

This paper has been accepted for publication at

Transportation Science

Profit Maximizing Distributed Service System Design with Congestion and Elastic Demand

Robert Aboolian* Oded Berman and Dmitry Krass†

Revised, May 2011

Abstract

In this paper we developed a service network design model that explicitly takes into account the elasticity of customer demand with respect to travel distance and congestion delays. The model incorporates a feedback loop between customer demand and congestion at the facilities. The problem is to determine the number of facilities, their locations, their service capacity and the assignment of customers to facilities so as to maximize the overall profit of the system. Two versions of the problem are presented. In one each facility is modeled as an $M/M/1$ queuing system where the service rate is a decision variable whereas in the other one the facility is modeled as an $M/M/k$ queuing model where the service rate is given but the number k is a decision variable. An exact algorithm and heuristics are developed and tested via computational experiments. While our model is of “directed choice” type where the assignment of customers to facilities is controlled by the decision-maker, computational results show that in the vast majority of the cases the customers are assigned to the utility-maximizing facility, indicating there is no conflict between the customers’ and decision-makers’ goals. A case study of locating preventive medicine clinics in Toronto, Ontario illustrates the model.

KEYWORDS: Service System Design, Elastic Demand, Congestion, Nonlinear Integer Program.

*College of Business Administration, California State University San Marcos, San Marcos, California 92096, USA

†Both authors at Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario, Canada M5S 3E6

1 Introduction

In this paper we address the problem of designing a service system consisting of a network of facilities facing stochastic demand from customers residing at the nodes of a network. The problem is to determine the number of facilities, their locations, as well as the service capacity of each facility and the assignment of customers to facilities, so as to maximize the overall profit (net revenue) of the system. It is assumed that the customer demand is elastic with respect to both the travel distance and service time at the facilities, where the stochasticity of demand creates congestion-related service delays.

We note that the two components affecting customer demand are quite different: while the travel distance can be estimated in advance and is invariant with respect to customers' behavior, the congestion delay is a direct result of the choices made by customers on whether to procure service from a given facility or not. As the customers facing congestion delays reduce their frequency of visits to a facility, the congestion delay itself is reduced. Thus, unlike the travel distance, the congestion delay is a result of a certain equilibrium behavior which is affected by both, the capacities of the facilities and by the sensitivity of customer demand to congestion delays. In this paper we develop a model that allows us to directly capture the feed-back loop between congestion delays and customer's reactions to these delays. The potential applications of this model can be found in both private and public sectors. We illustrate one such potential application with a case study of locating a network of preventive health care facilities. In the latter part of the paper we discuss how our model can be used to determine the number and locations of the facilities for the applications where facility locations are flexible.

The main contributions of the paper are:

1. *Explicit modeling of the interconnection between customer demand and congestion delays*

Most previous works in location literature do not consider the equilibrium behavior described earlier. Instead, a certain service level requirement is assumed at all facilities (e.g., expected waiting time not to exceed 10 mins 90% of the time), and the customer demand is assumed to be deterministic (inelastic) as long as this service requirement is met. In our case, while we assume certain minimum service level guarantee has to be enforced, the actual service level may be different at each facility - as dictated by the capacity costs and potential losses of customer demand due to congestion. The standard assumption with respect to distance is of cover type - as long as the customer's location is within a certain

coverage radius of the closest facility, full demand is captured. In our case we assume elasticity of demand with respect to distance with no fixed coverage radius.

2. *Profit optimization objective.* In most prior work the objective is either to capture all available demand with minimum cost or to capture the most available demand with a fixed number of facilities. However, due to the simplifying assumptions of inelastic demand and fixed service level (explained above), the expected profit cannot be computed and thus “true” profit optimization is impossible. In contrast, our model allows us to capture the direct trade-off between the cost of increasing the capacity in certain parts of the system and the resulting extra revenue, allowing us to optimize the profitability of the resulting system.
3. *Development of an efficient exact algorithm for finding the optimal solutions to the problem.* Our model belongs to the class models with equilibrium constraints for which few, if any, efficient optimization schemes are available; mostly one has to rely on heuristic approaches. In our case, by separating the capacity assignment from the customer assignment and location subproblems, we are able to find exact optimal solutions for fairly large-scale instances

Our model belongs to the class of location models with immobile servers (since customers are assumed to travel to the facilities to obtain service). The early work in this direction ignored congestion and capacity considerations altogether, focusing only on demand sensitivity to distance - leading to the development of the Uncapacitated Facility Location Problem (UFLP) - see Cornuejols, Nemhauser and Wolsey (1990). A notable attempt to incorporate congestion costs into a UFLP model is Desrocheers, Marcotte and Stan (1995), who assumed fixed (inelastic) demand and exogenously-specified congestion cost; the objective is to minimize the overall system cost consisting of fixed location costs, congestion cost (that incorporates the capacity cost) and travel costs. Incorporation of stochastic demand and the resulting queuing behavior (which yields endogenous congestion delays) originated with the papers by Marianov and Serra (1998) and Marianov and Rios (2000). In these papers the sensitivity of demand to both travel distance and congestion delays is modeled in a rather simple manner: the covering location model is used to represent distance sensitivity implying that customers are assumed to be not distance-sensitive at all within a certain distance of the facility (known as the “coverage radius”), with the demand falling to 0 outside of this distance. Similarly, a certain acceptable waiting time requirement is

enforced at the facilities (through a hard constraint), with the customers assumed to be insensitive to the congestion delays below this time. Thus, as long as a customer does not have to travel further than the coverage radius or wait longer than the waiting time requirement, 100% of the customer's demand is assumed to be captured.

A different median-type perspective was taken in several papers [Wang, Batta and Rump (2002), Berman and Drezner (2006,2007)] who seek to minimize the objective consisting of a weighted combination of the travel time and expected waiting time, thus taking the customer's point of view; in these papers the elasticity of demand is not captured explicitly - all customers attempt to visit the facilities and the resulting congestion delays are penalized. Wang, Batta and Rump also considered a different objective which minimizes the total service and location costs (the service provider's perspective), while several authors [Aboolian, Berman and Drezner (2008), Elhedehli (2006), Castillo, Ingolfsson and Sim (2010), Amiri (1997)] have taken the so-called "socially optimal" perspective, where the objective function includes all of the terms mentioned above.

The first paper to explicitly model demand losses resulting from the elasticity with respect to travel distance and congestion appears to be Berman and Kaplan (1987) who analyzed a one-facility system and, under fairly strong assumptions, were able to decouple the location and capacity decisions. Berman, Krass and Wang (2006) assumed a gradual loss of demand due to distance (essentially following the UFLP framework) and that the demand is lost when the waiting time exceeds certain threshold (thus the demand loss due to congestion is modeled as a step function while in the current paper we assume a general functional form of the elasticity of demand with respect to the congestion delay); no explicit equilibrium-type conditions were included in their model. Other models taking elastic demand into account but without explicit equilibrium feed-back constraints (and employing only heuristic approaches) are Marianov, Rios and Barros (2005) and Marianov, Rios, and Icaza (2008).

Very recently, Zhang, Berman and Verter (2009) have analyzed a multi-location model with elastic demand and congested-related delays; the customers select facilities that minimize the sum of travel and waiting time. This approach was further improved in Zhang, Berman, Marcotte, and Verter (2010) where explicit traffic-type equilibria were computed via convex programming (the initial paper used only heuristic estimates). There are several important differences between the last two papers and the current paper. First, we consider a profit-maximizing design, while these papers focused on maximizing accessibility. Second, we allow for a much more general

form of relationship between demand and waiting times (the previous papers assumed a linear relationship). Third, we assume the facility choice is independent of congestion delays (i.e., the customers are not aware of the delays when they plan their trips). Most importantly, we are able to develop an exact optimization approach for our model, while the previous papers relied entirely on heuristic methods (at least for location) decisions.

The remainder of the paper is organized as follows. In the next section we present a formal description of the problem and formulate our model. In Section 3 we show how, once the location decisions are made, the problem of determining optimal capacity of each facility can be solved efficiently. In Section 4 we discuss exact and approximate solution procedures for the problem. Results of a set of computational experiments and a case study are presented in Section 5. Concluding remarks and directions for future research are discussed in Section 6.

2 Model Description and Formulation

We assume that a discrete set $M = \{1, \dots, m\}$ of potential facility locations, a discrete set $N = \{1, \dots, n\}$ of customer locations, and a distance metric d_{ij} for $i, j \in M \cup N$ are specified. Without loss of generality, we assume $M \subset N$ since customer locations with 0 demand can always be added. Depending on the application, N could represent nodes of a network (in which case d_{ij} is the shortest path distance between $i, j \in N$), or a set of points on a plane. The facilities provide make-to-order service, i.e., each facility can be thought of as a queuing system.

It is easiest to initially assume that there is a facility at every $j \in M$, some of which may subsequently be closed if this improves the objective function value. We assume that customers at $i \in N$ generate a stream of Poisson demands with homogeneous rate $\lambda_i \geq 0$, where λ_i , *the demand rate of node i* , is determined as follows. Let the parameter $\lambda_i^{max} \geq 0$ represent the maximum demand rate that can be generated by node i - this can be thought of as the total number (or total purchasing power) of customers at i who could potentially be interested in the services offered by the facilities. Second, consider a facility at $j \in M$ and suppose that all customers from node i use the services of facility j . Then the actual demand from i seen by j is elastic with respect to two factors: the travel distance d_{ij} (used here as a proxy for the travel time) and the expected waiting time W_j at j . Specifically, we assume the following multiplicative

relationship:

$$\lambda_i = \sum_{j \in M} I\{i, j\} \lambda_i^{\max} \mathcal{F}(W_j) \mathcal{G}(d_{ij}), \quad (1)$$

where $\mathcal{F}(W_j) \in [0, 1]$ represents the sensitivity of demand to the waiting time (assumed to be non-increasing and continuous in W_j), $\mathcal{G}(d_{ij}) \in [0, 1]$ represents the sensitivity of demand to travel distance, and $I\{i, j\}$ is the indicator function taking on a value of 1 if customers from i are assigned to the facility j , and 0 otherwise.

It is implicitly assumed that the prices charged for service are uniform (i.e. identical for all customers) and are set exogenously. Therefore, the price is not a decision variable in the model and is not explicitly incorporated into the demand equation above (more precisely, it is assumed to be reflected in the λ_i^{\max} term). We note that the uniform pricing assumption is reasonable in many cases, as firms often choose to offer uniform prices in large geographical areas; this offers a number of benefits, including the ability to use mass advertising to support sales. The exogenous (or price-taking) environment is reasonable in cases of high competition. If the firm is a price-setter (e.g., a monopolist), our model can be solved parametrically for different price levels, and the price yielding the maximum expected profit can be selected. Another situation where prices do not need to be taken into account explicitly is the case of public service systems, where no explicit price for service is charged.

We note that equation (1) implicitly makes the “no demand splitting” assumption, i.e., that all customers from node $i \in N$ must be assigned to the same facility $j \in M$. While this assumption is very common in location models (see, e.g., Shen, Coullard and Daskin, 2003) and reflects practical considerations in many real-life systems, where solutions involving splitting the demand from a single node (geographic market) between several facilities may be difficult to implement, it may come at a cost: a solution that allows demand splitting may outperform one where no splitting is allowed. While relaxing this assumption by allowing continuous allocation of nodal demand to different facilities leads to significant technical difficulties and will not be attempted in the current paper, we note that it is always possible to use “discrete” node splitting in our model. This is accomplished by introducing $c \geq 1$ copies of each node $i \in N$ (at the same location as the original node) each containing $\frac{1}{c} \lambda_i^{\max}$ units of potential demand (e.g., for $c = 2$, node i is split into two copies each containing half of the original potential demand). Each copy of the original node can then be assigned to a different facility; this allows the model to split the demand from node i into c equal parts assigned to at up to c different facilities (at the cost of

increasing the dimensionality of the decision problem, of course). By increasing the value of c , a successively finer approximation of the "true" node splitting model can be obtained. We will illustrate this approach in Section 5.1.1.

Note that both, \mathcal{F} and \mathcal{G} can be interpreted as probabilities (or probabilistic filters) applied to the initial demand rate λ_i^{max} - thus, $\mathcal{F}(W_j)\mathcal{G}(d_{ij})$ is the probability that a potentially interested customer from i assigned to the facility j will actually employ the services of that facility in view of the travel distance and the expected waiting time, i.e., $1 - \mathcal{F}(W_j)\mathcal{G}(d_{ij})$ is the proportion of the potential demand at i that is lost either to competition or to substitutable services. Some additional assumptions about the \mathcal{F} term will be discussed in the next section.

While the distance sensitivity term $\mathcal{G}(d_{ij})$ is not affected by λ_i , such is clearly not the case for the congestion sensitivity, as the expected waiting time delay W_j is directly affected by the actual demand rate at facility j . Specifically, let

$$\lambda_{ij}^{max} = \lambda_i^{max} \mathcal{G}(d_{ij})$$

be the maximum potential demand from i that can be seen at j . Let $E_j = \{i \in N | I\{i, j\} = 1\}$ be the set of all demand points served by facility j . Then the total demand at facility j is given by

$$\Lambda_j = \sum_{i \in E_j} \lambda_i = \sum_{i \in E_j} \lambda_{ij}^{max} \mathcal{F}(W_j). \quad (2)$$

We will consider two potential queuing disciplines for the facilities: the $M/M/k$ queue with k identical parallel Markovian servers and the $M/M/1$ single-channel Markovian service queue. In the first case, which is appropriate when the capacity of a facility can only be expanded in discrete steps, each server is assumed to have a pre-defined service rate μ , while the number of servers k_j to be stationed at facility $j \in M$ is a decision variable. In the second case, which is more appropriate when the capacity level of a facility can be adjusted continuously, the service rate $k_j \geq 0$ of the single server at j is the decision variable (note that k_j need not be integer in this case). We note that in the model that follows, no significant complications arise when non-Markovian service is allowed (e.g., $M/D/k$ or $M/D/1$ disciplines), as long as formulas for the expected waiting time are available; we assume Markovian service mainly for the ease of exposition.

For the multi-server $M/M/k$ case, the expected waiting time at facility j is given by (see e.g., Winston (1994) for all relevant queuing formulas):

$$W_j = W(\Lambda_j, k_j) = \frac{1}{\mu} + \frac{P(\Lambda_j, \mu, k_j)}{\Lambda_j(1 - \rho_j)} \rho_j, \quad \rho_j < 1, \quad (3)$$

where $\rho_j = \Lambda_j/\mu k_j$, is the utilization rate, and $P(\Lambda_j, \mu, k_j)$ the probability that all the k_j servers are busy given by:

$$P(\Lambda_j, \mu, k_j) = \frac{(k_j \rho_j)^{k_j}}{(1 - \rho_j) k_j!} \left(\frac{(k_j \rho_j)^{k_j}}{(1 - \rho_j) k_j!} + \sum_{r=0}^{k_j-1} \frac{(k_j \rho_j)^r}{r!} \right)^{-1}.$$

An efficient way of calculating $W_j(\Lambda_j, \mu, k_j)$ is given in Pasternack and Drezner (1998). For the single-server $M/M/1$ case, the expected waiting time can be computed as follows:

$$W_j = W(\Lambda_j, k_j) = \frac{1}{k_j - \Lambda_j}, \quad \Lambda_j < k_j. \quad (4)$$

In both cases, the waiting time is a function of the demand rate, which, in turn, is affected by the waiting time, implying that the equation (2) above can be regarded as an equilibrium condition.

Throughout the paper, we consider a facility at $j \in M$ to be “open” if at least one customer node $i \in N$ is assigned to j . We assume that for each open facility $j \in M$ there is a minimal service capacity K_j^L that represents the minimum practically feasible facility design; thus the constraint $K_j^L \leq k_j$ must be satisfied for all open facilities. In addition, we assume a service-level constraint: a parameter $\varphi > 0$ such that the waiting time $W_j \leq \varphi$ must hold for all open facilities $j \in M$; this constraint is meant to provide some level of service guarantee to the customer - no matter which facility they use. Since W_j in our case includes the time in the system, it follows from (3) and (4) that $\varphi \geq \frac{1}{\mu}$ in the $M/M/k$ case and $\varphi \geq \frac{1}{K_j^L}$ in the $M/M/1$ case (recall that K_j^L represents the minimum allowable service rate in this case); otherwise the problem will be infeasible. Note also that the constraint on W_j implicitly requires that $W_j = W(\Lambda_j, k_j)$ exists, i.e., that the service utilization does not exceed 1 at any of the facilities.

We next turn our attention to the objective function. We assume a price-taking environment, i.e., a fixed, exogenously determined price \mathcal{P} is charged to each customer; this parameter can also be regarded as the benefit of participation in case of public service facilities. We assume a variable cost h per server (or per unit of capacity in the $M/M/1$ case). Thus, the cost of locating k_j servers (capacity of k_j for the $M/M/1$ case) at $j \in M$ is $h k_j$ (there is no difficulty in extending the model to the case where the variable cost is location-dependent, i.e., h_j instead of h). While explicit fixed cost for having an open facility at $j \in M$ can be added, we chose not to

do so to simplify the notation. We note that since we require minimal service capacities at all open facilities, the quantity hK_j^L represents an implicit fixed costs for having an open facility at $j \in M$.

We will use the following decision variable. In addition to the non-negative capacity allocation vector \mathbf{k} with components $k_j, j \in M$, we define an assignment vector \mathbf{y} with components $y_{ij} \in \{0, 1\}$ taking on a value of 1 if customers from $i \in N$ are assigned to facility $j \in M$ and 0 otherwise. We also define a binary indicator variable $x_j, j \in M$ which takes on a value of 1 if there is an open facility at j , i.e. if $\sum_{i \in N} y_{ij} > 0$. Note that the demand stream at location j is given by $\Lambda_j(\mathbf{y}, \mathbf{k})$ according to (2). The net revenue is now given by

$$Z(\mathbf{y}, \mathbf{k}) = \mathcal{P} \sum_{j \in M} \Lambda_j(\mathbf{y}, \mathbf{k}) - h \sum_{j \in M} k_j. \quad (5)$$

We can now state the mathematical programming formulation of the Distributed Service System with Elastic Demand (DSSSED) model as follows:

$$\max Z(\mathbf{y}, \mathbf{k}) = \mathcal{P} \sum_{j \in M} \Lambda_j - h \sum_{j \in M} k_j$$

Subject to

$$\sum_{j \in M} y_{ij} \leq 1, \quad i \in N \quad (6)$$

$$\Lambda_j \leq \sum_{i \in N} \lambda_{ij}^{max} y_{ij}, \quad j \in M \quad (7)$$

$$\Lambda_j = \sum_{i \in N} \lambda_{ij}^{max} \mathcal{F}(W(\Lambda_j, k_j)) y_{ij}, \quad j \in M \quad (8)$$

$$k_j \geq \Lambda_j / \mu, \text{ integer (for the } M/M/k \text{ case)} \quad j \in M \quad (9)$$

$$y_{ij} \leq x_j \quad j \in M, i \in N \quad (10)$$

$$K_j^L x_j \leq k_j, \quad j \in M \quad (11)$$

$$W(\Lambda_j, k_j) \leq \varphi, \quad j \in M \quad (12)$$

$$y_{ij} \in \{0, 1\}, \quad x_j \in \{0, 1\}, \quad \Lambda_j \geq 0 \quad j \in M, i \in N$$

Constraints (6) require that proportion of total demand from customer i assigned to all facilities does not exceed 1; in principle it is possible that less than 100% of customer's demand is assigned, even though we will show shortly that full demand from each customer will be assigned at an optimal solution. Constraints (7) place an upper bound on the demand rate

at each facility. They are not necessary in the current formulation, but will be useful later. Constraints (8) are the equilibrium constraints for Λ_j . They depend on both, the form of the \mathcal{F} function and the queuing assumption - the Equations (3, 4) should be used for the $W()$ term for the the $M/M/k$ and $M/M/1$ models, respectively. The next set of constraints (9) ensure the stability of the queuing system at each facility. These constraints are redundant in view of (8), but are useful to ensure the formulation is well-behaved. The requirement that k_j be integer only applies in the case of $M/M/k$ queues. The constraint (10) enforces the connection between y and x variables, ensuring that $x_j = 1$ if at least one customer node is assigned to a facility at j . The constraint (11) enforces the minimum capacity requirements at each open facility, and constraint (12) does the same for the maximum waiting time requirements.

The main difficulty in solving the model above is, clearly, the equilibrium constraint (8), which is highly non-linear, as well as constraints (12).

The formulation above belong to the “directed choice” class of location models since it assumes there is a certain central authority that can assign customers to facilities. In contrast, a “customer choice” formulation would require that each customer be assigned to the utility-maximizing facility for this customer. Note, however, that due to the loss of potential demand when the customer is assigned away from their most preferred facility, there is a strong incentive for the model to generally follow “customer choice” assignments. Indeed this appears to be the case most of the time in our computational experiments. We will come back to this issue in Section 5 below.

The formulation above can be viewed as a bi-level model, where at the upper level we determine the facility locations and the customer assignments, and at the lower level the optimal capacities (server allocations) are determined. We start by addressing the lower level problem in the next section.

3 Determining the Optimal Capacity of a Facility

In this section we assume that the customer assignment vector \mathbf{y} has have been fixed, which automatically induces the location vector $\mathbf{x}(\mathbf{y})$ (recall that a facility is open if at least one customer node is assigned to it). Our goal is to determine the corresponding optimal capacity assignment vector $\mathbf{k}(\mathbf{y})$. Let $S_y = \{j \in M \mid x_j(\mathbf{y}) = 1\}$ be the set of facility nodes under \mathbf{y} and

for $j \in S_y$ let

$$\Lambda_j^{max}(\mathbf{y}) = \sum_{i \in N} y_{ij} \lambda_{ij}^{max}$$

be the maximum possible demand rate at facility j . We also define $K_j^{max} = \min\{k \geq 1 | W(\Lambda_j^{max}, k) \leq \varphi\}$ - this is the number of servers (processing rate in the $M/M/1$ case) which assures that the maximum waiting time is not exceeded even if the arrival rate reaches its maximum level (note that $K_j^{max} = 0$ if the facility is not open. Let

$$K_j^U = \max\{K_j^{max}, \lfloor \frac{\mathcal{P}\Lambda_j^{max}}{h} \rfloor\}.$$

Note that this is the most capacity we will ever want to allocate to a facility at j : the first term ensures that the waiting time requirement is met and the second one is the highest capacity which results in non-negative profit. We now consider the following problem of determining optimal capacity at a facility $j \in S_y$:

$$\begin{aligned} \max C_j(\mathbf{x}, \mathbf{y}) &= \mathcal{P}\Lambda_j - hk_j & (13) \\ \Lambda_j &= \Lambda_j^{max} \mathcal{F}(W(\Lambda_j, k_j)) \\ W(\Lambda_j, k_j) &\leq \varphi \\ K_j^L &\leq k_j \leq K_j^U, \quad k_j \text{ integer for the } M/M/k \text{ case.} \end{aligned}$$

The following observation follows directly from the definitions above and allows us to solve the optimal capacity problem separately for each facility.

Observation 1 *Suppose that Problem (13) has an optimal solution k_j^* with optimal value C_j^* for each $j \in S_y$. For $j \notin S_y$ set $k_j^* = 0$. Then the capacity vector $\mathbf{k}^* = (k_1^*, \dots, k_m^*)$ is optimal with respect to $\mathbf{y}, \mathbf{x}(\mathbf{y})$.*

For the remainder of the current section we focus on solving problem (13) for a particular $j \in S_y$. To simplify the notation, we will generally skip the arguments \mathbf{x} and \mathbf{y} . The next result summarizes the properties of \mathcal{F} that simplify the solution of (13). Note that the waiting time $W(\Lambda_j, k_j)$ is strictly increasing in the demand rate Λ_j and strictly decreasing in k_j . This leads to the following result.

Lemma 1 *Suppose \mathcal{F} is non-increasing, non-negative and differentiable with respect to W with $\mathcal{F}(0) = 1$ and $\lim_{W \rightarrow \infty} \mathcal{F}(W) = 0$. Suppose $k_j \mu > \Lambda_j$ for the $M/M/k$ case and $k_j > \Lambda_j$ for the $M/M/1$ case. Then*

1. \mathcal{F} is differentiable and non-increasing with respect to Λ_j
2. \mathcal{F} is non-decreasing with k_j . For the $M/M/1$ case, it is differentiable with respect to k_j .

The following result ensures that a unique equilibrium arrival rate Λ_j always exists.

Theorem 1 For any $\Lambda_j^{max} \geq 0$ and $k_j \geq 0$, the equation

$$\Lambda_j = \Lambda_j^{max} \mathcal{F}(W(\Lambda_j, k_j)) \quad (14)$$

has a unique solution $\Lambda_j(\Lambda_j^{max}, k_j)$. Moreover, this solution is non-decreasing in Λ_j^{max} and k_j .

Proof: Since $\mathcal{F}(0) = 1$, the right-hand side of (14) is $\Lambda_j^{max} \geq 0$ for $\Lambda_j = 0$. On the other hand, since $W(\Lambda_j, k_j) \rightarrow \infty$ as $\Lambda_j \rightarrow k_j \mu$ in the $M/M/k$ case ($\Lambda_j \rightarrow k_j$ in the $M/M/1$ case), it follows that $\Lambda_j^{max} \mathcal{F}(W(\Lambda_j, k_j)) < \Lambda_j$ for Λ_j chosen so that the system utilization ρ is sufficiently close to 1. The existence of the solution now follows by the continuity of \mathcal{F} , and the uniqueness by the fact that \mathcal{F} is non-increasing.

Note that the right-hand side of (14) is increasing in Λ_j^{max} and non-decreasing in k_j (the latter follows since the expected waiting time is decreasing in k_j). The second statement of the theorem now follows. \square

The previous result has several important consequences. First, note that for $k_j \geq K_j^{max}$ the waiting time constraint is satisfied since $W(\Lambda_j(\Lambda_j^{max}, k_j), k_j) \leq W(\Lambda_j^{max}, k_j) \leq \varphi$, where the last inequality holds by the definition of K_j^{max} . The following corollary now follows since $K_j^U \geq K_j^{max}$ assures feasibility of Problem (13).

Corollary 1 Problem (13) has an optimal solution k_j^* .

Note that the preceding result implies that the condition of Observation 1 is automatically satisfied. The next corollary states that there will be no unassigned customers in the optimal solution to the (DSSSED). It follows since the optimal solution is non-decreasing in Λ_j , and thus, according to the previous result, is non-decreasing in Λ_j^{max} as well; clearly assigning a new customer to facility j will not decrease the latter quantity.

Corollary 2 If it is economical to open any facilities, there exists an optimal solution to the (DSSSED) model where $\sum_{j \in M} y_{ij} = 1$ for all $i \in N$.

Another consequence of Theorem 1 is that the optimal capacity level $k_j^*(\mathbf{x}, \mathbf{y})$ can be found by a simple algorithm (consisting of two line search procedures) described in the following result.

Corollary 3 *The following algorithm will find the optimal capacity k_j^* in at most $O(\log_2(K_j^U - K_j^L))$ steps.*

Step 1. *Apply a line search (e.g., bisection or golden rule search) procedure to the set $\{K_j^L, \dots, K_j^U\}$ to find the capacity \hat{k}_j that maximizes the objective function of Problem (13). If*

$W(\Lambda(\Lambda_j^{max}, \hat{k}_j), \hat{k}_j) \leq \varphi$, stop and report the optimal solution $k_j^ = \hat{k}_j$. Else, proceed to Step 2.*

Step 2. *Apply a line search procedure to find the first non-negative value of $[\varphi - W(\Lambda(\Lambda_j^{max}, k_j), k_j)]$ over the set $\{\hat{k}_j + 1, \dots, K_j^U\}$. Report this value as the optimal solution.*

Proof: First note that for a line search procedure in Step 1 to work, the objective function of (13) must be unimodal. This is clearly the case since the first term is non-decreasing in k_j and the second term is decreasing at a linear rate. Note also that the function $w(k_j) = W(\Lambda(\Lambda_j^{max}, k_j), k_j)$ is non-decreasing in k_j since, by Theorem 1, $\Lambda(\Lambda_j^{max}, k_j)$ is non-decreasing in k_j , and $F(W)$ is non-increasing in W . Thus, we know that either the waiting time constraint is satisfied at \hat{k}_j or else the optimal solution must be the minimal $k_j > \hat{k}_j$ for which the waiting time constraint holds. Moreover, since $w(k_j)$ is non-decreasing, this k_j can be found by a line search procedure on the set $\{\hat{k}_j + 1, \dots, K_j^U\}$ (the existence of such k_j is assured by Corollary 1 above). \square

The solution of Problem (13) is illustrated in the following example.

Example 1. Consider the congestion sensitivity function of the following form:

$$\mathcal{F}(W) = \frac{1}{1 + \alpha W}, \quad (15)$$

where the parameter $\alpha \geq 0$ represents sensitivity to waiting. This function satisfies the conditions of Lemma 1.

Assume the $M/M/k$ setting with processing rate of $\mu = 5$ per server per unit time, and the waiting sensitivity parameter $\alpha = 1$. Suppose that price $\mathcal{P} = 10$, server cost $h = 8$ and that under a certain location and assignment vectors the maximum arrival rate at the facility j is $\Lambda_j^{max} = 10$. We also assume that the maximum expected waiting time should not exceed $\varphi = .5$ and that the minimum number of servers $K_j^L = 1$.

Note that at the maximum arrival rate of 10, at least 3 servers are required for stability and

INSERT Figures 1(a) and 1(b) about here

Figure 1: Optimal capacity determination for Example 1. Figure(a) illustrates equilibrium arrival rates for $k=1,2,3$ servers. Figure (b) shows profit as a function of the number of servers for different maximum arrival rates.

INSERT Figure 2 about here

Figure 2: The optimal capacity function $k^*(\Lambda^{max})$ for Example 1.

$W(10, 3) = .09$, below the waiting time requirement. Thus $K_j^{max} = 3$. Since $\Lambda_j^{max} * \mathcal{P}/h = 12.5$, the number of servers is limited by $K_j^U = \max\{3, 12\} = 12$.

The procedure for finding the equilibrium demand rate for $k_j = 1, 2$ and 3 servers is illustrated on Figure 1 (a). Note that $\Lambda(10, 1) = 4.336$ and $W(4.336, 1) = 1.31 > \varphi$, and thus $k = 1$ is not feasible. On the other hand, $\Lambda(10, 2) = 7.72$ and $W(7.72, 2) = .29 < \varphi$, implying that any number of servers larger than 1 will satisfy the waiting time constraint. Therefore the optimal solution in this case can be found by optimizing the objective function of (13) via a line search procedure for $k \in \{2, \dots, 12\}$. The profit function for different number of servers is illustrated on Figure 1(b) (where the profit curves are also given for two other values of Λ_j^{max}). For $\Lambda_j^{max} = 10$ the optimal profit of 69.6 is achieved with 3 servers and with $\Lambda_j = 9.36$. Note that $W(9.36, 3) = .068 < \varphi$, confirming that the waiting time constraint is satisfied.

The previous results illustrate how the optimal server capacity can be efficiently computed for a given location and allocation vectors. In fact, further efficiency can be gained by observing that k_j^* only depends on \mathbf{y} through the maximum arrival rate $\Lambda_j^{max}(\mathbf{y})$. Since $\Lambda_j^{max}(\mathbf{y}) \in [0, \sum_{i \in N} \lambda_{ij}^{max}]$, we can construct a function $k^*(\Lambda^{max})$ before the assignment vector is known, specifying the optimal capacity of a facility for any possible value of Λ_j^{max} . It is obvious that this function must be non-decreasing in Λ^{max} (since any feasible solution associated with a lower Λ^{max} is available at the higher value as well). For the $M/M/k$ case, since the number of servers is discrete, the function $k^*(\Lambda^{max})$ has the form of a non-decreasing step function. For Example 1 above this function is illustrated on Figure 2 (under the assumption that $\Lambda^{max} \leq 20$). Once this function is computed during the pre-processing stage, the optimal capacity problem can be solved via a simple table look-up.

In some cases, the equilibrium condition (14) can be solved explicitly. For example, assume that for the $M/M/1$ case the function \mathcal{F} is given by (15) above. In that case, using (4) it is easy to show that the equilibrium demand is given by:

$$\Lambda_j(\Lambda_j^{max}, k_j) = \frac{1}{2} \left(k_j + \Lambda_j^{max} + \alpha - \sqrt{(k_j + \Lambda_j^{max} + \alpha)^2 - 4k_j\Lambda_j^{max}} \right),$$

from which the function $k_j^*(\Lambda_j^{max})$ can be computed numerically (since the capacity is continuous for the $M/M/1$ model, the function is a non-decreasing continuous function).

The previous discussion illustrates that for a specific assignment vector, the optimal capacity problem (13) is easy to solve using one-dimensional search procedures (which, in most cases, can be reduced to simple table lookups). In the next section we discuss how these properties can be applied to develop an optimal solution for the original problem (DSSSED).

4 Exact and Approximate Solution Algorithms for DSSSED

Our algorithms (both exact and approximate) are heavily based on obtaining efficient upper bounds for DSSSED. These are described in Section 4.1 below. The exact algorithm is based on successive improvement in the upper bounds and the corresponding lower bounds - it is described in Section 4.2. Finally we also present two heuristics that also utilize the structure of the upper bounds; these are presented in Section 4.3.

4.1 UFLP-based Upper Bounds for DSSSED

We start by computing an upper bound for the arrival rate at a facility $j \in M$ for a given customer assignment vector y_{ij} and assuming $\sum_{i \in N} y_{ij} \geq 1$.

By the definition of the equilibrium arrival rate, we have

$$\Lambda_j(\mathbf{y}) = \Lambda_j^{max}(\mathbf{y})\mathcal{F}(W(\Lambda_j(\mathbf{y}), k_j)) \geq \Lambda_j^{max}(\mathbf{y})\mathcal{F}(\varphi),$$

since φ is an upper bound on the waiting time at facility j . It now follows from the system stability conditions represented by the constraint (9) in the (DSSSED) formulation and the minimum capacity requirements that the lower bound on the required capacity at j is given by

$$k_j \geq \max \left\{ K_j^L, \frac{\Lambda_j^{max}(\mathbf{y})\mathcal{F}(\varphi)}{\mu} \right\}$$

where $\mu = 1$ for the $M/M/1$ case.

Now consider the following linear integer program that (due to its similarities to the uncapacitated facility location problem) we will call (UFLP):

$$\begin{aligned}
Z_*^U &= \max_{\mathbf{y}} Z^U(\mathbf{y}) = \mathcal{P} \sum_{i \in N} \sum_{j \in M} \lambda_{ij}^{max} y_{ij} - h \sum_{j \in M} z_j \\
&\text{Subject to} \\
&y_{ij} \leq x_j \text{ for } i \in N, j \in M \\
&\sum_{i \in N} y_{ij} = 1 \text{ for } j \in M \\
&z_j \geq \frac{\mathcal{F}(\varphi)}{\mu} \sum_{i \in N} \lambda_{ij}^{max} y_{ij} \text{ for } j \in M \\
&z_j \geq K_j^L x_j \text{ for } j \in M \\
&y_{ij}, x_j \in \{0, 1\}, z_j \geq 0 \text{ for } i \in N, j \in M.
\end{aligned} \tag{16}$$

Here the objective function represents the difference between the upper bound on the revenue and the lower bound on the facility costs, the variable z_j captures the lower bound on the processing capacity of facility j , the first two constraints are the standard uncapacitated facility location problem constraints, and the third and fourth two constraints define the value of z_j .

The following observation now follows:

Observation 2 *The optimal value Z_*^U of (UFLP) model (16) is a valid upper bound on optimal value of model (DSSED).*

The advantage of the (UFLP)-based upper bound above is that, unlike the (DSSED) model, the (UFLP) model is a linear IP for which many efficient algorithms (both exact and approximate) exist, which allows us to compute an upper bound efficiently. The disadvantage is that the estimate of the capacity of facility j used in the objective function may not be tight.

For the $M/M/k$ case, a better estimate can be obtained by using the $k_j^*(\Lambda^{max})$ function defined in the pervious section; we use the subscript j to designate the location of the facility. Recall that this function has the form of a non-decreasing step function, i.e., for each $j \in N$ there exist $R(j)$ breakpoints

$$0 = \Lambda_j^0 < \Lambda_j^1 < \dots < \Lambda_j^{R(j)} \leq \Lambda_j^M \equiv \sum_{i \in N} \sum_{j \in M} \lambda_{ij}^{max}$$

with corresponding values

$$K_j^L \leq K_j^1 < \dots < K_j^{R(j)} \leq K_j^U$$

such that $k_j^*(\Lambda^{max}) = K_j^r$ when $\Lambda_j^{max} \in (\Lambda_j^{r-1}, \Lambda_j^r]$. The breakpoints can be pre-computed for every $j \in M$ during the pre-processing step as illustrated in Section 3 above. This allows us to replace the objective $Z^U(\mathbf{y})$ above with a tighter bound

$$\hat{Z}(\mathbf{y}) = \mathcal{P} \sum_{i \in N} \sum_{j \in M} \lambda_{ij}^{max} y_{ij} - h \sum_{j \in M} \sum_{r=1}^{R(j)} K_j^r z_{jr}, \quad (17)$$

along with the first two constraints as in (16) above and the following additional constraints replacing the last two constraints above (along with the variables z_j which are no longer required):

$$\begin{aligned} \sum_{r=1}^{R(j)} z_{jr} &\leq 1, \quad j \in M \\ \sum_{i \in N} \lambda_{ij}^{max} y_{ij} &\leq \sum_{r=1}^{R(j)} \Lambda_j^r z_{jr}, \quad j \in M \\ y_{ij}, z_{jr} &\in \{0, 1\}, \quad i \in N, j \in M, r \in \{1, \dots, R(j)\}. \end{aligned} \quad (18)$$

Note that in the formulation above (which we will refer to as ‘‘Improved UFLP’’) the first term in the objective over-estimates the revenue from facility j , hence the optimal value is still an upper bound on the original model (even though the capacity cost term is now exact). A similar approach can be applied in the $M/M/1$ case as well after first constructing a step-function under-estimator of the continuous $k_j^*(\Lambda^{max})$ function. In the following sections we show how these bounds can be used to generate exact and approximate solutions for the DSSSED model.

4.2 An Exact Algorithm for DSSSED

The algorithm presented below is based on successive improvements to the upper and lower bounds. At each step we introduce a ‘‘cut’’ that eliminates the best solution found so far and improves the upper bound on the remaining solutions. The procedure terminates when the gap between the current upper bound and the lower bound (i.e., the best feasible solution found) is within the tolerance limits. A similar general idea was used for the solution algorithms in Aboolian, Sun and Koehler (2009) and Aboolian, Berman and Drezner (2008). We note that since some feasible solutions are cut off at each step, the procedure described below belongs to the class of ‘‘supervalid inequality-based’’ algorithms recently introduced by Israeli and Wood (2002).

The algorithm could be based on either the upper bound provided by the UFLP (16) or the Improved UFLP defined in (17-18); we will assume the latter for concreteness. Note that any feasible assignment vector \mathbf{y} , in particular the one produced in the course of solving the Improved

UFLP, allows us to produce a feasible solution to DSSED by first defining the induced location vector $\mathbf{x}(\mathbf{y})$ of facilities to which at least one customer node is assigned, and then computing the optimal capacity vector $\mathbf{k}^*(\mathbf{y})$ for each open facility as described in Section 3. The resulting value of the objective function $Z(\mathbf{y}, \mathbf{k}^*)$ provides a lower bound for DSSED; in fact, in view of Collorary 3, it provides the tightest possible lower bound for a given assignment vector.

Next, suppose a certain set of feasible assignment vectors Y has already been examined. We would like to add a constraint to the improved UFLP to ensure that all vectors in Y are no longer feasible. This is accomplished as follows. For each $\mathbf{y}' \in Y$ let $j(i, \mathbf{y}')$ be the facility to which customer node i is assigned under the assignment vector \mathbf{y}' . Then we add a new cut for each \mathbf{y}' that ensures that at least one customer node is assigned differently than under \mathbf{y}' :

$$\sum_{i \in N} y_{i,j(i,\mathbf{y}')} \leq n - 1. \quad (19)$$

We assume the tolerance level $\epsilon \geq 0$ is specified. The algorithm proceeds as follows.

Successive Improvement Algorithm

Step 0: Set $t = 0$, $g = \infty$, $UB^* = \infty$, $LB^* = 0$, $\mathbf{y}^* = \emptyset$, $A(t) = \emptyset$.

Step 1 Set $t = t + 1$. Solve the Improved UFLP augmented with the additional constraint set $A(t - 1)$. If the problem is feasible, let \mathbf{y}^t be the optimal assignment vector, UB^t be the corresponding objective function value and proceed to Step 2. Else, STOP and report the optimal assignment vector \mathbf{y}^* and optimal value LB^* .

Step 2 If $UB^t < UB^*$, set $UB^* = UB^t$. Find the optimal capacity assignment vector $\mathbf{k}^t = \mathbf{k}^*(\mathbf{y}^t)$ and the corresponding objective function value $LB^t = Z(\mathbf{y}^t, \mathbf{k}^t)$. If $LB^t > LB^*$ set $LB^* = LB^t$, $\mathbf{y}^* = \mathbf{y}^t$, and $g = UB^* - LB^*$. If $g \leq \epsilon$ Stop and report \mathbf{y}^* and LB^* . Else go to Step 3.

Step 3 For each $i \in N$ and each $j \in M - \{j(i, \mathbf{y}^t)\}$ Do

Step 3.1 Generate a new assignment vector $\mathbf{y}^t(i, j)$ by re-assigning customer i to facility j .

Step 3.2 Determine new optimal capacities for facilities j and $j(i, \mathbf{y}^t)$ (the other components of the capacity vector \mathbf{k}^t are unchanged). Compute the new lower bound $LB^t(i, j) = Z(\mathbf{y}^t(i, j), \mathbf{k}^t(i, j))$ where $\mathbf{k}^t(i, j)$ is the capacity vector with two recomputed components. If $LB^* < LB^t(i, j)$ set $LB^* = LB^t(i, j)$ and go to Step 3.3. Else Repeat Step 3.

Step 3.3 Set $g = UB^* - LB^*$. If $g \leq \epsilon$, Stop and report $\mathbf{y}^t(i, j)$ and LB^* . Else Repeat Step 3.

Step 4 Generate a new cut $\sum_{i \in N} y_{i,j(i,\mathbf{y}^t)} \leq n - 2$ and repeat Step 1.

In Step 1 we generate a new customer assignment vector by solving the Improved UFLP. This also provides an Upper Bound for the problem. We then generate the corresponding optimal capacity vector in Step 2, obtaining a lower bound, and check whether the gap g between the current best upper and lower bounds is within the tolerance ϵ . If so, we stop and report the current solution. Else, we use a one-opt procedure in Step 3 by trying to reassign each customer to a different facility, updating the lower bounds and testing whether the tolerance level has been reached. Once all re-assignments have been tested we add a new cut to the Improved UFLP formulation in Step 4. Note that the new cut is identical to (19) except for the right-hand side, which is set to $n - 2$ instead of $n - 1$. This is because all simple reassignments around the current solution \mathbf{y}^t have already been tried in Step 3, and thus we need to reassign at least two customers to get a new assignment vector. Under normal circumstances, the algorithm terminates in Step 2 or Step 3 when the gap between the current upper and lower bound falls below the specified tolerance level. The algorithm can also terminate in Step 1 if all potential location vectors have been examined. It is obvious that the algorithm terminates in a finite number of steps.

Note that the Upper Bounds $UB^t, t = 0, 1, \dots$ constructed by the algorithm form a non-increasing sequence. Since at iteration t , LB^* represents the value of the best-found feasible solution, while UB^* is a valid upper bound on all unexamined location vectors, the algorithm is guaranteed to find the optimal solution to DSSSED.

The full set of the computational results for the Successive Improvement Algorithm is reported in Section 5.1 below. The number of iterations and the running time of the algorithm depend critically on the quality of the initial upper bounds (or, more precisely, on the initial gap between the upper and lower bounds). Since the Improved UFLP tends to produce significantly tighter bounds than the original UFLP, we found that the extra computational burden involved in solving the Improved UFLP is worthwhile. In addition, if a heuristic solution is available, its value can be used as the initial lower bound LB^* at the initialization step. Several heuristic approaches are described in the following section.

We also note that another advantage of the algorithm above is that at each iteration the bounds on the optimality gap (both the absolute bound g computed by the algorithm and the relative bound $r = LB^*/UB^*$) with respect to the best found solution are available. Thus, a user

can stop the algorithm whenever the current estimated optimality gap is judged to be sufficiently small.

4.3 Heuristic Approaches for DSSED

While a number of heuristics could be specified for our model, we note that two simple (but effective) heuristics are available as a result of computing the upper bound using either the UFLP or the Improved UFLP models defined earlier. Since either model produces a feasible assignment vector \mathbf{y} we can use this vector to compute the optimal capacity assignment vector $\mathbf{k}(\mathbf{y})$, yielding the objective function value $Z(\mathbf{y}, \mathbf{k})$ - i.e., completing the first iteration of the Successive Improvement Algorithm above. As noted earlier, this yields both, a feasible solution to DSSED and an estimate of the optimality gap for this solution. We call this approach the “UFLP Heuristic”. We use the original UFLP model (16) to obtain the assignment vector, as it is easier to solve. Note that due to the structural properties of the UFLP problem, all customers are assigned to the closest open facility.

An alternative approach is to use the feasible location vector above as a starting point of a procedure where at each step the current location vector is improved by a set of “allowable moves”. In our case, the allowable moves are reassigning a customer from one facility to a different one; if this results in the original facility having no assigned customers, than that facility is closed. At each step all possible moves are examined and a move that provides the largest improvement to the current solution is selected (the procedure terminates when no improving moves are available). This is similar to the “ascent” algorithm described in Cooper (1964), we therefore call it the “Ascent Heuristic”. We note that a Variable Neighborhood Search heuristic (Hansen and Mladenovic, 2001) and similar techniques are often used as extensions of the above procedure. However, as seen in the following section, such extensions are likely not needed in our case as the Ascent Heuristic already appears to be highly effective for DSSED.

5 Computational Results and Solution Structure

In this section we analyze the performance of the DSSED model in two ways. In Section 5.1 we review the results of a series of computational experiments. The main goal of these experiments is to assess the performance of the Successive Improvement Algorithm, as well as the UFLP

and Ascent heuristics and to understand which problem parameters make the DSSED harder or easier to solve. We also examine the impact of node splitting (i.e., allocating splitting customer demand between several facilities) on the quality of the solutions (Section 5.1.1). In Section 5.2 we describe the results of a case study where DSSED model is applied to the problem of locating a set of preventive medical clinics in Toronto, ON. The main goal is to understand how the structure of the optimal solutions depends on the problem parameters.

5.1 Results of Computational Experiments

We conducted a series of computational experiments where the (exact) Successive Improvement Algorithm, the UFLP heuristic and the Ascent heuristic were applied to a set of problem instances described below.

The problem instances were generated as follows. We used the networks from the standard m -median test problems of Beasley (1990); the value of m (the number of facilities) was not used as it is a decision variable in our model. Only the networks with 100 nodes were used, i.e., the first 5 instances in Beasley (1990), as the Successive Improvement (SI) Algorithm tended to exceed the set time limit of 3600 seconds for larger instances. We set $M = N$, i.e., the set of possible facility locations was initially set to the set of nodes (the problem solvability is critically affected by $|M|$; with smaller $|M|$ it is possible to solve much larger instances within the same time limit).

The server cost h was set to \$80 per unit time (hour), the revenue \mathcal{P} to \$100 per customer, and the service time limit φ was set to 2 time units. We used the form (15) of the congestion sensitivity function, where parameter α represents the sensitivity of customer demand to waiting. We used the following values for α : .25, .5, .75, with higher values representing more sensitive customers. We assumed $M/M/k$ queuing model at the facilities and used the following values for the service rate μ of each server: 2, 10, 20. Note that higher service rates represent less flexibility on the part of the decision-maker to adapt the capacity of each facility to the expected demand. For the minimum number of servers at an open facility we used the values 2, 5, 9 (recall that the product of h and the minimum number of servers represents the fixed cost of opening the facility in our case). We left the maximum number of servers at a facility unconstrained. Thus, there were a total of 27 combinations of the different values of the parameters, resulting in $5 * 27 = 135$ problem instances solved.

Insert Table 1 about here

Table 1: Overall experimental results. Note that all ratios were computed vs. the best solution found for each instance.

All problem instances were solved on a machine with Intel Core2 Duo 2.67 ghz CPU with 4 GB RAM running Windows Vista. As noted earlier, we set the time limit of 3600 seconds for each instance for the SI algorithm and used the tolerance limit of $\epsilon = .001$. If the algorithm failed to converge within the time limit, the best feasible solution found was used. The time limit was not relevant for the two heuristic procedures (the solution times were well below 3600 seconds in all cases). All UFLPs and LPs were solved using the CPLEX package, version 11.2 (we note that more efficient special-purpose solvers are available for UFLP and could be used in practical problems to improve the solution times). Full set of experimental results are available from the authors; summarized results can be found on Table 1. Each row represents the average for the 5 Beasley problems with given values of the parameters. The first five columns represent provide the values of the parameter settings. The next column specifies for how many instances (out of 5) was the SI algorithm able to find the optimal solution within the time limit. The next four columns provide information about the solution quality: the ratio of the objective value vs. the best found solution, as well as the gap between the best found solution and the final upper bound (this gap is only relevant for the instances for which the optimal solution was not found). The next three columns provide the solution times for the two heuristic procedures and the SI algorithm. The next to last column gives the number of open facilities in the best found solution. The last column specifies the percentage of customers assigned to their utility-maximizing facility in the best found solution.

As can be seen from Table 1, the service rate μ has the largest effect on the difficulty of a given problem instance: the SI algorithm was not able to solve any of the instances with $\mu = 2$ to optimality within the time limit, but was able to solve nearly all instances with $\mu = 10$ or $\mu = 20$. This is due to the fact that the solution space is much larger when the service rate of an individual server is low (as more servers can be located at a facility). On the other hand, the SI algorithm produced the “best found” solution for every instance solved, including instances with low service rates. Indeed, we observed that the algorithm tends to find the best solution

quite quickly; the rest of the time is spent “proving” its optimality.

Another parameter having an important effect on the solvability of a problem instance is the minimum number of servers K^L (the same value was used for all facility nodes): for instances with $\mu > 2$, the largest solution times occurred for the cases with $K^L = 2$, while the cases with $K^L = 9$ were solved quite quickly. Not surprisingly, this parameter also had the largest impact on the number of open facilities in the best found solution, with this number approaching $|M|$ for the cases where the minimum number of servers (and thus the fixed cost) is low and dropping sharply as the minimum number of servers is increased.

On the other hand, the sensitivity to waiting, α appears to have no impact on either the solvability of a problem instance or the number of open facilities.

It is also interesting to note that nearly 100% of the time the customers were assigned to their utility-maximizing facility. Thus, even though the model is of “directed choice” and, in principle, a customer can be assigned to any facility, the elasticity of demand with respect to service quality (comprised of both the travel distance and the waiting time) ensures that most of the time the interests of the customer and of the decision-maker coincide: customers are assigned to the utility-maximizing facility - i.e., the facility they would choose under the “user choice” setting.

Finally we note that performance of the UFLP heuristic was generally quite poor - with large deviations from the best found solution in most cases. This is because the customer assignments are very naive here - under UFLP the customer is always assigned to the closest facility, which is often quite suboptimal.

On the other hand, the performance of the Ascent heuristic was excellent - in the vast majority of the cases it was able to find the best found solution; in the remaining cases the deviation from the best found solution was well under 1%. Coupled with low solution times this suggests that the procedure of choice is to run both, the Ascent heuristic and the SI algorithm for some reasonable amount of time, and then select the best solutions found by these two procedures.

5.1.1 Evaluating the Impact of Node Splitting

To examine the potential improvement in the solution quality from being able to split customer demand from a single node between several facilities we conduct the following set of experiments. We start from an instance with 10 customer node (generated by randomly selecting 10 nodes

Insert Table 2 about here

Table 2: Profit improvement for $c = 10$ case versus the original no-splitting solution for different values of model parameters.

from one of the 100-node instances used in the previous section). Each customer node has an initial potential demand $\lambda_i^{max} = 10$) for $i = 1, \dots, 10$. We then create c copies of each customer node with each copy receiving $\frac{1}{c}$ portion of the demand from the original node; the values of $c = 2, 5, 10$ were used. The copies are located at a distance of 0 from each other and have the same distances to all other nodes as the original node. As explained earlier, this procedure allows for the demand from the original node to be split between several facilities since not all copies need to be assigned to the same facility. We then find the optimal solution to the resulting network (having 20, 50 or 100 customer node) and compare the total profit to that of the original network. Through numerical experimentation we found that server cost h and sensitivity to waiting α have the largest impact on the solutions. We have thus fixed the values of other model parameters as follows: $\beta = .05, \mu = 20$, and revenue per customer per unit time of 100. The values of the two key parameters were varied as follows: $\alpha = .05, .25, .5, .75, 1$ and $h = 200, 400, \dots, 1800$. At $h = 200$ the server cost is low enough so that a server is located at every node in the original network, and thus no improvement from node splitting is possible. On the other hand, at $h = 1800$, there is only a single server on the network, so no splitting can take place. The intermediate cases are the most interesting. The results are displayed on Table 2 where we show the relative improvement in profit for $c = 10$ case over the original $c = 1$ case for different values of model parameters.

It can be observed that in most cases there is no improvement at all - the “node splitting” solution is identical to the “non-splitting” one. The largest profit gaps occur when server costs are high and sensitivity to waiting is low. Still the relative improvements are very low - under 2% in all cases and well under 1% in most cases. These results suggest that the potential improvements from allowing the customer demand to be split between several facilities may not be very significant.

5.2 Case Study: Locating Preventive Medical Facilities in Toronto, Canada

The medical needs of the residents of the province of Ontario, Canada (of which Toronto is the capital and the major city) are covered by the government of Ontario through OHIP (Ontario Health Insurance Plan). The rate of increase of the costs of OHIP, which far exceeds the rate of inflation, has led to an increased interest in preventive medicine programs. In this case study we analyze a hypothetical program of locating a set of preventive medicine clinics in Toronto, Ontario. We patterned our study on the program developed and implemented in the province of Quebec, Canada, for creating the network of mammography clinics. This program used the model developed by Verter and Lapierre (2002) to determine the location of the clinics; this model was later extended in Zhang, Berman and Verter (2009) and Zhang *et al* (2010). We have used the basic assumptions from Verter and Lapierre (including clinic capacity, demand sensitivity to distance, etc.) applied to the geography, demographic distribution and real estate costs of Toronto, Ontario. It should be noted that under OHIP, the actual medical services are provided by private clinics (that act as independent profit-maximizing businesses) with the government providing remuneration at a set price. Thus, since the goal of the clinics is to maximize profit, the demand is elastic with respect to travel distance and waiting time (as discussed in Verter and Lapierre, 2002), and the prices will be fixed in advance by OHIP, we regard the DSSED model as a good fit for this setting. From the point of view of OHIP, it is interesting to predict what kind of service network might appear in response to a given price level; OHIP can then select the price that achieves the best balance between program participation and program costs.

The case study was performed as follows. We assumed that the new service will be targeted to females aged 50-59, who make up roughly .071% of Canadian population (according to Statistics Canada data) and who will require service once a year. Following Verter and Lapierre, we assumed the maximum probability of participation to be 95% (i.e., this is the probability of participation by the member of the target demographic assuming no demand loss due to travel or waiting time), that the distance sensitivity is a piece-wise linear function with no demand loss within 2 miles of the clinic and a linear decay to 0% participation at a distance of 7 miles. We assumed that each clinic will be open 52 weeks per year, 5 days per week and 10 hours per day. Each clinic will house a certain number of mammography units (servers), with each unit having the processing rate of one patient per 29 minutes and the annual operating costs

per unit are \$1 Million per year (see Coelli *al*, 2007). The city of Toronto was divided into 95 regions called “FSA’s” - using the Canada Post classification (each region has roughly the same number of households) - these regions are roughly equivalent to the 5-digit ZIP codes in the US. Euclidean travel distance between the region centroids was assumed. Demographic data at the FSA level was available from Statistics Canada. Finally, in accordance with the current practice, we assumed that the clinics would be located within some of the 21 general service hospitals in the city of Toronto. Thus the set of potential locations M was the set of FSAs containing a hospital, with $|M| = 21$. Finally, the waiting time function (15) was used with the waiting time sensitivity set to $\alpha = .25$ (as in the case of the computational experiments reported earlier, this parameter did not appear to have a significant impact on the results). We allowed the minimum of one server at a facility and a maximum of 10 servers. All problem instances were solved using the SI algorithm with a time limit of 3600 seconds (if the algorithm failed to converge within the time limit, we used the best feasible solution found) and the Ascent Heuristic; the best solution found by these two methods was recorded and used for further analysis.

We assumed that OHIP will remunerate the clinics at the rate of P dollars per patient visit and wanted to see for different remuneration levels (a) how many facilities will be open and (b) what will be the total participation rate of the target demographic group. Clearly, if the remuneration rate is set below the break-even level, no clinics will be open. Through trial and error, we found the break-even level to be \$ 202.63. We thus used the remuneration rate of \$202.63, \$250, \$500, . . . , \$1500. The results can be found on Table 3. It can be seen that the maximum level of participation that can be achieved is around 72.3%, moreover this rate is achieved when the remuneration is set to \$1,000 per patient visit - i.e., about 5 times higher than the break-even rate. This shows that under the public-private partnership system employed by OHIP, the remuneration rates must be set to quite high levels if high degree of participation by the target demographic group is the desired goal. The results also show that inducing a solution where a facility is open at every hospital may be impossible - due to the close proximity of some of the hospitals to each other, there is simply not enough demand to open facilities at some of the adjoining hospitals; 14 facilities appears to be the maximum that can be expected. Finally, we note that all of the open facilities had either one or two servers. Thus, the profit-optimizing solution opts for a dispersed network of relatively small facilities rather than a few centralized large facilities.

Insert Table 3 about here

Table 3: Impact of the remuneration rate on the overall participation and number of facilities open.

6 Concluding Remarks and Future Research

In this paper we developed a service network design model that explicitly takes into account the elasticity of customer demand with respect to travel distance and congestion delays. In particular, the model incorporates a feedback loop between customer demand and congestion at the facilities. We feel that this careful modeling of the interactions between facility locations and capacities is particularly important when customer sensitivity to waiting time and travel distance is high and/or when the capacity is flexible and can be precisely adjusted to meet the expected demand; in these cases failure to account for the effects described above can lead to large departures from optimality. An interesting feature of our model is that we are able to derive exact optimal solutions for fairly large instances - in most models incorporating equilibrium-type constraints, only heuristic solutions are available.

A natural extension of our model is to allow facilities to set optimal prices and for customers to adjust their facility choice (as well as demand) based on the total utility of service (i.e., to switch to a customer choice environment). This extension, which introduces a traffic equilibrium-like behavior leads to serious additional complications, particularly in the multi-facility case. The resulting model is explored in Berman, Krass and Tong (2009). Another natural extension is to allow non-nodal demand. While assuming that customer demand is concentrated at a set of discrete points (nodes) is a standard assumption in many location models, in practice it often means aggregating geographically dispersed demands to relatively few points (see Francis, Lowe and Tamir (2000) for an exploration of the resulting aggregation errors). For many applications it may, in fact, be more realistic to assume that demand originates from a continuous distribution over a set of links of a network or over a region of the plane. The resulting models are typically more challenging than the discrete demand models; this is likely to be the case for the model proposed in the current paper as well.

Yet another interesting extension is to analyze the leader-follower structure of the game described in the case study in Section 5.2 where the leader (e.g., a government agency) sets the

price and the follower (the company creating the service system) responds to it by determining the number, locations and capacities of facilities.

References

- A. Amiri (1997) "Solution procedures for the service system design problem", *Computers and Oper. Res.*, 24(1), 49-60
- Aboolian R., O. Berman and Z. Drezner (2008) "Location-Allocation of Service Units on a Congested Network," *IIE Transactions*, 40, 112.
- Aboolian R., Y. Sun and G. Koehler (2009) "A Location-Allocation Problem for a Web Services Provider in a Competitive Environment," *European Journal of Operational Research*, 194, 64-77.
- Beasley J.E. (1990) "OR Library-Distruting Test Problems by Electronic Mail," *Journal of the Operational Research Society*, 41, 1069-1072.
- Berman O., E. Kaplan (1987) "Facility Location and Capacity Planning with delay-Dependent Demand", *Int. Jour. Production Research*, 25(12), 1773-1780.
- Berman O., D. Krass and D. Tong (2009) "Facility Location and Revenue Optimization under Equilibrium-Driven Demand and Congestion", working paper.
- Berman O., D. Krass and J. Wang (2006) "Locating Service Facilities to Reduce Lost Demand," *IIE Transactions*, 38, 933-946.
- Berman O. and Z. Drezner (2007) "The Multiple Server Location Problem," *Journal of the Operational Research Society*, 58, 91-97.
- Berman, O. and Z. Drezner (2006) "Location of Congested Capacitated Facilities with Distance Sensitive Demand," *IIE Transactions*, 38, 213-221.
- Cooper L. (1964) "Heuristic Methods for Location - Allocation Problems," *SIAM Review*, 6, 37-53.
- Castillo, I., A. Ingolfsson and T. Sim (2010) "Socially Optimal Location of Facilities with Fixed Servers, Stochastic Demand and Congestion," *Production and Operations Management*, 18(6), 721-736.

- Cornuejols, M.L., G. L. Nemhauser and L.A. Wolsey (1990) “The Uncapacitated Facility Location Problem,” *Discrete Location Theory*, R.L. Francis and P. Mirchandani (eds.) Wiley Interscience, New York.
- Coelli F., Ferreira R., Almeida R., and W. Pereira (2007), “Computer simulation and discrete-event models in the analysis of a mammography clinic patient flow”, *Computer Methods and Programs in Biomedicine*, 87(3): 201-207.
- M. Desroches, P. Marcotte and M. Stan, (1995) “The Congested Facility Location Problem”, *Location Science* 3, 9-23.
- S. Elhedehli (2006) “Service System Design with Immobile Servers, Stochastic Demand, and Congestion ”, *MSOM* 8(1), 92-97.
- R.L. Francis, T.J. Lowe, and A. Tamir (2000) “Aggrgation Error Bounds for a Calss of Location Models”, *Operaitons Research* 48(2), 294-307.
- P. Hansen and N. Mladenovic. (2001) “Variable Neighborhood Search: Principles and Applications”, *European J. Oper. Res.* 130, (3), 449-467.
- Israeli, E. and R.K. Wood (2002) “Shortest-path network interdiction,” *Networks*, 40, 97-111.
- Marianov, V. and M. Rios (2000) “A Probabilistic Quality of Service Constraint for a Location Model of Switches in ATM Communications Networks,” *Annals of Operations Research*, 96, 237-243.
- Marianov, V., M. Rios and F. J. Barros (2005) “Allocating servers to facilities, when demand is elastic to travel and waiting times”, *RAIRO Operations Research*, 39, 143-162.
- Marianov, V., M. Rios, and M. J. Icaza (2008) “Facility location for market capture when users rank facilities by travel and waiting times,” *European Journal of Operational Research*, 191(1), 32-44.
- Marianov V. and D. Serra (1998) “Probabilistic Maximal Covering Location—Allocation for Congested Systems,” *Journal of Regional Science*, 38, 401-424.
- Pasternack B.A. and Z. Drezner (1998) “A Note on Calculating Steady State Results for an $M/M/k$ Queuing System When the Ratio of the Arrival Rate to the Service Rate is Large,” *Journal of Applied Mathematics and Decision Sciences*, 2, 133-135.

Z.-J.M. Shen, C.R. Coullard, M.S. Daskin (2003). “A joint location-inventory model”, *Transportation Science*, 37(1), (2003), 40-55.

Wang Q., R. Batta and C.M. Rump (2002) “Algorithms for a Facility Location Problem with Stochastic Customer Demand and Immobile Servers,” *Recent Developments in the Theory and Applications of Location Models Part II, Annals of Operations Research*, O. Berman and D. Krass (eds.), Kluwer Academic Publishers, 111, 17-34.

Winston W. L. (1994) *Operations Research: Applications and Algorithms*, third edition, Duxbury Press, Belmont, CA.

Zhang, Y., Berman, O., and V. Verter (2009) “Incorporating Congestion in Healthcare Facility Network Design”, *European Journal of Operations Research*, 198, 922-935.

Zhang, Y., Berman, O., Marcotte, P., and V. Verter, (2010) “A Bilevel Model for Preventive Healthcare Facility Network Design with Congestion”, *IIE Transactions*, forthcoming.