OPERATIONS RESEARCH CENTER

working paper

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LOCATING MOBILE SERVERS ON A NETWORK WITH MARKOVIAN PROPERTIES

by

Oded Berman

and

Amedeo R. Odoni

OR 083-78 October 1978

Abstract

The median problem has been generalized to the case in which facilities can be moved, at a cost, on the network in response to changes in the state of the network. Such changes are brought about by changes in travel times on the links of the network due to the occurrence of probabilistic events. For the case examined here, transitions among states of the network are assumed to be Markovian. The problem is examined for an objective which is a weighted function of demand travel times and of facility relocation costs. It is shown that when these latter costs are a concave function of travel time, an optimal set of facility locations exists solely on the nodes of the network. The location-relocation problem is formulated as an integer programming problem and its computational complexity is discussed. An example illustrates the basic concepts of this paper.

Introduction

The problem of locating facilities on a network is one that has attracted an enormous amount of attention over the last fifteen years. The two classical problems are the p-median (or minisum) problem and the p-center (or minimax) problem.

The problem that we shall examine here is in many ways similar to the p-median problem. Specifically, we shall consider the situation in which demand for a service is generated only at the set (or a subset of the set) of nodes of an undirected network. A newly-generated demand will travel to its nearest server to obtain the service in question or vice versa. A number p of servers is available on the network and we shall seek to find locations for these servers such that the long-term total "cost" of offering the service is minimized. We shall introduce two important new elements, however:

First, we shall assume that, for the network at hand, the link travel times are not constant but undergo random changes. The travel time for each given link can take any one of a finite number of finite values associated with that link. As a result of this assumption the network itself, at any given time instant, can be in any one of a finite number of states, with each state differing from all others by a change in at least one link travel time. The network, moreover, makes transitions from one state to another dynamically. It will be assumed that there is a Markovian dependence among the states of the network.

Second, whereas in the p-median problem the p facilities are to be located once and for all at the points on the network which are deemed optimal, in our problem we have the option of relocating, at a cost, one or more of our p servers in reaction to changes in the state of the network.

The motivation for introducing these two new elements has been provided by our desire to develop a version of the p-median problem which is more realistic for the urban setting of applications with which we were concerned. The probabilistic transitioning of the network among states with different link travel times is intended to reflect the fact that link travel times in an urban environment do vary widely and depend on such factors as the time of day, the weather conditions, the day of the week, the occurrence of accidents, etc. The Markovian dependence among states is due to the certain degree of "predictability" and interdependence that exists with regard to changes in the state of the network as a result of the fact that certain of the factors mentioned above are predictable (e.g., time of day). The motivation for having mobile servers, as well, is obvious since many urban services involve servers which are dynamically "pre-positioned" at well-chosen locations and wait to respond to randomly occurring incidents in the areas they serve.

The problem of locating permanent (stationary) facilities on networks whose link travel times are discrete random variables, as in our case, has already been examined by Mirchandani [12] and by Mirchandani and Odoni [13]. Two observations made in these papers apply to the present problem as well. These observations also offer an insight as to why the problem of finding optimal locations on such networks is a computationally complex problem. First, it should be noted that the shortest travel time and, indeed, the shortest path itself between two given points on the network may change with the state of the network. The second observation--following directly from the last one--is that, when there are two or more servers, which demand points on the network are assigned to which particular server will depend on the state of the network (assuming that a demand is always served by its closest facility). This is true even when the servers by assumption are stationary.

An additional complication arises in our problem when the number, p, of mobile servers is greater than 1. For in this case, in addition to determining an optimum set of locations for the p servers for each and every state of the network, it is also necessary to find which mobile server is assigned to what location each time that a relocation of servers takes place as a result of a change of network state.

In the following sections we shall begin by stating our problem and assumptions in more formal terms and by defining the relevant quantities. We shall then present our main result, namely that under our set of assumptions and if the cost of relocating servers is a non-decreasing concave function of travel time, a Hakimi-type [8,9] property holds for our problem. That is, at least one set of optimal locations exists on the set of nodes of the network. (Note, however, that this still requires searching for the optimal set of p nodes corresponding to <u>each</u> state of the network and that location choices for different states are inter-dependent due to the Markovian transition probabilities among states).

The second result to be presented is that the whole problem, including the problem of which servers to relocate where, can be formulated as an Integer Programming problem whose computational complexity, however, grows explosively. The number of possible solutions, in terms of sets of locations, is equal to $\binom{n}{p}^m$, where n is the number of nodes of the network, p the number of servers and m the number of distinct states of the network. In addition, each one of these possible solutions is associated with an Assignment Problem aimed at identifying the associated optimal server-relocation strategy. An example will be used to illustrate these ideas as well as to suggest a procedure that simplifies somewhat the search for optimal locations for problems of modest size.

The paper will conclude with a discussion of the relationship of our approach to work in related areas of research. As added encouragement to the reader, we also note that, while the notation in the remainder of this paper tends to become onerous, the basic concepts are relatively simple.

The Problem, Notation and Assumptions

Let G(N,L) be an undirected network with N the set of nodes ($|N| = n$) and L the set of links. At constantly spaced time intervals (epochs) G(N,L) undergoes changes of <u>state</u>. If r and s are two distinct states of the network and if $t_r(i,j)$ indicates the travel time on link $(i,j)\varepsilon L$ (for $i\varepsilon N$, $j\varepsilon N$) then $t_r(i,j) \neq t_s(i,j)$ for at least one link $(i,j)\varepsilon L$. Let M be the set of all possible states of G, $|M| = m$.

Transitions between network states at the epochs are governed by an ergodic Markov transition matrix P with $p_{rs}\varepsilon P$ being the probability of a transition from a state r to a state s ($r\varepsilon M$, $s\varepsilon M$). We also denote the steady-state probability vector of the matrix P as π ($\pi P = \pi$, $\sum_{r=1}^{m} \pi_r = 1$).

A total of p mobile servers are to be located on the network. The servers serve demands which are generated exclusively at the nodes of G(N,L) with h_i being the conditional probability that a demand comes from node i ($i\varepsilon N$) given that a demand was generated (h_i can be viewed as the "normalized weight" of node i).

The servers are operated as follows: Whenever there is a demand for service, that demand is assigned and travels to the server closest to it, in terms of travel time. Whenever there is a change of state of the network, the operator of the service has the option of relocating one or more of the servers. A relocation of a server is associated with a cost, which we shall choose to express in units of travel time. The operator's objective is to minimize the long-term expected cost (again expressed in terms of units of travel time) of providing the service. That long term cost will be a weighted sum of the total expected travel time of demands to the servers (under all states of the network) and of the expected cost of all the server relocations that take place per unit of time. We now define some additional quantities needed to express this problem in quantitative terms:

Let $K(r) = \{K_1(r), K_2(r) \ldots, K_p(r)\}$ be a set of p points where the p servers are located when the network is in state r. We shall denote: the shortest travel time between any two points x and y on G when the network is in state r as $d_r(x,y)$; the shortest travel time between any point in the set K(r) and a specific point x on G when the network is in state s as $d_s(K(r),x)$; the shortest travel time between the α-th point in the set K(r) and the γ-th point in the set K(s) (for α and $\gamma = 1,2,\ldots,p$) when the network is in state ℓ as $d_\ell(K_\alpha(r), K_\gamma(s))$; and the shortest travel time between sets K(r) and K(s) with the network in state ℓ as $d_\ell(K(r),K(s))$.

The cost (in units of travel time) of relocating the α-th server in K(r) to the γ-th location in K(s) with the network in state s is given by $f[d_s(K_\alpha(r), K_\gamma(s))]$. We also define binary variables $W_s(K_\alpha(r), K_\gamma(s))$ as follows: if the server at $K_\alpha(r)$ is relocated to the location $K_\gamma(s)$ when the state of the network changes from r to s, then $W_s(K_\alpha(r), K_\gamma(s)) = 1$; otherwise it is equal to 0.

Finally, we define as a <u>strategy</u>, any vector $K = (K(1), K(2), \ldots, K(m))$ with m elements, where each element K(r), $r \varepsilon M$, provides the set of p locations where the servers will be placed when the network is in state r. A <u>simple strategy</u> then is any strategy with $K(1) = K(2) = \ldots = K(m)$, i.e., a strategy in which servers remain stationary under all states of the network.

We shall now state the assumptions under which the results of this paper have been derived:

1. The travel times $t_r(i,j)$ for all $r \varepsilon M$ and all $(i,j) \varepsilon L$ are finite.
2. The time required to travel a fraction θ $(0 \leq \theta \leq 1)$ of any link $(i,j) \varepsilon L$ for any state $r \varepsilon M$ is equal to $\theta \cdot t_r(i,j)$.
3. The current state of the network is known to the service operator at all times.

4. Time intervals between changes of state are much longer than trip times on the network.
5. No demands or further state changes occur while servers are being relocated after a change of network state.
6. All demands are served by the closest (in terms of travel time) servers and all servers are available whenever a demand occurs.
7. The relocation cost function, $f(\cdot)$, is non-decreasing concave.

Assumption 1 assures connectivity of the network under all states. A less restrictive version of Assumption 1 (one that allows some of the $t_r(i,j)$ to be infinite while still leading to the same results) is given in Mirchandani [12] and in Berman [1]. Assumption 2 concerning uniformity of travel time on any given link will be used in the proof of Theorem 1 in the Appendix. Assumption 3 allows the operator of the service to always choose the path with the shortest travel time when directing a server from its location to a demand point. Assumption 4 renders negligible the probability that link travel times will change while a server is travelling to a demand. (Were this to happen the server might no longer be travelling on a shortest travel time path.)

Assumptions 5 and 6 are the major simplifying assumptions in our analysis. Both Assumptions would be approximately true, in practice, if, for instance, the average time interval between generation of demands on the network was much longer than travel times on the network (assuming that demands are generated according to a stationary renewal process independently of the state of the servers). Assumption 6 would also be true if each mobile server could simultaneously serve any number of incidents (e.g., mobile libraries or "bloodmobiles"). It is worth noting at this point that, although seldom stated explicitly, Assumption 6 is fundamental to the work that has been done to date on the p-median and p-center problems. The implicit assumption in this area of research has been that facilities (medians or centers) are always available to serve demands--perhaps by having unlimited service capacity--and are not subject to queueing type of phenomena. We shall return to this particular point at the conclusion of our paper.

Finally, Assumption 7 is necessary for Theorem 1 to hold. It implies "economies of scale" for the cost of travel times--a reasonable hypothesis in most practical contexts. (Obviously the family of acceptable functions, $f(\cdot)$, also includes the linear cost function.)

The Problem

We can now express our objective function in terms of the quantities that we have defined. For any given strategy K = (K(1),K(2),...,K(m)) the quantity,

$$A = \sum_{r=1}^{m} \sum_{i=1}^{n} \pi_r h_i d_r(K(r),i) \tag{1}$$

gives the long term ("steady-state") expected travel time to servers on the network per transition epoch. Similarly the quantity

$$B = \sum_{r=1}^{m} \sum_{\substack{\ell=1 \\ \ell \neq r}}^{m} \pi_r \cdot p_{r\ell}\left[\sum_{\alpha=1}^{p} \sum_{\gamma=1}^{p} W_\ell(K_\alpha(r),K_\gamma(\ell)) \cdot f[d_\ell(K_\alpha(r),K_\gamma(\ell))]\right] \tag{2}$$

represents the long-term expected cost of server relocations per transition epoch, taking into account all possible changes of state from any possible state.

Our problem is to minimize the weighted average

$$Z = A + c_1 B \tag{3}$$

of the two quantities above with c_1 being the weight assigned to server relocation costs per transition by comparison to expected travel time (for convenience we will assume from now on that $c_1 = 1$). Two sets of contraints apply to our problem:

$$\sum_{\alpha=1}^{p} W_\ell(K_\alpha(r),K_\gamma(\ell)) = 1 \qquad \text{for } \gamma = 1,2,\ldots,p,;\forall r,\ell \in M;\ r \neq \ell \tag{4}$$

$$\sum_{\gamma=1}^{p} W_\ell(K_\alpha(r),K_\gamma(\ell)) = 1 \qquad \text{for } \alpha = 1,2,\ldots,p;\ \forall r,\ell \in M;\ r \neq \ell \tag{5}$$

The following theorem can now be proved.

Theorem 1 At least one set of optimal locations for the problem above exists on the nodes of the network.

A full proof of the theorem is included in the Appendix. As a result of this theorem our location-relocation problem has been reduced from optimization over an infinite set of points to an optimization over a finite set of nodes. The total number of strategies when we consider locations only on nodes is $\binom{n}{p}^m$. Note also that the number of simple strategies is $\binom{n}{p}$. In the next section we present a simple numerical example in which we use Theorem 1.

A Numerical Example

Let us consider the simple network of Figure 1. This network can be in one of two states, 1 and 2. The numbers next to the links of the network

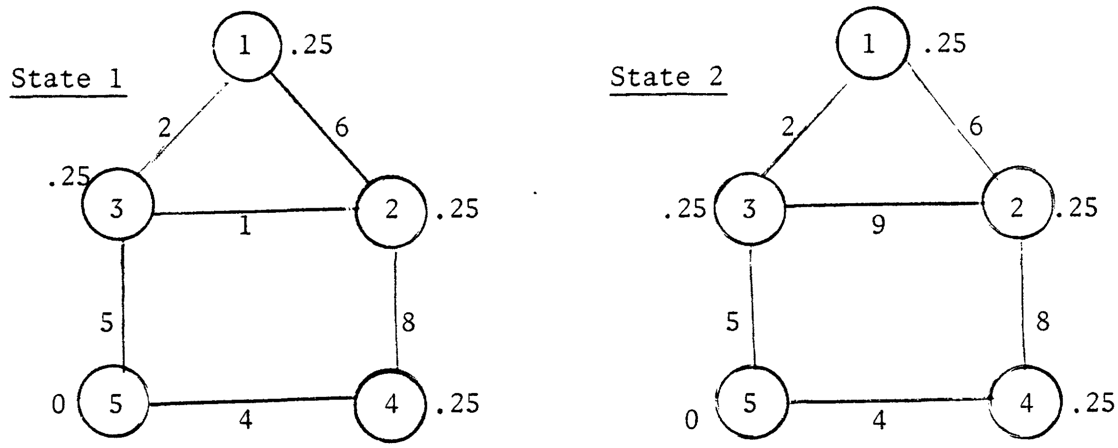

FIGURE 1: The Network Under States 1 and 2

represent lengths (travel times) whereas the numbers next to the nodes are the weights, h_i. Obviously, the only difference between the two states is the travel time on link (3,2) which is equal to 1 under state 1 and to 9 under state 2.

The Markovian transition matrix, P, that describes the statistical dependence between the two states is shown below.

P =

state \ state	1	2
1	.25	.75
2	.5	.5

The steady state probabilities associated with P are $\pi_1 = .4$ and $\pi_2 = .6$. Let $f(.) = (.1)\sqrt{(.)}$ be the relocation cost function, an increasing concave one.

Suppose that we wish to locate two servers on our network. Using the result of Theorem 1 we have to find an optimal strategy $K^* = (K^*(1),K^*(2))$ where $K^*(1) = \{i,j\}, K^*(2) = \{k,\ell\}$, with $i,j,k,\ell \varepsilon \{1,2,3,4,5\}$ as well as an optimal relocation plan $W_2(K^*_\alpha(1),K^*_\gamma(2))$, $W_1(K^*_\alpha(2),K^*_\gamma(1))$; $\alpha,\gamma = 1,2$. The number of possible strategies now is $\binom{5}{2}^2 = 100$ and the number of simple strategies is $\binom{5}{2} = 10$.

The shortest distance matrices for the two states of the network can be derived easily by inspection and are shown in Table 2, as D_r, $r = 1,2$. We also include in the table two other distance matrices $\hat{D}_1$ and $\hat{D}_2$. These matrices give the shortest distance to any one of the ten possible sets of facility locations from each one of the nodes of the network for the two states, 1 and 2, respectively.

The problem can now be solved easily by hand. By performing the operation $\pi_r h\hat{D}_r$, $r = 1,2$ we obtain all the possible components of the term A (the long term expected travel time to a random demand point) in the objective function (3). The results of this operation are given in Table 1 below.

TABLE 1: The Long Term Expected Travel Time to a Random Demand Point

Location	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
$\pi_1 \sum_{i=1}^{5} h_i d_1(K(1),i)$	.9	1	.5	.9	1	.4	.8	.3	.7	1.8
$\pi_2 \sum_{i=1}^{5} h_i d_2(K(2),i)$	1.5	2.25	1.2	1.8	1.5	2.1	2.25	1.5	2.1	3
Total	2.4	3.25	1.7	2.7	2.5	2.5	3.05	1.8	2.8	4.8

Some helpful observations can now be made. First, the last row in Table 1 provides us with the value of the objective function (3) for all ten

TABLE 2: Distance Matrices

D_1 =

Node \ Node	1	2	3	4	5
1	0	3	2	11	7
2	3	0	1	8	6
3	2	1	0	9	5
4	11	8	9	0	4
5	7	6	5	4	0

D_2 =

Node \ Node	1	2	3	4	5
1	0	6	2	11	7
2	6	0	8	8	12
3	2	8	0	9	5
4	11	8	9	0	4
5	7	12	5	4	0

$\hat{D}_1$ =

Node \ K(1)	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
1	0	0	0	0	2	3	3	2	2	7
2	0	1	3	3	0	0	0	1	1	6
3	1	0	2	2	0	1	1	0	0	5
4	8	9	0	4	8	0	4	0	4	0
5	6	5	4	0	5	4	0	4	0	0

$\hat{D}_2$ =

Node \ K(2)	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
1	0	0	0	0	2	6	6	2	2	7
2	0	6	6	6	0	0	0	8	8	8
3	2	0	2	2	0	8	5	0	0	5
4	8	9	0	4	8	0	4	0	4	0
5	7	5	4	0	5	4	0	4	0	0

simple strategies. Among them the simple strategy K(1) = K(2) = (1,4) is the best, and can serve as the "incumbent"--the best strategy obtained so far. All other simple strategies can be eliminated from further consideration. Second, the value of A in (3) for all remaining 90 (non-simple) strategies can be obtained very easily from Table 1, just by adding relevant terms from the first two rows of the table. Third, since all the terms in the objective function (3) are always nonnegative, many other strategies can be eliminated, merely by inspection. Actually, any strategy with a value of A greater than or equal to 1.7 (the value of the objective function for the incumbent strategy (1,4)), can be immediately eliminated. This leaves us with only two non-simple strategies, $K_1 = \{(3,4),(1,4)\}$ and $K_2 = \{(2,4),(1,4)\}$, while eliminating all the other 88 non-simple strategies.

Let us consider now the first strategy K_1. We solve the following two assignment problems (for relocations when the network changes from state 1 to state 2 and vice versa) for K_1:

$$\min\{w_2(3,1)(.141) + w_2(3,4)(.3) + w_2(4,1)(.331) + w_2(4,4)(0)\}$$

s.t.

$$w_2(3,1) + w_2(3,4) = 1$$

$$w_2(4,1) + w_2(4,4) = 1$$

$$w_2(3,1) + w_2(4,1) = 1$$

$$w_2(3,4) + w_2(4,4) = 1$$

$$w_2(3,1), w_2(3,4), w_2(4,1), w_2(4,4) = 0,1$$

and

$$\min\{w_1(1,3)(.141) + w_1(1,4)(.331) + w_1(4,3)(.3) + w_1(4,4)(0)\}$$

s.t.

$$
\begin{aligned}
w_1(1,3) + w_1(1,4) \qquad\qquad\qquad\qquad &= 1 \\
w_1(4,3) + w_1(4,4) &= 1 \\
w_1(1,3) \qquad + w_1(4,3) \qquad\qquad &= 1 \\
w_1(1,4) \qquad\qquad + w_1(4,4) &= 1 \\
w_1(1,3), w_1(1,4), w_1(4,3), w_1(4,4) &= 0,1
\end{aligned}
$$

(The co-efficients in the objective functions are obtained from the cost function $f(\cdot)$, e.g., $0.1\ (\sqrt{2}) = .141$). We then obtain: (a) the value of the term B (the long term expected cost of server relocations) in (3) is .0846 and thus the value of the objective function (3) is 1.5846; (b) the relocation solution is:

$$w_2(3,1) = w_2(4,4) = w_1(1,3) = w_1(4,4) = 1 \quad \text{and}$$

$$w_2(3,4) = w_2(4,1) = w_1(4,3) = w_1(1,4) = 0.$$

Since $1.5846 < 1.7$, K_1 becomes the new incumbent.

By returning to Table 1 again, we can now eliminate strategy K_2 since the value of A for K_2 is $1.6 > 1.5846$. Therefore the non-simple strategy $K_1 = \{(3,4),(1,4)\}$ is the optimal strategy.

The optimal solution to our problem thus is: the two servers should be located on nodes {3,4} if the network is at state 1 and on nodes {1,4} if the network is at state 2. When the network changes from state 1 to state 2, the server on node 4 should stay at its location ($w_2(4,4) = 1$) whereas the server on node 3 should be relocated to node 1, and vice-versa when the change of states is from 2 to 1. The long-term cost of this policy is 1.5846 units of time per transition epoch.

It is worth noting that if the relocation cost function is $f(.) = C\sqrt{(.)}$ where $C \geq .2357$, the simple strategy K(1) = K(2) = (1,4) is the optimal strategy. This is hardly surprising, since relocations for large C become expensive, and, therefore, it will eventually become optimal to make the servers stationary.

Finally, it should be clear that our approach to solving this example can be applied only when the problem is of very modest size. In the next section we present a mathematical programming formulation that hopefully can be used to obtain solutions to larger problems.

Mathematical Programming Formulation of the Problem

The problem of locating mobile servers on a network with Markov properties can be formulated as an integer programming problem. Let us define the following three sets of binary variables:

$$Y_{i,j,r} = \begin{cases} 1 & \text{if node } i \text{ is served by the server at node } j \text{ when the network is at state } r \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{j,r} = \begin{cases} 1 & \text{if the server is located at node } j \text{ when the network is at state } r \\ 0 & \text{otherwise} \end{cases}$$

$$X_{u(r),v(\ell)} = \begin{cases} 1 & \text{if the server in node } u \text{ is assigned to node } v \text{ when the network changes from state } r \text{ to state } \ell \\ 0 & \text{otherwise} \end{cases}$$

for $i,j,u,v \varepsilon N;\ r,\ell \varepsilon M,\ r \neq \ell$

Now we can write the problem as:

$$\min\{\sum_{r=1}^{m} \pi_r \sum_{i=1}^{n} \sum_{j=1}^{n} h_i d_r(i,j) Y_{i,j,r} + c_1 \sum_{r=1}^{m} \pi_r \sum_{\substack{\ell=1 \\ \ell\neq r}}^{m} p_{r\ell} \sum_{u=1}^{n} \sum_{v=1}^{n} X_{u(r),v(\ell)} \cdot f[d_\ell(u,v)]\} \tag{6}$$

subject to:

$$Y_{i,r} + \sum_{\substack{j=1 \\ j\neq i}}^{n} Y_{i,j,r} = 1 \qquad \forall i \varepsilon N;\ \forall r \varepsilon M \tag{7}$$

$$Y_{j,r} \geq Y_{i,j,r} \qquad \forall i,j \varepsilon N;\ \forall r \varepsilon M \quad i \neq j \tag{8}$$

$$\sum_{v=1}^{n} X_{u(r),v(\ell)} = Y_{u,r} \qquad \forall u \varepsilon N;\ \forall r,\ell \varepsilon M \quad r \neq \ell \tag{9}$$

$$\sum_{u=1}^{n} X_{u(r),v(\ell)} = Y_{v,\ell} \qquad \forall v \varepsilon N;\ \forall r,\ell \varepsilon M \quad r \neq \ell \tag{10}$$

$$\sum_{j=1}^{n} Y_{j,r} = p \qquad \forall r \varepsilon M \tag{11}$$

$$Y_{j,r}, Y_{i,j,r}, X_{i(r),j(\ell)} = 0,1 \qquad \forall i,j \varepsilon N \quad \forall r,\ell \varepsilon M \quad r \neq \ell \tag{12}$$

Constraints (7) make sure that every node is served by a server. Constraints (8) limit the assignment of demands to only those nodes at which servers are located. Constraints (9) and (10) limit the relocations to only those nodes that are also actual locations. Constraints (11) restrict the number of servers to p and constraints (12) restrict all variables to be zero or one.

Obviously, the size of this I.L.P. problem grows very quickly as n, p, and m increase. We have no computational experience with problems of this type to date. However, due to considerable similarity with the formulation of the deterministic median problem, the recent research of Erlencotter [5], Garfinkel, Neebe and Rao [7], Revelle and Swain [14], Galvao [6], Jarvinen, Rajala and Sinervo [10] and Cornuejols, Fisher, and Nemhauser [4] can be of help in solving our problem as well. Therefore we believe that (at least heuristic) solutions to our problem with $n \leq 50$ and $m \leq 5$ (for any value of p) are well within possibility.

Discussion

The mobile server location presented here is a very general one, in the sense that most known versions of minisum facility location problems on networks can be viewed as special cases of our problem. For instance, the problem discussed by Mirchandani and Odoni [13] in which stationary facilities must be located on a network that undergoes probabilistic transitions among m states is a special case of our problem in which only simple strategies are permitted. This eliminates the need to examine assignments of servers since the variables $W_s(K_\alpha(r),K_\gamma(s))$ are equal to 1 for $\alpha = \gamma$ and to 0 otherwise for all values of α, γ, r and s. If in the Mirchandani-Odoni problem, we further allow the number of states m to be reduced to 1, we obtain the classical p-median problem [9] and, naturally, with p = 1 we are back to the original single median ("minisum facility") location problem.

Be that as it may, it is clear that even our model, if it is to be used for applications in the context of some urban services, still suffers from some major oversimplifications. Foremost among them is the assumption that no queueing phenomena occur at the service locations, i.e., that servers are always available as demands are generated. In this respect, our model can be viewed as the opposite of such well-known, urban service system models as Larson's hypercube [11] or Carter, Chaiken and Ignall's [3]: the latter are primarily concerned with problems caused by "congestion" and unavailability of servers (and less concerned about good server placement in anticipation of demands) while our model reverses these priorities. Recently Berman [1] and Berman and Larson [2] combined the hypercube with a model similar to the one presented in this paper to formulate a model which, for at least some cases, can be used to make both server location decisions and resource allocation (to prevent facility congestion) decisions, simultaneously.

The discussion above, incidentally, indicates as well how inadequate the traditional p-median model is in the applications context of urban services. That model implicitly assumes no queueing phenomena whatsoever at the facilities and makes no allowance for system response (by relocating servers) to the dynamics of the network (changes in travel times) or of the demands.

Before closing this discussion, we note that, while our analysis here assumed a discrete time Markovian model for transitions between network states, i.e., constantly spaced transition intervals, we could, as well--with only minor modifications--analyze a continuous time Markovian model or a semi-Markovian model (in which an embedded Markovian probability matrix dictates transitions between states at the transition epochs). In the latter case we would have to assign a weight to each state in the quantities A and B in (3), proportional to the expected duration of the state, i.e., use the limiting probabilities of the Markov matrix (see, for instance, Ross [15], page 104).

Finally, we also note that our analysis above can be applied just as well to directed networks. The optimal location-relocation strategies would then clearly depend on whether demands are assumed to travel to the servers or vice versa.

References

[1] Berman, O., "Dynamic Positioning of Mobile Servers on Networks," Technical Report No. 144, Operations Research Center, M.I.T.

[2] Berman, O. and Larson, R.C., "The Congested Median Problem," Working Paper OR 076-78, M.I.T., submitted to Operations Research.

[3] Carter, G.M., Chaiken, J.M. and Ignall, E., "Response Areas for Two Emergency Units," Operations Research 20, No. 3, May-June 1972, pp. 571-594.

[4] Cornuejols, G., Fisher, M.L. and Nemhauser, G.L., "Location of Bank Accounts to Optimize Float," Management Science 23, No. 8, April 1977, pp.

[5] Erlenkotter, D., "A Dual-Based Procedure for Uncapacitated Facility Location," Working Paper 261, Western Management Science Institute, U.C.L.A., California.

[6] Galvao, R.D., "A Dual-Bounded Algorithm for the P-Median Problem," International Symposium on Locational Decisions, April 24-28, 1978, Banff, Alberta.

[7] Garfinkel, R.S., Neebe, A.W. and Rao, M.R., "An Algorithm for the M-Median Plant Location Problem," Transportation Science 8, No. 3, August 1974, pp. 217-231.

[8] Hakimi, S.L., "Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph," Operations Research 12, 1964, pp. 450-459.

[9] Hakimi, S.L., "Optimum Distribution of Switching Centers on a Communication Network and Some Related Graph Theoretic Problems," Operations Research 13, 1965, pp. 462-475.

[10] Jarvinen, P., Rajala, J. and Sinvero, H., "A Branch and Bound Algorithm for Seeking the P-Median," Operations Research 20, No. 1, Jan.-Feb. 1972, pp. 173-182.

[11] Larson, R.C., "A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services," Computers and Operations Research 1, March 1974, pp. 67-95.

[12] Mirchandani, P.B., "Analysis of Stochastic Networks in Emergency Service Systems," IRP-TR-15-75, Operations Research Center, M.I.T.

[13] Mirchandani, P.B. and Odoni, A.R., "Locations of Medians on Stochastic Networks," Technical Report OR 065-77, M.I.T. Operations Research Center, Cambridge, MA, October 1977. Submitted to Transportation Science.

[14] Revelle, C. and Swain, R., "Central Facilities Locations," Geographical Analysis 2, No. 1, January 1970, pp. 30-42.

[15] Ross, S.M., Applied Probability Models with Optimization Applications, Holden-Day, San Francisco.

Appendix 1: Proof of Theorem 1

Let us define (for convenience),

$$g_\ell(K(r),K(\ell)) = \sum_{\alpha=1}^{P} \sum_{\gamma=1}^{P} w_\ell(K_\alpha(r),K_\gamma(\ell))f[d_\ell(K_\alpha(r),K_\gamma(\ell))].$$

Let $K^*(r), w_\ell(K^*_\alpha(r),K^*_\gamma(\ell))$, $\forall\, r,\ell \in M$; $\alpha,\gamma = 1,2,\ldots,P$ be the optimal solution to our problem. Suppose that for state $s \in M$, $K^*_\beta(s)$ is interior to the link $(a,b) \in L$. By Assumption 2 we can define

$$\frac{t_r(a,K^*_\beta(s))}{t_r(a,b)} = \theta \quad \text{with} \quad 0 < \theta < 1 \quad \forall r \in M \tag{13}$$

Now we can rewrite (3) as

$$\pi_s \sum_{i=1}^{n} h_i d_s(K^*(s),i) + \pi_s \sum_{\substack{\ell=1\\ \ell\neq s}}^{m} p_{s\ell} g_\ell(K^*(s),K^*(\ell))$$

$$+ \sum_{\substack{r=1\\ r\neq s}}^{m} \pi_r p_{rs} g_s(K^*(r),K^*(s)) + A \tag{14}$$

where the term A includes only states $r \neq s$:

$$A = \sum_{\substack{r=1\\ r\neq s}}^{m} \pi_r \sum_{i=1}^{n} h_i d_r(K^*(r),i) + \sum_{\substack{r=1\\ r\neq s}}^{m} \pi_r \sum_{\substack{\ell=1\\ \ell\neq s,r}}^{m} p_{r\ell} g_\ell(K^*(r),K^*(\ell)) \tag{15}$$

Let $N[K^*_\alpha(s)]$ be the set of all the nodes that will be visited by the facility located at $K^*_\alpha(s)$, in case of a random incident when the network is at state s; $\alpha = 1,\ldots,P$.

Therefore:

$$\pi_s \sum_{i=1}^{n} h_i d_s(K^*(s),i) = \pi_s \sum_{\alpha=1}^{p} \sum_{i\varepsilon N[K^*_\alpha(s)]} h_i d_s(K^*_\alpha(s),i)$$

$$= \pi_s \sum_{i\varepsilon N[K^*_\beta(s)]} h_i d_s(K^*_\beta(s),i) + B \qquad (16)$$

where the term B does not include any node of the set $N[K^*_\beta(s)]$

$$B = \pi_s \sum_{\substack{\alpha=1 \\ \alpha\neq\beta}} \sum_{i\varepsilon N[K^*_\alpha(s)]} h_i d_s(K^*_\alpha(s),i) \qquad (17)$$

Let $N_a[K^*_\beta(s)] \subset N[K^*_\beta(s)]$ be the set of all nodes that communicate most efficiently with $K^*_\beta(s)$ via a when the system is at state $s\varepsilon M$ (the term "communicate" implies minimal travel time), and let $N_b[K^*_\beta(s)] = N[K^*_\beta(s)] - N_a[K^*_\beta(s)]$. If a node communicates equally efficiently with $K^*_\beta(s)$ via nodes a and b we include that node either in $N_a[K^*_\beta(s)]$ or in $N_b[K^*_\beta(s)]$, but not in both. Therefore we can write (16) as:

$$\pi_s \sum_{i\varepsilon N_a[K^*_\beta(s)]} h_i d_s(K^*_\beta(s),i) + \pi_s \sum_{i\varepsilon N_b[K^*_\beta(s)]} h_i d_s(K^*_\beta(s)] + B \qquad (18)$$

By (13) for r = s we get that: $\forall i\varepsilon N_a[K^*_\beta(s)], d_s(K^*_\beta(s),i) = d_s(i,a) + \theta t_s(a,b)$ and $\forall i\varepsilon N_b[K^*_\beta(s)], d_s(K^*_\beta(s),i) = d_s(i,b) + (1-\theta)t_s(a,b)$.

Therefore we can write (18) as:

$$\pi_s H(s,a) \cdot \theta + \pi_s H(s,b) \cdot (1-\theta) + C + B \qquad (19)$$

where the terms H(s,a), H(s,b) and C do not include travel time from the point $K^*_\beta(s)$:

$$H(s,a) = \sum_{i \varepsilon N_a[K^*_\beta(s)]} h_i t_s(a,b) \tag{20}$$

$$H(s,b) = \sum_{i \varepsilon N_b[K^*_\beta(s)]} h_i t_s(a,b) \tag{21}$$

$$C = \pi_s \sum_{i \varepsilon N_a[K^*_\beta(s)]} h_i d_s(i,a) + \pi_s \sum_{i \varepsilon N_b[K^*_\beta(s)]} h_i d_s(i,b) \tag{22}$$

The term $g_\ell(K^*(s),K^*(\ell))$, $\ell \neq s$ can be written as:

$$\sum_{\gamma=1}^{p} w_\ell(K^*_\beta(s),K^*_\gamma(\ell)) f[d_\ell(K^*_\beta(s),K^*_\gamma(\ell)] + D \tag{23}$$

where the term D, does not include relocations from the point $K^*_\beta(s)$:

$$D = \sum_{\substack{\alpha=1 \\ \alpha \neq \beta}}^{p} \sum_{\gamma=1}^{p} w_\ell(K^*_\alpha(s),K^*_\gamma(\ell)) f[d_\ell(K^*_\alpha(s),K^*_\gamma(\ell))] \tag{24}$$

Again by using (13) for $r = \ell$ we can write $d_\ell(K^*_\beta(s),K^*_\gamma(\ell))$ as:

$$\min\{\theta t_\ell(a,b) + d_\ell(a,K^*_\gamma(\ell)); (1-\theta) t_\ell(a,b) + d_\ell(b,K^*_\gamma(\ell))$$

Therefore, $d_\ell(K^*_\beta(s),K^*_\gamma(\ell))$ is a minimum of two linear functions of θ and therefore concave. But since f is non decreasing and concave, $f[d_\ell(K^*_\beta(s),K^*_\gamma(\ell))]$ is also concave. Therefore, $g_\ell(K^*(s),K^*(\ell))$ is concave too since it is a linear combination of concave functions. In the same way:

$g_s(K^*(r),K^*(s))$, $r \neq s$ is also concave.

Using our discussion so far we can write now (3) as:

$$\pi_s H(s,a)\theta + \pi_s H(s,b)(1-\theta) + \pi_s \sum_{\substack{\ell=1 \\ \ell \neq s}}^{m} p_{s\ell} g_\ell(K^*(s),K^*(\ell))$$

$$+ \sum_{\substack{r=1 \\ r \neq s}}^{m} \pi_r p_{rs} g_s(K^*(r),K^*(s)) + A + B + C \tag{25}$$

But since π_s, $p_{s\ell}$, p_{rs}, H(s,a), H(s,b), A, B, C are all constants with respect to θ, (25) is also concave in θ. Therefore the value of the objective function (3) cannot increase when taking either $\theta = 0$ or $\theta = 1$ (but not both) corresponding to location at a or b, respectively (we keep the relocation variables $w_\ell(K^*_\alpha(r), K^*_\gamma(\ell))$ as before with a (or b) replacing $K^*_\beta(s)$). Clearly the node a is optimal if the coefficient of θ in (25) is larger than the coefficient of $(1-\theta)$. Otherwise b is optimal or a tie exists, in which case either is optimal. Once the node a or b is reached, members of the route partitioning sets may have to be interchanged and new relocation solutions may be obtained, to improve the value of (3) achieved with the original route-partitioning sets and relocation solutions. Moreover, the same proof with the new route-partitioning sets and relocation solutions demonstrates the nonoptimality of moving away from the node.