AN APPROXIMATION ALGORITHM FOR THE AT LEAST VERSION OF THE GENERALIZED MINIMUM SPANNING TREE PROBLEM

P. C. POP,[*] A. HORVAT-MARC[*] and CORINA POP SITAR[†]

Abstract. We consider the at least version of the Generalized Minimum Spanning Tree Problem, denoted by L-GMSTP, which consists in finding a minimum cost tree spanning at least one node from each node set of a complete graph with the nodes partitioned into a given number of node sets called clusters. We assume that the cost function attached to edges satisfies the triangle inequality and the clusters have sizes bounded by ρ. Under these assumptions we present a 2ρ approximation algorithm.

The algorithm works by rounding an optimal fractional solution to a linear programming relaxation. Our technique is based on properties of optimal solutions to the linear programming formulation of the minimum spanning tree problem and the parsimonious property of Goemans and Bertsimas.

MSC 2000. 05C05, 68R10, 90C05, 90C10, 90C27, 90C39.
Keywords. Approximation algorithms, minimum spanning tree, generalized minimum spanning trees, integer programming, linear relaxation.

1. INTRODUCTION

The minimum spanning tree (MST) problem can be generalized in a natural way by considering instead of nodes node sets (clusters) and asking for a minimum cost tree spanning *exactly* one node from each cluster. This problem is called the generalized minimum spanning tree problem (GMSTP) and it was introduced by Myung *et al.* [6].

Meanwhile, the GMSTP have been studied by several authors w.r.t. heuristics and metaheuristics, LP-relaxations, polyhedral aspects and approximability, cf., e.g. Feremans, Labbe, and Laporte [3], Feremans [2], Pop, Kern and Still [10, 11] and Pop [7, 8].

Two variants of the generalized minimum spanning tree problem were considered in the literature: one in which in addition to the cost attached to the edges, we have costs attached also to the nodes, called the *prize collecting generalized minimum spanning tree problem*, see [9] and the second one consists

[*]North University of Baia Mare, Faculty of Sciences, Department of Mathematics and Computer Science, Victoriei 76, 430122 Baia Mare, Romania, e-mail: pop_petrica@yahoo.com, andrei.horvatmarc@gmail.com.
[†]"Babeş-Bolyai" University, Faculty of Economics, 3400 Cluj-Napoca, Romania.

in finding a minimum cost tree spanning *at least* one node from each cluster, denoted by L-GMSTP and introduced by Dror *et al.* [1]. The same authors have proven that the L-GMSTP is *NP-hard*.

The theory of *NP-completeness* has reduced hopes that the *NP-hard* problems can be solved within polynomially bounded computation times. At the expense of reducing the quality of the solution by relaxing some of the requirements, for example we can relax the requirement that the algorithm always finds an optimal solution, we can get considerable speed-up in complexity. Consequently, there is much interest in approximation, heuristic and meta-heuristic algorithms.

An algorithm is an α-*approximation algorithm* for an optimization problem if

1. The algorithm runs in polynomial time.
2. The algorithm produces a solution which is within a factor of α of the value of the optimal solution.

Approximation algorithms have been around since 1966, but in the last years there has been a great deal of research, and two different strands converged: complexity theorists have developed powerful tools for showing that no α-approximation algorithms can exist unless $P = NP$ and algorithm designers have developed techniques that apply to a wide range of problems.

The aim of this paper is to describe an approximation algorithm for the at least version of the generalized minimum spanning tree problem under some special assumptions.

2. DEFINITION OF THE PROBLEM

The at least version of the generalized minimum spanning tree problem (L-GMSTP) is defined on an undirected graph $G = (V, E)$ with nodes partitioned into m clusters. Let $|V| = n$ and $K = \{1, 2, \ldots, m\}$ be the index set of the node sets (clusters). Then, $V = V_1 \cup V_2 \cup \ldots \cup V_m$ and $V_l \cap V_k = \emptyset$ for all $l, k \in K$ such that $l \neq k$. We assume that the graph G is complete and each edge $e = \{i, j\} \in E$ has a nonnegative cost denoted by c_{ij}.

The L-GMSTP is the problem of finding a minimum-cost tree spanning a subset of nodes which includes at least one node from each cluster.

3. INTEGER PROGRAMMING FORMULATIONS

The L-GMSTP can be formulated as an integer program in many different ways. For example, introducing the variables $x_e \in \{0, 1\}$, $e \in E$ and $y_i \in \{0, 1\}$, $i \in V$, to indicate whether an edge $'e'$ respectively a node $'i'$ is contained in the spanning tree, a feasible solution to the L-GMSTP can be seen as a connected subgraph with at least one node selected from every cluster and connecting all the clusters. Therefore the L-GMSTP can be formulated as the
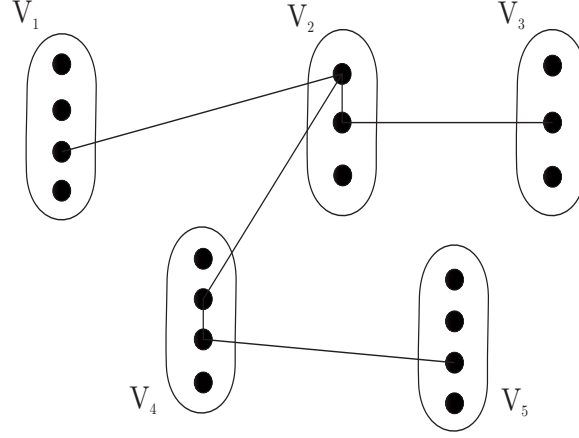
Fig. 1.   Example of a feasible solution of the L-GMSTP problem, where *at least* one node
from each cluster is selected.

following 0-1 integer programming problem:

$$\min \quad \sum_{e \in E} c_e x_e$$

$$
\begin{aligned}
&(1) &s.t. \quad &y(V_k) \geq 1, &&\forall\, k \in K = \{1, ..., m\} \\
&(2) & &x(\delta(S)) \geq y_i + y_j - 1, &&\forall\, i \in S \subset V, j \notin S \\
&(3) & &x(E) = y(V) - 1 \\
&(4) & &x_e \in \{0, 1\}, &&\forall\, e \in E \\
&(5) & &y_i \in \{0, 1\}, &&\forall\, i \in V,
\end{aligned}
$$

where for $S \subseteq V$, the *cutset*, denoted by $\delta(S)$, is defined as usually:

$$\delta(S) = \{e = (i, j) \in E \mid i \in S,\ j \notin S\}.$$

In the above formulation, we use the standard shorthand notations:

$$x(F) = \sum_{e \in F} x_e,\ F \subseteq E \text{ and } y(S) = \sum_{i \in S} y_i,\ S \subseteq V.$$

In the integer programming formulation of the L-GMSTP, constraints (1) guarantee that from every cluster we select at least one node, constraints (2) guarantee that the selected subgraph is connected and finally constraint (3) guarantees that the selected subgraph has $y(V) - 1$ edges.

An equivalent integer programming formulation of the L-GMSTP used in developing an approximation algorithm for the problem is described in what

it follows:

$$Z_{IP} = \min \quad \sum_{e \in E} c_e x_e$$
$$s.t. \quad y(V_k) \geq 1, \qquad \forall\, k \in K = \{1, ..., m\}$$
$$x(\delta(S)) \geq y_i, \qquad \forall\, S \subset V \setminus \{i\}$$
$$x_e \in \{0, 1\}, \qquad \forall\, e \in E$$
$$y_i \in \{0, 1\}, \qquad \forall\, i \in V.$$

In this new integer programming formulation, the constraint (3) is omitted, because it is redundant under the assumption that the edges have a non-negative cost.

We consider the linear programming (LP) relaxation of this integer program obtained by replacing the integrality constraints $x_e \in \{0, 1\}$, for all $e \in E$ and $y_i \in \{0, 1\}$, for all $i \in V$ by the constraints $x_e \in [0, 1]$, for all $e \in E$ and $y_i \in [0, 1]$, for all $i \in V$.

4. AN APPROXIMATION ALGORITHM FOR THE L-GMSTP

In this section we present an approximation algorithm for the L-GMSTP under the following two assumptions:

1. The cost function $c : E \to \mathbb{R}_+$ attached to the edges of G satisfies the triangle inequality:

$$c_{ij} \leq c_{ik} + c_{kj} \qquad (i, j, k \in V)$$

2. The clusters are bounded:

$$|V_k| \leq \rho, \qquad k \in K$$

for some $\rho > 0$.

For this class of problem instances we can efficiently construct a solution with cost at most 2ρ times the optimum. The design of the approximation algorithm is based on solving the LP relaxation of the integer programming formulation of the L-GMSTP and round the fractional solution to a nearby integral one. Our technique of rounding is based on properties of optimal solutions to the linear programming formulation of the minimum spanning tree problem and the parsimonious property of Goemans and Bertsimas, see [4].

Let $(y^*, x^*, Z^*_{LP}) = ((y^*_i)^n_{i=1}, (x^*_e)_{e \in E}, Z^*_{LP})$ be the optimal fractional solution of the LP relaxation. Obviously, the following inequality between the value of the optimal fractional solution of the LP relaxation and the value of the optimal solution of the IP formulation holds:

$$Z^*_{LP} \leq Z_{IP}.$$

The LP relaxation can be solved in polynomial time (relative to the input size of the problem) using the ellipsoid method [5] or interior point methods [12].

Because the cluster sizes are assumed to be bounded by ρ, i.e. $|V_k| \leq \rho$, $k \in K$, there exists in each cluster V_k at least one node $v \in V_k$ such that

$$y^*(v) \geq \frac{1}{\rho}.$$

Let $W = \{v_1, ..., v_m, ..., v_p\} \subseteq V$ denote the set of chosen nodes. We now compute the minimum cost tree spanning the nodes of W and claim that this tree, which is a feasible (approximate) solution of the L-GMSTP, denoted $T(W)$, has the cost at most 2ρ times the optimum of the IP formulation of L-GMSTP, Z_{IP}. More precisely, we show that

THEOREM 1. *The performance ratio for approximating the optimum solution to the L-GMSTP satisfies:*

$$c(T(W)) \leq (2 - \tfrac{2}{n})\rho Z_{IP}.$$

5. PROOF OF CORRECTNESS

The crucial argument in providing the approximation algorithm is based on the parsimonious property, see [4].

Given a complete undirected graph $G = (V, E)$. We associate with each edge $(i, j) \in E$ a cost c_{ij} and for any pair (i, j) of vertices, let r_{ij} be the connectivity requirement between i and j (r_{ij} is assumed to be symmetric, i.e. $r_{ij} = r_{ji}$). A network is called *survivable* if it has at least r_{ij} edge disjoint paths between any pair (i, j) of vertices.

The *survivable network design problem* consists in finding the minimum cost survivable network. This problem can be formulated by the following integer program:

$$
\begin{aligned}
(IP_\emptyset(r)) \quad IZ_\emptyset(r) = \quad &\min \quad \sum_{e \in E} c_e x_e \\
(6) \qquad\qquad &s.t. \quad x(\delta(S)) \geq \max_{(i,j) \in \delta(S)} r_{ij}, \ S \subset V, \ S \neq \emptyset \\
& \qquad\quad\, 0 \leq x_e, \qquad\qquad e \in E, \\
& \qquad\quad\, x_e \text{ integral}, \qquad e \in E.
\end{aligned}
$$

We denote by $IZ_\emptyset(r)$ the optimal value of the above integer program. Let $(P_\emptyset(r))$ denote the linear programming relaxation of $(IP_\emptyset(r))$ obtained by dropping the integrality restrictions and let $Z_\emptyset(r)$ be its optimal value.

By definition the degree of vertex $i \in V$ is $d_x(i) = x(\delta(i))$, for any feasible solution x, either to $(IP_\emptyset(r))$ or to $(P_\emptyset(r))$. Because of constraints (6) for $S = \{i\}$, the degree of vertex i is at least equal to $\max_{j \in V \setminus \{i\}} r_{ij}$. If $d_x(i) = \max_{j \in V \setminus \{i\}} r_{ij}$, then we say that x is *parsimonious* at vertex i. If we impose that the solution x is parsimonious at all the vertices of a set $D \subseteq V$, we get some interesting variations of $(IP_\emptyset(r))$ and $(P_\emptyset(r))$, denoted by $(IP_D(r))$ and

$(P_D(r))$, respectively. The formulation of $(IP_D(r))$ as an integer program is:

$$(IP_D(r)) \quad IZ_D(r) = \quad \min \quad \sum_{e \in E} c_e x_e$$

$$\text{s.t.} \quad x(\delta(S)) \geq \max_{(i,j) \in \delta(S)} r_{ij}, \ S \subset V, \ S \neq \emptyset$$

$$x(\delta(i)) = \max_{j \in V \setminus \{i\}} r_{ij}, \ i \in D, \ D \subseteq V$$

$$0 \leq x_e, \qquad e \in E$$

$$x_e \text{ integral}, \qquad e \in E.$$

We denote by $IZ_D(r)$ the optimal value of the above integer program. Let $(P_D(r))$ denote the linear programming relaxation of $(IP_D(r))$ obtained by dropping the integrality restrictions and let $Z_D(r)$ be its optimal value.

THEOREM 2. (parsimonious property, Goemans and Bertsimas [4])
If the costs c_{ij} satisfy the triangle inequality, then

$$Z_\emptyset(r) = Z_D(r)$$

for all subsets $D \subseteq V$.

The proof of this theorem is based on a result on connectivity properties of Eulerian multigraphs.

Let now $W \subseteq V$ and consider the following linear program:
Problem LP2:

$$Z_2^*(W) = \quad \min \quad \sum_{e \in E} c_e x_e$$

(7) $\text{s.t.} \quad x(\delta(S)) \geq 1, \ S \subset V, \text{ s.t. } W \cap S \neq \emptyset \neq W \setminus S$

(8) $x(\delta(i)) = 0, \ \ i \in V \setminus W,$

(9) $0 \leq x_e \leq 1, \ e \in E.$

Replacing constraints (9) with the integrality constraints $x_e \in \{0,1\}$, the formulation obtained is the formulation of the minimum tree spanning the subset of nodes $W \subset V$.

Consider the following relaxation of the problem LP2.
Problem LP3:

$$Z_3^*(W) = \quad \min \quad \sum_{e \in E} c_e x_e$$

$$\text{s.t.} \quad x(\delta(S)) \geq 1, \ S \subset V, \text{ s.t. } W \cap S \neq \emptyset \neq W \setminus S$$

(10) $0 \leq x_e, \qquad e \in E.$

Thus we omitted constraint (8) and relaxed constraint (9).

The following result is a straightforward consequence of the parsimonious property, if we choose $r_{ij} = 1$, if $i, j \in W$, and 0 otherwise, and $D = V \setminus W$.

LEMMA 3. *The optimal solution values to linear programming problems* LP2 *and* LP3 *are the same, that is*

$$Z_2^*(W) = Z_3^*(W).$$

Consider the following integer programming problem:

Problem IP4:

$$Z_4 = \min \sum_{e \in E} c_e x_e$$

(11) $s.t. \quad x(\delta(S)) \geq 1, \; S \subset V, \; \text{s.t.} \; S \neq \emptyset \neq V$

(12) $x_e \in \{0,1\}, \qquad e \in E.$

Clearly, it is the integer programming formulation of the MST (minimum spanning tree) problem. Let LP4 be the LP relaxation of this formulation, that is, we simply replace the constraint (12) by the constraint $0 \leq x_e \leq 1$, for all $e \in E$.

Denote by Z_4^* the value of the optimal solution of the LP4. The following known result for minimum spanning trees holds:

PROPOSITION 4.

$$c(T(V)) \leq (2 - \tfrac{2}{|V|})Z_4^*,$$

where $c(T(V))$ denotes the cost of the minimum spanning tree on V.

Proof. See for example [4].   □

Let $W \subseteq V$, then Proposition 4 can be easily modified to obtain:

PROPOSITION 5.

$$c(T(W)) \leq (2 - \tfrac{2}{|W|})Z_2^*(W).$$

Proof. Let (x_e) be a feasible solution to linear program LP2. If

$$e \notin E(W) = \{(i,j) \mid i,j \in W\}$$

implies that $x_e = 0$ and using Proposition 4 we prove the inequality.

Consider that (y^*, x^*, Z_{LP}^*) is the optimal solution to the linear programming relaxation of the L-GMSTP, then we define

$$\widehat{x}_e = \rho x_e^*$$

$$\widehat{y}_i = \begin{cases} 1 & \text{if } y_i^* \geq \tfrac{1}{\rho} \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$W = \left\{ i \in V \mid y_i^* \geq \tfrac{1}{\rho} \right\} = \left\{ i \in V \mid \widehat{y}_i = 1 \right\}.$$

Now, let us show that $(\widehat{x}_e)_{e \in E}$ is a feasible solution to LP3. Indeed, $\widehat{x}_e \geq 0$ for all $e \in E$, hence condition (10) is satisfied. Let $S \subset V$ be such that

$W \cap S \neq \emptyset \neq W \setminus S$ and choose some $i \in W \cap S$. Hence $\widehat{y}_i = 1$ and $y_i^* \geq \frac{1}{\rho}$. Then we have

$$\widehat{x}(\delta(S)) = \sum_{e \in \delta(S)} \widehat{x}_e = \rho \sum_{e \in \delta(S)} x_e^* \geq \rho y_i^* \geq \rho \frac{1}{\rho} = 1,$$

by definition of $\widehat{x}_e$ and the fact that the (x_e^*) solve the linear programming relaxation of the L-GMSTP. Therefore $(\widehat{x}_e)$ satisfy constraint (7) in LP3.

Now we are able to prove the performance bounds for the approximation algorithm that we proposed for the L-GMSTP:

$$
\begin{aligned}
c(T(W)) &\leq (2 - \tfrac{2}{|W|})Z_2^*(W) = (2 - \tfrac{2}{|W|})Z_3^*(W) \\
&\leq (2 - \tfrac{2}{|W|}) \sum_{e \in E} c_e \widehat{x}_e = (2 - \tfrac{2}{|W|})\rho \sum_{e \in E} c_e x_e^* = (2 - \tfrac{2}{|W'|})\rho Z_1^* \\
&\leq (2 - \tfrac{2}{|W|})\rho Z_{LP} = (2 - \tfrac{2}{|W|})\rho Z_{IP}.
\end{aligned}
$$

And since $W \subseteq V$, that is, $m \leq |W| \leq |V| = n$, we have proved the following:

$$c(T(W)) \leq (2 - \tfrac{2}{n})\rho Z_{IP} \leq 2\rho Z_{IP}.$$

$\square$

6. CONCLUSIONS

For a special class of the L-GMSTP instances, i.e. graphs with the cost function attached to the edges of the graph satisfying the triangle inequality and the clusters having bounded sizes by ρ, we provided an approximation algorithm for the problem which delivers a solution with cost at most 2ρ times the optimum.

REFERENCES

[1] DROR, M., HAOUARI, M. and CHAOUACHI, J., *Generalized Spanning Trees*, European Journal of Operational Research, **120**, pp. 583–592, 2000.

[2] FEREMANS, C., *Generalized Spanning Trees and Extensions*, PhD thesis, Universite Libre de Bruxelles, Belgium, 2001.

[3] FEREMANS, C., LABBE, M. and LAPORTE, G., *A Comparative analysis of several formulations of the generalized minimum spanning tree problem*, Networks, **39**, no. 1, pp. 29–34, 2002.

[4] GOEMANS, M.X. and BERTSIMAS, D.J., *Survivable networks, linear programming relaxations and parsimonious property*, Mathematical programming, **60**, pp. 145–166, 1993.

[5] GRÖTSCHEL, M., LOVÁSZ, L. and SCHRIJVER, A., *Geometric Algorithm and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.

[6] MYUNG, Y.S., LEE, C.H. and TCHA, D.W., *On the generalized minimum spanning tree problem*, Networks, **26**, pp. 231–241, 1995.

[7] POP, P.C., *The Generalized Minimum Spanning Tree Problem*, PhD thesis, University of Twente, The Netherlands, 2002.

[8] POP, P.C., *New models of the generalized minimum spanning tree problem*, Journal of Mathematical Modelling and Algorithms, **3**, no. 2, pp. 153–166, 2004.

[9] POP, P.C., *On the Prize-Collecting Generalized Minimum Spanning Tree Problem*, to appear in Annals of Operations Research.

[10] POP, P.C., KERN, W. and STILL, G., *Approximation theory in combinatorial optimization. Application to the generalized minimum spanning tree problem*, Rev. Anal. Numér. Theor. Approx., **34**, no. 1, pp. 93–102, 2005.

[11] POP, P.C., KERN, W. and STILL, G., *A new relaxation method for the generalized minimum spanning tree problem*, European Journal of Operational Research, **170**, pp. 900–908, 2006.

[12] ROOS, C., TERLAKY, T. and VIAL, J.-Ph., *Theory and Algorithms for Linear Optimization. An Interior Point Approach*, John Wiley & Sons, 1997.