establish the subadditivity property of the PMST.

Claim. $f(X^{(n)})$ is subadditive, i.e., if Q_i, $i=1, \dots m^d$ is a partition of the unit square in m^d subsquares, then

$$f(x^{(n)} \cap [0, r]^d) \leq \sum_{i=1}^{m^d} f(X^{(n)} \cap rQ_i) + crm^{d-1}$$

Proof. Since the functional is Euclidean, we can restrict our attention to the case $r=1$. Consider the following algorithm:

1. For every nonempty subsquare Q_i, construct the PMST T_i for the points $X^{(n)} \cap Q_i$.
2. Select a point in each subsquare which is a leaf in T_i. Call these points representatives. Consider the representatives as points always present ("black" points).
3. Construct a MST T^* among the representatives.
4. The trees T_i and T^* create a tree T, which connects all the points $X^{(n)}$.

The expected length of the tree T is

$$E[L_T] = \sum_{i=1}^{m^2} f_1(X^{(n)} \cap Q_i) + L_{T^*}$$

where $f_1(X^{(n)} \cap Q_i)$ is the expected length of T_i in which one point, the representative, is always present (it is a "black" node) and all the others have probability p for being present. If we turn a "black" node to a "white" node (a node that has probability p of being present), the expected length of T_i decreases. The resulting tree has expected length not smaller than $E[L_{Tp}]$, since by definition, T_p is the PMST. Then

$$E[L_{Tp}] \leq \sum_{i=1}^{m^2} f_1(X^{(n)} \cap Q_i) + L_{T^*} \tag{23}$$

It is well known (Eilon et al. [3]) that

$$L_{T^*} \leq b_d m^{d-1}$$

that is, the MST of $l = m^d$ points is less than $bl^{(d-1)/d}$ for some constant b. The question now is to relate $f_1(X^{(n)} \cap Q_i)$ with $f(X^{(n)} \cap Q_i)$ or equivalently $E[L_{T_i}]$ with $E[L_{T_i} \mid \text{a leaf is a black node}]$. The expected length $E[L_{T_i}]$ is given by

$$E[L_{T_i}] = \sum_{e \in T_i} c(e)\{1-(1-p)^{|K_e|}\}\{1-(1-p)^{n_i-|K_e|}\}$$

where K_e is defined to be the set of nodes that the component not containing the black node i has, if the edge e is deleted from the tree (see Fig. 7) and n_i is the number of points in $X^{(n)} \cap Q_i$. The above equation is derived by considering the contribution of every edge of the tree T_i in the expected length of T_i. Note that with this definition $|K_e|$ is not restricted to be less than $[n_i/2]$. Similarly

$$E[L_{T_i} \mid i \text{ is black}] = \sum_{e \in T_i} c(e)\{1-(1-p)^{|K_e|}\}$$

Then

$$\begin{aligned} E[L_{T_i}] &= \sum_{e \in T_i} c(e)\{1-(1-p)^{|K_e|}\} - \sum_{e \in T_i} c(e)(1-p)^{n_i-|K_e|}\{1-(1-p)^{|K_e|}\} \\ &= E[L_{T_i} \mid i \text{ is black}] - \sum_{e \in T} c(e)(1-p)^{n_i-|K_e|}\{1-(1-p)^{|K_e|}\}. \end{aligned}$$

As a result

$$\begin{aligned} E[L_{T_i} \mid i \text{ is black}] &\le E[L_{T_i}] + \sum_{e \in T_i} c(e)(1-p)^{n_i-|K_e|} \\ &\le E[L_{T_i}] + c_{\max} \cdot \sum_{e \in T_i} (1-p)^{n_i-|K_e|} \end{aligned}$$

We need to bound the term $\Sigma_{E \in T_i} (1-p)^{n_i-|K_e|}$.

Claim. $s(T_i) \triangleq \Sigma_{e \in T_i} (1-p)^{n_i-|K_e|} \le (1-p)/p[1-(1-p)^{n_i-1}]$. We will prove the claim by induction on n_i. For $n_i = 3$, $\Sigma_{e \in T_i} (1-p)^{3-|K_e|} = (1-p) + (1-p)^2 = (1-p)/p[1-(1-p)^{3-1}]$.

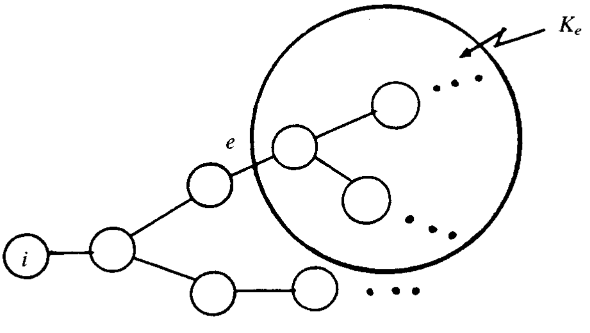

FIG. 7. The set K_e.

Suppose the claim is true for $n_i - 1$. Consider a tree $T_i^{n_i}$ with n_i nodes. There exists a second leaf j other than i. Deleting j we obtain a tree $T_i^{n_i - 1}$ with $n_i - 1$ nodes. Then

$$s(T_i^{n_i}) = s(T_i^{n_i-1}) + (1-p)^{n_i-1} \le (1-p)/p[1-(1-p)^{n_i-2}] + (1-p)^{n_i-1}$$
$$= (1-p)/p[1-(1-p)^{n_i-1}]$$

by the induction hypothesis, and thus the claim is proved.

As a result of the claim we find that

$$E[L_{T_i} \mid i \text{ is black}] \le E[L_{T_i}] + c_{\max}(1-p)/p$$

Since $c_{\max} \le \sqrt{d}/m$, we find that

$$f_1(X^{(n)} \cap Q_i) \le f(X^{(n)} \cap Q_i) + \sqrt{d}\,\frac{1-p}{mp} \tag{24}$$

From (23) and (24), we then conclude that

$$E[L_{T_p}] = f(X^{(n)} \cap [0,1]^d) \le \sum_{i=1}^{m^d} f(X^{(n)} \cap Q_i) + \left(b_d + \sqrt{d}\,\frac{1-p}{p}\right) m^{d-1}$$

which means that the PMST problem is subadditive. ■

Let $m_n = E_X E[L_{T_p^n}(X^{(n)})]$, where E_X denotes the expectation taken over all random sequences $X^{(n)}$. To extract the asymptotics of m_n, we will first prove that

$$n m_{n-1} \le \left(n + \frac{4}{p}\right) m_n$$

Consider the optimal PMST T_p^n on $X^{(n)}$. Let x' be an element x_j of the neighborhood $N(i)$ of x_i in T_p^n such that $|x_i - x_j|$ is minimum. By taking the edges of T_p^n, deleting all the edges incident to x_i and adding the set of edges which join x' to other neighbors of x_i, we get a spanning tree T_i on $X^{(n)} - \{x_i\}$. Then

$$E[L_{T_i}] \le E[L_{T_p^n}] + 2 \sum_{j \in N(i)} |x' - x_j|$$

since the weight of each edge in $E[T_i]$ not adjacent to x_i, x' is less than the corresponding weight in T and the weight of the edges that are adjacent to x' are less than 1. From the triangle inequality $|x' - x_j| \le |x' - x_i| + |x_i - x_j| \le 2|x_i - x_j|$, we find that

$$E[L_{T_p^{n-1}}(X^{(n)} - \{x_i\})] \le E[L_{T_i}] \le E[L_{T_p^n}] + 2 \sum_{j \in N(i)} |x_i - x_j|$$

Adding the for all $1 \le i \le n$ and taking expectation (E_X), we find that

$$nm_{n-1} \le (nm_n + 4E_X[L_{T_p^n}] \le \left(n + \frac{4}{p}\right)m_n$$

since any tree T satisfies $p(1-(1-p)^{n-1})L_T \le E[L_T]$, which implies that

$$n^k m_n \ge (n-1)^k m_{n-1}$$

$$\text{for } k \ge \frac{4}{p(1-(1-p)^{n-1})}$$

Having established the subadditivity and the approximate monotonicity of the PMST, we follow the same techniques as in Steele [12] and we prove that with Probability 1

$$\lim_{n\to\infty} \frac{E[L^n_{T_p}(X^{(n)})]}{n^{(d-1)/d}}$$

exists and it is equal to a constant $c(p, d)$ that depends only on the dimension d and the coverage probability p. Since the proof is very similar to [12], we omit the details.

Since for every tree $p(1-(1-p)^{n-1})L_T \le E[L_T] \le L_T$ and $E[\Sigma_{\text{STEINER}}] \le E[L_T]$, we use (22) for the case of f being uniform and theorem 14 to find the following bounds on $c(p, d)$:

$$\max\{\beta_{\text{MST}}(d)p, \beta_{\text{STEINER}}(d)p^{(d-1)/d}\} \le c(p,d) \le \beta_{\text{MST}}(d) \qquad \blacksquare$$

6.3. The PMST and MST Reoptimization Strategy in the Random-Length Model

In this model, we are given a complete graph with the costs $c(i, j)$ being independently and uniformly distributed in $[0, 1]$ and the coverage probability p fixed. We want to compare the MST reoptimization strategy and the PMST strategy asymptotically. We base our analysis on the following quite remarkable result, proved by Frieze [4].

In the random-length model, the MST converges in probability to

$$\lim_{n\to\infty} \mathrm{L}^n_{\text{MST}} = \zeta(3) = \sum_{k=1}^{\infty} \frac{1}{k^3} \simeq 1.202 \qquad (25)$$

$\blacksquare$

Remark. Frieze proved that the above theorem holds in expectation and Steele strengthened the result to prove convergence in probability.

Based on this result, we prove the following theorem about the behavior of the MST reoptimization strategy.

Theorem 16. In the random-length model, for all p such that $\lim_{n\to\infty} np = \infty$, the strategy of MST reoptimization converges in probability to

$$\lim_{n\to\infty} E[\Sigma^n_{\mathrm{MST}}] = \zeta(3)$$

Proof. The proof follows along the same lines of Theorem 14. The idea is that the asymptotically important terms in $E[\Sigma_{\mathrm{MST}}]$ are the ones that correspond to the number of points present within ϵ of np. ■

For the PMST strategy, we will need only an easy bound in order to compare this strategy with the MST reoptimization strategy. Since $pL_T(1-(1-p)^{n-1}) \leq E[L_T] \leq L_T$ and using (25) we obtain:

Proposition 17. The following inequality holds in probability

$$p\zeta(3) \leq \liminf_{n\to\infty} \mathrm{E}[\mathrm{L}^n_{T_p}] \leq \limsup_{n\to\infty} E[L^n_{T_p}] \leq \zeta(3)$$

We conjecture that

Conjecture 18. In the random-length model, the PMST converges in probability to

$$\lim_{n\to\infty} E[L^n_{T_p}] = p\zeta(3) \qquad \blacksquare$$

One can observe that we have not discussed the Steiner reoptimization strategy in this model. The reason is that in contrast with the MST there do not exist sharp theorems characterizing the asymptotic behavior of the deterministic Steiner tree problem in the random-length model. Since the Steiner reoptimization strategy is always better than is the PMST strategy, the following ordering holds asymptotically is probability:

$$E[\Sigma^n_{\mathrm{STEINER}}] \leq E[L^n_{T_p}] \leq E[\Sigma^n_{\mathrm{MST}}] = \zeta(3)$$

7. CONCLUDING REMARKS

We have seen that a natural probabilistic variation of a classical combinatorial problem has the potential to model various practical situations, offers an alternative way to update solution to problem instances that are modified probabilistically, and leads to very different properties in comparison with its deterministic counterpart. The simple possible version of the PMST problem was proved to be *NP-complete*, in sharp contrast with the fact that the MST problem is solved by a greedy, most straightforward algorithm. We have examined, however, some special cases in which the PMST can be solved in polynomial time.

Surprisingly, our analysis of the combinatorial properties of the problem established some interesting connections with the network design problem and naturally with the MST and the Steiner tree. In particular, as the probability of presence p tends to 0, the PMST approaches the solution to the network design problem. This limiting behavior suggests the idea of solving the network design problem as a sequence of PMST problems, which is a topic of future research.

Finally, we compared the PMST updating strategy with the MST and the Steiner reoptimization strategies. The PMST strategy has the property that it finds a solution to the modified instance very quickly (in linear time) and so it can be used in real time, but it is suboptimal on the worst case. It should also be emphasized that it uses Steiner points. The reoptimization strategies on the contrary find optimal solutions at every instance, but they need exponential time for the Steiner reoptimization strategy and quadratic time for the MST reoptimization strategy. It is quite surprising to find that the PMST strategy is asymptotically at least as good in terms of performance as the MST reoptimization strategy in the random-length model and within constant factors of the MST and the Steiner reoptimization strategies in the random Euclidean model.

As a general conclusion, probabilistic variations of classical combinatorial optimization problems raise interesting and entirely new questions compared with their deterministic counterparts and, in addition, understanding of the properties of the probabilistic problem can add insight to determinsitic problems, as it was the case with the network design problem. Our results add evidence that **a priori** strategies may offer a useful and practical method for resolving combinatorial optimization problems on modified instances.

I would like to thank my thesis advisor Professor Amedeo Odoni for several useful comments and interesting discussions and Dr. Patrick Jaillet for his constructive remarks. This research was partially supported by the National Science Foundation Grant ECS-8717970.

References

[1] J. Beardwood, J. Halton, and J. Hammersley, The shortest path through many points. *Proc. Camb. Phil. Soc.* **55**, (1959) 299–327.

[2] D. Bertsimas, Probabilistic combinatorial optimization problems, Ph.D. Thesis, Technical Report No. 194, Operations Research Center, Massachusetts Institute of Technology, Cambridge (1988).

[3] S. Eilon, C. Watson-Gandy, and N. Christofides, *Distribution Management*, Griffin, London (1971).

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

[5] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. F. Freeman, San Francisco (1979).

[6] P. Jaillet, Probabilistic traveling salesman problems, Ph.D. Thesis, Technical Report No. 185, Operations Research Center, Massachusetts Institute of Technology, Cambridge (1985).

[7] D. Johnson, J. K. Lenstra, and A. Rinnooy Kan, The complexity of the network design problem. *Neworks* **8** (1978) 279–285.

[8] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ (1982).

[9] P. Raghavan, Probabilistic construction for deterministic algorithms approximating packing integer programs. *J. Comp. Syst. Sci.* **36** (1988) 991–1003.

[10] J. M. Steele, Subadditive euclidean functionals and nonlinear growth in geometric probability. *Ann. Prob.* **9** (1981) 365–376.

[11] J. M. Steele, On Frieze's $\zeta(3)$ limit for lengths of minimal spanning trees. *Disc. Appl. Math.* **18**, (1987) 99–103.

[12] J. M. Steele, *Growth Rates of Euclidean Minimal Spanning Trees with Power Weighted Edges*, technical report, Princeton University, (1988).

[13] J. M. Steele, Probabilistic and Worst Case Analyses of Classical Problems of Combinatorial Optimization in Euclidean Space, technical report, Princeton University, (1988).

Received August 1988
Accepted August 1989