SHORTEST PATHS, NETWORK DESIGN, AND ASSOCIATED POLYHEDRA

Thomas L. Magnanti
Sloan School of Management, MIT, Cambridge

and

Prakash Mirchandani
Katz Graduate School of Business, University of Pittsburgh, Pittsburgh

April 1990

Shortest Paths, Network Design, and Associated Polyhedra

by

Thomas L. Magnanti & Prakash Mirchandani

OR 215-90 April 1990

Abstract

We study a specialized version of network design problems that arise in telecommunication, transportation and other industries. The problem, a generalization of the shortest path problem, is defined on an undirected network consisting of a set of arcs on which we can install (load), at a cost, a choice of up to three types of capacitated facilities. Our objective is to determine the configuration of facilities to load on each arc that will satisfy the demand of a single commodity at the lowest possible cost.

Our results (i) demonstrate that the single-facility loading problem and certain "common breakeven point" versions of the two-facility and three-facility loading problems are polynomially solvable as a shortest path problem; (ii) show that versions of the two-facility loading problem are strongly NP-hard, but that a shortest path solution provides an asymptotically "good" heuristic; and (iii) characterize the optimal solution (that is, specify a linear programming formulation with integer solutions) of the common breakeven point versions of the two-facility and three-facility loading problems. In this development, we introduce two new families of facets, give geometric interpretations of our results, and demonstrate the usefulness of partitioning the space of the problem parameters to establish polyhedral integrality properties. Generalizations of our results apply to (i) multicommodity applications and (ii) situations with more than three facilities.

Keywords: Shortest paths, multiple capacitated facilities, polyhedral structure, convex hull.

Abbreviated title: SHORTEST PATHS AND NETWORK DESIGN.

1. INTRODUCTION

Although we don't typically think of it this way, the shortest path problem is a special case of a more general fixed charge network design problem. Consider a network flow problem with a single commodity, single source, and single destination. Suppose we wish to design a network to send a unit of flow of this commodity from the source to the destination nodes. Moreover, suppose we can install (load) integer multiples of a unit capacity facility on each arc {i,j} of the network, incurring a per unit cost a_{ij} for each unit of the facility. We wish to design a network at minimal cost that has the capacity to meet the flow requirements for the given commodity. This design problem is easy to solve: we solve a shortest path problem from the commodity's source node to its destination node with respect to the arc costs a_{ij} and load one unit of the facility on each arc of the shortest path.

Now suppose that we need to send d (an integer) units of the commodity from its source to its destination and that we can load multiple types of facilities on each arc; that is, arc capacities are now available in several base capacities and we can install integer multiples of any base capacity, at a per unit cost, on each arc. We refer to this generalization of the shortest path problem as the *single commodity network loading problem* since we are loading the network with facilities, at a cost, to carry the required flow. The transition from the original shortest path problem to this more general setting raises several questions. Is the problem still easily solvable? Can we solve it as a shortest path problem? Can we formulate a linear programming model whose extreme point solutions satisfy the integrality restrictions of the problem? In this paper we consider these issues. In particular, we consider three versions of the problem: (i) one with a single base level of capacity equal to C units, (ii) one with two base levels of capacity, a low level (LC) equal to one unit and a medium level (MC) equal to C units, and (ii) one with three base levels of capacity, low (LC), medium (MC) and high (HC) equal to one unit, C units and λC units. We assume that both C and λ are integers greater than one, and refer to these three versions of the problem as the one-facility (1F), two-facility (2F), and three-facility (3F) loading problems, respectively. As we will see, certain versions the two-facility and three-facility problems are strongly NP-hard, and other versions can be solved efficiently. Moreover, for these efficiently solved versions of the problem we are able to offer affirmative answers to the questions we have posed.

This study is motivated by a set of network design problems we have encountered in the telecommunications industry (see Magnanti, Mirchandani, and Vachani [12]) and related problems that arise in transportation freight flow planning (for example, see Leung, Magnanti, and Singhal [11]). In the telecommunications industry, data transmission lines are available in several service types, for example, digital service type zero, or DS0, lines and digital service type one, or DS1, lines whose capacity is 24 times that of a DS0 line. In freight flow applications, trucks on any transportation link might be available in multiple capacities: for example, 24 foot trailers or 48 foot trailers. These applications typically have many commodities, so the problem we are considering in this paper arises as a subproblem.

More general fixed charge network design problems arise in many application contexts, notably telecommunications, computer networking, facility location, production planning, and transportation. (For examples and for a discussion of the underlying methodology, we refer the reader to surveys by Magnanti and Wong [14] and Minoux [17].) Furthermore, some classical combinatorial optimization problems such as the traveling salesman problem, the minimum spanning tree problem, and the Steiner tree problem are special cases of the general network design problem. Consequently, the study of generic network design problems could yield theoretical, algorithmic, and practical insight that might cut across a wide variety of problem domains. Our hope is that the results presented in this paper might not only resolve (partially) the questions we have posed, but might also contribute to a better understanding of more general design models.

Because of its importance, the network design problem has attracted substantial attention in recent years. In a more general form than we are considering, the problem associates two kinds of costs with each arc: (i) variable (flow) costs that depend upon total arc flow volume, and (ii) fixed charges that determine the level of installed capacity on the arc. Researchers have assumed a variety of functional forms for either cost, including models with one of the costs equal to zero. Many researchers have focused on a flow cost function that is concave and nondecreasing. A concave functional form, which reflects efficiencies of scale and volume discounts, arises often in the transportation and telecommunication industries. Zangwill [23] has studied the minimum concave cost flow problem and demonstrated how it captures the concave cost warehouse location problem, the single and multiproduct production and inventory models and the plant location problem. Yaged [21], Zadeh [22], and Minoux [16], among others, have studied this problem in the context of the telecommunication industry, and Balakrishnan and Graves [2]

have studied the problem in the context of freight flow planning. (See also Powell and Sheffi [19].)

The fixed charge network flow problem associates both fixed costs with installing capacity on the arcs and linear flow costs. Balinski [4] and Gray [8] have studied the specialized fixed charge transportation version of the problem. Balakrishnan, Magnanti, and Wong [3] have suggested a dual ascent approach that has been successful in solving large scale uncapacitated fixed charge transshipment problems.

Our model assumes a piecewise staircase form for fixed costs and no flow costs. This cost function is closely related to the ones considered by Goldstein and Rothfarb [7], and Magnanti, Mirchandani, and Vachani [12]. The first set of these authors have studied the single-source multiple-destination problem and discussed properties of the optimal solution. The model considered by the second set of authors is more general and allows for commodity demand between every pair of nodes. These authors have developed a polyhedral based approach for solving this problem. Padberg, Van Roy, and Wolsey [18] have studied the polyhedral properties of a core single node fixed charge problem. LeBlanc and Simmons [10] have assumed nonzero flow costs in their model, but allowed capacity to be available at continuous levels.

We focus on a single commodity version of the problem for situations with up to three types of facilities. The economies of scale in the tariff structure of these facilities implies that the cost function is neither convex nor concave; moreover, as we will see, the optimal solution does not inherit the nice extremal flow property that characterizes models with concave cost flows.

We assume our problem is defined over a network G = (N,A) with node set N and undirected arc set A. Let **a**, **b** and **c** be real vectors of dimension |A|, whose components equal, respectively, the cost of loading each unit of the LC, MC and HC facilities. The first breakeven point of arc {i,j}, $m_{ij}^1 = b_{ij}/a_{ij}$, is the ratio of the cost of loading a MC facility to the cost of loading a LC facility. Similarly, the second breakeven point, $m_{ij}^2 = c_{ij}/a_{ij}$, is the ratio of the cost of loading a HC facility to the cost of loading a LC facility. For the most part, we assume that $1 < m_{ij}^1 < C$ and $m_{ij}^1 < m_{ij}^2 < \lambda m_{ij}^1$; otherwise, the optimal solution need not consider any MC and/or HC facilities.

This paper studies the one-facility, two-facility, and three-facility variations of the single commodity loading problem in increasing order of difficulty. Section 2 introduces our notation and model formulation and Sections 3, 4, and 5 study the three loading problems. In Section 3, we tighten the original formulation of the 1F problem by adding a class of facets and show that this enhanced linear programming formulation describes the convex hull of feasible solutions: in the enhanced problem formulation, the extreme points corresponding to the arc loading variables are all integer, and in a projected lower-dimensional space containing only these variables, the formulation completely describes the convex hull of the integer feasible solutions. We also relate this polyhedron to the shortest path solution to the problem. Section 4 describes a heuristic for the 2F loading problem that generates "good" solutions in the sense that the relative error of the heuristic solution goes to zero as the demand, d, approaches infinity. We next discuss some variations of the 2F loading problem that are strongly NP-hard. In Section 5, we introduce two new classes of facets for the 3F loading problem. These facets are useful for generating a linear program that has an optimal solution with integer values for the arc design variables $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ when the breakeven points m^1_{ij} and m^2_{ij} are the same on all arcs $\{i,j\}$. In our proof, we demonstrate the use of partitioning the space of problem parameters to identify optimal primal and dual solutions. We also show how to generalize these classes of facets for a broader class of problems. Section 6 concludes the paper with some possibilities for future research directions.

2. Notation and Model Formulations

This section introduces our notation and describes the basic ingredients of our model. Let f_{ij} (f_{ji}) denote the flow of the commodity from i to j (j to i) on arc $\{i,j\}$ and let x_{ij}, y_{ij} and z_{ij} denote the number of LC, MC and HC facilities loaded on arc $\{i,j\}$. In principle, the design variables x_{ij}, y_{ij} and z_{ij} could be unbounded; however, in practice we can bound the feasible set of design vectors by a sufficiently large integer, say L, without altering the problem in any essential way.

We define an (undirected) cutset $\{S,T\}$ by a partitioning of the node set N into two nonempty disjoint sets $S \subset N$ and $T = N \backslash S$. An arc $\{i,j\}$ belongs to cutset $\{S,T\}$ if nodes i and j belong to different sets S and T. If the origin node O and the destination node D belong to different sets S and T, we refer to $\{S,T\}$ as an O-D cutset . We also define

aggregate design variables for each cutset {S,T}: $X_{S,T}$ equals the total number of LC facilities loaded on the cutset arcs, i.e., $X_{S,T} = \Sigma_{\{i,j\} \in \{S,T\}} x_{ij}$, and $Y_{S,T}$ and $Z_{S,T}$ equal the total number of MC and HC facilities loaded on this cutset. $D_{S,T}$ denotes the total demand from the set S to the set T; note that $D_{S,T}$ equals either d or 0 depending on whether {S,T} is an O-D cutset or not. We let $r_{S,T} = D_{S,T} \bmod(C)$ and suppress the subscripts when doing so would not seem to cause any confusion. We adopt the convention that $r_{S,T}$ equals C when $D_{S,T}$ is a multiple of C.

Lastly, for $\mu \in \Re^1$, we define $\mu^+ = \max(\mu,0)$ and $\mu^- = \min(\mu,0)$.

Using this notation we can formulate the 3F loading problem as the following mixed-integer program.

[Problem P(IP3)]:

$$\text{minimize} \sum_{\{i,j\} \in A} (a_{ij}x_{ij} + b_{ij}y_{ij} + c_{ij}z_{ij})$$

subject to:

$$\sum_{j \in N} f_{ji} - \sum_{j \in N} f_{ij} = \begin{cases} -d \text{ if } i = O \\ d \text{ if } i = D \\ 0 \text{ otherwise} \end{cases}$$

$$f_{ij} + f_{ji} \le x_{ij} + Cy_{ij} + \lambda Cz_{ij} \quad \text{for all } \{i,j\} \in A$$

$$\left.\begin{array}{l} x_{ij} \le L \\ y_{ij} \le L \\ z_{ij} \le L \end{array}\right\} \quad \text{for all } \{i,j\} \in A$$

$$x_{ij}, y_{ij}, z_{ij} \in Z_+^1 , \quad f_{ij}, f_{ji} \in \Re_+^1 \text{ for all } \{i,j\} \in A.$$

In this formulation, the objective is to minimize the total cost incurred in loading the LC, MC and HC facilities on all the arcs. The first set of constraints are the usual *flow conservation* constraints: they ensure that the flow conforms to the mass balance requirements of each node. The second set of constraints guarantee that the total flow on an arc does not exceed the total installed capacity on that arc. We call these constraints the *capacity* constraints. Finally, in addition to the *upper bounding* constraints, we require that the design variables, x_{ij}, y_{ij} and z_{ij} are nonnegative integers, and the flow variables f_{ij} are nonnegative.

In the 1F specialization of this problem, the formulation does not contain either of the x_{ij} and z_{ij} variables, and in the 2F specialization, it does not contain the z_{ij} variables. (We refer to these formulations as P(IP1) and P(IP2) respectively.)

The max flow-min cut theorem permits us to recast the formulation P(IP3) by projecting it into the subspace of design variables. Given values for the vectors, **x**, **y** and **z**, the problem has a feasible flow of d units from O to D if and only if the capacity of every O-D cutset is at least d. In other words, the problem has a feasible flow if and only if the design variables satisfy the *aggregate-capacity demand inequality* $X_{S,T} + CY_{S,T} + \lambda CZ_{S,T} \geq d$ for all O-D cutsets $\{S,T\}$. In this alternate formulation, the model becomes

[Problem P(CUT3)]:

$$\text{minimize} \sum_{\{i,j\} \in A} (a_{ij}x_{ij} + b_{ij}y_{ij} + c_{ij}z_{ij})$$

subject to:

$$X_{S,T} + CY_{S,T} + \lambda CZ_{S,T} \geq d \text{ for all O-D cusets } \{S,T\}$$

$$\left.\begin{array}{l} x_{ij} \leq L \\ y_{ij} \leq L \\ z_{ij} \leq L \end{array}\right\} \text{ for all } \{i,j\} \in A$$

$$x_{ij},\ y_{ij},\ z_{ij} \in Z_+^1 \text{ for all } \{i,j\} \in A.$$

We refer to the one-facility version of this formulation without the variables **x** and **z** as P(CUT1) and the two-facility version without the variables **z** as P(CUT2).

For the 1F loading problem, the aggregate-capacity demand inequality reduces to $Y_{S,T} \geq D_{S,T}/C = d/C$ if $\{S,T\}$ is an O-D cutset. But since the left hand side of this inequality is an integer, the inequality remains valid if we round the right hand side up to the next nearest integer as well. This integrality argument implies that the *cutset* inequality $Y_{S,T} \geq \lceil D_{S,T}/C \rceil$ is valid for P(IP1). A special case of a more general result of Magnanti, Mirchandani, and Vachani [12] shows that this inequality in fact defines a facet of formulation P(IP1) whenever $D_{S,T} > 0$ and the subgraphs defined by S and T are connected. Note that if d mod(C) equals C, then $Y_{S,T} \geq \lceil D_{S,T}/C \rceil$ is redundant since it is implied by the flow conservation and the capacity constraints. In Section 3 we show that we can use this inequality to define the convex hull of the formulation P(CUT1) in the

subspace of the design variables. A generalization of the cutset inequalities applies to the 2F and the 3F cases, and define facets of the underlying polyhedra (see Magnanti, Mirchandani, and Vachani [12]). Sections 4 and 5 study these inequalities and show how they tighten the formulations P(IP2) and P(IP3).

To this point we have formulated the network loading problem in two ways: one model contains both the flow variables **f** and arc design variables **x**, **y** and **z**, and one model contains only the design variables. It is also instructive to model the problem with only the flow variables as well.

[Problem P(**f**)]:

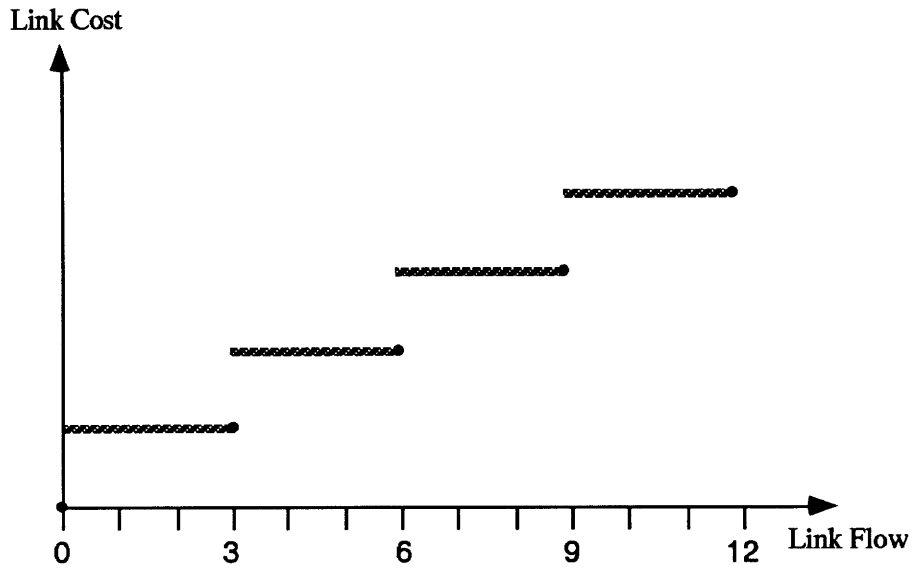
$$\text{minimize} \sum_{\{i,j\} \in A} \phi_{ij}(f_{ij} + f_{ji})$$

subject to:

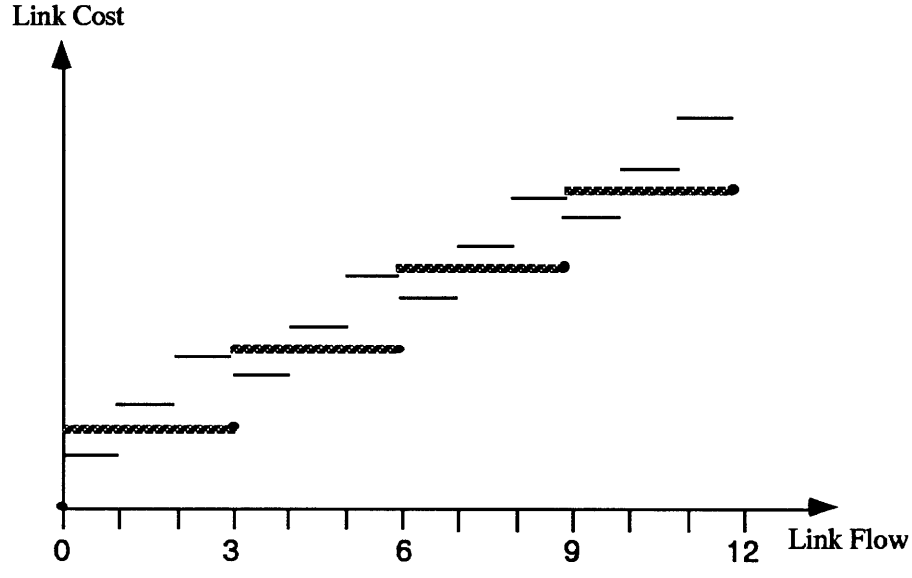
$$\sum_{j \in N} f_{ji} - \sum_{j \in N} f_{ij} = \begin{cases} -d \text{ if } i = O \\ d \text{ if } i = D \\ 0 \text{ otherwise} \end{cases}$$

$$f_{ij}, f_{ji} \in \Re_+^1 \text{ for all } \{i,j\} \in A.$$

In this formulation $\phi_{ij}(\cdot)$ denotes the cost function on arc {i,j} as a function of the total flow on that arc. For the network loading problem, this cost function is neither convex or concave. Figures 1(i) and 1(ii) show the structure of this cost function for the one-facility and two-facility versions of the problem (with C=3).



(i) Single Facility



(ii) Two Facilities

Figure 1. Arc Cost for Network Loading

Notice from Figure 1 that for the one-facility loading problem and any nonnegative integer k, the cost function is constant on the interval $kC<f\leq(k+1)C$. For the two-facility problem, the cost structure over the interval $0<f\leq C$ is a step function of unit width for the small facility and a single step of the cost function of the medium facility. The optimal cost on this interval is the lower envelop of these two cost functions. At the end of the interval $0<f\leq C$, we can again begin to use the small capacity facility, but now together with a single unit of the medium facility. Therefore, the cost function on the interval $C<f\leq 2C$ replicates the cost function on the interval $0<f\leq C$, except that it has the added cost of a single medium capacity facility. Similarly, the cost function self-replicates itself on each subsequent interval of size C.

Observe that this flow formulation of the problem places all the model complexities in the objective function; the constraints are very simple: they define a shortest path polytope whose extreme points correspond to spanning trees in the underlying network with a flow of d units on the unique path in the tree joining the source and destination nodes

3. THE ONE-FACILITY LOADING PROBLEM

We next consider the polyhedral structure of the 1F model obtained by appending the cutset inequalities $Y_{S,T} \geq \lceil D_{S,T}/C \rceil$ to the formulation P(IP1). As we will see, the one-facility version of the problem is easy to analyze since it essentially is a shortest path problem. To solve the problem we simply solve a shortest path problem from the source to destination nodes with respect to the arc costs $\mathbf{b}^+$. Let $\text{len}(\mathbf{b}^+)$ denote the shortest path length. We next load $\lceil d/C \rceil$ MC facilities on all arcs of this path with $b_{ij} \geq 0$ and L MC facilities on all arcs of the network with $b_{ij} < 0$. We also send a flow of d units on the facilities loaded on the shortest path. We refer to this solution as the *loaded shortest path solution*. The cost of this solution is $\lceil d/C \rceil * \text{len}(\mathbf{b}^+) + L\Sigma_{\{i,j\} \in A} b^-_{ij}$.

We will argue for the validity of this approach in two ways: a direct approach based upon redirecting flow onto a shortest path and then an indirect approach via linear

programming duality. The indirect approach will set the stage for our subsequent analysis of the two-facility and three-facility versions of the problem.

We first might make a preliminary observation. Consider the problem P(IP1) assuming that C=1, that is, the formulation P(IP3) without the **y** or **z** variables. Notice that it is easy to remove the **x** variables from the linear programming relaxation of this formulation: if $a_{ij} < 0$, we set $x_{ij} = L$, and if $a_{ij} \geq 0$ we set $x_{ij} = f_{ij} + f_{ji}$. By making these associations, we can eliminate the **x** variables and if d = 1 the resulting formulation in the **f** variables becomes the standard formulation of the shortest path problem. Therefore, for any choice of the cost vector **a**, the problem has a solution with a 0-d flow vector **f**. Setting $x_{ij} = L$ or $x_{ij} = f_{ij} + f_{ji}$, depending upon the sign of a_{ij}, we see that if d and L are integral, then the linear programming relaxation of the problem always has an integral solution in the **x** variables. Consequently, the values of the **x** variables are integral in every extreme point to the polyhedron defined by relaxing the integrality constraints of the problem. This same argument applies to the formulation P(IP1) in the **y** variables whenever d is a multiple of C and both are positive integers since in this case $y_{ij} = (f_{ij} + f_{ji})/C$ and $f_{ij} + f_{ji} = 0$ or d in the solution to the shortest path problem that arises after we eliminate the **y** variables. Since the **y** components of every extreme point in this polyhedron are integer, its projection into the space of the **y** variables has integer extreme points. Let us record this result formally for later reference.

Lemma 3.1. *If d is a multiple of C and both C and d are integers, then the linear programming relaxation of the cutset formulation P(CUT1) has integer extreme points and the values of the variables* **y** *are integer in every extreme point to the polyhedron defined by the linear programming relaxation of P(IP1).*

Direct Approach for Establishing Optimality of the Loaded Shortest Path Solution

We wish to show that for any possible arc cost vector **b**, the one-facility problem has a loaded shortest path solution.

First, note that for any feasible solution **f** to the flow balance equations of P(IP1), it always is cost effective to set $y_{ij} = L$ on every arc {i,j} with $b_{ij} < 0$, and $y_{ij} = \lceil (f_{ij} + f_{ji})/C \rceil$ for every arc {i,j} with arc cost $b_{ij} \geq 0$. By always choosing the arc design variables **y** in this way, we can consider the problem as formulated solely in terms of the flow variables **f**.

Recall that the flow decomposition property of network flows (e.g., see Ahuja, Magnanti, and Orlin [1]) implies that it is possible to express the flow variables **f** in any feasible solution to the flow formulation P(**f**) as the sum of flows on paths from the source to the destination nodes plus flows on cycles. We claim that, without loss of generality, we can assume that

(i) no two paths in this flow decomposition carry C or more units of flow; and

(ii) the subgraph G′ of G corresponding to those arcs whose flow value is not a multiple of C contains at most one path joining the source and the destination nodes.

To establish property (i), suppose that the flow decomposition contained two paths carrying flow of C or more units. By reassigning C units flow from the more costly to the less costly of any two such paths (or between any two paths that tie in cost) with respect to the cost vector $\mathbf{b}^+$ (and by redefining the design variables as described in the last paragraph), we obtain a solution whose cost is at least as small as that of the given solution. Therefore, we can assume that only one path has a flow of C or more units. This path must be a shortest path with respect to the arc costs $\mathbf{b}^+$, otherwise we could define a more cost effective solution.

To establish property (ii), suppose that for a given solution **f**, the network G′ contains two paths P_1 and P_2 which we will view as directed from O to D. Suppose that the length of path P_1 with respect to the cost vector $\mathbf{b}^+$ is less than or equal to the length of path P_2. Let $\alpha = \min \{f_{ij}: (i,j) \in P_1 \backslash P_2\}$ and let $\beta = \min \{C - f_{ij} \bmod(C): (i,j) \in P_2 \backslash P_1\}$, and $\gamma = \min\{\alpha,\beta\}$. If we redefine the flow on the paths P_1 and P_2 as $g_{ij} = f_{ij} - \gamma$ for all $(i,j) \in P_1 \backslash P_2$ and $g_{ij} = f_{ij} + \gamma$ for all $(i,j) \in P_2 \backslash P_1$, then the resulting solution has a cost no more than the cost of the original solution **f**. With respect to the new solution **g**, the network G′ contains at least one fewer arc. Therefore, by repeating this argument we can find a solution whose cost is no more than that of the original solution **f** and for which the network G′ contains at most one path joining the source and destination nodes.

Next note that properties (i) and (ii) imply that we can assume that any candidate optimal flow decomposes into at most two paths, one P_1 whose flow is a multiple of C and one P_2 whose flow is not a multiple of C. We can however reassign the flow of both paths onto the path P_1 or P_2 which has the lower cost with respect to the arc cost vector $\mathbf{b}^+$.

This conclusion shows that the loaded shortest path solution is optimal. This result, in turn, provides us with a complete polyhedral description of the convex hull of the integer solutions of cutset formulation P(CUT1).

Proposition 3.2. *The convex hull of the solutions to P(CUT1) is defined by the inequalities* $Y_{S,T} \geq \lceil d/C \rceil$ *for all O-D cutsets* $\{S,T\}$, *and, the upper bounding and the nonnegativity constraints on the arc design variables. For all cost vectors* $\mathbf{b} \in \Re^{|A|}$, *the linear programming model*

[Problem P(FACET1)]:

$$\text{minimize} \sum_{\{i,j\} \in A} b_{ij} y_{ij}$$

subject to:

$$\sum_{j \in N} f_{ji} - \sum_{j \in N} f_{ij} = \begin{cases} -d \text{ if } i = O \\ d \text{ if } i = D \\ 0 \text{ otherwise} \end{cases}$$

$$f_{ij} + f_{ji} \leq C y_{ij} \text{ for all } \{i,j\} \in A$$

$$Y_{S,T} \geq \lceil d/C \rceil \text{ for all O-D cutsets } \{S,T\}$$

$$y_{ij} \leq L \text{ for all } \{i,j\} \in A$$

$$y_{ij},\ f_{ij},\ f_{ji} \in \Re^{1}_{+} \text{ for all } \{i,j\} \in A.$$

has an optimal solution with integer values for all the $\mathbf{y}$ *variables.*

Proof. Since the loaded shortest path solution is optimal, in terms of the design variables $\mathbf{y}$ and optimal value of the objective function, the problem remains unchanged if we replace the demand d by $C\lceil d/C \rceil$. But then d is a multiple of C and the result is a consequence of Lemma 3.1. ⊗

An Indirect Approach

Recall from the theory of blocking polyhedra (Fulkerson [6]) that the length of the shortest O-D path using the vector $\mathbf{u} \in \Re^{|A|}_{+}$ as arc lengths equals the maximum number of

O-D cuts that can be packed into the vector **u**. In other words, the length of the shortest O-D path equals

$$\text{maximize } \{\theta^t \mathbf{1} : \theta^t \mathbf{D} \le \mathbf{u},\ \theta \in \Re_+^{2^{**}(|N|-2)}\}.$$

In this expression, each component of the column vector θ corresponds to an O-D cutset, **D** is a matrix whose rows are the arc incidence vectors of O-D cuts, and **1** is a column vector of ones of the appropriate dimension. Notice that the dual to this linear program is the weighted minimum cut problem:

$$\text{minimize } \{\mathbf{u}\mathbf{y} : \mathbf{D}\mathbf{y} \ge \mathbf{1},\ \mathbf{y} \in \Re_+^{|A|}\}.$$

Nemhauser and Wolsey [20] show how to use Dijkstra's algorithm to assign optimal values to the components of θ. Although θ is exponential in size, the algorithm assigns values so that at most $|N|-1$ of its components are strictly positive. We will use this procedure to give an alternate proof of Proposition 3.2 and present this algorithm here, as algorithm SPP, for the sake of completeness. Balakrishnan, Magnanti, and Wong [3] start from the "flow" formulation of the shortest path problem as a network design problem (that is, P(IP1) with C=1) and develop a similar dual ascent method for assigning node potentials that equal the shortest path lengths from some source node. We might view these two procedures as alternate interpretations of Dijkstra's algorithm for different formulations of the shortest path problem.

Algorithm SPP operates as follows. It initializes all variables $\theta_{S,T}$ to 0 and starts with a set Q equaling the origin node O. It then finds the minimum cost arc, on say arc $\{i^*,j^*\}$, across the cutset $\{Q, N\backslash Q\}$ with node $i^* \in Q$. Next, it increases the variable corresponding to this cutset by the cost of $\{i^*,j^*\}$ and reduces the costs of all arcs across the cutset by the same quantity. Next, the algorithm transfers node j^* to the set Q and repeats this process until it has assigned the destination node D to the set Q. At this stage, the sum of all $\theta_{S,T}$ variables equals the shortest path length from O to D.

Algorithm SPP

Given a graph G=(N,A) and a nonnegative cost vector **u** associated with the arcs.

Step 0:

Initialize $Q = \{O\}$.

Set $\theta_{S,T} = 0$ for all cutsets $\{S,T\}$.

For any $Q \subset N$, $j \in N \backslash Q$; define $u_{Qj} = \min_{\{i,j\} \in \{Q, N\backslash Q\}} u_{ij}$.

Step 1:

Until $D \in Q$

do

(1a) determine $j^* = \text{argmin}_j u_{Qj}$;

(1b) If $u_{Qj^*} = \propto$, then print "no path from O to D"; stop;

(1c) $u_{ij} = u_{ij} - u_{Qj^*}$; for all $\{i,j\} \in \{Q, N\backslash Q\}$;

(1d) $\theta_{Q, N\backslash Q} = u_{Qj^*}$;

(1e) $Q = Q \cup \{j^*\}$;

end.

Step 2:

The shortest path length from O to D with components of **u** as the arc lengths =

$$\sum_{\{S,T\}} \theta_{S,T}.$$

We will use this algorithm and a primal-dual linear programming argument to provide an alternate proof of Proposition 3.2.

Alternate Proof of Proposition 3.2. Let $(\mathbf{y}^*, \mathbf{f}^*)$ be a loaded shortest path solution to the integer problem P(IP1). We will show that $(\mathbf{y}^*, \mathbf{f}^*)$ is an optimal solution to the linear program P(FACET1) by constructing a feasible solution to its dual D(FACET1) that has the same objective function value as P(FACET1). Let us denote the optimal value of the problems P(IP1), P(FACET1) and D(FACET1) by v^{IP1}, v^{FACET1} and $v^{DFACET1}$ respectively. Furthermore, let Conv(.) denote the convex hull of feasible solutions to problem P(.). Since $\text{Conv(IP1)} \subseteq \text{P(FACET1)}$, $v^{IP1} \geq v^{FACET1}$. By construction, we will show that $v^{IP1} \leq v^{DFACET1}$, and by duality, we know that $v^{DFACET1} \leq v^{FACET1}$. These inequalities will imply that $v^{IP1} = v^{FACET1}$.

We first construct a dual solution having this same objective value as the loaded shortest path solution. The dual to problem P(FACET1) is given by

[Problem P(DFACET1)]:

$$\text{maximize } dv_D + \sum_{O\text{-}D \text{ cutsets } \{S,T\}} \lceil d/C \rceil \mu_{S,T} - L \sum_{\{i,j\} \in A} \pi_{ij}$$

subject to:

$$\begin{aligned} v_i - v_j &\le w_{ij} \\ v_j - v_i &\le w_{ij} \end{aligned} \tag{3.1}$$

$$Cw_{ij} + \sum_{\substack{O\text{-}D \text{ cutsets } \{S,T\}, \\ \{i,j\} \in \{S,T\}}} \mu_{S,T} - \pi_{ij} \le b_{ij} \text{ for all } \{i,j\} \in A \tag{3.2}$$

$$\begin{aligned} &v_i \text{ u.i.s.} \quad \text{for all } i \in N \\ &\pi_{ij},\ w_{ij} \ge 0 \quad \text{for all } \{i,j\} \in A \\ &\mu_{S,T} \ge 0 \quad \text{for all O-D cutsets } \{S,T\}. \end{aligned}$$

In this formulation, the dual variable v_i corresponds to the flow conservation constraint for node i, and w_{ij} corresponds to the capacity constraint for arc $\{i,j\}$. Since one of the flow conservation constraints is redundant, we have assigned a value of 0 to v_O. Moreover, $\mu_{S,T}$ is the dual variable corresponding to the cutset constraint for cutset $\{S,T\}$, and π_{ij} is the dual variable corresponding to the upper bounding constraints $y_{ij} \le L$ on the design variables.

Apply Algorithm SPP with $\mathbf{b}^+$ as the arc lengths to obtain the nonnegative variables $\theta_{S,T}$ for all O-D cutsets $\{S,T\}$. To construct the dual solution, set $v_i=0$ for all $i \in N$, $w_{ij}=0$ for all $\{i,j\} \in A$, $\mu_{S,T} = \theta_{S,T}$ and $\pi_{ij} = -b^-_{ij}$. This solution satisfies the nonnegativity constraints and constraints (3.1). Since

$$\sum_{\substack{O\text{-}D \text{ cutsets } \{S,T\}, \\ \{i,j\} \in \{S,T\}}} \mu_{S,T} = \sum_{\substack{O\text{-}D \text{ cutsets } \{S,T\}, \\ \{i,j\} \in \{S,T\}}} \theta_{S,T} \le b^+_{ij} = b_{ij} - b^-_{ij} = b_{ij} + \pi_{ij},$$

the solution also satisfies constraints (3.2). It is easy to see that the objective function value of this dual solution equals the objective function value of the loaded shortest path solution. ⊗

As shown by simple examples, Proposition 3.2 is valid only for the special form of the objective function we have considered with the zero cost coefficients for the flow variables. That is, if the flow costs are nonzero, the problem can have an extreme point solution with nonintegral values for the **y** variables. Barany, Van Roy, and Wolsey [5] have also considered special objective function structures in their study of the convex hull of solutions to uncapacitated lot-sizing problems.

The proofs of Proposition 3.2 show that we can determine the optimal solution to the 1F loading problem by solving a single shortest path problem and that all the demand flows over a single path in the optimum solution. Thus, even though our cost function has a staircase form, and is not convex or concave, the optimum solution occurs at an extreme point of the shortest path polytope in formulation P(**f**) just as it does for concave cost single commodity flow problems (Zangwill [23]). This result implies that we can scale the demand and cost figures to obtain an equivalent loading problem that has a unit demand.

Observe that in the proof we have just given, the optimal dual variables for the flow conservation and capacity constraints in the flow model P(IP1) are zero: thus these constraints are not "critical" at the optimal primal solution. Indeed, the cutset inequalities, along with the upper bounding and the nonnegativity constraints, are sufficient for describing the convex hull of the projection of P(IP1) into the subspace of **y** variables and so we have an alternate proof for the first statement in Proposition 3.2 as well.

4. The Two-Facility Loading Problem

For situations with two instead of one facility, the network loading problem becomes more difficult and the results of Section 3 no longer apply. The problem is more complex for several reasons:

(i) Adding a generalized version of the cutset inequalities $Y_{S,T} \geq \lceil D_{S,T}/C \rceil$ to the linear programming relaxation of the formulation P(IP2) of two-facility case is not sufficient for generating integer optimal **x** and **y** solutions.

(ii) A heuristic that is a natural generalization of the shortest path algorithm for the 1F case, while generating "good" solutions, does not necessarily generate an optimal solution.

(iii) The optimum flow need not be an extreme flow in the shortest path polyhedron defined by the formulation P(**f**). Thus, 2F loading problem does not inherit the nice characterization of the optimal solution to the 1F loading problem.

(iv) Variations of the 2F loading problem are strongly NP-hard.

In this section and the next subsection, we consider these four features of the two-facility problem.

Property (i). The generalized cutset inequality for any cutset {S,T} is $X_{S,T} + r_{S,T} Y_{S,T} \geq r_{S,T} \lceil D_{S,T}/C \rceil$. Note that if $X_{S,T} = 0$, then this inequality reduces to the cutset inequality $Y_{S,T} \geq \lceil D_{S,T}/C \rceil$ for the one-facility problem. Moreover if $Y_{S,T} = \lceil D_{S,T}/C \rceil - 1$, then the inequality states that the unit capacity facility must provide at least $X_{S,T} \geq r_{S,T}$ units of capacity so that the total capacity across the cutset {S,T} is at least $D_{S,T} = C\left(\lceil D_{S,T}/C \rceil - 1\right) + r_{S,T}$. Magnanti, Mirchandani, and Vachani [12] establish the validity of this generalized cutset inequality and show that it typically is a facet of the integer polyhedron defined by the formulations P(IP2) and P(CUT2). The example in Figure 2 shows, however, that the linear program consisting of the flow conservation constraints, the capacity constraints, and the generalized cutset inequalities does not necessarily have integral **x** and **y** solutions. In this example, we can lease LC and MC (with capacity C equal to 10 units) facilities on the five network arcs. The numbers next to the arcs in Figure 2(i) represent the costs for leasing these facilities; the first number indicates the LC cost and the second number the MC cost. A flow of 12 units must be sent from node 1 to node 4. The optimal solution to the problem shown in Figure 2(ii) satisfies all these constraints and has fractional values for some of the **x** and **y** variables.

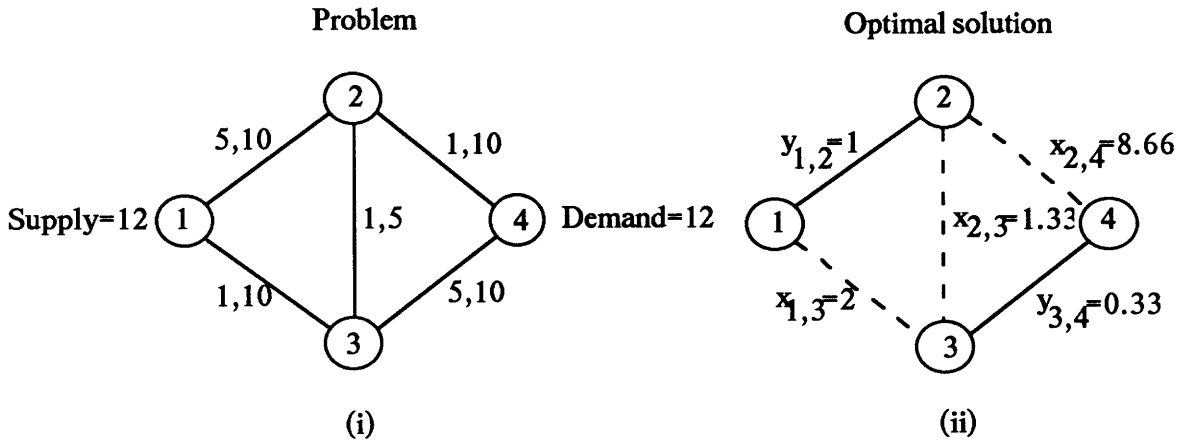


Figure 2. Numbers on arcs in figure (i) indicate the cost of installing LC and MC facilities. The fractional optimal solution shown in figure (ii) satisfies the flow conservation, capacity and cutset inequalities.

Property (ii) and (iii). The following heuristic is a natural generalization of the shortest path algorithm for the 1F problem, and even though it generates "good" solutions, it does not necessarily generate an optimal solution to the problem. For reasons that will become apparent, we refer to this heuristic as the *two-path heuristic.*

The two-path heuristic

Step 0: Let r = d mod(C) and d = qC + r.

Step 1: Find the shortest path form node O to node D using b^+_{ij} as the arc lengths. Load q MC facilities on this path and send a flow of qC on this path.

Step 2: For each arc {i,j}, if $a^+_{ij} = 0$, set $e_{ij} = 0$; otherwise, calculate the (adjusted for sign) breakeven point $m^+_{ij} = b^+_{ij}/a^+_{ij}$. If $r \geq m^+_{ij}$, set $e_{ij} = b^+_{ij}/r$. Otherwise, set $e_{ij} = a^+_{ij}$.

Step 3: Find the shortest path length using e_{ij} as the arc lengths. Load 1 MC facility on all arcs of this path satisfying $r \geq m^+_{ij}$ and r LC facilities on those arcs that satisfy $r < m^+_{ij}$. Send a flow of r on this path.

Step 4: Load the required number of MC facilities on each arc with $b^-_{ij} < 0$ so that these arcs have a total of L MC facilities. Similarly, load the appropriate number of LC facilities on arcs with $a^-_{ij} < 0$ to obtain a total of L LC facilities on these arcs.

This solution costs $\text{len}(\mathbf{b}^+)*q + \text{len}(\mathbf{e})*r + L(a^-_{ij} + b^-_{ij})$.

The two-path heuristic need not generate the optimal solution to the problem. The example in Figure 3 illustrates this fact. The two-path heuristic finds the shortest MC path – that is, the shortest path using MC facilities as the arc lengths – from node 1 to node 4 and loads $\lfloor 12/10 \rfloor = 1$ MC facility on all the arcs of this path. We can send a flow of 10 units on path 1-2-4 at a cost of 55. To send the remaining flow of 2 units, we find the shortest LC path, making appropriate adjustments for those arcs that have a breakeven point of 2 or less (that is, arcs {2,3} and {3,4}). The corresponding shortest path is 1-3-2-4, and the cost of sending the remaining 2 units is 28; thus the total cost of this heuristic solution is 83. However, the optimal solution places 1 MC facility on each of the arcs {1,2},{2,3} and {3,4}, and 2 LC facilities on arcs {1,3} and {2,4} at a total cost of 78.

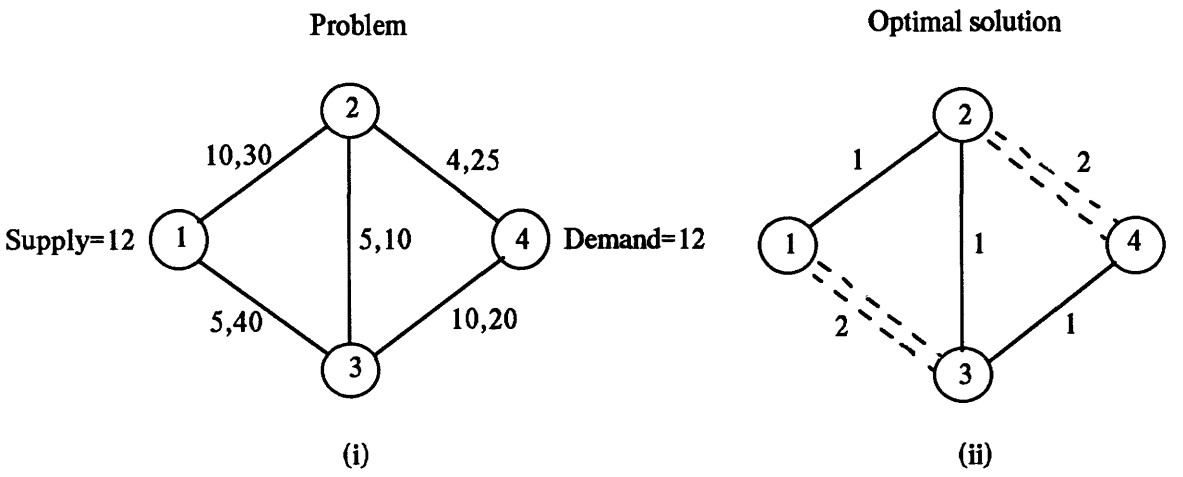


Figure 3. Numbers next to the arcs in figure (i) indicate LC, MC costs for problem definition, and those in figure (ii) specify the number of LC or MC facilities used in the optimal design. Dashed lines in the optimal design indicate LC facilities and solid lines indicate MC facilities.

This example also shows that the optimum solution need not necessarily be an extreme flow in the flow formulation P(**f**) of the 2F loading problem.

We might note that the error introduced by the solution generated by the two-path heuristic is bounded. To consider the most interesting cases, assume that **a, b** > **0** and let $\varepsilon = (v(H)-v^*)/v^*$ denote the relative error measuring the difference between the optimal value v^* of the 2F loading problem and the value $v(H)$ of the solution generated by the two-path heuristic. Then ε tends to 0 as d approaches infinity.

Proposition 4.1. *Let v^* denote the optimal solution to the 2F loading problem and $v(H)$ denote the value of the heuristic solution provided by the two-path heuristic. Then*

$$\lim_{d\to\infty} \frac{v(H)-v^*}{v^*} = 0.$$

Proof. $v(H) = \text{len}(\mathbf{b}^+)*q + \text{len}(\mathbf{e})*r$

$$\leq \text{len}(\mathbf{b}^+)*q + \text{len}(\mathbf{b}^+)$$

$$= \text{len}(\mathbf{b}^+)*(q + 1).$$

Since the LP relaxation of P(IP2) will use only MC facilities, $v^* \geq \text{len}(\mathbf{b}^+)*d/C$. Therefore, since $q+1 = \lceil d/C \rceil \geq d/C$, $v(H) - v^* \leq \text{len}(\mathbf{b}^+)$ which implies the result . ⊗

In a special case, the addition of the cutset inequalities to P(IP2) is sufficient to guarantee integer **x** and **y** optimal solutions. This situation arises when the breakeven points of all the arcs are equal.

Proposition 4.2. *Suppose that the value of the breakeven point* $m_{ij}^1 = b_{ij}/a_{ij}$ *is the same for every arc* $\{i,j\}$ *of the network G. Then*

(i) every solution to P(CUT2) is an extreme point solution to the linear program defined by the cutset inequalities $X_{ST} + r_{S,T}\, Y_{ST} \geq r_{S,T}\lceil D_{S,T}/C \rceil$ *for all O-D cutsets* $\{S,T\}$*, the aggregate-capacity demand constraints, and, the upper bounding and the nonnegativity conditions* $x_{ij} \geq 0$ *and* $y_{ij} \geq 0$ *for all arcs* $\{i,j\}$*, and*

(ii) if we append the cutset inequalities to the formulation P(IP2), then the values of the variables **x** *and* **y** *are integer in every extreme point solution to the linear programming relaxation of the problem formulation.*

Rather than proving this result at this point, we will prove a more general result for the 3F loading problem in Section 5. A modification of same proof technique will prove the result for the 2F case. The proof will also imply that the version of the 2F loading problem with the same breakeven point on each arc is polynomially solvable. In the next section, we describe a variation of the problem and discuss its complexity.

4.1 A Strongly NP-Hard Variation

Suppose we are given an existing network with specified capacities on some of the arcs that can be used to satisfy the demand requirements. We want to load additional capacity on the arcs in order to send a flow of d from node O to D. An equivalent way of viewing this problem is to assume we can obtain the first unit of capacity at zero cost on some arcs. We wish to study the computational complexity of this variation of the 2F loading problem.

Recall that a recognition problem is said to be strongly NP-complete if the existence of a pseudopolynomial algorithm for it implies that P = NP. Furthermore, an optimization problem is said to be strongly NP-hard if some strongly NP-complete recognition problem can be polynomially reduced to it.

We shall use the three exact cover problem (3XC) – which is strongly NP-complete – to prove that the 2F loading problem problem with existing arc capacities is strongly NP-hard. The 3XC can be described as follows (Garey and Johnson [9]): Given a set P with cardinality equal to p, and a collection S of 3 element subsets of P, does there exist a subcollection $S' \subseteq S$ with the property that every element of P occurs in exactly one member of S'?

Proposition 4.3. *The 2F loading problem with existing arc capacities is strongly NP-hard.*

Proof. See Appendix I.

Corollaries 4.4 and 4.5 are immediate consequences of Proposition 4.3.

Corollary 4.4. *The 2F loading problem with upper bounds on arcs flows is strongly NP-hard.*

Proof. Refer to Figure I.1. Impose upper bounds on the flow variables as follows: (i) on arcs $\{1,\sigma_i\}$, the upper bound is 3, and (ii) on arcs $\{\pi_j,n\}$, this bound is 1. ⊗

Corollary 4.5. *The 2F loading problem with upper bounds on the design variables is strongly NP-hard.*

5.0 THE THREE-FACILITY LOADING PROBLEM

We now consider the 3F loading problem. A straightforward generalization of the two-path heuristic for the 3F loading problem generates "good" (i.e., asymptotically optimal) solutions which, again, are not necessarily optimal. Furthermore, since λ of the MC facilities are equivalent to an HC facility, the two-facility cutset inequality also generalizes to the 3F case to give the inequality

$$X_{S,T} + r_{S,T}Y_{S,T} + \lambda r_{S,T}Z_{S,T} \geq r_{S,T} \left\lceil \frac{D_{S,T}}{C} \right\rceil. \qquad (5.1)$$

for all cutsets {S,T}. Magnanti, Mirchandani, and Vachani [12] show that these inequalities are valid, and that they are facet defining under fairly mild restrictions. However unlike the 2F case, the linear program obtained by adding the 3F cutset inequalities to P(IP3) does not guarantee integer **x**, **y** and **z** optimal extreme point solutions to its linear programming relaxation even if $m^1_{ij} = m^1$ and $m^2_{ij} = m^2$ for all arcs {i,j}. We require two additional classes of facets for this result to be true.

To describe these facets, we introduce some further notation. Let $d = p\lambda C + q C + r$ with $r = d \bmod(C)$ and $q C + r = d \bmod(\lambda C)$. By appending subscripts S,T to p, q and r, we can write a similar expression to represent any cutset demand $D_{S,T}$, even for problems with multiple commodities. For notational convenience, in the expressions to follow we will not state these subscripts explicitly.

Note that we could view the inequality (5.1) as a "lifting" of its two-facility analog without the variables **z** (that is, the inequality obtained by setting **z = 0**). (See Figure 4 for an example.) Similarly, suppose that we started with two-facility problems containing only the variables **x** and **z**, or **y** and **z**. The resulting two-facility cutset inequalities are

$$X_{S,T} + (qC+r)Z_{S,T} \geq (qC+r)\left\lceil \frac{D_{S,T}}{\lambda C} \right\rceil$$

and

$$Y_{S,T} + (q+1)Z_{S,T} \geq (q+1)\left\lceil \frac{D_{S,T}}{\lambda C} \right\rceil.$$

Figure 4 shows an example. We can interpret the last inequality as follows. In the space of the **y** and **z** variables, the aggregate-capacity demand constraint is $Y_{S,T} + \lambda Z_{S,T} \geq D_{S,T}/C$. But since the left hand side of this expression is an integer, for any O-D cutset {S,T}, we can replace the right hand side by $\lceil d/C \rceil = \lceil (p\lambda C + q C + r)/C \rceil = p\lambda + (q+1)$. So now if we view λ as the capacity of the higher capacity facility, the remainder on dividing the demand by λ is (q+1) and so the given inequality is just a version of the usual two-facility cutset inequality.

If we "lift" these inequalities to the full space of **x**, **y**, and **z** variables, they become

$$X_{S,T} + \min(qC+r, C)Y_{S,T} + (qC+r)Z_{S,T} \geq (qC+r)\left\lceil \frac{D_{S,T}}{\lambda C} \right\rceil \tag{5.2}$$

and

$$X_{S,T} + rY_{S,T} + r(q+1)Z_{S,T} \geq r(q+1)\left\lceil \frac{D_{S,T}}{\lambda C} \right\rceil. \tag{5.3}$$

Proposition 5.1 shows that the inequalities (5.2) and (5.3) are valid and Proposition 5.2 discusses necessary and sufficient conditions for ensuring that these inequalities are facet defining.

Example 1: Suppose d = 74, C = 10, λ = 3 so that d = (2)λC + (1) C + 4, p = 2, q = 1, and r = 4. The the three inequalities (5.1), (5.2), and (5.3) become (we will subsume the subscripts {S,T}).

$$X + 4Y + 12Z \geq 32$$

$$X + 10Y + 14Z \geq 42$$

and

$$X + 4Y + 8Z \geq 24.$$

Figure 4 shows the polyhedron defined by these inequalities and the nonnegative orthant. Note that the extreme points of this polyhedron have integer values for all the variables. Proposition 5.3 shows that (under certain circumstances) this integrality result applies more generally, not only for the aggregate variables across any cutset, but also in the space of full variables for the entire network. ⊗

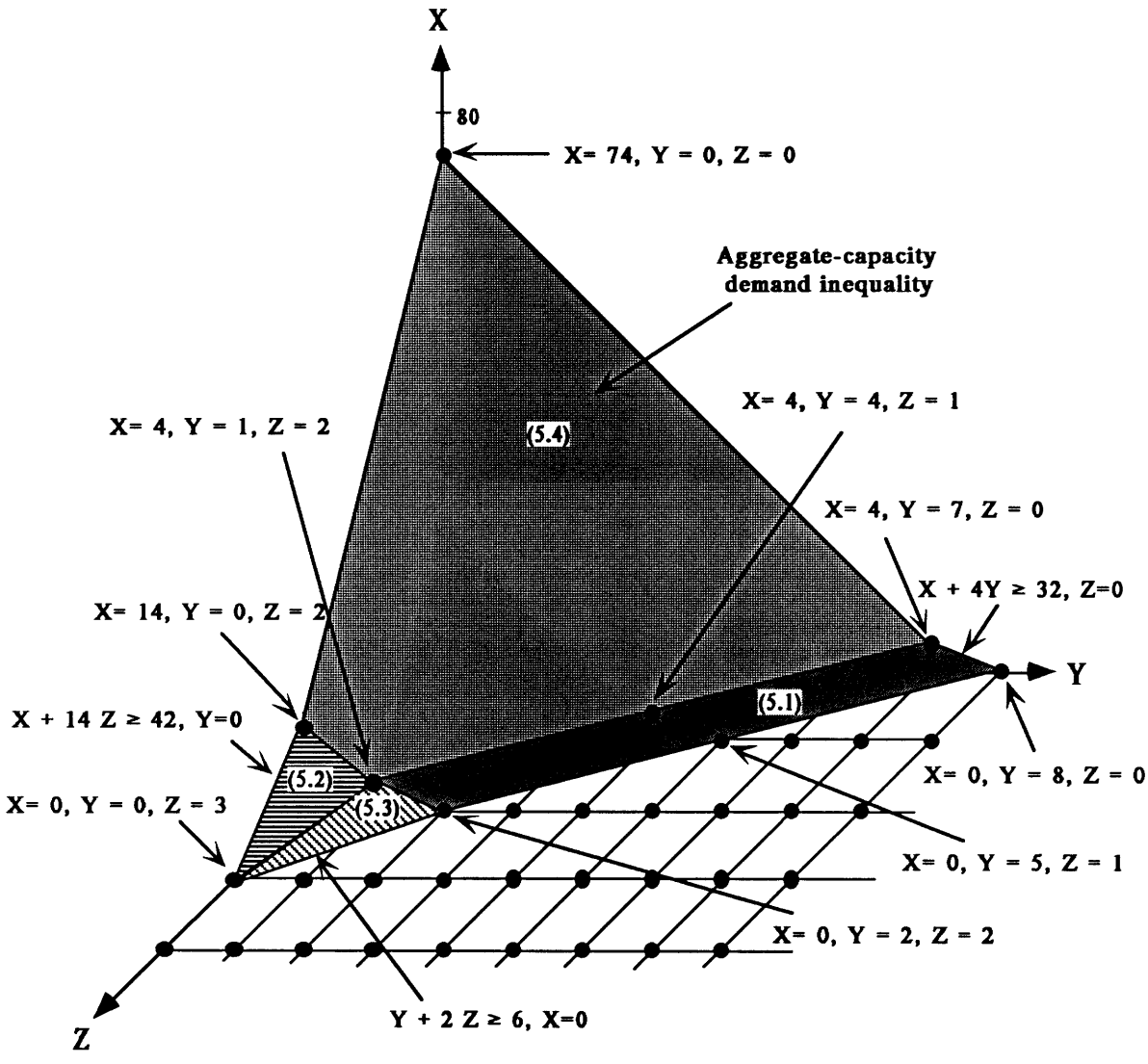


Figure 4. Facets corresponding to a cutset

We might note that the inequalities (5.1)-(5.3) are related for the boundary values of parameters r and q. First, if $q = \lambda - 1$, then inequality (5.1) is equivalent to inequality (5.3) and therefore for a fixed value of λ, the faces defined by these two inequalities alternately coalesce and dissociate as we increase the value of d. Second, when $q = 0$, or when $r = C$, then inequality (5.2) is equivalent to inequality (5.3). In particular, when $\lambda = 2$ then $q = \lambda - 1$ or 0 and one of the inequalities (5.1) or (5.2) subsumes (5.3). Moreover, when $\lambda = 1$ (and so only two types of facilities are available), then q always equals zero and inequalities (5.1), (5.2) and (5.3) are equivalent.

Finally, all three inequalities are equivalent when $q = \lambda - 1$ (for an arbitrary value of λ) and $r = C$ (i.e., when $D_{S,T}$ is a multiple of λC). In fact, in this case, these inequalities are equivalent to the aggregate-capacity demand inequality, $X_{S,T} + CY_{S,T} + \lambda CZ_{S,T} \geq D_{S,T}$.

The validity proof of inequality (5.1) given by Magnanti, Mirchandani, and Vachani [12] is based on the Chvàtal-Gomory (C-G) procedure. This procedure repeatedly considers nonnegative combinations of already known valid inequalities and uses integrality arguments to generate new stronger valid inequalities. Similarly, we use the C-G procedure on the aggregate-capacity demand inequality to establish the validity of the new inequality (5.2). For showing the validity of the other new inequality (5.3), we use an algebraic argument.

Proposition 5.1. *For all nonempty sets $S \subset N$ and $T = N \backslash S$, the inequalities (5.2) and (5.3) are valid for the convex hull of feasible solutions to the formulations P(IP3) and P(CUT3).*

Proof.

(a) Since the inequality (5.2) is redundant if $D_{S,T} = 0$, assume $D_{S,T} = d$. We start with the aggregate-capacity demand inequality:

$$X_{S,T} + CY_{S,T} + \lambda CZ_{S,T} \geq D_{S,T} = p\lambda C + qC + r \tag{5.4}$$

and use an induction argument to establish (5.2). Consider the inequality

$$X_{S,T} + CY_{S,T} + (\lambda C - k)Z_{S,T} \geq p(\lambda C - k) + qC + r \tag{5.5}$$

for integer values of k between 0 and $\lambda C - \max\{C, qC+r\}$. For $k = 0$, this inequality is the aggregate-capacity demand inequality (5.4); if $q \geq 1$, then $\max\{C, qC+r\} = qC+r$ and for $k = \lambda C - (qC+r)$ the inequality (5.5) becomes the inequality (5.2) since $\lceil D_{S,T}/\lambda C \rceil = p+1$. Although the inequality is not yet of the form of (5.2) if $q = 0$ and $k = \lambda C - C$, by further arguments we will later show how this inequality with $k = \lambda C - C$ implies (5.2).

Suppose (5.5) is valid for $k = l$, $0 \leq l \leq \lambda C - \max\{C, qC+r\} - 1$. We wish to show that it is also valid for $k = l + 1$.

The nonnegativity of $X_{S,T}$ and $Y_{S,T}$ implies that $\frac{1}{[\lambda C - (I+1)]} X_{S,T} \geq 0$ and $\frac{C}{[\lambda C - (I+1)]} Y_{S,T} \geq 0$. Adding these two inequalities and (5.5) with k = I, we obtain

$$\frac{(\lambda C - I)}{[\lambda C - (I+1)]} X_{S,T} + \frac{(\lambda C - I)}{[\lambda C - (I+1)]} C\, Y_{S,T} + (\lambda C - I) Z_{S,T} \geq p\,(\lambda C - I) + qC + r\,.$$

Multiplying this inequality by $\frac{[\lambda C - (I+1)]}{(\lambda C - I)}$ gives

$$X_{S,T} + C\, Y_{S,T} + [\lambda C - (I+1)] Z_{S,T} \geq p\,[\lambda C - (I+1)] + (qC + r) \frac{[\lambda C - (I+1)]}{(\lambda C - I)}$$
$$= p\,[\lambda C - (I+1)] + (qC + r) - \frac{(qC + r)}{(\lambda C - I)}.$$

Now, since qC+r < λC−I and the left hand side is necessarily integer, integrality arguments permit us to round up the right hand side to obtain inequality (5.5) with k = I+ 1. So by induction, we have established (5.5) with k = λC − max{C, qC+r}. If q > 0, or if q = 0 and r = C, this inequality is (5.2) and we are done. Otherwise, q = 0, r < C and we obtain the inequality

$$X_{S,T} + (C-w)\, Y_{S,T} + (C-w) Z_{S,T} \geq p(C-w) + r$$

with w = 0. We now use the previous argument, inducting on the value of w between 0 and C−r. (We use the fact that r < C in the integer rounding argument.). At the value w = C−r, the inequality becomes (5.2), completing the proof of the first part of the proposition. ⊗

(b) Consider inequality (5.3). If $Z_{S,T} \geq \lceil D_{S,T}/\lambda C \rceil$, then inequality (5.3) is clearly valid. Otherwise, let $Z_{S,T}$ equal $\lceil D_{S,T}/\lambda C \rceil - u$, for some u ≥ 1. Using these facilities, we can send a maximum flow of $(\lceil D_{S,T}/\lambda C \rceil - u)\lambda C$ from the set S to the set T. Therefore, we need to load additional MC and LC facilities to send the remaining flow of $D_{S,T} - (\lceil D_{S,T}/\lambda C \rceil - u)\lambda C = C(u\lambda - \lambda + q) + r$. To establish the validity of the inequality (5.3), we have to show that

$$X_{ST} + rY_{S,T} \geq r(q+1)u. \qquad (5.6)$$

If $Y_{S,T} \geq (u-1)\lambda + (q+1)$, then the left hand side of inequality (5.6) is at least $r(q+1)u$ since $\lambda \geq (q+1)$. So assume that $Y_{S,T} = (u-1)\lambda + (q+1) - v$ for some $v \geq 1$. In addition to the $(\lceil D_{S,T}/\lambda C \rceil - u)$ HC and the $[(u-1)\lambda + (q+1) - v]$ MC facilities so loaded, we must load at least $C(u\lambda-\lambda+q) + r - C[(u-1)\lambda + (q+1) - v]$ LC facilities, i.e., we must load at least $C(v-1) + r \geq vr$ LC facilities in order to satisfy the remaining demand.

However, in order to prove the validity of inequality (5.3), we need to show that $X_{S,T} \geq r(q+1)u - r[(u-1)\lambda + (q+1) - v] = r[(u-1)(q+1-\lambda) + v]$ which is no larger than vr since $\lambda \geq q+1$. Hence, the inequality (5.3) is valid. ⊗

Proposition 5.2 describes necessary and sufficient conditions for inequalities (5.2) and (5.3) to define facets of the underlying polyhedron. These conditions are similar to the conditions required for inequality (5.1) to be facet defining (see Magnanti, Mirchandani, and Vachani [12]).

Proposition 5.2. *The following conditions are necessary and sufficient for the inequalities (5.2) and (5.3) to be facet defining for P(IP3):*

1. *$D_{S,T} > 0$. (This condition implies that {S,T} is an O-D cutset.)*

2. *The subgraphs defined by S and T are connected.*

Proof of necessity of the conditions. The necessity part of the proposition is easy to establish. If $D_{S,T} = 0$, then both the inequalities are simply aggregations of the nonnegativity constraints. If S is not connected and consists of two components S_1 and S_2 with $O \in S_1$, it is easy to see that the inequalities corresponding to S_1 are stronger than the inequalities corresponding to S; therefore, S must be connected for (5.2) and (5.3) to be facet defining. Analogously, T must also be connected. ⊗

Remarks.

(i) Recall that if d is a multiple of λC, then both the inequalities (5.2) and (5.3) are equivalent to the aggregate-capacity demand inequality. Thus, although they still define facets, they do not add to the formulation P(IP3).

(ii) The sufficiency part of the proof is lengthy and technical, but essentially follows the argument of the cutset inequality proof given by Magnanti, Mirchandani, and Vachani [12]. We will not provide the proof here.

Some natural generalization of these facet inequalities apply to (i) multicommodity situations, as well as (ii) situations with more than three facilities. The extension for multicommodity situations is straightforward. We just let $D_{S,T}$ denote the total demand between nodes in S and nodes in T for all commodities. To illustrate the multiple-facility extensions, let us generalize inequality (5.3). We first need to define additional notation for the new facilities. Suppose facilities LC, HC(0), HC(1), HC(2),..., and HC(p) are available with capacities, respectively, of 1, C and $\lambda_k C$ units for k = 1 to p. Define $\lambda_0 = 1$ and reindex the facilities if necessary so that $\lambda_k < \lambda_{k+1}$ for i = 0, 1, ..., (p−1). Let the number of facilities of each type on arc {i,j} be x_{ij}, z_{ij}^k with the obvious aggregation over any {S,T} cutset. Let $d = p_t\lambda_t C + p_{t-1}\lambda_{t-1}C + \ldots + p_0 C + r$ with $r = d \bmod (C)$ and let

$$p_k = \left\lfloor d - \left(\sum_{i=k+1}^{t} p_i\lambda_i C + r \right) / \lambda_k C \right\rfloor \quad \text{for } k = 0, 1, \ldots, t.$$

The following inequality is a generalization of (5.3) (with $\prod_{i=0}^{-1} (p_i+1) = 1$):

$$X_{S,T} + r \sum_{k=0}^{t} \prod_{i=0}^{k-1} (p_i+1)\, Z_{S,T}^k \geq r \prod_{i=0}^{t-1} (p_i+1) \left\lceil \frac{d}{\lambda_t C} \right\rceil.$$

Therefore, although we have discussed inequalities (5.1), (5.2) and (5.3) only for the single-commodity three-facility case, they are applicable in more general settings. We will not explore these generalizations in this paper; our objective here is to show that the addition of these inequalities to formulation P(IP3) produces a linear program that has integer optimal **x**, **y**, and **z** solutions whenever the breakeven points m_{ij}^1 equal m^1 and m_{ij}^2 equal m^2 for all arcs {i,j}. In fact, we can obtain the optimal solution to the single commodity problem with equal breakeven points on all the arcs by solving a single shortest path problem.

Theorem 5.3 *For all* $a \in \Re^{|A|}$, *the following optimization problem P(FACET3) has an optimal solution with integer values for the* x, y *and* z *variables.*

[Problem P(FACET3)]:

$$\text{minimize} \sum_{\{i,j\} \in A} \left(a_{ij}x_{ij} + m^1 a_{ij}y_{ij} + m^2 a_{ij}z_{ij}\right)$$

subject to:

$$\sum_{j \in N} f_{ji} - \sum_{j \in N} f_{ij} = \begin{cases} -d & \text{if } i = O \\ d & \text{if } i = D \\ 0 & \text{otherwise} \end{cases}$$

$$f_{ij} + f_{ji} \le x_{ij} + C\,y_{ij} + \lambda C\,z_{ij} \quad \text{for all } \{i,j\} \in A$$

$$X_{S,T} + r\,Y_{S,T} + \lambda r\,Z_{S,T} \ge r\,\lceil d/C \rceil \quad \text{for all } O\text{-}D \text{ cutsets } (S,T)$$

$$X_{S,T} + \min(qC+r,\, C)\,Y_{S,T} + (qC+r)\,Z_{S,T} \ge (qC+r)\lceil d/\lambda C \rceil \quad \text{for all } O\text{-}D \text{ cutsets } (S,T)$$

$$X_{S,T} + r\,Y_{S,T} + r(q+1)\,Z_{S,T} \ge r(q+1)\lceil d/\lambda C \rceil \quad \text{for all } O\text{-}D \text{ cutsets } (S,T)$$

$$\left.\begin{array}{l} x_{ij} \le L \\ y_{ij} \le L \\ z_{ij} \le L \end{array}\right\} \quad \text{for all } \{i,j\} \in A$$

$$x_{ij},\, y_{ij},\, z_{ij},\, f_{ij}, f_{ji} \ge 0 \quad \text{for all } \{i,j\} \in A.$$

Proof. We first assume that $1 < m^1 < C$, and $m^1 < m^2 < \lambda m^1$ so that HC facilities are more cost effective per unit than either LC or MC facilities, and MC facilities are more cost effective per unit than LC facilities. We present the proof only for $r < C$. As in the proof of Proposition 3.2, we will construct a primal feasible solution that is integer and construct a dual feasible solution whose objective function value equals that of the primal solution.

We will consider the four cases for developing the primal and dual feasible solutions. These cases exhaust all possible values for the parameters of the problem. For Case 1, the optimal primal solution will use only HC facilities; for Cases 2, 3, or 4, the optimal primal solution will use a combination of HC and MC, or HC and LC, or HC, MC, and LC facilities respectively. Note that if we scale the cost on each arc by a_{ij} and $a_{ij} \ge 0$, and the arc carries the d units of demand, then given our cost assumptions, it is cost effective to load the arc with $p = \lfloor d/\lambda C \rfloor$ HC facilities, incurring a cost of $p\,m^2$, together with a combination of LC, MC, and HC facilities. The additional cost beyond $p\,m^2$ is

$(q+1)m^1$ if we install only MC additional facilities

qm^1+r if we install MC and LC additional facilities

m^2 if we install only a single HC additional facility.

The following cases define outcomes for distinguishing between these solutions.

Case 1. $\min[(q+1)m^1, qm^1+r] \geq m^2$.

Subcase 1(a). $m^2C < m^1(qC+r)$.

Subcase 1(b). $m^2C \geq m^1(qC+r)$.

Case 2. $(q+1)m^1 < m^2$ and $r \geq m^1$.

Case 3. $q = 0$, $r < m^1$.

Subcase 3(a). $\lambda r < m^2$.

Subcase 3(b). $\lambda r \geq m^2$.

Case 4. $q \geq 1$, $r < m^1$, $m^1q+r < m^2$.

Let $\Theta = (C-r)[m^2-(m^1q+r)] + r(m^1-C)[\lambda-(q+1)]$.

Subcase 4(a). $\Theta < 0$.

Subcase 4(b). $\Theta \geq 0$.

Some comments about this classification are in order. First, for Case 1 to be true, the value of q must be at least 1 since $m^1 < m^2$. Second, Case 2 can occur both when $q = 0$ or when $q \geq 1$. Finally, Θ consists of two terms: (i) $(C-r)[m^2-(m^1q+r)]$ and (ii) $r(m^1-C)[\lambda-(q+1)]$. The first term is nonnegative and the second term is nonpositive. Their sum can be either positive or negative, differentiating cases 4(a) and 4(b).

Construction of a feasible solution to P(IP3)

Step 1: Calculate $r = d \bmod(C)$ and $q = [(d - r) \bmod(\lambda C)]/C$.

Step 2: Find a shortest path, P_{OD} from the origin O to the destination D using a_{ij}^{+} as the arc lengths. Denote this path length by $\text{len}(\mathbf{a}^{+})$.

Step 3: Construct a feasible primal solution by

(i) loading the configuration of facilities as shown in Table I on each arc of path P_{OD} and

(ii) sending a flow of d on these facilities.

Table I. Configuration of facilities for the primal solution

Condition	Number of facilities loaded		
	HC	MC	LC
Case 1	$\lceil d/\lambda C \rceil$	0	0
Case 2	$\lfloor d/\lambda C \rfloor$	q+1	0
Case 3	$\lfloor d/\lambda C \rfloor$	0	r
Case 4	$\lfloor d/\lambda C \rfloor$	q	r

Step 4: Augment the solution obtained in Step 3 so that each arc $\{i,j\}$ satisfying $a_{ij} < 0$ has L of each of the LC, MC and HC facilities.

Before showing how to generate a feasible solution to the dual, D(FACET3), of problem P(FACET3), let us formulate the dual problem.

[Problem D(FACET3)]:

$$\text{maximize } d\,v_D + r\left\lceil\frac{d}{C}\right\rceil \sum_{O\text{-}D\ cutsets\ \{S,T\}} \mu_{S,T} + (qC+r)\left\lceil\frac{d}{\lambda C}\right\rceil \sum_{O\text{-}D\ cutsets\ \{S,T\}} \sigma_{S,T} \qquad (5.7a)$$
$$r(q+1)\left\lceil\frac{d}{\lambda C}\right\rceil \sum_{O\text{-}D\ cutsets\ \{S,T\}} \theta_{S,T} - L \sum_{\{i,j\}\in A} (\tau_{ij} + \pi_{ij} + \phi_{ij})$$

subject to:

$$\left.\begin{aligned} v_j - v_i - w_{ij} &\le 0 \\ v_i - v_j - w_{ij} &\le 0 \end{aligned}\right\} \quad \text{for all } \{i,j\} \in A \qquad (5.7b)$$

$$w_{ij} + \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \mu_{S,T} + \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \sigma_{S,T} + \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \theta_{S,T} - \tau_{ij} \le a_{ij} \quad \text{for all } \{i,j\} \in A \qquad (5.7c)$$

$$Cw_{ij} + r \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \mu_{S,T} + \min(qC+r, C) \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \sigma_{S,T} + r \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \theta_{S,T} - \pi_{ij} \le m^1 a_{ij} \quad \text{for all } \{i,j\} \in A \qquad (5.7d)$$

$$\lambda C w_{ij} + \lambda r \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \mu_{S,T} + (qC+r) \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \sigma_{S,T} + r(q+1) \sum_{\substack{O\text{-}D\ cutsets\ \{S,T\},\\ \{i,j\}\in\{S,T\}}} \theta_{S,T} - \phi_{ij} \le m^2 a_{ij} \quad \text{for all } \{i,j\} \in A \qquad (5.7e)$$

$$\begin{aligned} &v_i \text{ u.i.s.} \quad \text{for all } i \in N \\ &w_{ij},\ \tau_{ij},\ \pi_{ij},\ \phi_{ij} \ge 0 \quad \text{for all } \{i,j\} \in A \\ &\mu_{S,T},\ \sigma_{S,T},\ \theta_{S,T} \ge 0 \quad \text{for all O-D cutsets } \{S,T\}. \end{aligned} \qquad (5.7f)$$

Notice that this formulation contains only the dual variables corresponding to O-D cutsets. We will not explicitly specify this restriction in the expressions below.

We first note that the first four terms of the objective function equal

$$\left\lfloor \frac{d}{\lambda C} \right\rfloor \left(\lambda C v_D + \lambda r \sum_{\{S,T\}} \mu_{S,T} + (qC+r) \sum_{\{S,T\}} \sigma_{S,T} + r(q+1) \sum_{\{S,T\}} \theta_{S,T} \right) +$$
$$q \left(C v_D + r \sum_{\{S,T\}} \mu_{S,T} + \min(qC+r, C) \sum_{\{S,T\}} \sigma_{S,T} + r \sum_{\{S,T\}} \theta_{S,T} \right) + \qquad (5.8)$$
$$r \left(v_D + \sum_{\{S,T\}} \mu_{S,T} + \sum_{\{S,T\}} \sigma_{S,T} + \sum_{\{S,T\}} \theta_{S,T} \right).$$

We will find it more convenient to work with the objective function in this form. If the primal solution uses all three types of facilities (i.e., Case 4 applies), then this representation of the objective function is more intuitive since, as we will see, we can assign dual variables so that on some subset {i,j} of the arcs

(i) $$\left(\lambda C w_{ij} + \lambda r \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \mu_{S,T} + (qC+r) \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \sigma_{S,T} + r(q+1) \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \theta_{S,T} \right) = m^2 a_{ij}^+$$

(ii) $$\left(C w_{ij} + r \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \mu_{S,T} + \min(qC+r, C) \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \sigma_{S,T} + r \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \theta_{S,T} \right) = m^1 a_{ij}^+,$$

(iii) $$\left(w_{ij} + \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \mu_{S,T} + \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \sigma_{S,T} + \sum_{\substack{\{S,T\}, \\ \{i,j\} \in \{S,T\}}} \theta_{S,T} \right) = a_{ij}^+,$$

(iv) $\phi_{ij} = -m^2 a_{ij}^-$, $\pi_{ij} = -m^1 a_{ij}^-$ *and* $\tau_{ij} = -a_{ij}^-$, *and*

(v) v_D equals the shortest path length from O to D using w_{ij} as the arc lengths.

If the dual variables were to satisfy these conditions, then the objective function value of this dual solution equals the objective function of the integer primal solution and the proof would be complete for this case. The representation (5.8) of the objective function is useful in other cases as well.

Construction of a feasible solution to D(FACET3)

Step 1: Calculate $r = d \bmod(C)$ and $s = [(d - r) \bmod(\lambda C)]/C$.

Step 2: Set $\tau_{ij} = -\bar{a}_{ij}$, $\pi_{ij} = -m^1\bar{a}_{ij}$ and $\phi_{ij} = -m^2\bar{a}_{ij}$.

Step 3: Set α_{ij}, β_{ij}, γ_{ij} and δ_{ij} as described in the Table II.

Step 4: Using any shortest path algorithm, find the shortest path from node O to every other node with α_{ij} as the arc costs. Let v_i equal the shortest path to node i. Set $w_{ij} = \alpha_{ij}$.

Step 5: Use Algorithm SPP with β_{ij}, γ_{ij}, and δ_{ij} as the arc lengths. (Note that these costs are nonnegative, so we can apply the algorithm.) This computation gives us values for the $\mu_{S,T}$, $\sigma_{S,T}$, and $\theta_{S,T}$ variables respectively.

Table II. Arc costs used for calculating the dual variables

Condition	α_{ij}	β_{ij}	γ_{ij}	δ_{ij}
Subcase 1(a)	0	0	$\frac{m^2}{(qC+r)}\, a_{ij}^+$	0
Subcase 1(b)	0	0	$\frac{[m^1(q+1)-m^2]}{(C-r)}\, a_{ij}^+$	$\frac{[m^2C-m^1(qC+r)]}{r(C-r)}\, a_{ij}^+$
Case 2	0	$\frac{[m^2-(q+1)m^1]}{r[\lambda-(q+1)]}\, a_{ij}^+$	0	$\frac{(\lambda m^1-m^2)}{r[\lambda-(q+1)]}\, a_{ij}^+$
Subcase 3(a)	$\frac{(m^2-\lambda r)}{\lambda(C-r)}\, a_{ij}^+$	$\frac{(\lambda C-m^2)}{\lambda(C-r)}\, a_{ij}^+$	0	0
Subcase 3(b)	0	$\frac{(m^2-r)}{r(\lambda-1)}\, a_{ij}^+$	$\frac{(\lambda r-m^2)}{r(\lambda-1)}\, a_{ij}^+$	0
Subcase 4(a)	0	$\frac{[m^2-(m^1q+r)]}{r[\lambda-(q+1)]}\, a_{ij}^+$	$\frac{(m^1-r)}{(C-r)}\, a_{ij}^+$	$\frac{-\Theta}{r[\lambda-(q+1)](C-r)}\, a_{ij}^+$
Subcase 4(b)	$\frac{\Theta}{[\lambda C-(qC+r)](C-r)}\, a_{ij}^+$	$\frac{(C-m^1)}{(C-r)}\, a_{ij}^+$	$\frac{(\lambda m^1-m^2)}{[\lambda C-(qC+r)]}\, a_{ij}^+$	0

We now show that the dual solution obtained in Steps 2, 4 and 5 is feasible and has a solution value equal to that to that of the primal solution we have found.

(i) Constraints (5.7b): The dual solution satisfies constraints (5.7b) as a result of the shortest path optimality conditions.

(ii) Constraints (5.7c): The dual solution satisfies constraints (5.7c) if

$$w_{ij} + \sum_{\substack{O\text{-}D\ cutsets\ (S,T),\\ \{i,j\}\in\{S,T\}}} \mu_{S,T} + \sum_{\substack{O\text{-}D\ cutsets\ (S,T),\\ \{i,j\}\in\{S,T\}}} \sigma_{S,T} + \sum_{\substack{O\text{-}D\ cutsets\ (S,T),\\ \{i,j\}\in\{S,T\}}} \theta_{S,T} \le a_{ij} + \tau_{ij} = a_{ij}^+.$$

Since $w_{ij} = \alpha_{ij}$, $\sum_{\substack{\text{O-D cutsets } \{S,T\},\\ \{i,j\}\in\{S,T\}}} \mu_{S,T} \le \beta_{ij}$, $\sum_{\substack{\text{O-D cutsets } \{S,T\},\\ \{i,j\}\in\{S,T\}}} \sigma_{S,T} \le \gamma_{ij}$ and

$\sum_{\substack{\text{O-D cutsets } \{S,T\},\\ \{i,j\} \in \{S,T\}}} \theta_{S,T} \le \delta_{ij}$, dual feasibility will follow if we can show

$$\alpha_{ij}+\beta_{ij}+\gamma_{ij}+\delta_{ij} \le a_{ij}^{+}. \tag{5.9a}$$

Similarly, the variables w_{ij}, $\mu_{S,T}$, $\sigma_{S,T}$, and $\theta_{S,T}$ will satisfy constraints (5.7d) and (5.7e), if we can show that

$$C\alpha_{ij}+r\beta_{ij}+min\,(qC+r,\ C)\gamma_{ij}+r\delta_{ij} \le m^{1} a_{ij}^{+} \tag{5.9b}$$

and

$$\lambda C\alpha_{ij}+r\lambda\beta_{ij}+(qC+r)\gamma_{ij}+r(q+1)\delta_{ij} \le m^{2} a_{ij}^{+} \tag{5.9c}$$

respectively.

The solutions given in Table II satisfy these inequalities and the nonnegativity restrictions. As an illustration, we consider Subcase 4(a). Notice in this case that $\delta_{ij} = -\beta_{ij} - [(m^1-C)/(m^1-r)]\,\gamma_{ij}$. Therefore,

$$\alpha_{ij} + \beta_{ij} + \gamma_{ij} + \delta_{ij} = \gamma_{ij}\left[1 - \frac{(m^1-C)}{(m^1-r)}\right] = \gamma_{ij}\frac{(C-r)}{(m^1-r)} = a_{ij}^{+}$$

and so the solution satisfies (5.9a).

Substituting in (5.9b) gives

$$C\alpha_{ij} + r\beta_{ij} + C\gamma_{ij} + r\delta_{ij} = \gamma_{ij}\left[C - \frac{r(m^1-C)}{(m^1-r)}\right] = \gamma_{ij}\frac{m^1(C-r)}{(m^1-r)} = m^1 a_{ij}^{+}$$

and substituting in (5.9c) gives

$$\lambda C\alpha_{ij} + \lambda r\beta_{ij} + (qC+r)\gamma_{ij} + r(q+1)\delta_{ij} = [\lambda r - r(q+1)]\beta_{ij} +$$

$$\gamma_{ij}\left[(qC+r) - \frac{r(q+1)(m^1-C)}{(m^1-r)}\right] = \{[m^2 - (m^1q+r)] + (m^1q+r)\}\, a_{ij}^{+} = m^2 a_{ij}^{+}.$$

The nonnegativity of the dual variables follows from the conditions of Subcase 4a. Finally, expression (5.8), or a complementary slackness argument, shows that the objective function value of this dual solution equals the objective value of the primal solution described in Table I.

Similar algebraic computations complete the proof for the other cases.

This proof works with minor modifications for situations with $\lambda=1$ (and so the inequalities (5.1), (5.2), and (5.3) are equivalent); in this case the problem has an integer **x** and **y** optimal solution whenever the (first) breakeven points for all the arcs are equal. The proof also applies with minor modifications to situations when the data do not satisfy the conditions $1 < m^1 < C$, and $m^1 < m^2 < \lambda m^1$. ⊗

Remarks.

(i) The primal solution to P(FACET3) and the solution of its dual depends on the objective function coefficients and the commodity demand value, i.e., on the problem instance. Our proof partitions the space of the problem parameters into seven regions, and identifies region specific primal and dual solutions. This approach of segmenting the parameter space is potentially useful for establishing convex hull proofs as well (see Magnanti, Mirchandani, and Vachani [13]).

(ii) For the special case when the network contains only a single arc (and when $q > 1$), the convex hull of the feasible region has the form shown in Figure 4. In this case, the values of the dual variables α_{ij}, β_{ij}, γ_{ij} and δ_{ij} given in Table II specify the multiples of the constraints (5.4), (5.1), (5.2), and (5.3), respectively, needed to represent the objective function as a nonnegative weighted combination of these constraints. That is, these variables specify the values of the optimal dual variables in the linear program defined over the polyhedron shown in the figure.

(iii) Note from Figure 4, that when the optimal design uses only HC facilities (the solution $X = 0$, $Y = 0$, and $Z = 3$), or when it uses a combination of all three facilities (the solution $X = 4$, $Y = 1$, and $Z = 2$), then the polyhedron is degenerate since four of the defining inequalities meet at the optimal solution (in the case of only HC facilities, two of the four inequalities are the nonnegativity conditions $X \geq 0$ and $Y \geq 0$). As reflected by Cases 1 and 4, in these instances the choice of which dual variables belong to the optimal

basis of the linear programming dual problem (and hence which primal constraints have positive weights) depends upon the coefficients of the objective function.

(iv) The choices of the values of in Table II might appear to be ad hoc. One way to see how to generate these values is to work backwards. Assuming we know which of the variables are positive in any case or subcase, we solve for the values of the variables by solving the appropriate system of equations obtained by setting a subset of (5.9a), (5.9b), and (5.9c) as equalities. That is, in Case 1 we formulate only (5.9c) as an equality; in Case 2, we formulate (5.9b) and (5.9c) as equalities; in Case 3, we formulate (5.9a) and (5.9c) as equalities; and in Case 4, we formulate all three of the inequalities as equalities.

To conclude this section, we note that projecting variables would permit us to use Theorem 5.3 to establish the following results.

Theorem 5.4. *For all* $\mathbf{a} \in \mathcal{R}^{|A|}$, *if the breakeven points* m_{ij}^1 *and* m_{ij}^2 *are the same for all arcs {i,j}, then with the addition of the inequalities (5.1) - (5.3), the cutset formulation P(CUT3) always has an optimal solution with integer values for the* $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{z}$ *variables.*

Theorem 5.5. *For all* $\mathbf{a} \in \mathcal{R}^{|A|}$, *if the breakeven points* m_{ij}^1 *and* m_{ij}^2 *are the same for all arcs {i,j}, then the flow formulation P(f) always has an optimal extreme point solution, that is, a solution in which all d units of demand flow on the same OD path.*

6.0 CONCLUSIONS

We have discussed several variations of a network design problem that we refer to as the single commodity network loading problem, showing that one variation is strongly NP-hard while others are polynomially solvable. In particular, we identified several families of facets for the problem and have shown that

(i) when the economic breakeven points between the LC and MC facilities and between the LC and HC facilities are the same on all the arcs (the common breakeven point problem), then (assuming all costs are nonnegative) it is always optimal to load facilities and send all the flow on a shortest path;

(ii) for the common breakeven point version of the problem, when augmented by the addition of the new facets, the linear programming relaxation of a problem formulation in the space of flow and design variables and of a cutset formulation in the space of the design variables always has an integer solution in the facility design variables;

(iii) for the common breakeven point version of the problem, the formulation of the problem in the space of the flow variables has an extreme point solution;

(iv) the general (with economic breakeven points that vary by arc) version of the problem need not have an extreme point solution in the flow space and the set of facets we have identified do not assure that the linear programming relaxation has an integer solution in the design variables; and

(v) a version of the general problem with existing arc capacities is strongly NP-hard.

This research raises several theoretical and algorithmic issues. In our model, we have assumed that the flow costs are zero. If the flow costs are not zero, then the results of this paper are no longer valid. A result similar to Proposition 4.2 with nonzero flow costs would provide us with the convex hull of feasible solutions to the problem. Moreover, it would be useful to investigate further the polyhedral structure and the computational complexity of the general 2F design problem. This investigation might yield additional families of facets. Finally, we could study situations with a larger number of available facilities and attempt to identify formulations of these problems that would yield integer design variables.

We might view this paper as a study of a single commodity specialization of a broader multicommodity network loading problem that arises in the telecommunication and transportation industries. While our study of the single commodity problem was not motivated by any direct practical applications, the investigation of this problem can help in determining the polyhedral properties of the multicommodity case. Another possible specialization would be to retain the multicommodity nature of the problem, but consider specialized network topologies. In a companion paper (Magnanti, Mirchandani, and Vachani [13]), we have identified the convex hull of solutions to a single arc (with both flow costs and design costs) and three node versions of the problem. These results, in turn, provide a set of inequalities that are valid across a generic cut and across a three partition of the general problem and, together with the results of this paper, furnish the

essential ingredients for a cutting plane approach to the general multicommodity flow problem (see Magnanti, Mirchandani, and Vachani [12]). The resulting computational results are promising and show that the solution of large-scale applications, if not yet a reality, may be within reach.

APPENDIX I

Proposition 4.2. *The 2F loading problem with existing arc capacities is strongly NP-hard.*

Proof. Let $P = \{\pi_1, \pi_2,...,\pi_p\}$ and $S = \{\sigma_1, \sigma_2,...,\sigma_s\}$ in the definition of the 3XC problem given in Section 4.1. We can assume that $\sigma_1, \sigma_2,...,\sigma_s$ are unique. We wish to define a network loading problem with existing arc capacities, which we denote as ec_{ij}, on some of the arcs $\{i,j\}$ whose solution will correspond to a solution to the 3XC problem. Construct a network G = (N, A) as follows.

$N = \{1\} \cup \{\pi_1, \pi_2,...,\pi_p\} \cup S = \{\sigma_1, \sigma_2,...,\sigma_s\} \cup \{n\}.$

$A = \{\{1,\sigma_i\} : i=1,2,...,s\} \cup \{\{\sigma_i,\pi_j\}, i,j : \pi_j \in \sigma_i\} \cup \{\{\pi_j,n\} : j=1,2,...,p\}.$

origin = 1, destination = n, demand = p, C = 3.

Assign the costs and existing capacities on the arcs as follows (see Figure I.1).

Level 1.

$$\left.\begin{array}{l} a_{1\sigma_i} = 1 \\ b_{1\sigma_i} = 3 - \delta \\ ec_{1\sigma_i} = 0 \end{array}\right\} \; i = 1, 2, 3,..., s;\ \delta > 0 \text{ and sufficiently small.}$$

Level 2.

$$\left.\begin{array}{l} a_{\sigma_i\pi_j} = 1 \\ b_{\sigma_i\pi_j} = 3 \\ ec_{\sigma_i\pi_j} = 1 \end{array}\right\} \; i,j : \pi_j \in \sigma_i.$$

Level 3.

$$\left.\begin{array}{l} a_{\pi_j n} = 1 \\ b_{\pi_j n} = 3 \\ ec_{\pi_j n} = 1 \end{array}\right\} \; j = 1, 2, \ldots, p.$$

Claim I.1. *The data P and S is a Yes instance of the 3XC problem if and only if the 2F loading problem we have defined has a solution that costs p/3*(3-δ).*

Proof. If we have a Yes instance of the 3XC problem, then it is trivial to obtain a feasible solution to the loading problem that costs p/3*(3-δ). So assume that we have a Yes instance of the loading problem. Then, a solution that costs p/3*(3-δ) satisfies the following properties.

1. It uses exactly p/3 MC facilities loaded on p/3 different $\{1,\sigma_i\}$ edges.
2. The flows on arcs $\{\sigma_i,\pi_j\}$ are either 0 or 1.
3. The flows on the arcs $\{\pi_j,n\}$ are exactly one.

It is then easy to derive an exact cover for the 3XC problem. ⊗

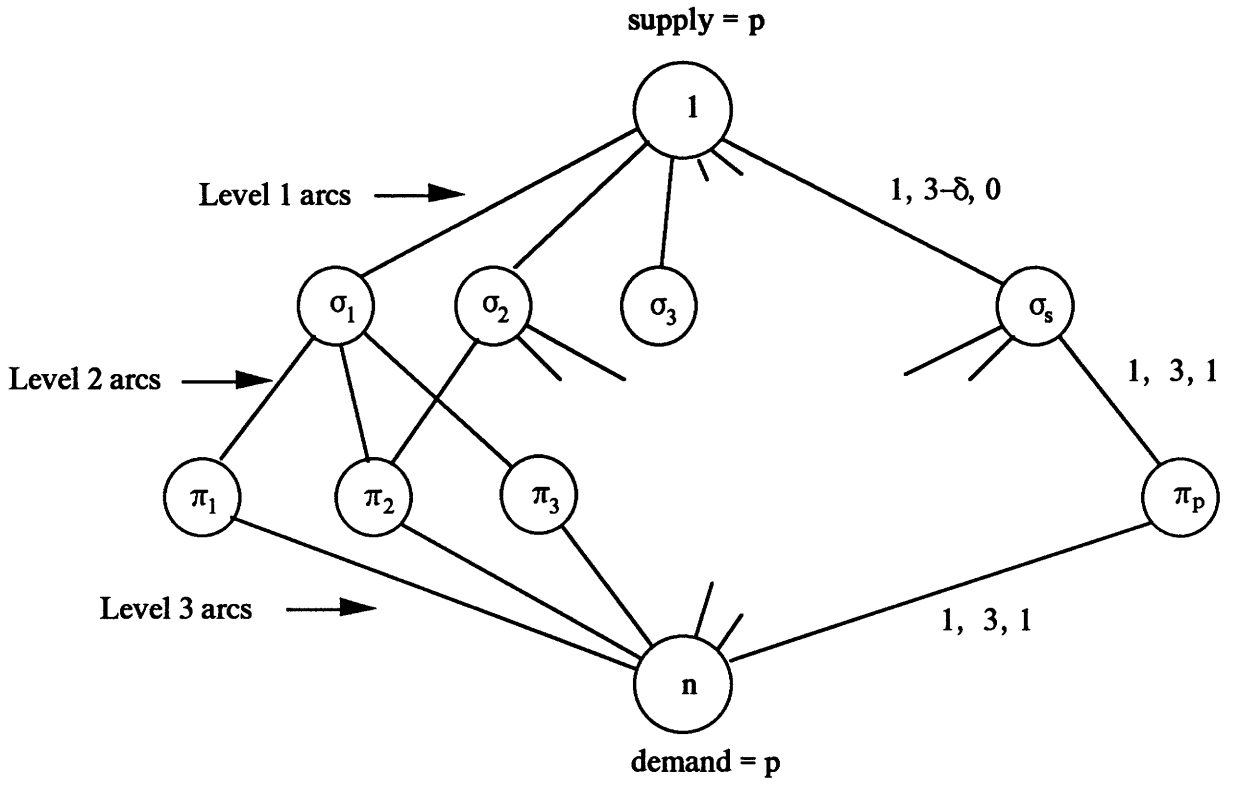


Figure I.1 Transformation of 3XC into the single commodity loading problem

REFERENCES

[1] R. Ahuja, T. L. Magnanti, and J. B. Orlin, "Network flows", in: G. L. Nemhauser, A. H. G. Rinnooy Kan and M. J. Todd, eds., *Handbooks of Operations Research and Management Science, Volume I, Optimization* (North Holland 1989) pp. 211-369.

[2] A. Balakrishnan and S. C. Graves, "A composite algorithm for a concave-cost network flow problem", *Networks* **19** (1989) 175-202.

[3] A. Balakrishnan, T. L. Magnanti, and R. T. Wong, "A dual-ascent procedure for large scale uncapacitated network design", *Operations Research* **37** (1989) 716–740.

[4] M. L. Balinski, "Fixed cost transportation problems", *Naval Research Logistics Quarterly* **8** (1961) 41–54.

[5] I. Barany, T. J. Van Roy, and L. A. Wolsey, "Uncapacitated lot-sizing: The convex hull of solutions", *Mathematical Programming Study* **22** (1984) 32–43.

[6] D. R. Fulkerson, "Blocking and antiblocking pairs of polyhedra", *Mathematical Programming* **1** (1971) 168–194.

[7] M. Goldstein and B. Rothfarb, "The one terminal telepack problem" *Operations Research* **19** (1971) 156–169.

[8] P. Gray, "Exact solution for the fixed charge transportation problem", *Operations Research* 19 (1971) 1529–1528.

[9] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Fransisco, 1979).

[10] L. J. LeBlanc and R. V. Simmons, "Continuous models for capacity design of large packet-switched telecommunications networks", *ORSA Journal of Computing* **1** (1989) 271–286.

[11] J. Leung, T. L. Magnanti, and V. Singhal, "Routing in point to point delivery systems", Working Paper OR 174-88, Operations Research Center, M.I.T., January 1988 (to appear in *Transportation Science*).

[12] T. L. Magnanti, P. Mirchandani, and R. Vachani, "The capacitated network loading problem", (1990a). In preparation.

[13] T. L. Magnanti, P. Mirchandani, and R. Vachani, "The convex hull of two core capacitated network loading problems", (1990b). In preparation.

[14] T. L. Magnanti and R. T. Wong, "Network design and transportation planning: Models and algorithms", *Transportation Science* **18** (1984) 1–55.

[15] P. Mirchandani, "Polyhedral structure of a capacitated network design problem with an application to the telecommunication industry", Unpublished Ph. D. Dissertation, MIT (Cambridge, MA, 1989).

[16] M. Minoux, "Multiflots de côut minimal avec fonctions de côut concaves", *Annales des Télélcommunications* **1** (1976) 77–92.

[17] M. Minoux, "Network synthesis and optimum network design problems: Models, solution methods and applications", *Networks* **19** (1989) 313–360.

[18] M. W. Padberg, T. J. Van Roy, and L. A. Wolsey, "Valid inequalities for fixed charge problems", *Operations Research* **33** (1985) 842–861.

[19] W. Powell and Y. Sheffi, "The load planning problem of motor carriers: Problem description and proposed solution approach", *Transportation Science* **17** (1983) 471–480.

[20] G. L. Nemhauser and L. A. Wolsey, *Integer and combinatorial optimization* (Wiley-Interscience, New York, 1988).

[21] B. Yaged, "Minimum cost routing for static network models", *Networks* **1** (1971) 139–172.

[22] N. Zadeh, "On building minimum cost communication networks", *Networks* **3** (1973) 315–331.

[23] W. I. Zangwill, "Minimum concave cost flows in certain networks", *Management Science* **14** (1968) 429–450.