Network design for a service operation with lost demand and possible disruptions

Opher Baron, Oded Berman, Yael Deutsch

Joseph L. Rotman School of Management, University of Toronto, 105 St. George St., Toronto, Ontario, M5S 3E6, Canada

Opher.Baron@rotman.utoronto.ca; Berman@rotman.utoronto.ca; Yael.Deutsch@rotman.utoronto.ca.

Abstract

We consider the problem of designing a network for service operation - choosing the optimal number, locations, and capacities of service facilities - recognizing that customers may be blocked due to a finite waiting room and facilities may fail due to disruptions. The goal is to minimize the total cost consisting of traveling, blocking, service capacity, disruptions, and setup costs. We derive structural results when facilities are on a single edge network and use these results to investigate the problem on a general network. For this general problem we prove that when facilities are reliable, the service capacity and blocking costs are independent of the number of opened facilities, and we conjecture that when facilities are unreliable, the service capacity and blocking costs are minimized with equitable demand rates. We use these results to develop algorithms for efficiently solving the problem on a general network. We observe that considering reliability ends up in increasing the number of facilities and reducing their capacities. Finally, we present results of a case study applying our models to the network design of drive-through branches of McDonald's in Toronto. This study suggests that (i) the actual solution may reduce McDonald's profit by $6.66\% - 28.8\%$; (ii) considering unreliability can be done with a relatively low added cost; and (iii) the corresponding solution is robust with respect to the failure probability.

Key words: Service network design, congestion, reliability, optimization.

1 Introduction and literature review

Motivation and research problem: In many settings, finite capacity and failures affect the service level provided to customers. In the health care industry, restricted capacity and its implications are often discussed in the media. For example, CBC reports [11] that (for hospital services in Canada): "The number of beds may not be enough or may be blocked for budget reasons." Similarly, the practice of ambulance diversion, where a hospital asks the emergency medical services to block additional ambulances from being sent to this hospital when it is overcrowded (see [18] and reference within) is also affected by the number of beds available. In call centers, the *grade of service* measures the probability of a call being blocked or delayed for more than a specified interval of time (see [23]). The size of the waiting room is also important in the drive-through windows of a fast food restaurant, once the lane for entering the drive-through is full of cars, arriving cars are blocked from joining the queue and thus the restaurant loses business. From these examples, it is clear that the measure of probability of blocked customers that end up lost to the system is important in many services.

Furthermore, there are many reasons why facilities can and do fail, e.g., equipment breakdowns, industrial accidents, natural disasters, labor strikes, and malicious terrorists attack. Specific examples include the SARS disease outbreak of 2003, which closed down hospitals in Toronto, Ontario; the flu pandemic outbreak of 2009 in the US, which closed down schools; the Ontario doctors' strike of 1986, which closed down hospitals and emergency rooms; and the Anthrax attacks of 2001 in the US, which, besides other effects, forced the closure of the postal facility in Washington D.C. for more than two years. In the fast food example, failures of facilities may be attributed to factors such as extreme weather, fire, sick shift manager, food poisoning, etc.

In this paper we consider two new network design models, focusing on the blocking probability as their main service measure, and on cost minimization. Both models consider $M/M/c/K$ facilities. The first model assumes that facilities are perfectly reliable, and the second model extends the first one by considering the possibility of failures. For each model we optimize the number of facilities to be opened, their location, and their service capacity. We consider costs of traveling, blocking, service capacity, fixed operating, and in the model with failures also penalties for failures. We compare the solutions of both models to improve our understanding of the effect of failures on these decisions. We note that other formulations such as profit maximization, with constraints on budget or service level, and with different queueing models and corresponding service levels can be pursued using a similar line of analysis and are left for future research.

Brief literature review: Network design and facility location are of great importance for a wide range of public and private firms. Both congestion and failures have been studied in the literature but independently of each other. Specifically, in location problems with congestion customers generate streams of uncertain demand and service times are uncertain. As a result of congestion some of the demand may be lost. We note that there are many models that take into account service capacity limitations in a deterministic way (e.g., models with a requirement that there is sufficient service capacity to provide adequate service, such as [32] and references therein).

We distinguish between location models with congestion where servers are mobile (e.g., emergency services) and where they are immobile. The models of mobile servers are more complicated since traveling to the facilities is a part of service, and service times are not independently and identically distributed, see the surveys [4] and [22]. For location models with congestion and immobile servers, see [6]. We note that the main service measure performance in the literature on immobile facilities is the expected time of customers in queue or in the system. Another service level that is used is the probability of waiting longer than a predetermined threshold.

In recent years, a great deal of research has been devoted to facility location problems with failures. The effects of failures can sometimes be significant and cause major disruption to regular operations, and to higher costs. The initial paper in this area is [15]. Then, [28] and [7] analyze the problem on a network assuming that failures are independent. In [5], the failures of facilities are not assumed independent. Further, [5] provides references to other related problems of unreliable facilities. We note that when facilities fail, it is reasonable to assume that some of their demand is redirected to other facilities. In our setting - the other facilities face a surge in demand resulting in a significantly higher blocking probability. We therefore believe that it is important to consider failures in the network design for services.

Methodology: The m-median problem that aims to locate m facilities to minimize the sum of the weighted shortest travel distances from customers on the nodes of a network to the facilities is a well known NP-hard problem. Because the m-median is a special case of the two models we consider, both are NP-hard problems as well. In addition to the complexity of the m-median problem, our models also choose the number of facilities to locate ("m"), which requires solving the m-median problem for different values of m and the service capacity at each facility. The consideration of service capacity and resulting blocking probability makes our problems' formulations nonlinear. Moreover, as we demonstrate below, in the model with failures, expressing the exact cost given a solution (i.e., given the number, location, and capacities of facilities) alone requires an exponential number of computations (that is, the number

of elements in the cost function increases exponentially with the number of facilities).

To overcome these difficulties we use several methods: exact non linear integer optimization, leading to closed form solutions; queueing theory, to express the blocking probabilities and their dependence on the service capacity and demand; and bounds and heuristic approaches to solve both models (implemented using MATLAB).

We start by solving simple cases of our models over the single edge network when facilities are modeled as $M/M/1/1$ queues, derive several results for these simple cases, investigate how these results extend to more general cases, and finally use these results and the insights they generate to derive efficient solution algorithms for the general problems. To demonstrate the applicability of our results and tractability of our algorithms, we end the paper with a case study, where we consider the drive-through operation of McDonald's in the Toronto metropolitan area. For this, a 96 demand nodes network is considered, and we discuss the solution of the different models, and what we learn from them.

Summary of results: For the simple model of demand in two nodes over the single edge network without failures we establish (see Theorem 1): (i) the exact conditions for the optimality of opening a single, or two facilities, or not-operating at all; (ii) closed form expressions for the exact service capacity of the opened facilities; and (iii) that the total service capacity and blocking costs are independent of the number of opened facilities. We establish for this model with failures (see Theorems 4-7), (i) the exact conditions for the optimality of opening a single, or two facilities, or not-operating at all (these conditions are, obviously, less insightful than the ones for the model without failures); (ii) closed form expressions of the service capacity of opened facilities (albeit, it is more effective to solve for the service capacity of the two facilities numerically); (iii) that in contrast to the model with no failures, the total service capacity and blocking costs depend on the number of opened facilities; and (iv) that if it is possible to allocate the demand to facilities by a fair allocation, i.e., allocation of demand as equal as possible , the blocking and service capacity costs are minimized (see [1, 2] for additional models where fair allocation is optimal).

We then study how and which of these results extend to the general models with demand over a network and $M/M/c/K$ facilities. The analysis of this extension leads to several results and insight: first, for the general models with no failures, the service capacity and blocking costs are independent of the number of opened facilities. Thus, the problem can be decomposed into three phases: (I) for any number of facilities, m, solve the m-median problem, which gives optimal location that minimizes the traveling cost; (II) find the optimal number of facilities to locate, m^*, by considering the traveling and setup costs given m; and (III) determine the optimal capacities given the demand faced by each facility

when there are m^* facilities and calculate the total cost, including service capacity and blocking costs. Algorithm 1 uses these three phases to effectively solve such general problems.

Second, for general models with failures, we conjecture (see Conjecture 1 and its justification) that a fair allocation of demand reduces the service capacity and blocking costs. As the m-median solution minimizes the traveling and fixed operating costs; the optimal facility location should balance the m-median type solution with a solution creating as equal demands as possible among the different facilities. While such problems can no longer be decomposed to three phases, Algorithm 2, which uses several heuristics, can be effectively used to solve them.

Third, comparing the models with and without failures, it is preferable to open more facilities with lower capacities when failures are considered. And, contrary to our initial intuition, even with the same number of opened facilities, the optimal capacities are lower for models with failures. Note that there are two effects on the optimal capacities as a result of failures. On the one hand, failures cause surges of high demand pushing to increase capacities. On the other hand, failures increase the cost of effective capacities (when facilities are operational), pushing to decrease capacities. It turns out that the latter effect dominates the choice of optimal capacities.

Finally, the case study reveals that (i) the current solution of McDonald's may be far from optimal, and may reduce the net profit by $6.66\% - 28.8\%$; (ii) considering unreliability can be done with a relatively low added cost; and (iii) when unreliability is considered, the resulted solution is robust with respect to the failure probability. These results support our theoretical results on the importance of considering congestion and failures when designing a service network.

The remainder of this paper is organized as follows. In Section 2, we formulate the general problem, where we distinguish between the case when facilities are perfectly reliable, and the case where they may fail. In sections 3 and 4 we model and solve the problem with reliable and unreliable facilities, respectively. Subsection 4.3 summarizes the main results for general models when facilities are unreliable. In Section 5 we apply our results to the problem of McDonald's drive-through facilities in Toronto, Ontario. And Section 6 summarizes the paper. All proofs are given in the online Appendix A. Algorithms 1 and 2 are given in the online Appendix B. Online Appendix C include numerical examples for $M/M/c/K$ facilities.

2 Location and service capacity choice in congested and unreliable facility system - problem formulation

Here we present the formulation of the problem. We first provide general notation. Then, we distinguish between the case when facilities are perfectly reliable, and the case when they may fail.

2.1 General notation

Let $G = (I, E)$ be an undirected, connected network, with a node set I of size $|I|$ and an edge set E. Each node $i \in I$ has a demand-rate $\lambda_i \geq 0$, of a Poisson distribution. We assume that the demand processes are all independent, with $\lambda \equiv \sum_{i \in I} \lambda_i$. Let $d[i, j]$ be the shortest distance between any two nodes i and j. The decision-maker wants to open several facilities in the network so as to minimize the total cost of operating the network. The total cost includes the costs of traveling, blocking, service capacity, setup and failure when facilities are unreliable as detailed below. Note that by [19] for the reliable case and [7] for the unreliable one, in order to minimize the total (expected) traveling cost it is optimal to locate facilities at the nodes. Also, in [7] it was proved that for the unreliable case when the probability of failure p is sufficiently large, close to 0.8, co-location can occur. Since in practice the value of p is relatively small, co-location is not expected to occur. In addition, in order to reduce the possibility of cannibalization among facilities belonging to the same franchise, e.g., McDonald's, some distance between facilities may be exogenously required. We thus focus on locating facilities on the nodes ignoring co-location. Therefore, the optimal number of opened facilities will satisfy $m^* \leq |I|$.

Let J be the set of potential locations. Without loss of generality, we assume $J \subseteq I$. Throughout the paper we denote by facility j the facility located at node j. The service facilities are modeled as $M/M/c/K$ queues, with K the total buffer size, including the number of customers in service ($c \leq K$). The arrivals are governed by a Poisson process with demand rates λ^j for each facility $j \in J$, service times are exponentially distributed with service rates μ^j per server. The service rates are decision variables in our model.

Due to the facilities' finite buffer, customers may be blocked. Let $\rho^j \equiv \lambda^j/(c\mu^j)$ and $r^j \equiv \lambda^j/\mu^j$. Then, the blocking probability of facility j, $P_B^j(\lambda^j)$, is given by (see, e.g., [29] Chapter 2, page 55):

$$P_B^j(\lambda^j) = \frac{(\rho^j)^K P_0}{c^{K-c} c!}, \tag{1}$$

where

$$P_0 \equiv \begin{cases} (\sum_{n=0}^{c-1} \frac{(r^j)^n}{n!} + \frac{(r^j)^c}{c!} \frac{1-(\rho^j)^{K-c+1}}{1-\rho^j})^{-1} & \text{if } \rho^j \neq 1, \\ (\sum_{n=0}^{c-1} \frac{(r^j)^n}{n!} + \frac{(r^j)^c}{c!}(K-c+1))^{-1} & \text{if } \rho^j = 1 \end{cases} \tag{2}$$

(when no confusion arises, we omit the dependency of the blocking probability on the demand rates). Note that (1) holds also for facilities modeled as $M/G/c/c$ queues, so our results below remain valid also when facilities are modeled as $M/G/c/c$.

The facilities are unreliable, and may fail with a known, identical, failure probability, $p \geq 0$. We assume that facilities' failure events are independent of each other. We further assume that customers always travel to their closest operational facility, if such exists (i.e., if some facilities are opened).

The cost of traveling from node i to node j is normalized to be $d[i,j]$, the blocking cost per unit of demand rate is C_B, the cost of service rate per unit of capacity is C_μ (when $c > 1$, the cost per server is C_μ/c), and opening a facility requires a setup cost of C_K per time period. Further, if all facilities fail, the decision-maker has to pay a penalty (which we call the failure cost) that may be equal for example to the cost customers incur from receiving the service outside the system. There are two different reasons why customers may not be served by the system: failures and blocking, however, the actual reason for not receiving service is often immaterial. Thus, we assume that the failure cost per unit of demand rate is also C_B.

Any parameter (or value) Z, that is related to demand node i and facility j when m facilities are opened, is denoted by $Z_i^{m,j}$ (when no loss of clarity follows, we omit part of these indexes). For example, when m facilities are opened, facility j's service rate per server will be denoted by $\mu^{m,j}$.

2.2 Perfectly reliable facilities

With perfectly reliable facilities, the objective is to minimize:

$$\textit{Total Cost} \equiv \textit{Traveling Cost} + \textit{Blocking Cost} + \textit{Service Capacity Cost} + \textit{Setup Cost}. \tag{3}$$

To formulate the decision-maker's problem as an integer programming problem, let X be the facilities' location vector, Y the customers' allocation matrix, and μ the service rate per server vector in the different facilities. Further, for each node $i \in I$, let S_i denote the list of all nodes in I, ordered by non-decreasing distance from it (in case of ties the choice of the order is random), and S_i^h denote the hth closest node

to i. Let,

$$x_j = \begin{cases} 1 \text{ if a facility is opened at node j} \\ 0 \text{ otherwise,} \end{cases} \tag{4}$$

and

$$y_{ij} = \begin{cases} 1 \text{ if demand node i is assigned to facility j} \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

With this notation, the problem formulation, which we call Problem (6) is:

$$\min_{X,Y,\mu} \quad \sum_{j\in J} \lambda^j d_{ij} + \sum_{j\in J} C_B \lambda^j P_B^j(\lambda^j) + \sum_{j\in J} C_\mu \mu^{m,j} + mC_K \tag{6a}$$

$$\text{s.t.:} \quad (1), (2) \ \forall \ j \in J, \tag{6b}$$

$$y_{ij} \leq x_j \ \forall \ i \in I, j \in J, \tag{6c}$$

$$\sum_{j\in J} y_{ij} = 1 \ \forall \ i \in I, \tag{6d}$$

$$\sum_{k=1}^{h} y_{iS_i^k} \geq x_{S_i^h} \ \forall \ i \in I, h = 1, \ldots, |I| - 1 \tag{6e}$$

$$\lambda^j = \sum_{i\in I} \lambda_i y_{ij} \ \forall \ j \in J, \tag{6f}$$

$$y_{ij} \in \{0, 1\} \ \forall \ i \in I, j \in J, \tag{6g}$$

$$x_j \in \{0, 1\} \ \forall \ j \in J, \tag{6h}$$

$$m = \sum_{j\in J} x_j. \tag{6i}$$

The first term in the objective function (6a) is the traveling cost, the second is the blocking cost, the third is the service capacity cost, and the fourth is the facilities setup cost. Constraints (6b) provide the blocking probability formula used in the calculation of the blocking cost in (6a). Constraints (6c) guarantee that customers are assigned only to locations where a facility is established. Constraints (6d) ensure that each demand node is allocated to a facility. Constraints (6e) ensure that nodes are assigned to their closest facility. Constraints (6f) guarantee that the total demand rate that goes to facility j, λ^j, is used in the calculation of facility j's blocking cost and the travel cost. Constraints (6g) and (6h) are binary restrictions of the y_{ij} variables and the x_j variables, respectively. Constraints (6i) calculates the number of opened facilities, m, to simplify the setup cost evaluation (note that m is a decision variable).

Note that the classical uncapacitated facility location problem, which is an NP-hard combinatorial optimization problem [24], is a special case of our problem (obtained when C_B and C_μ are equal to zero),

therefore, problem (6) is also NP-hard.

2.3 Unreliable facilities

Recall that we assume that the probability of failure for each facility p is relatively small, and all facilities failures are independent of each other and other events (such as demand occurrences). With unreliable facilities, the objective is to minimize:

$$Total\ Cost \equiv Traveling\ Cost + Blocking\ Cost + Service\ Capacity\ Cost + Setup\ Cost + Failure\ Cost. \tag{7}$$

The main differences of (7) and (6a) are the additional failure cost, and the dependence of the traveling and blocking costs in (7) on the operating facilities. We next discus how to express these costs. Given m facilities, there are $2^m - 1$ different realizations (which we call events) of the operating facilities (excluding the one when all facilities fail), and thus each facility has possibly 2^{m-1} different demands. The demand rate that facility j faces given any failure event can be calculated, by adding the demands from all nodes for which this facility is the closet operational to them. Mathematically, let L denote the set of all possible failure events. Let $l \in L$ represent the nodes' status vector at the failure event l, i.e.:

$$l_j = \begin{cases} 1 \text{ if facility } j \text{ is opened and operational at } l, \\ 0 \text{ if either facility } j \text{ is opened and unoperational at } l, \text{ or there is no facility opened at } j. \end{cases} \tag{8}$$

Let P_l denote the probability of occurrence of event $l \in L$. Note that there are different events with the same probability of occurrence (e.g., there are m events with a single facility failure, each occur with probability $(1-p)^{m-1}p$). Let the demand rate that facility j faces at failure event $l \in L$ be λ_l^j and let Y^l denote the matrix of assignments of demand nodes to facilities at event $l \in L$, where

$$y_{ij}^l = \begin{cases} 1 \text{ if demand from node } i \text{ goes to facility } j \text{ at event } l \\ 0 \text{ otherwise.} \end{cases} \tag{9}$$

Then, the total demand at facility j in event l is given by

$$\lambda_l^j = \sum_{i \in I} \lambda_i y_{ij}^l \ \forall l \in L, j \in J. \tag{10}$$

Given (9) and (10) the rate of blocked customers at facility j at event l is $\lambda_l^j P_B^j(\lambda_l^j)$, where $P_B^j(\lambda_l^j)$ is

calculated by (1) and (2).

Let the matrix $Y^L(m)$ include the collection of 2^{m-1} assignments matrices for $m = 1, \ldots, |I|$, Y^l (ordered in some predetermined fashion), and Y^L the collection of all $Y^L(m)$ assignments. With this notation, the problem formulation, which we call Problem (11) is:

$$\min_{X,Y^L,\mu} \quad \sum_{j\in J}\sum_{l\in L} \lambda_l^j d_{ij} P_l + \sum_{j\in J}\sum_{l\in L} C_B \lambda_l^j P_B^j(\lambda_l^j) P_l + \sum_{j\in J} C_\mu \mu^{m,j} + C_K m + C_B \lambda p^m \tag{11a}$$

$$\text{s.t.:} \quad (1), (2), (10) \ \forall\ j \in J, l \in L, (6h), (6i), \tag{11b}$$

$$y_{ij}^l \leq x_j l_j \ \forall\ i \in I, j \in J, l \in L, \tag{11c}$$

$$\sum_{j\in J} y_{ij}^l = 1 \ \forall i \in I, l \in L, \tag{11d}$$

$$\sum_{k=1}^{h} y_{iS_i^k}^l \geq x_{S_i^h} l_{S_i^h} \ \forall\ i \in I, h = 1, \ldots, |I|-1, l \in L, \tag{11e}$$

$$y_{ij}^l \in \{0,1\} \ \forall\ i \in I, j \in J, l \in L. \tag{11f}$$

The objective function (11a) consists of the five terms in (7) The constraints are similar to the ones in Problem (6), with the obvious corrections to only consider assignments to operating facilities at each event l.

Clearly, Problem (11) is NP-hard, as it is extension of Problem (6). Moreover, even the complexity of evaluating the objective function in (11a) (in particular, evaluating the expected traveling and blocking costs), when all decision variables are known, increases exponentially with m because the size of the set L increases exponentially with m. In fact, as the assignments in the optimization are done over $Y^L(m)$, it is clear that the number of decisions variables in this problem is exponentially increasing in the number of nodes of the network.

3 Solution for perfectly reliable facilities

In this section, we consider the problem with perfectly reliable facilities. We first model and solve this problem on a single edge network where facilities are modeled as $M/M/1/1$ queues. Then, using the insights obtained from this analysis, we solve the general problem.

3.1 The problem on the single edge network

3.1.1 Problem formulation on the single edge network

Let $G = (I, E)$ be a network with a single edge and two nodes, i.e., let $I \equiv \{1, 2\}$ and $E \equiv \{[1, 2]\}$. Each node $i \in I$ has a demand rate λ_i. We assume, without loss of generality, that $\lambda_1 \leq \lambda_2$. The total demand at the network is $\lambda = \lambda_1 + \lambda_2$. The decision-maker needs to determine how many facilities to open, where to locate them, and what should be their service capacities.

The blocking probability of facility $j \in J$, which faces an arrival rate λ^j, and has a service rate μ^j is given by (1),(2) with $c = K = 1$. Specifically:

$$P_B^j(\lambda^j) = \frac{\lambda^j}{\lambda^j + \mu^j} \tag{12}$$

Let $Cost^m$ be the total cost of operating m reliable facilities. As explained in Subsection 2.1, it is optimal to locate facilities at the nodes so $m^* \leq |I|$, i.e., $m^* \in \{0, 1, 2\}$. It is easy to verify that when $\lambda_1 \leq \lambda_2$, the optimal location of the single facility is node 2, and the optimal locations of the two facilities are nodes 1 and 2. Let C_K^* denote the facility setup cost which makes the cost of operating two facilities identical to that of operating a single one.

3.1.2 Solution on the single edge network

From (6a), the total cost when a single facility is opened at node 2 is

$$Cost^1 \equiv \lambda_1 + \frac{C_B \lambda^2}{\lambda + \mu^{1,2}} + C_\mu \mu^{1,2} + C_K, \tag{13}$$

and the total cost when two facilities are opened (clearly, there are no traveling costs) is

$$Cost^2 \equiv \frac{C_B(\lambda_1)^2}{\lambda_1 + \mu^{2,1}} + \frac{C_B(\lambda_2)^2}{\lambda_2 + \mu^{2,2}} + C_\mu(\mu^{2,1} + \mu^{2,2}) + 2C_K. \tag{14}$$

The next theorem provides a solution for the problem on the single edge network.

Theorem 1. *For the single edge network, we have*

1. *If $C_B < C_\mu$ it is optimal not to operate. If*

$$C_B \geq C_\mu. \tag{15}$$

Then,

$$\mu^{1,2} = \lambda(\sqrt{\frac{C_B}{C_\mu}} - 1), \tag{16}$$

and

$$\mu^{2,j} = \lambda_j(\sqrt{\frac{C_B}{C_\mu}} - 1) \quad j = 1, 2. \tag{17}$$

So that

$$\mu^{1,2} = \mu^{2,1} + \mu^{2,2} \tag{18}$$

2. *The total service capacity and blocking costs are independent of* $m = 1, 2$.

3.

$$Cost^1 \leq Cost^2 \Leftrightarrow \lambda_1 \leq C_K^* \tag{19}$$

Note that if (15) does not hold, it is cheaper to pay the blocking costs per customer than to purchase a unit of service capacity to serve it. Thus, it is best not to operate at all. For simplicity of the exposition we assume that (15) is satisfied below. Part 1 of Theorem 1 implies the optimal capacities are only dictated by the blocking and service capacity costs. Specifically, the optimal service capacity safety is the difference between the square root of the ratio between these costs and 1. Because this safety capacity is independent of the demand rate, the total service capacity required is independent of the number of facilities opened, as indicated in (18). The implication of (18) is that the service capacity and blocking costs are independent of m, as is indicated in part 2 of the theorem. With these results what dictates the difference between opening 1 and 2 facilities is the trade off between the lower traveling cost when $m = 2$ and the lower setup cost when $m = 1$, as given in (19).

The important implication of Theorem 1 is that the problem over the single edge network when facilities operate as $M/M/1/1$, can be decomposed into three phases: at the first phase for any number of facilities, m, we solve a *Location problem* that gives optimal location that minimize the traveling cost, at the second phase, we *find* m^* by considering the traveling and setup cost given m. Finally, at the third phase, we express the *optimal capacities* given the demand faced by each facility when there are m^* facilities and calculate the total cost, including service capacity and blocking cost.

In the next section we show that the main results of Theorem 1 extend to more general settings of both a general network (rather than a single edge network) and when facilities modeled as $M/M/c/K$ (rather than $M/M/1/1$). Therefore, the solution approach with the three phases can be used there as

well.

3.2 Solution on a general network

In this section we solve the problem with perfectly reliable facilities on a general network when facilities operate as $M/M/c/K$. The next theorem generalizes the results of Theorem 1 to these settings.

Theorem 2. *For any network $G = (I, E)$, when facilities operate as $M/M/c/K$ queues, for given c, K, C_B and C_μ, when it is optimal to operate, we have:*

1. *The optimal capacities are given by*

$$\mu^{m,j} = \lambda^{m,j}\eta,$$

   where $\lambda^{m,j}$ is the demand rate that facility j faces when m facilities are opened, and η is independent of m, the demand rate, the traveling, and the setup costs.

2. *The total service capacity and blocking costs are independent of m.*

3. *The choice between opening m and $m+1$ facilities depends on the trade off between the lower traveling cost of $Cost^{m+1}$ and the lower setup cost of $Cost^m$.*

4. *The optimal location and number of facilities, m^*, are independent of the parameters c and K of the $M/M/c/K$ queue, and the costs C_B and C_μ.*

Part 1 of Theorem 2 implies that in the general case, the optimal capacities are only dictated by the blocking and service capacity costs. As the safety capacity, η, is independent of the demand rate, the total service capacity required is independent of the number of facilities opened. Thus, the service capacity and blocking costs are independent of m, as is indicated in part 2 of the theorem, and the choice between opening m and $m+1$ facilities is the trade off between the lower traveling cost when $m+1$ facilities are opened and the lower setup cost when m facilities are opened, as stated in part 3 of the theorem.

Note that while the blocking and service capacity costs depend on c and K, these parameters have no effect on the location and number of facilities. When facilities are reliable, their number and location is only affected by the network topology and demand rates.

Therefore, Theorem 2 implies that the three phases approach discussed above can be used to minimize the total cost. Specifically, in the first phase we solve a *Location problem*. For any given m, the facilities should be located according to the m-median solution that minimizes the traveling cost (for the integer

programming formulation of the m-median problem, see, e.g., [19]). In the second phase we *find* m^*, the optimal number of facilities to open in accordance to the trade off between the traveling cost, calculated at the first phase, and the setup cost. Finally, the third phase determines the *optimal capacities*. The service capacity of each facility is calculated using the demand it faces; note that the demands are based upon the solution of the first phase.

Theorem 2 also implies that there is no need to include close assignment constraints (6e) in the formulation (6). Because we first solve a Location problem, without considering the blocking and service capacity costs, these constraints will be the result of the traveling cost minimization in (6a).

Algorithm 1 that solves the reliable problem on a general network is given in the online Appendix B. The input of the algorithm is the network G, the demand rates λ_i for $i \in I$, the costs C_B, C_μ, C_K, and the parameters c, K of the $M/M/c/K$ queue. The Algorithm includes four procedures, one for an initial guess (discussed below), and three corresponding to the phases for the solution of the problem.

Note that at a general network, if we use the three phases described above without any changes, at Phase I we need to solve the m-median problem $|I|$ times, which may be very time consuming. We therefore first execute Procedure 0, the *initial guess* procedure. Procedure 0 finds a candidate $\hat{m}$ as an initial number to start the search for the optimal number of facilities to open. For each $m \in I$, *initial guess* finds the optimal location and the resulted traveling cost according to the solution of the relaxed m-median problem, evaluates the sum of the traveling and setup cost of operating m facilities, and determines the number of facilities that minimizes this sum, $\hat{m}$.

Then, Phase I solves exact m-median problems, and evaluates the resulted exact traveling and setup costs. Phase II finds the optimal number of facilities that should be opened, m^*, by a search in a close neighborhood around $\hat{m}$ (the stopping rule we implement for the procedure assumes convexity in m of the sum of the traveling and setup cost. However other stopping rules, including a branch and bound based on the lower bound of the traveling costs found by Procedure 0, can be used as well). Finally, Phase III calculates numerically the optimal capacities, and evaluates the total costs when m^* facilities are opened.

4 Solution for unreliable facilities

Here we consider the problem with unreliable facilities. In Subsection 4.1 we overcome the complexity of evaluating the objective function (11a). In Subsection 4.2 we model and solve the problem over the

single edge network. In Subsection 4.3 we summarize our findings and results for the single edge network and general case with unreliable facilities. The detailed results are then demonstrated in Subsection 4.4.

4.1 Bounds on the objective function

Traveling cost:

Note that $\sum_{u=1}^{h-1} x_{S_i^u}$ represents the number of opened facilities that are closer to node i than the hth closest node to node i. Therefore, if there is a facility at this hth closest node, it will serve demand from node i if (i) the facility is operational, with probability $(1-p)$, and (ii) all closer facilities are unoperational, with probability $p^{\sum_{u=1}^{h-1} x_{S_i^u}}$:

Observation 1. *The hth closest node to node i hosts the closest operational facility to node i with probability $(1-p)p^{\sum_{u=1}^{h-1} x_{S_i^u}} x_{S_i^h}$.*

Therefore, the traveling cost from node i to facility h is

$$(1-p)p^{\sum_{u=1}^{h-1} x_{S_i^u}} x_{S_i^h} \lambda_i d_{iS_i^h} \tag{20}$$

(the traveling cost per unit of demand is 1), so that the total expected traveling cost is given (and can be evaluated in polynomial time) by:

$$\sum_{i\in I}\sum_{h\in I}(1-p)p^{\sum_{u=1}^{h-1} x_{S_i^u}} x_{S_i^h} \lambda_i d_{iS_i^h}, \tag{21}$$

Blocking cost:

Because exact evaluation of the blocking cost in the objective function (11a) is possible only for a rather small number of facilities, we provide upper and lower bounds for it for any number of facilities.

Suppose that m identical $(M/M/c/K)$ facilities are opened and located, and their service rates are given. Let Π denote the exact total expected blocking cost. Let $v \in \{0,\ldots,m-1\}$, let $\underline{\Pi^v}$ denote the exact expected blocking cost for the events where $v' = 0,\ldots,v$ facilities fail. Let $j_{\min}$ denote the facility with the lowest service rate. Let $\Pi^{j_{\min},\lambda}$ denote the blocking cost of facility $j_{\min}$ which faces a demand rate of λ (recall, λ is the total demand at the network). The logic behind these bounds is that the possibility of failure of more than v facilities is of order p^{v+1} when there are $m > v$ facilities.

For $0 \le v \le m-1$, let $P^{v^+} \equiv \sum_{i=v+1}^{m-1} \binom{m}{i}(1-p)^{m-i}p^i$ be the probability that more than v facilities fail, and let $\overline{\Pi^v} \equiv \underline{\Pi^v} + P^{v^+}\Pi^{j_{\min},\lambda}$. Then,

Theorem 3.

$$\underline{\Pi^v} \leq \Pi \leq \overline{\Pi^v} \text{ for every } v = 0, \ldots m-1 \tag{22}$$

The lower bound $\underline{\Pi^v}$ and the upper bound $\overline{\Pi^v}$ get closer to each other as v increases. Moreover, as the failure probability is assumed to be rather low, the difference between the upper and lower bound, is of order of P^{v^+}, and thus is small. For example, for a failure probability $p = 0.02$, as we use in our case study, and $m = 15$ facilities, the probability that 4 and more facilities will fail is $8.12 \cdot 10^{-6}$. Also, note that exact calculation of Π^v for a small v is reasonable. For example, for $m = 15$ and $v = 3$, there are 576 different failure events that their blocking cost is calculated exactly.

4.2 The problem on the single edge network

Consider the simplified setting of section 3.1, but with unreliable facilities. Specifically, each facility may fail independently with probability p. Let $Cost^m(p)$ be the total cost of operating m unreliable facilities. From (11a), the total cost when a single facility is opened is

$$Cost^1(p) = \lambda_1(1-p) + \frac{C_B\lambda^2(1-p)}{\lambda + \mu^{1,2}} + C_\mu\mu^{1,2} + C_K + C_B\lambda p, \tag{23}$$

and the total cost when two facilities are opened is

$$\begin{aligned} Cost^2(p) \quad = \quad & \lambda p(1-p) + C_B((1-p)^2(\frac{(\lambda_1)^2}{\lambda_1 + \mu^{2,1}} + \frac{(\lambda_2)^2}{\lambda_2 + \mu^{2,2}}) + (1-p)p(\frac{\lambda^2}{\lambda + \mu^{2,1}}) + \\ & + C_B(1-p)^2\frac{\lambda^2}{\lambda + \mu^{2,2}} + C_\mu(\mu^{2,1} + \mu^{2,2}) + 2C_K + C_B\lambda p^2. \end{aligned} \tag{24}$$

Let $\gamma \equiv C_B(1-p)/C_\mu$. The next theorem determines the condition for operating facilities, and the optimal capacities of the single and two facilities.

Theorem 4. *For the single edge network, with unreliable facilities, we have*
If $C_B < \frac{C_\mu}{1-p}$ *it is optimal not to operate the system. If*

$$C_B \geq \frac{C_\mu}{1-p}, \tag{25}$$

i.e., $\gamma \geq 1$,

$$\mu^{1,2} = \lambda(\sqrt{\gamma} - 1), \tag{26}$$

and for $j = 1, 2$, $\mu^{2,j}$ *is the solution of*

$$\frac{(\lambda_j)^2(1-p)}{(\lambda_j+\mu^{2,j})^2} + \frac{(\lambda)^2 p}{(\lambda+\mu^{2,j})^2} = \frac{1}{\gamma} \tag{27}$$

Note that although the decision-maker pays C_μ per each unit of service capacity, this service capacity is meaningful only if the facility is operational, i.e., with probability $(1-p)$. If (25) is not satisfied, then it is optimal not to operate the system at all. Further, for $p = 0$, (25) becomes (15), (26) becomes (16), and (27) becomes (17), as expected. Finally, comparing (26) to (16), we observe that the service capacity of a single facility when unreliability is considered is lower than that when failures are ignored.

4.3 Summary of results

Optimal capacities: For two and more opened facilities, the capacities depend on m, the demand rates, C_B, C_μ, and p, and there is no explicit expression for them. Even when two facilities are opened over the single edge network, the capacities are given implicitly (see Theorem 4). These *capacities can be bounded as given in Theorem 5, and approximated as in (29).*

Blocking and service capacity costs: The total blocking and service capacity costs depend on m. Using the upper bounds (29), the total blocking and service capacity costs when two facilities are opened over the single edge network are higher than these when a single facility is opened (see Theorem 6). Further, we conjecture that on a general network, locating facilities such that their demand rates are equal to each other minimizes the total service capacity and blocking costs (see Conjecture 1). We observe that the *service capacity of opened facilities in the unreliable case is lower than that in the reliable one* (see Example 3).

Traveling cost: The *total traveling cost may increase or decrease with* m, depending on the value of the failure probability. For example, the total traveling cost over the single edge network when two facilities are opened is lower than the traveling cost when a single facility is opened if $p < \frac{\lambda_1}{\lambda}$ (see Theorem 6).

Choice between opening m **and** $m+1$ **facilities:** The choice between opening m and $m+1$ facilities is a trade off among all the costs (see Theorem 6). Moreover, comparing (33) with (19), the choice of opening more facilities happens "earlier" (for a smaller threshold value) than in the reliable case. That is, *the optimal solution with unreliability opens more smaller facilities with lower capacities.* Thus, as the m-median solution minimizes the total traveling cost, and as the fair-m-median solution minimizes the total blocking and service capacity costs, in order to solve the unreliable problem we can use these two

solutions as possible locations (see Algorithm 2 in the online Appendix B).

4.4 Demonstration of the results

4.4.1 Bounds on the optimal capacities

In Theorem 4, the expressions for the optimal $\mu^{2,j}$s are given implicitly by (27), as the exact expressions are cumbersome (solutions of a quartic equation). Though (27) can be efficiently solved numerically, we next provide upper and lower bounds for the optimal $\mu^{2,j}$s, in order to be able to gain some intuition and structural results. The upper bounds provide intuitive approximations for the optimal capacities. The bounds are based on the following two intuitive ideas:

1. The decision-maker pays C_μ per each unit of service capacity for facility $j = 1, 2$, but, "enjoys" this service capacity only when the facility is operational, that is, during $(1 - p)$ of the time. Thus, it is as if the service capacity cost is $C_\mu/(1 - p)$.

2. With failures, facility j faces a demand-rate of λ (when it is the only operational facility). Thus, facility $j = 1, 2$ receives an expected demand rate of

$$\lambda_j(p) \equiv \lambda_j(1-p)^2 + \lambda p(1-p) = (1-p)(\lambda_j + p\lambda_{3-j}). \tag{28}$$

We substitute these cost and demand in the optimal service capacity for the model with perfectly reliable facilities (17) to approximate the optimal capacities by

$$\overline{\mu}^{2,j}(p) \equiv (\lambda_j + p\lambda_{3-j})(\sqrt{\gamma} - 1). \tag{29}$$

In Theorem 5 below we establish that $\overline{\mu}^{2,j}(p)$ is an upper bound on the optimal service capacity of facility j. Note also that (29) agrees with (26), which gives the optimal service capacity for the single unreliable facility.

Notice that the bound in (29) grows linearly with λ and λ_j and depends on the ratio C_B/C_μ.

Let

$$T \equiv -\frac{\left(\frac{(\lambda_j)^2(1-p)}{(\lambda_j+\overline{\mu}^{2,j}(p))^2} + \frac{\lambda^2 p}{(\lambda+\overline{\mu}^{2,j}(p))^2} - \frac{1}{\gamma}\right)}{\left(\frac{2(\lambda_j)^2(1-p)}{(\lambda_j+\overline{\mu}^{2,j}(p))^3} + \frac{2\lambda^2 p}{(\lambda+\overline{\mu}^{2,j}(p))^3}\right)}, \tag{30}$$

and

$$\underline{\mu}^{2,j}(p) \equiv \overline{\mu}^{2,j}(p) - T, \tag{31}$$

then, we have

Theorem 5. *The optimal service capacity* $(\mu^{2,j})^*$ *is bounded by*

$$\underline{\mu}^{2,j}(p) \leq (\mu^{2,j})^* \leq \overline{\mu}^{2,j}(p) \text{ for } \ j = 1, 2. \tag{32}$$

While both bounds in Theorem 5 are given in closed form, the upper bound is much more intuitive. However, from the proof, it is clear that the lower bound is much closer to the optimal service capacity than the upper bound. Moreover, the same procedure used to express the lower bound can be used again to come up with a tighter upper bound on the optimal service capacity. Thus, the logic behind the proof provides an alternative numerical method to solve for the optimal service capacity; however, directly solving (27) numerically is immediate and we thus do not pursue this alternative any further.

To demonstrate the accuracy of the bounds from Theorem 5, consider the example with $C_K = 35, C_B = 10, C_\mu = 5, \lambda = 100, \lambda_1 = 20, \lambda_2 = 80$, and $p \in [0, 0.5]$ (the chosen interval of p follows from (25), where we compare the optimal solutions and the bounds for facility 1's capacities for different λ_1 values (it can be shown that the differences between the optimal service capacity and its bounds are higher for facility 1 than for facility 2). Nevertheless, the example shows that these differences are rather small (unless stated otherwise, the data of the numerical examples in the rest of the paper for the $[1, 2]$ network is as in this example).

Example 1. ***Accuracy of the bounds:*** *Let* $\lambda_1 = 20, 30, 50$ *(*$\lambda_2 = 100 - \lambda_1$*). Figure 1 presents* $(\mu^{2,1})^*, \overline{\mu}^{2,1}(p)$*, and* $\underline{\mu}^{2,1}(p)$ *for these three* $\lambda_1 s$ *as a function of* p*. The exact, upper, and lower bound solutions are drawn by the solid, dotted, and dashed line, respectively. The lower, middle, and upper curves are the solutions for* $\lambda_1 = 20, 30,$ *and* 50*, respectively.*

As the lower bounds are close to the exact solutions, the dashed line for $\lambda_1 = 50$ *is not seen in the figure. As the difference* $(\lambda_2 - \lambda_1)$ *decreases when* λ_1 *grows to 50, the difference between the exact capacities and their bounds decreases.*

4.4.2 A comparison between $Cost^1(p)$ and $Cost^2(p)$

Given the relative accuracy and simplicity of the upper bounds, we use them here and in the following subsection to analytically compare the operating cost of the single unreliable facility with that of the two unreliable facilities, and the performance of the reliable solution when facilities are unreliable. Specifically, we use the approximations for $\mu^{2,j}$, $j = 1, 2$ and the exact solution of $\mu^{1,2}$ given by (26)).

Theorem 6.

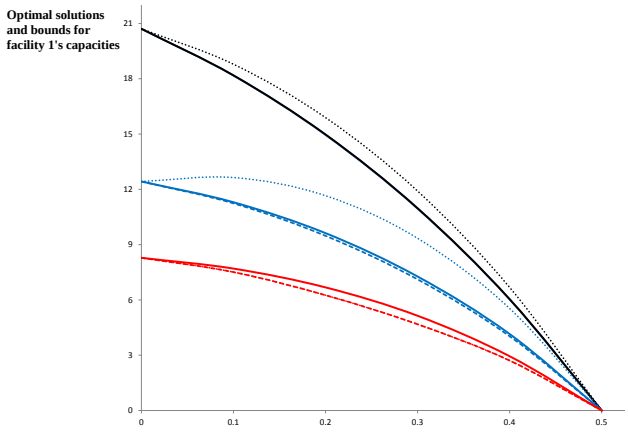


Figure 1: $(\mu^{2,1})^*$ (solid), $\overline{\mu}^{2,1}(p)$ (dotted), $\underline{\mu}^{2,1}(p)$ (dashed) for $\lambda_1 = 20, 30, 50$

1. *Let*

$$\hat{\lambda_1} \equiv (\lambda_1 - \lambda p)(1-p) + C_\mu \lambda p(\sqrt{\gamma} - 1) + C_B(\lambda(1-p)p + \frac{\lambda(1-p)}{\sqrt{\gamma}} - \frac{(\lambda_1)^2(1-p)^2}{(\lambda_1 + \lambda_2 p)\sqrt{\gamma} - \lambda_2 p} - \frac{(\lambda_2)^2(1-p)^2}{(\lambda_2 + \lambda_1 p)\sqrt{\gamma} - \lambda_1 p} - \frac{(1-p)p\lambda^2}{(\lambda_1 + \lambda_2 p)\sqrt{\gamma} + \lambda_2(1-p)} - \frac{(1-p)p\lambda^2}{(\lambda_2 + \lambda_1 p)\sqrt{\gamma} + \lambda_1(1-p)}). \tag{33}$$

Then,

$$Cost^1(p) \leq Cost^2(p) \Leftrightarrow \hat{\lambda_1} \leq C_K^* \tag{34}$$

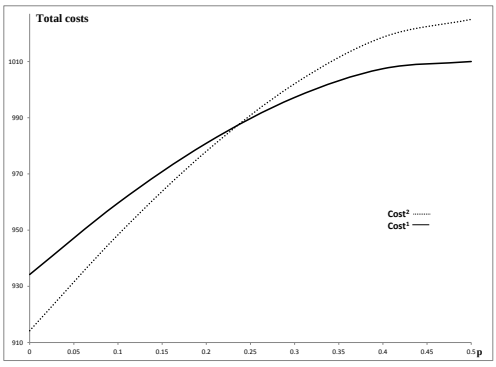
2. *The following conditions are satisfied:*

   (a) *If* $p < \frac{\lambda_1}{\lambda}$, *the traveling cost of* $Cost^2(p)$ *is lower than the traveling cost of* $Cost^1(p)$.

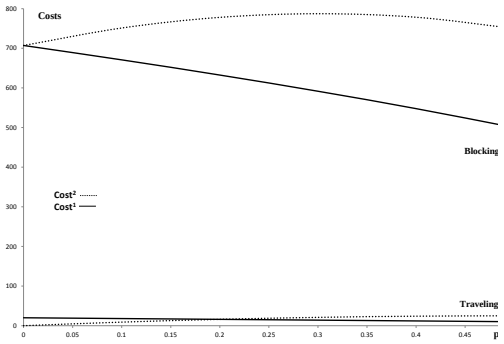
   (b) *The blocking, service capacity, and setup costs of* $Cost^2(p)$ *are higher than or equal to these of* $Cost^1(p)$.

   (c) *The failure cost of* $Cost^2(p)$ *is lower than that of* $Cost^1(p)$.

Note that $\hat{\lambda_1}$ for this model is clearly more elaborate than λ_1 for the model with fully reliable facilities. Moreover, for $p = 0$, $\hat{\lambda_1} = \lambda_1$ and thus (34) is the same as (19). However, for $p > 0$, C_K^* is smaller than λ_1. The following example demonstrates parts 2a for the case $p < \frac{\lambda_1}{\lambda} = 0.2$, and part 2b of Theorem 6.



(a) Total costs



(b) Traveling and blocking costs

Figure 2: Costs of a single vs. two unreliable facilities

Example 2. ***Costs of a single vs. two unreliable facilities:*** *Let $C_K = 0$. Figure 2a presents the total costs (excluding the setup costs) of $Cost^1(p)$ (solid line) and $Cost^2(p)$ (dashed line), using the* ***exact*** *solutions of capacities. The value of C_K^* is the difference between the two curves. It is decreasing from $C_K^* = \lambda_1 = 20$ for $p = 0$ to $C_K^* = 0$ for $p = 0.24$. For $p > 0.24$, opening a single facility is always optimal. Figure 2b presents the traveling and blocking costs $Cost^1(p)$ (solid line) and $Cost^2(p)$ (dashed line). It is clear that Theorem 6, parts 2a and 2b, are satisfied.*

Performance of the reliable solution when facilities are unreliable

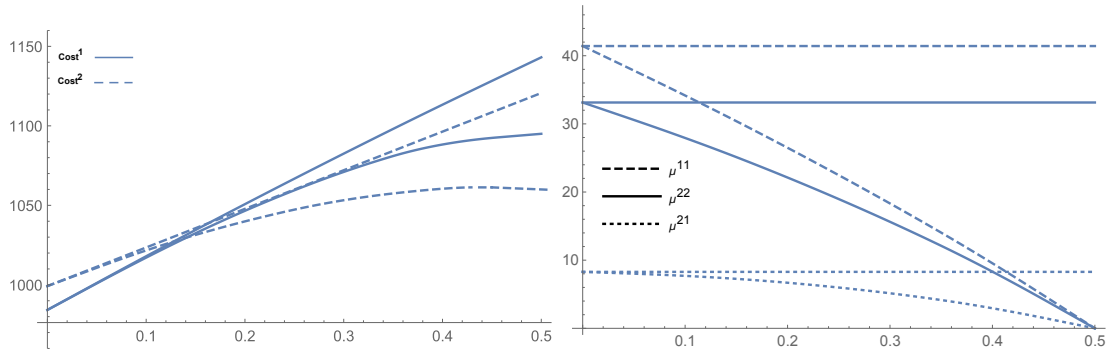
Suppose the decision-maker does not consider the possibility that facilities may fail, and chooses the capacities according to (16) and (17). Let the capacities in this case be denoted by $\mu^{1,2}(0)$ and $\mu^{2,j}(0)$, for $j = 1, 2$. So, using (23) and (24),

$$Cost^1(p, \mu^{1,2}(0)) = \lambda_1(1-p) + 2\sqrt{C_B C_\mu}\lambda - p\sqrt{C_B C_\mu}\lambda + C_B\lambda p - C_\mu\lambda + C_K, \tag{35}$$

and

$$\begin{aligned} Cost^2(p, \mu^{2,j}(0)) \quad = \quad & \lambda p(1-p) + C_B(1-p)^2\frac{\lambda\sqrt{C_\mu}}{\sqrt{C_B}} + \frac{C_B(1-p)p(\lambda)^2}{\lambda_2 + \lambda_1\sqrt{\frac{C_B}{C_\mu}}} + \\ & \frac{C_B(1-p)p(\lambda)^2}{\lambda_1 + \lambda_2\sqrt{\frac{C_B}{C_\mu}}} + C_\mu\lambda(\sqrt{\frac{C_B}{C_\mu}} - 1) + C_B\lambda p^2 + 2C_K. \end{aligned} \tag{36}$$

The following example shows the effect of ignoring failures.



(a) Total costs (b) Capacities

Figure 3: Total costs and capacities, considering and not considering failures

Example 3. ***Wrong choices when failures are ignored:*** *Figure 3a presents* $Cost^1(p)$ *(solid line) and* $Cost^2(p)$ *(dashed line) as a function of* p*, when failures are considered and when they are not. When failure are ignored, the costs are linearly increasing with* p*. When failure are considered, the costs are concave increasing with* p *in a rate lower than linear.*

When failure are ignored, operating a single facility seems better than operating two facilities for $p \in [0, 0.165]$*. However, when failure are considered, operating a single facility is better than operating two facilities only for* $p \in [0, 0.145]$*.*

Figure 3b presents $\mu^{1,2}$ *(dashed line),* $\mu^{2,1}$ *(dotted line), and* $\mu^{2,2}$ *(solid line), when failures are considered, and when they are ignored. When failures are ignored, the optimal capacities are independent of* p*. When failures are considered, the capacities are decreasing with* p*. That is, the capacities when failures are considered are lower than those when failures are ignored.*

To summarize, ignoring failures, may result in two wrong choices, one with respect to the number of facilities to operate, and one with respect to their capacities. In example 3, for reasonable (low) failure probabilities, the main difference in cost results from wrong capacities' choice, and the resulting cost difference are up to 5.3% (for $p \leq 0.5$).

4.4.3 $M/M/c/K$ Facilities

In the online Appendix C, we (numerically) consider unreliable facilities that are modeled as $M/M/c/K$ queues with $1 \leq c \leq K$ over the single edge network. In Example 5 we show that the optimal capacities increase with K and decrease with c, and in Example 6 we show that (i) the blocking cost of $Cost^2(p)$ is higher than or equal to that of $Cost^1(p)$, as stated in Theorem 6 part 2b for the $M/M/1/1$ facilities, and (ii) the value of C_K^* for $p = 0$ is λ_1, as in the $M/M/1/1$ facilities case. However, *in contrast* to the

case with $M/M/1/1$ facilities, the value of C_K^* is not necessarily decreasing with p.

4.5 Demand allocation among facilities

4.5.1 The importance of demand allocation to facilities

Note that when facilities are perfectly reliable, according to Theorem 2, the service capacity and blocking costs are independent of m, and so the decision of opening m or $m+1$ facilities trades off the traveling and setup costs. Therefore, the problem is separable, and we can solve it using the three phases approach discussed in Subsection 3.2., i.e., first to solve a *Location problem*, then to *finds* m^* and finally to determine the *optimal capacities*.

In contrast, when facilities are unreliable, according to Theorem 6, the service capacity and blocking costs depend on m, and the decision between opening m and $m+1$ facilities trades off all costs. Therefore, the problem when facilities are unreliable is no longer separable, and is thus much harder to solve. Specifically, the blocking, service capacity, and setup costs of operating $m+1$ facilities are higher than those of operating m facilities, the failure cost of operating $m+1$ is lower (but with a low failure probability this cost may have a limited effect), and the traveling cost can be lower or higher, depending on the failure probability. In a general network, the facilities' location influences the demand they face as well as the traveling, blocking, and service capacity costs. Therefore, in this setting, the question *"what is the cost of minimizing allocation of demand to facilities?"* becomes important.

We show that with unreliable facilities, *a fair allocation of the demand is best* in terms of minimizing the blocking and service capacity costs. We prove this for the single edge network with $M/M/1/1$ facilities, establish sufficient conditions for this to hold for general network with $M/M/c/K$ facilities, show that these conditions are satisfied when using approximations for the service capacities for the $M/M/1/1$ case, and demonstrate this for an example over the single edge network with $M/M/c/K$ facilities. Based on these, we conjecture that for a general network with $M/M/c/K$ facilities, a fair allocation of demand minimizes the blocking and service capacity costs. This result will support a heuristic developed in section 4.6, that will locate facilities in a general network such that their demand rates when there are no failures will be as equal as possible.

4.5.2 Fair allocation

As in a network settings, the choice of location affects the demands that the facilities face, we study what are good allocations of the total demand to facilities. We first show that over the single edge network

with $M/M/1/1$ facilities, a fair allocation minimizes the blocking and service capacity costs:

Theorem 7. *$Cost^2(p)$ is minimized when $\lambda_1 = \lambda_2 = 0.5\lambda$.*

We then observe that a sufficient condition for a fair allocation to minimize the blocking and service capacity costs for general networks is that the sum of the blocking and service capacity costs of facility j is convex in λ_j:

Observation 2. *Consider the following optimization problem:*

$$\min \quad F = \sum_{j=1}^{m} f(\lambda_j) \tag{37a}$$

$$S.t. \quad \sum_{j=1}^{m} \lambda_j = \lambda, \tag{37b}$$

where $f(\lambda_j)$ is a convex function in λ_j. Then, the optimal solution is $\lambda_j = \frac{\lambda}{m}$ $\forall j \in \{1, \ldots, m\}$.

To understand observation 2 intuitively, consider the single edge setting and define $f(\lambda_j) \equiv C_B(1-p)(\frac{(1-p)(\lambda_j)^2}{\lambda_j+\mu^{2,j}} + \frac{p\lambda^2}{\lambda+\mu^{2,j}}) + C_\mu \mu^{2,j}$ for $j \in J$ (the blocking and service capacity costs of facility j at $Cost^2(p)$). Observation 2 implies that if $f(\lambda_j)$ is convex in λ_j, then $Cost^2(p)$ is minimized if $\lambda_j = \frac{\lambda}{2}$ for $j = 1, 2$ (this can provide an alternative proof for Theorem 7).

We next use the same ideas behind (29) (see Section 4.4.1), and extend (29) to any number $m > 2$ of $M/M/1/1$ facilities. That is, we approximate the service capacity of the j^{th} facility when there are m opened facilities on the network by

$$\overline{\mu}^{m,j}(p) \equiv \frac{\lambda_j(p)}{1-p}(\sqrt{\gamma} - 1), \tag{38}$$

where $\lambda_j(p)$ is the expected demand rate of facility $j \in J$. As this expected demand depends also on the structure of the network, we cannot provide an explicit form for $\lambda_j(p)$. Nevertheless, it is important to note that the dependence of these bounds on λ_j is linear and increasing. This observation is used to prove Lemma 1, which shows that the $f(\lambda_j)$s for m facilities are convex in λ_j, when the approximations (38) are used for the optimal service capacities.

Lemma 1. *For m $M/M/1/1$ facilities, and using the approximations (38) for the optimal capacities, the $f(\lambda_j)$ are convex in λ_j.*

Combining Lemma 1 with Observation 2 implies that a fair allocation of demand to facilities minimizes the blocking and service capacity costs in these settings.

The next example demonstrates that in accordance with Theorem 7, a fair allocation of demand minimizes the blocking and service capacity costs also for (some) $M/M/c/K$ facilities.

Example 4. ***Fair allocation of demand:*** *Let* $\lambda_1 = 20, 30, 50$ $(\lambda_2 = 100 - \lambda_1), c = 2$, *and* $K = 3$. *Figure 4 shows* $Cost^2(p)$ *as a function of* p *for these three different divisions cases.* $Cost^2(p)$ *when* $\lambda_1 = 20$ *and* 30 *are drawn by a solid line, and* $Cost^2(p)$ *when* $\lambda_1 = 50$ *is drawn by a dashed line (note that the curves for* $\lambda_1 = 20$ *and* 30 *are almost identical). There is almost no difference in* $Cost^2(p)$ *when* $\lambda_1 = 20$ *and* 30. *However, it is clear that* $Cost^2(p)$ *is minimized when* $\lambda_1 = \lambda_2 = 50$.

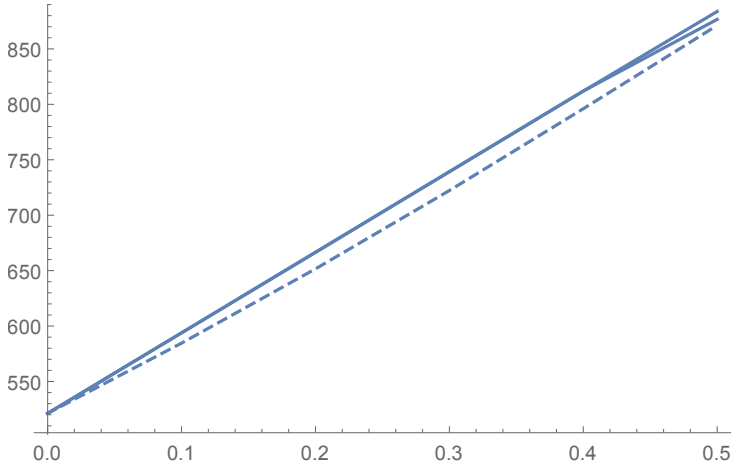


Figure 4: $Cost^2(p)$ for $\lambda_1 = 20, 30$, and 50

Given Theorem 7, Lemma 1, and Example 4, we conjecture that in a general network G with $M/M/c/K$ facilities, with the same cost structure as (11a), there are m convex functions $f(\lambda^j)$s, which are composed of the blocking and service capacity costs of the opened facilities. Thus, together with Observation 2, we conjecture:

Conjecture 1. *For any network* $G = (I, E)$, *with* m $M/M/c/K$ *facilities, locating facilities such that their demand rates are equal to each other minimizes the total blocking and service capacity cost.*

4.6 Solution on a general network

In this section we develop two heuristics to solve the problem with unreliable facilities on a general network when facilities operate as $M/M/c/K$ queues. We base these heuristics on our observations from the analysis of the problem with reliable facilities, and on the problem on the single edge network with unreliable facilities. Note that the problem with unreliable facilities is inseparable (as opposed to the problem with reliable facilities), so each heuristic has to calculate the total expected cost for a given

number of facilities. In the first heuristic we locate facilities according to the m-median solution, because this is the optimal location for the problem with reliable facilities (see Theorem 2). In the second heuristic we locate facilities such that each one faces as equitable demand as possible, because this is the optimal location for minimizing the total blocking and service capacity costs (see Conjecture 1). Both heuristics start their search at an initial guess as detailed below.

The input of Algorithm 2, given in the online Appendix B, is the network G, the demand rates λ_i for $i \in I$, the costs C_B, C_μ, C_K, the parameters c, K of the $M/M/c/K$ queue, and the failure probability p.

The algorithm starts with Procedure 0 that finds an *initial guess* $\hat{m}$ and is based on the relaxation of the m-median problem. Procedure 0 is followed by Procedure I, which is based on the exact m-median problem, and by Procedure II, which is based on the fair-m-median problem, discussed in [3].

Procedure 0**: *initial guess*:** similar to Procedure 0 of Algorithm 1 (and following the same reasoning), Procedure 0 finds a candidate $\hat{m}$ as a starting point of the search for the optimal number of facilities to open, m^*. For each $m \in I$, it finds the optimal location of the relaxed m-median problem, calculates a lower bound on the *total* expected cost (as detailed below), and determines the number of facilities that minimizes this cost, $\hat{m}$.

The lower bound on the total expected cost when m facilities are opened is obtained by bounding the traveling and blocking costs. For the traveling cost, let $Travel(h)$ denote the traveling cost of the relaxed h-median solution. Then, the bound $Travel(m)_{LOW}$, is the sum of $Travel(h)$ multiplied by the probability that h facilities are operational, for $h \in \{m, m-1, \ldots, 1\}$, that is,

$$Travel(m)_{LOW} = \sum_{h=1}^{m} Travel(h) \binom{m}{h} p^{m-n} (1-p)^h.$$

The lower bound on the expected blocking cost is obtained by assuming that each operating facility always faces an equal demand rate. For example, when there are h operational facilities, each face a demand rate λ/h. Summing the blocking cost when h facilities are operational, multiplied by the probability that h facilities are operational,

$$Block(m)_{LOW} \Leftarrow m \sum_{h=1}^{m} C_B(\lambda/h) P_B^j(\lambda/h, \mu^{h,j}) \binom{m}{h} p^{m-h} (1-p)^h$$

(with $P_B^j(\cdot)$ given by (1)-(2)). The optimal capacities are evaluated numerically using the FOC of $Block(m)_{LOW} + mC_\mu \mu^{m,j}$ with respect to $\mu^{m,j}$ (the assumption of equitable demand rates implies that facilities are identical). Finally, Procedure 0 determines the number of facilities that minimizes the total

cost, denoted by $\hat{m}$.

Procedure I: m-*median*: Procedure I determines the optimal number of facilities to be opened by searching over a close neighborhood around $\hat{m}$ using the exact m-median problem. The stopping rule we implement for the procedure assumes convexity of the total cost in m, and the total cost are approximated by their lower bound developed in Theorem 3 with $v = 3$ (i.e, the exact calculation is based on at most 3 failed facilities). For each m in this neighborhood, after finding the optimal facilities location, and the resulted traveling cost of the m-median problem, Procedure I determines $\underline{\Pi^3}$, and calculates numerically the "optimal" capacities using the FOC of $\underline{\Pi^3}$ and the service capacity cost. Then, it evaluates $\overline{\Pi^3}$, and the service capacity cost. Finally, it calculates the expected traveling cost as described in Subsection 2.3.2., and the total cost of operating m^* facilities, including the setup and failure costs.

Procedure II: *fair*-m-*median*: Procedure II repeats Procedure I, but locates facilities according to the fair-m-median heuristic from [3].

5 Case study

Our results can be applied to systems of identical service facilities with finite buffers, such as restaurants, emergency departments and post-offices. For each application, appropriate choice of the parameters C_B, C_μ, C_K, C_T, c, and K are required. To demonstrate our results, we next apply them to the McDonald's drive-through restaurants in Toronto, Canada.

5.1 McDonald's drive-throughs

A drive-through allows customers to purchase products without leaving the car. Generally, the orders are placed using a microphone, and are collected in person at the pick-up window. The cars move in one direction in the drive-through lane toward the foods request window. The food is brought to the window by a server and customers can remain in the car and eat. As the lane has a finite length, some potential customers may see the entire lane busy, and are thus blocked (see, e.g., Figure 5, [14]). We believe that the finite buffer queueing model is appropriate for this application.

The McDonald's Corporation is the world's largest chain of fast food restaurants. It is closely associated with its drive-through service, which aims at being available 24/7. However, customers often are not satisfied with it, mostly because of the long wait and service times, and the number of orders that are served incorrectly (the order accuracy). For example, according to [31]: "The drive-through is

Figure 5: A McDonald's drive-through

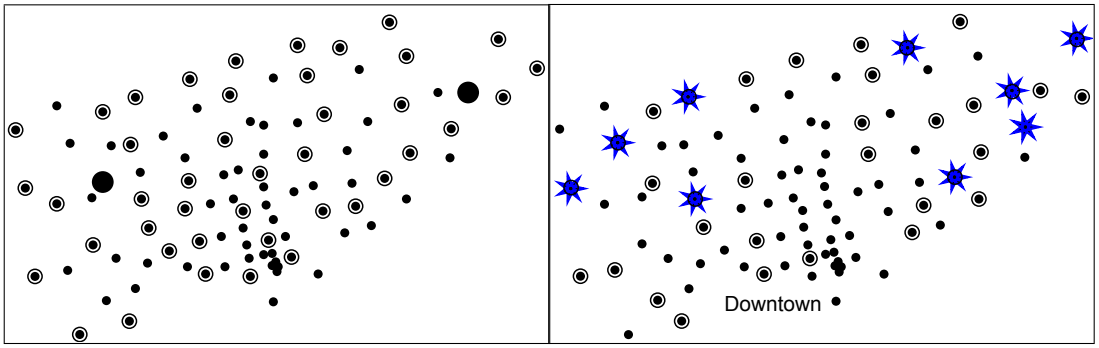
getting slower. OK, so maybe this is one part of the drive-through experience that doesn't come as a total surprise. But the latest version of an annual study from QSR (2011) Magazine confirms that wait times at the drive-through are on the rise. Last year study suggests that the average drive-through wait time hit 181 seconds" (note that the "wait time" usage here probably relates to the sojourn times, i.e., the service plus waiting time in queuing theory), and according to [26], the order accuracy of McDonald's is 88.3% (note that the order accuracy and the service times are correlated - longer service times imply a higher order accuracy).

We utilize the same data used in [8] to represent the Toronto metropolitan area in Canada. Specifically, the area is partitioned into 96 FSAs areas (defined by the first three digits of the Canadian postal code), and each such FSA is represented as a node in the network. Each one of the nodes has a demand that represents the population that resides and work at the corresponding area. There is an edge between two nodes only if the corresponding FSAs share a boundary.

We next solve the McDonald's drive-through facility problem when minimizing the daily costs. We model the drive-through facilities as $M/M/1/15$ queues. We assume that $K = 15$ because on average there is a place for 15 cars in the drive-through lane. We assume that $c = 1$ because the window is a single server. And its service capacity is dictated by a combination of several factors such as the number of grills, so a continuously adjustment of service capacity is sensible.

The traveling cost is assumed to equal the distance that the customers travel, i.e., $C_T = \$1$ per unit distance. The setup cost C_K we consider is \$800 dollars per day. This corresponds to a planning period of five years to return to equity and a typical McDonald's facility setup cost of about $\$1,500,000$, see, e.g., [16].

We assume that it costs \$1 to serve a customer on average, so the service capacity cost is $C_\mu = \$1$. Further, we assume that a typical order at McDonald's drive-through is \$7.5, and that the costs of food



(a) Theoretical solutions (b) Actual

Figure 6: McDonald's drive through' locations

and packaging cost is \$4. So, with the \$1 service capacity cost, the blocking cost is $C_B = \$2.5$ (profit lost per unsatisfied order on average and additional loss of goodwill, which is harder to quantify and we thus ignore). We consider failure probabilities of 0.01 and 0.02, corresponding to a failure of about three days and a week per year, respectively.

Finally, there are 2.5 million visits to McDonald's restaurants across Canada every day [30]. With 35.16 million people living in Canada [10], this implies that about 7% of the entire population in Canada visits a McDonald's restaurant every day. As according to [25], about 57% of fast food restaurants customers use the drive-through service; we assume that 3.5% of the population in Toronto visits McDonald's drive-throughs every day. Thus, demand for each FSA is multiplied by 0.035.

Given this data and assumptions, we used Algorithm 1 to solve the Reliable facilities problem, Algorithm 2 to solve the Unreliable facilities problem, and the actual locations (given in [17]) assuming an optimal service capacity is opened in each location to estimate the actual cost (thus, these results serve as lower bounds on the actual cost). The results are presented in Figure 6 and Table 1 and are discussed below.

Case	m	Cost when $p = 0$	Cost when $p = 0.01$	Cost when $p = 0.02$	Service capacity and setup costs	Average service time (minutes)
Reliable	42	107,374	109,975	115,219	73,850	1.5
Unreliable $p = 0.01$	42	107,438	109,325	115,075	73,998	1.5
Unreliable $p = 0.02$	44	107,603	109,910	110,323	74,695	1.6
Actual	43(34)	136,821	138,783	142,502	74,481	1.22

Table 1: Number of facilities, costs, and average service times of the theoretical and actual solutions.

5.2 Locations and capacities

The optimal facilities' locations, and the actual locations are presented in Figure 6. Panel (a) includes the theoretical solutions. For a failure probability of 0.01, the solution of the unreliable model suggests

the same number and location of facilities as in the reliable case. For a failure probability of 0.02, the solution of the unreliable model suggests opening two additional facilities (the two larger black circles), without changing optimal locations of the reliable model.

The fact that the theoretical locations of the 42 facilities in the reliable case are identical to these in the unreliable cases is not that surprising. Considering that the network has 96 nodes, the reliable problem solution already locates at almost 50% of the nodes. Because the failure probability is low, there is no co-location, and so there are not many different options for locating the facilities.

The actual solution, depicted in panel (b), has stars that indicate that there are two facilities located at the same FSA (the phenomena of the apparent co-location of the actual situation indicates that the data we use may be too aggregated, and we could have divided Toronto into smaller demand areas defined, e.g., by the entire postal code).

Comparing the theoretical locations with the actual ones, we observe that in practice there are fewer restaurants in the downtown area than in our theoretical solutions. This change can be attributed to two factors. First, the fixed setup cost, C_K is location dependent, and the real estate prices in downtown Toronto are the highest in the city, so there is a preference for opening restaurants in other parts of the city. Second, as the traffic in downtown Toronto is quite heavy, many people prefer walking or using public transportation to driving - reducing the demand for drive-throughs.

We note that the locations of the theoretical models correspond to the m-median solutions rather than to the problem with fair allocation of demand. The network structure for the Toronto metropolitan area has a highly uneven demand among the different FSAs, and thus the search for a location resulting in equitable demand rates with a closest assignment restriction is not fruitful.

Table 1 presents the number of facilities, costs and average service times of the different solutions. Rows $2, 3, 4$ report those values for the theoretical solutions assuming full reliability, unreliability with probability 0.01 and unreliability with probability 0.02, respectively. Row 5 reports those values for the actual situation. For example, according to Algorithm 2, with a failure probability of 0.01, it is optimal to open 42 facilities, with a total cost of $109,325$ (row 3, columns 2 and 4, respectively). The cost of this solution when the actual probability of failure is 0 and 0.02, is $107,438$ and $115,075$, respectively (row 3, columns 3 and 5).

The last column of Table 1 presents the average service times of all facilities for the different solutions. The theoretical solutions offer similar average service times, of $1.5 - 1.6$ minutes, whereas the solution of the actual situation offers a smaller average service time, of 1.22 minutes. We note that because of

the higher service times of the theoretical solutions, their order accuracy should be higher than that of the actual situation. Moreover, as the theoretical average service times in the different cases are close to each other, there is no strict preference (in terms of service times) among them.

To summarize, the theoretical models imply that when unreliability increases (as it is also shown in Example 3 for the single edge network), it is optimal to open more facilities with lower capacities.

5.3 Costs

We observe that when $p = 0$ and $p = 0.01$, using either the reliable problem solution or any one of the two unreliable solutions gives similar costs (Table 1, rows 2-4, columns 3 and 4). However, when $p = 0.02$, the optimal solution has a significant cost reduction of about 4.2% comparing to the costs of the reliable solution and unreliable solution with $p = 0.01$ (Table 1, rows 2-4, column 5). So, it appears as if the solution of the unreliable case with $p = 0.02$ is robust with respect to the failure probability.

There seems to be big differences between the costs of the theoretical solution and actual situation, of 21% on average. The first reason for this high difference is the fixed setup cost. As mentioned above, opening a facility in downtown Toronto is expensive, so, the actual setup costs of the theoretical solutions is likely underestimated in our models. The second reason is the total cost we consider. This cost includes "opportunity cost" represented by traveling and blocking costs, which are not out of pocket expenses of McDonald's. The out of pocket expenses (service capacity and setup costs) are reported in the sixth column of Table 1. Considering only these costs, the theoretical solutions save only $100(74,481 - (73,850 + 73,998 + 74,695)/3)/74,481\% = 0.4\%$ of the total daily cost (but, as we evaluate the costs of the actual solutions assuming optimal capacities, this saving is likely higher).

Further, note that while 0.4% may seem like a small saving, with a profit margins of 6% for McDonald's [27], 0.4% cost reduction represents a 6.66% profit improvement. So, the improvement in McDonald's out of pocket cost may be significant.

The difference between the costs of the actual and theoretical models discussed above may suggest that either we overestimate the opportunity costs or that McDonald's underestimate them. We further study this in the next subsection.

5.4 Effect of the opportunity costs

We consider lower values of C_T and C_B, in order to understand better the reasons for the (relatively high) differences between the theoretical and actual costs. Specifically, we assume that $C_T = \$0.55$ and

$C_B = \$1.2$ (note that $C_T = \$0$ cannot be used in our theoretical models because the reliable model trades off the traveling and setup costs. Further, C_B cannot be reduced below \$1, as it must be higher than $C_\mu = \$1$ according to (25)). The results for this case are presented in Table 2.

Case	m	Cost when $p = 0$	Cost when $p = 0.01$	Cost when $p = 0.02$	Service capacity and setup costs	Average service time (minutes)
Reliable	32	89,366	90,439	92,232	58,980	1.38
Unreliable $p = 0.01$	32	89,375	90,429	92,204	58,705	1.4
Unreliable $p = 0.02$	33	90,726	90,454	92,161	58,840	1.4
Actual	43(34)	106,811	107,779	109,489	67,879	1.46

Table 2: Number of facilities, costs, and average service times of the theoretical and actual solutions.

As for the location of facilities, the results are similar to these of the previous case: for a failure probability of 0.01, the solution of the unreliable model suggests the same number and location of facilities as in the reliable case. For a failure probability of 0.02, the solution of the unreliable model suggests opening an additional facility, without changing optimal locations of the reliable model. Again, we observe that when unreliability increases, it is optimal to open more facilities with lower capacities. An interesting observation is that the cost of the theoretical solution with $p = 0.02$ when the actual failure probability is $p = 0.01$ is lower than its cost with actual $p = 0$ whereas the costs of the other solutions increase with p.

In contrast to the previous settings, here for all failure probabilities the costs are close to each other, and there is no strict preference among the three theoretical solutions. Moreover, the percent of extra cost (in comparison to the lowest feasible cost) of the actual solution is decreasing with the probability of failure. Overall, it appears also here that the solution with $p = 0.02$ is relatively robust with respect to lower unreliability parameters.

The differences between the theoretical and actual solutions total costs are, on average, 16% for the total costs and 13.33% for the out of pocket expenses. Note that the big difference of the out of pocket expenses is due also to the increase in the setup cost of the 10 additional facilities opened in the actual solution, which is \$8,000. Without this cost, the out of pocket cost reduction is $100((67,879 - 8,000) - (58,840 + 58,705 + 58,980)/3)/(67,879 - 8,000) = 1.73\%$, representing a 28.8% profit improvement.

Finally, even after substantially reducing the opportunity costs, the current solution appears far from optimal and there may be a place for improving the overall profitability of McDonald's drive-through restaurants in Toronto by improving their locations and considering the possibility of failure.

6 Summary

We considered two cost minimization facility location problems. The first is a location problem with congestion, where facilities are perfectly reliable and modeled as $M/M/c/K$ queues. The service level measure we considered is the probability of blocked customers, i.e, of lost demand. The second problem extends the first one by considering unreliable facilities. We assume customers are aware of the operational status of the facilities, and always travel to their closest operational facility. The cost includes traveling, blocking, service capacity, fixed operating, and in the model with failures also penalties for failures. For each problem, the number of facilities to be opened, their location, and their capacities are optimized.

We first analyzed the problems over the a single edge network. Then, using results and insights from this analysis, we showed how such problems can be solved on a general network. We developed two efficient algorithms to solve both problems. We demonstrated the applicability of our results and our algorithms with a case study that considered drive-through operation of McDonald's in the Toronto metropolitan area. The case study revealed that (i) the actual solution may reduce McDonald's profit by $6.66\% - 28.8\%$; (ii) considering unreliability can be done with a relatively low added cost; (iii) considering unreliability provides a rather robust solution with respect to the failure probability; and that (iv) considering reliability ends up in opening more facilities with lower capacities.

References

[1] Baron, O., Berman, O., Krass, D. 2008. Facility location with stochastic demand and constraints on waiting time. *Manufacturing and Service Operations Management*, 10(3), 484–505.

[2] Baron, O., Berman O., Krass D., Wang, Q. 2007. The equitable location problem on the plane. *European Journal of Operations Research*, 183(2), 578–590.

[3] Berman O., Drezner Z., Tamir A., Wesolowsky, G. O. Optimal location with equitable loads. 2009. *Annals of operations research*, 167, 307–326.

[4] Berman,O. Krass,D. 2002 Facility location problems with stochastic demands and congestion, in Facility Location: Applications and Theory, Drezner, Z. and Hamacher, H.W. (eds.), Springer-Verlag, New York, Chapter 11, 329–371.

[5] Berman O., Krass D. 2011. On n-facility median problem with facilities subject to failure facing uniform demand. *Discrete Applied Mathematics*, 159(6), 420–432.

[6] Berman, O. Krass D. 2015. Stochastic location models with congestion, Chapter 17 Location and Logistics. Laporte, G., Nickel, S., Saldanha-da-Gama, F. Springer.

[7] Berman O., Krass D., Menezes M. B. C. 2007. Facility reliability issues in network p-median problems: strategic centralization and co-location effects. *Operations Research*, 55(2), 332–350.

[8] Berman O., Krass D., Wang J. 2006. Locating facilities to reduce lost demand. *IIE Transactions*, 38, 933–946.

[9] Boyd S., Vandenberghe L. 2004. Convex Optimization. Cambridge University Press.

[10] Canada population 2014 (October 19, 2014). http://worldpopulationreview.com/countries/canada-population/.

[11] CBC News-Health (June 03, 2014). Medical wait times up to 3 times longer in Canada. http://www.cbc.ca/news/health/medical-wait-times-up-to-3-times-longer-in-canada-1.2663013.

[12] Chang C. S., Chao X., Pinedo M., Shanthikumar J. G. 1991. Stochastic convexity for multidimensional processes and its applications. *Automatic Control, IEEE Transactions on*, 36(12), 1347–1355.

[13] Cornes R. 1992. Duality and modern economics. Cambridge: Cambridge University Press

[14] Daily mail reporter (February 05, 2012). She's not loving it: Woman tasered in McDonald's drive-thru after cutting line. http://www.dailymail.co.uk/news/article-2096662/Evangeline-Lucca-tasered-stun-gun-McDonalds-drive-cutting-line.html.

[15] Drezner Z. 1987. Heuristic solution methods for two location problems with unreliable facilities. *Journal of the Operational Research Society*, 509–514.

[16] Entrepreneur. About McDonald's. http://www.entrepreneur.com/franchises/mcdonalds/282570-0.html.

[17] Find a McDonald's near you. http://www.mcdonalds.ca/ca/en/restaurant_locator.html.

[18] Gurvich I., Sarang D. 2011. Centralized vs. decentralized ambulance diversion: a network perspective. *Management Science*, 57(3): 1300–1319.

[19] Hakimi S.L. 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3), 450–459.

[20] Hakimi S.L. 1965. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13(3), 462–475.

[21] Harel A. 1990. Convexity properties of the Erlang loss formula. *Operations Research*, 38(3), 499–505.

[22] Ingolfsson A. 2013. EMS planning and management. *Operations Research and Health Care Policy*. Springer New York, 105–128.

[23] Jagerman D. L. 1975. Nonstationary blocking in telephone traffic. *Bell System Technical Journal*, 54(3), 625–661.

[24] Krarup, J., Pruzan, P.M. 1983. The simple plant location problem: survey and synthesis. *European Journal of Operational Research*, 12(1), 36–81.

[25] McDonnell S. What percentage of sales are from drive through windows at fast food restaurants? http://smallbusiness.chron.com/percentage-sales-drive-through-windows-fast-food-restaurants-75713.html.

[26] QSR. 2011 Drive-thru study: historical order accuracy. http://www.qsrmagazine.com/content/2011-drive-thru-study-historical-order-accuracy.

[27] Quora. What is the profit margin for fast food franchises like McDonald's and Domino's Pizza? http://www.quora.com/What-is-the-profit-margin-for-fast-food-franchises-like-McDonalds-and-Dominos-Pizza.

[28] Snyder L. V, Daskin M. S. 2005. Reliability models for facility location: the expected failure cost case. *Transportation Science*, 39(3):400–416.

[29] Sztrik J. 2011. Basic queueing theory. University of Debrecen, Faculty of Informatics.

[30] The history of the golden arches. http://www.mcdonalds.ca/ca/en/our_story/our_history.html.

[31] Tuttle B. 10 Things you didnt know about the fast food drive-thru (October 8, 2014). `http://time.com/money/3478752/drive-thru-fast-food-fast-casual/`.

[32] Verter, V., Lapierre S.D. 2002. Location of preventive health care Facilities. *Annals of Operations Research*, 110, 123–132.

[33] Yao D. D., Shanthikumar J. G. 1986. The optimal input rates to a system of manufacturing cells. Produced and distributed by Center for Research in Management, University of California, Berkeley Business School.

A Appendix

Proof of Theorem 1

Part 1: from the first order conditions (FOC) with respect to the service capacity of (13) (second order conditions, SOC, can be verified as well): $\frac{\partial Cost^1}{\partial \mu^{1,2}} = -\frac{C_B(\lambda)^2}{(\lambda+\mu^{1,2})^2} + C_\mu = 0$, implying (16). Similarly, the FOC (SOC can be verified as well) with respect to (14) imply (17). For the capacities to be real and positive, we require (15) and (18) is immediate.

Part 2: using (18) and the linearity of the service capacity cost, the total service capacity cost is independent of $m(=1,2)$. Further, the blocking cost of $Cost^1$ is $\frac{C_B(\lambda)^2}{\lambda+\mu^{1,2}} = \frac{C_B(\lambda)^2}{\lambda+\lambda(\sqrt{\frac{C_B}{C_\mu}}-1)} = \frac{C_B\lambda}{\sqrt{\frac{C_B}{C_\mu}}} = \lambda\sqrt{C_BC_\mu}$, and that of $Cost^2$ is $\frac{C_B(\lambda_1)^2}{\lambda_1+\mu^{2,1}} + \frac{C_B(\lambda_2)^2}{\lambda_2+\mu^{2,2}} = \frac{C_B\lambda_1}{\sqrt{\frac{C_B}{C_\mu}}} + \frac{C_B\lambda_2}{\sqrt{\frac{C_B}{C_\mu}}} = \frac{C_B\lambda}{\sqrt{\frac{C_B}{C_\mu}}} = \lambda\sqrt{C_BC_\mu}$. That is, this cost is also independent of m.

Part 3: given part 2, we are left with the traveling and setup costs. From (13), these costs for $Cost^1$ are $\lambda_1 + C_K$, and from (14), the setup cost for $Cost^2$ (there is no traveling cost) is $2C_K$, leading to (19).

Proof of Theorem 2

Part 1: We first show that the optimal capacities of $M/M/c/K$ facilities are given by

$$\mu^{m,j} = \lambda^{m,j}\eta, \tag{39}$$

where, as before $\lambda^{m,j}$ is the arrival rate that facility j faces when m facilities are opened. Note that given m facilities and their location, the service capacity only affects the blocking and service capacity costs. That is, using (1-2) and (6a) the optimal service capacity, $\mu^{m,j}$, minimizes the cost (we focus on the case $\lambda^{m,j}/\left(c\mu^{m,j}\right) \neq 1$, the proof for this limiting case follows by a similar argument)

$$\begin{aligned} F(\mu) &= C_B\lambda^{m,j}P_B\left(\lambda^{m,j}\right) + C_\mu\mu \\ &= \frac{\frac{C_B\left(\lambda^{m,j}\right)^{K+1}}{\mu^K c^{K-c}c!}}{\sum_{n=0}^{c-1}\frac{(\lambda^{m,j})^n}{\mu^n n!} + \frac{(\lambda^{m,j})^c}{\mu^c c!}\left(\frac{1-(\lambda^{m,j}/(c\mu))^{K-c+1}}{1-\lambda^{m,j}/(c\mu)}\right)} + C_\mu\mu. \end{aligned} \tag{40}$$

And note that $F(\mu)$ is convex in μ because the blocking probability is convex in μ (see, [21]) and the service capacity cost is linear in μ. Thus, the optimal service capacity, $\mu^{m,j}$, is given by the solution of the FOC. Consider a change of variables, i.e., let $\mu = \lambda^{m,j}\eta$ in (40) and note that $d\mu = \lambda^{m,j}d\eta$ then

$$\begin{aligned} F(\eta) &= \frac{\frac{C_B\left(\lambda^{m,j}\right)^{K+1}}{(\lambda^{m,j}\eta)^K c^{K-c}c!}}{\sum_{n=0}^{c-1}\frac{(\lambda^{m,j})^n}{(\lambda^{m,j}\eta)^n n!}+\frac{(\lambda^{m,j})^c}{(\lambda^{m,j}\eta)^c c!}\left(\frac{1-(\lambda^{m,j}/(c\lambda^{m,j}\eta))^{K-c+1}}{1-\lambda^{m,j}/(c\lambda^{m,j}\eta)}\right)}+C_\mu\lambda^{m,j}\eta \\ &= \lambda^{m,j}\left(\frac{\frac{C_B}{\eta^K c^{K-c}c!}}{\sum_{n=0}^{c-1}\frac{1}{\eta^n n!}+\frac{1}{\eta^c c!}\left(\frac{1-(1/(c\eta))^{K-c+1}}{1-1/(c\eta)}\right)}+C_\mu\eta\right). \end{aligned}$$

And it is clear that $dF(\eta)/d\mu = dF(\eta)/d\eta * d\eta/d\mu = d\left(F(\eta)/\lambda^{m,j}\right)/d\eta$ is fixed in $\lambda^{m,j}$, and thus the η that solves the FOC is independent of $\lambda^{m,j}$. Thus, the function in (39) solves the FOC for each $\lambda^{m,j}$. Moreover, the conditions for the implicit function theorem to hold for the FOC around any $\lambda \geq 0$ are easily verified, thus for any λ there is a *unique* function $\mu^{m,j}(\lambda)$ that holds in the neighborhood of λ and solve the FOC there and thus (39) is the unique function of the optimal service capacity in $\lambda^{m,j}$.

Next because all facilities have the same η, the total service capacity cost is given by $\sum_{j\in J} C_\mu\mu^{m,j} = \sum_{j\in J} C_\mu\lambda^{m,j}\eta = C_\mu\lambda\eta$, implying that the total service capacity cost is independent of $m \in I$.

Now, the blocking cost of facility j is given by (using (1-2)):

$$C_B\lambda^{m,j}P_B(\lambda^{m,j}) = C_B\lambda^{m,j}\frac{(\lambda^{m,j}/(c\mu^{m,j}))^K}{c^{K-c}c!(\sum_{n=0}^{c-1}\frac{(\lambda^{m,j}/\mu^{m,j})^n}{n!}+\frac{(\lambda^{m,j}/\mu^{m,j})^c}{c!}\frac{1-(\lambda^{m,j}/(c\mu^{m,j}))^{K-c+1}}{1-\lambda^{m,j}/(c\mu^{m,j})})}.$$

Substituting (39), we get:

$$C_B\lambda^{m,j}\frac{(1/(c\eta))^K}{c^{K-c}c!(\sum_{n=0}^{c-1}\frac{(1/\eta)^n}{n!}+\frac{(1/\eta)^c}{c!}\frac{1-(1/(c\eta))^{K-c+1}}{1-1/(c\eta)})}.$$

So, the total blocking cost is

$$\begin{aligned} \sum_{j\in J} C_B\lambda^{m,j}P_B(\lambda^{m,j}) &= \sum_{j\in J} C_B\lambda^{m,j}\frac{(1/(c\eta))^K}{c^{K-c}c!(\sum_{n=0}^{c-1}\frac{(1/\eta)^n}{n!}+\frac{(1/\eta)^c}{c!}\frac{1-(1/(c\eta))^{K-c+1}}{1-1/(c\eta)})} \\ &= C_B\lambda\frac{(1/(c\eta))^K}{c^{K-c}c!(\sum_{n=0}^{c-1}\frac{(1/\eta)^n}{n!}+\frac{(1/\eta)^c}{c!}\frac{1-(1/(c\eta))^{K-c+1}}{1-1/(c\eta)})}, \end{aligned}$$

that is, it is independent of $m \in I$.

Parts 2 and 3 are direct results of part 1.

Proof of Theorem 3

First, note that Π^v is increasing in v and $\Pi = \Pi^{m-1}$. So, for any $v = 0, \ldots m-1$, $\Pi \geq \underline{\Pi^v}$. Second, for $0 \leq v \leq m-1$, let Π^v_{Exact} be the exact blocking cost when v facilities fail, and let $P^v \equiv \binom{m}{v}(1-p)^{m-v}p^v$ be the probability that v facilities fail. With these notation, $\Pi = \Pi^{m-1} = \Pi^v + \sum_{u=v+1}^{m-1} P^v \Pi^v_{Exact} \leq \Pi^v + \sum_{i=v+1}^{m-1} P^i \Pi^{j_{\min},\lambda} = \Pi^v + P^{v^+}\Pi^{j_{\min},\lambda} = \overline{\Pi^v}$ (the last inequality follows from the monotonicity properties of the blocking probability formula (1-2), which is increasing in the demand rate (see, e.g., [33]), and decreasing in the service rate (see, e.g., [12])).

Proof of Theorem 4

The FOC with respect to the service capacity of (23) (SOC can be verified), $\frac{\partial Cost^1(p)}{\partial \mu^{1,2}} = -\frac{C_B(\lambda)^2(1-p)}{(\lambda+\mu^{1,2})^2} + C_\mu = 0$, imply (26). And (27) is the FOC with respect to the capacities in (24) (SOC can be verified). Finally, for $\mu^{1,2}$ to be real and positive, we require (25). For $\mu^{2,j}$ to be real and positive, we require $C_B \geq \frac{C_\mu}{2(1-p)^2}$, which for every $p \leq 0.5$, is less strict than (25).

Proof of Theorem 5:

For simplicity of the exposition, and without loss of generality, we prove the results for facility 1 (the analysis for facility 2 is similar). We first show that the function $F(\mu^{2,j})$, defined by the left side of (27), is decreasing and convex in $\mu^{2,j}$.

We will prove that $F(\overline{\mu}^{2,j}(p))$ is convex in p and both $F(\overline{\mu}^{2,j}(0)) = \frac{1}{\gamma}$ and $F(\overline{\mu}^{2,j}(1-\frac{C_\mu}{C_B})) = \frac{1}{\gamma}$ that implies the line segment between $(0, F(\overline{\mu}^{2,j}(0)))$ and $(1-\frac{C_\mu}{C_B}, F(\overline{\mu}^{2,j}(1-\frac{C_\mu}{C_B})))$ lies above the graph of $F(\mu^{2,j}(p))$, so that $F(\overline{\mu}^{2,j}(p)) < \frac{1}{\gamma}$ for $p \in (0, 1-\frac{C_\mu}{C_B})$. This inequality, together with that $F(\mu^{2,j})$ is decreasing in $\mu^{2,j}$ implies $(\mu^{2,j})^* \leq \overline{\mu}^{2,j}(p)$.

We next prove that:

(a) $F(\mu^{2,1})$ is decreasing and convex in $\mu^{2,1}$.

(b) $F(\overline{\mu}^{2,1}(p))$ is convex in p.

(c) $\Delta(F) \equiv F(\overline{\mu}^{2,1}(p)) - \frac{1}{\gamma} \leq 0$ for $p \in [0, 1-\frac{C_\mu}{C_B}]$.

Part a: the first and second derivatives of $F(\mu^{2,1})$ with respect to $\mu^{2,1}$ are given by:

$$\frac{\partial F(\mu^{2,1})}{\partial \mu^{2,1}} = -2\left(\frac{(\lambda_1)^2(1-p)}{(\lambda_1+\mu^{2,1})^3} + \frac{\lambda^2 p}{(\lambda+\mu^{2,1})^3}\right) \leq 0,$$

and

$$\frac{\partial^2 F(\mu^{2,1})}{\partial {\mu^{2,1}}^2} = 6\left(\frac{(\lambda_1)^2(1-p)}{(\lambda_1+\mu^{2,1})^4} + \frac{\lambda^2 p}{(\lambda+\mu^{2,1})^4}\right) \geq 0,$$

both inequalities are immediate. Thus, $F(\mu^{2,1})$ is decreasing and convex in $\mu^{2,1}$.

Part b: we will show that $\frac{\partial^2 F(\overline{\mu}^{2,1}(p))}{\partial p^2} \geq 0$. Recall,

$$\begin{aligned} F(\overline{\mu}^{2,1}(p)) &= \frac{(\lambda_1)^2(1-p)}{(\lambda_1+(\lambda_1+p\lambda_2)(\sqrt{\gamma}-1))^2} + \frac{\lambda^2 p}{(\lambda+(\lambda_1+p\lambda_2)(\sqrt{\gamma}-1))^2} \\ &= \frac{(\lambda_1)^2(1-p)}{((\lambda_1+\lambda_2 p)\sqrt{\gamma}-\lambda_2 p)^2} + \frac{\lambda^2 p}{((\lambda_1+\lambda_2 p)\sqrt{\gamma}-\lambda_2 p+\lambda_2)^2} \\ &= \frac{1-p}{((1+Ap)\sqrt{\gamma}-Ap)^2} + \frac{(A+1)^2 p}{((1+Ap)\sqrt{\gamma}-Ap+A)^2} \\ &= \frac{1-p}{Z^2} + \frac{(A+1)^2 p}{(Z+A)^2}, \end{aligned} \tag{41}$$

with $A = \frac{\lambda_2}{\lambda_1} (\geq 1)$, and $Z = (1+Ap)\sqrt{\gamma} - Ap$. Note that $Z > 0$, because $[(1+Ap)\sqrt{\gamma} - Ap \geq 0] \Leftrightarrow [\sqrt{\gamma} \geq \frac{Ap}{1+Ap}]$, which follows because $\gamma \geq 1$ by (25).

The second derivative with respect to p is given by (note this includes $\frac{\partial Z}{\partial p} = -A + A\sqrt{\gamma} - \frac{C_B(1+Ap)}{2C_\mu\sqrt{\gamma}}$, because $\gamma = \frac{C_B(1-p)}{C_\mu}$):

$$\begin{aligned} \frac{\partial^2 F(\overline{\mu}^{2,1}(p))}{\partial p^2} &= \left(\frac{3(1+Ap)^2 C_B}{2C_\mu Z^4} + \frac{(1+Ap)C_B^2(1-p)}{2C_\mu^2 Z^3 \gamma^{3/2}} - \frac{2(1+Ap)C_B}{C_\mu Z^3\sqrt{\gamma}}\right) \\ &+ \frac{2(1+A)^2(1+Ap)C_B}{C_\mu(A+Z)^3\sqrt{\gamma}} + \frac{3(1+A)^2(1+Ap)^2 C_B p}{2C_\mu(A+Z)^4(1-p)} + \frac{(1+A)^2(1+Ap)C_B^2 p}{2C_\mu^2(A+Z)^3\gamma^{3/2}}. \end{aligned}$$

The last three terms are clearly positive. Summing the three terms in the first bracket:

$$\begin{aligned} &\frac{3(1+Ap)^2 C_B}{2C_\mu Z^4} + \frac{(1+Ap)C_B^2(1-p)}{2C_\mu^2 Z^3\gamma^{3/2}} - \frac{2(1+Ap)C_B}{C_\mu Z^3\sqrt{\gamma}} \\ = &\frac{3(1+Ap)^2 C_B}{2C_\mu Z^4} - \frac{3(1+Ap)C_B}{2C_\mu Z^3\sqrt{\gamma}} \\ = &3(1+Ap)C_B\frac{(1+Ap)\sqrt{\gamma}-Z}{2C_\mu Z^4\sqrt{\gamma}} \\ = &3(1+Ap)C_B\frac{(1+Ap)\sqrt{\gamma}-(1+Ap)\sqrt{\gamma}+Ap}{2C_\mu Z^4\sqrt{\gamma}} \\ = &3(1+Ap)C_B\frac{Ap}{2C_\mu Z^4\sqrt{\gamma}} \geq 0. \end{aligned}$$

Part c: By (27), $F((\mu^{2,1})^*) = \frac{1}{\gamma}$, thus, given (41) the gap between $F(\overline{\mu}^{2,1}(p))$ and $F((\mu^{2,1})^*)$ is

$$\Delta(F) \equiv F(\overline{\mu}^{2,1}(p)) - \frac{1}{\gamma} = \frac{(\lambda_1)^2(1-p)}{(\lambda_1+(\lambda_1+p\lambda_2)(\sqrt{\gamma}-1))^2} + \frac{\lambda^2 p}{(\lambda+(\lambda_1+p\lambda_2)(\sqrt{\gamma}-1))^2} - \frac{1}{\gamma}. \tag{42}$$

Now, for $p=0$, $\gamma = \frac{C_B}{C_\mu}$, so $F(\overline{\mu}^{2,1}(0)) = \frac{(\lambda_1)^2}{(\lambda_1+\lambda_1(\sqrt{\frac{C_B}{C_\mu}}-1))^2} = \frac{1}{(1+\sqrt{\frac{C_B}{C_\mu}}-1)^2} = \frac{1}{\gamma}$, and $\Delta(F) = 0$.

And for $p = 1 - \frac{C_\mu}{C_B}$, $\gamma = 1$, so $F(\mu^{2,1}(1 - \frac{C_\mu}{C_B})) \equiv \frac{(\lambda_1)^2(\frac{C_\mu}{C_B})}{(\lambda_1)^2} + \frac{\lambda^2(1-\frac{C_\mu}{C_B})}{\lambda^2} = \frac{1}{\gamma}$, and $\Delta(F) = 0$.

(a)-(c) prove the right inequality in (5).

We next calculate the error $\overline{\mu}^{2,1}(p) - (\mu^{2,1})^*$. Recall that a continuously differentiable function of one variable is convex on an interval if and only if the function lies above all of its tangents [9]. As $F(\overline{\mu}^{2,1}(p))$ is convex in p, the gap between the approximation $\overline{\mu}^{2,1}(p)$ and the exact solution $(\mu^{2,1})^*$ is smaller than $T \equiv \frac{\Delta(F)}{\frac{\partial F(\overline{\mu}^{2,1}(p))}{\partial \overline{\mu}^{2,1}(p)}} > 0$, where T is given in (30) the inequality follows because the enumerator $\Delta(F)$ is negative by (c) and the denominator is negative because $\lambda_1, \lambda, \overline{\mu}^{2,1}(p), p, 1-p$ are all positive, establishing the left inequality in (5).

Proof of Theorem 6

Part 1: this part is obtained by comparing $Cost^1(p)$ and $Cost^2(p)$, given by (23) and (24), respectively, and using the definition of $\hat{\lambda_1}$ in (33).

Part 2a: **Traveling costs:** the traveling cost of $Cost^1(p)$ is higher than the traveling cost of $Cost^2(p)$ if $[\lambda_1(1-p) \geq \lambda p(1-p)] \Leftrightarrow [p \leq \frac{\lambda_1}{\lambda}]$.

Part 2b: **Blocking costs:** substituting the optimal $\mu^{1,2}$ (16), the blocking cost of $Cost^1(p)$ is

$$\frac{C_B\lambda^2(1-p)}{\lambda + \mu^{1,2}} = \frac{C_B\lambda(1-p)}{\sqrt{\gamma}}$$

Substituting the approximations (29), the blocking cost of $Cost^2(p)$ is

$$\frac{C_B(\lambda_1)^2(1-p)^2}{\lambda_1 + (\lambda_1 + \lambda_2 p)(\sqrt{\gamma} - 1)} + \frac{C_B(\lambda_2)^2(1-p)^2}{\lambda_2 + (\lambda_2 + \lambda_1 p)(\sqrt{\gamma} - 1)} + \frac{C_B\lambda^2(1-p)p}{\lambda + (\lambda_1 + \lambda_2 p)(\sqrt{\gamma} - 1)} + \frac{C_B\lambda^2(1-p)p}{\lambda + (\lambda_2 + \lambda_1 p)(\sqrt{\gamma} - 1)}.$$

According to Theorem 7 (which its proof does not require the results of Theorem 6), the blocking cost of $Cost^2(p)$ is minimized when $\lambda_j = \frac{\lambda}{2}$.

In this case, the blocking cost of $Cost^2(p)$ is:

$$C_B(1-p)\lambda(\frac{1-p}{(1+p)\sqrt{\gamma} - p} + \frac{4p}{(1+p)\sqrt{\gamma} + 1 - p}).$$

We next show that this minimal blocking cost is higher than that in $Cost^1(p)$ (that is independent of the division of demand between the nodes), and so the blocking cost of $Cost^1(p)$ is always lower than the blocking cost of $Cost^2(p)$:

$$[\frac{C_B\lambda(1-p)}{\sqrt{\gamma}} \leq^? C_B(1-p)\lambda(\frac{1-p}{(1+p)\sqrt{\gamma}-p} + \frac{4p}{(1+p)\sqrt{\gamma}+1-p})]$$
$$\Leftrightarrow [\frac{1}{\sqrt{\gamma}} \leq^? \frac{1-p}{(1+p)\sqrt{\gamma}-p} + \frac{4p}{(1+p)\sqrt{\gamma}+1-p}]$$
$$\Leftrightarrow [((1+p)\sqrt{\gamma}+1-p)((1+p)\sqrt{\gamma}-p) \leq^? \sqrt{\gamma}((1+p)\sqrt{\gamma}+1-p)(1-p) + \sqrt{\gamma}((1+p)\sqrt{\gamma}-p)4p]$$
$$\Leftrightarrow [-2p(p+1)\gamma + p(p+1)\sqrt{\gamma} + p(p-1) \leq^? 0]$$
$$\Leftrightarrow [-2(p+1)\gamma + (p+1)\sqrt{\gamma} + (p-1) \leq^? 0]$$

Note that $-2(p+1)\gamma + (p+1)\sqrt{\gamma} + (p-1) \leq^? -2(p+1)\gamma + (p+1)\sqrt{\gamma} + (p+1)$, so we will show that $-2(p+1)\gamma + (p+1)\sqrt{\gamma} + (p+1) \leq 0$:

$$[-2(p+1)\gamma + (p+1)\sqrt{\gamma} + (p+1) \leq^? 0]$$
$$\Leftrightarrow [-2\gamma + \sqrt{\gamma} + 1 \leq^? 0],$$

which follows because $\gamma \geq 1$ by (25).

Service capacity costs: substituting the optimal $\mu^{1,2}$ (16), the service capacity cost of $Cost^1(p)$ is

$$C_\mu\lambda(\sqrt{\gamma}-1).$$

Substituting the approximations (29), the service capacity cost of $Cost^2(p)$ is

$$C_\mu\lambda(1+p)(\sqrt{\gamma}-1),$$

implying that the service capacity cost of $Cost^1(p)$ is lower than that of $Cost^2(p)$.

Finally, the setup cost of Part 2b, and Part 2c are immediate.

Proof of Theorem 7:

We will show that $Cost^2(p)$ is minimized when $\lambda_1 = \lambda_2 = \frac{\lambda}{2}$ using the first and second order conditions. In particular, we will show that when $\lambda_1 = \lambda_2 = 0.5\lambda$, and when the optimal capacities are inserted into $Cost^2(p)$, then the first derivative of $Cost^2(p)$ with respect to λ_1 is 0, and the second derivative of $Cost^2(p)$ with respect to λ_1 is positive.

As the division of λ into λ_1 and λ_2 affects only the blocking and service capacity costs of $Cost^2(p)$,

we will focus on them, and we will recall that the capacities are also a function of the demand rates, i.e., $\mu^{2,1}[\lambda_1]$ and $\mu^{2,2}[\lambda_2]$. We will denote λ_2 as $\lambda - \lambda_1$, and define:

$$\begin{aligned} f(\lambda_1, \mu^{2,1}[\lambda_1], \mu^{2,2}[\lambda - \lambda_1]) \quad \equiv \quad & (C_B(1-p)^2 \frac{(\lambda_1)^2}{\lambda_1 + \mu^{2,1}[\lambda_1]} + C_B(1-p)p\frac{\lambda^2}{\lambda + \mu^{2,1}[\lambda_1]} + C_\mu \mu^{2,1}[\lambda_1]) + \\ & (C_B(1-p)^2 \frac{(\lambda - \lambda_1)^2}{\lambda - \lambda_1 + \mu^{2,2}[\lambda - \lambda_1]} + C_B(1-p)p\frac{\lambda^2}{\lambda + \mu^{2,2}[\lambda - \lambda_1]} + \\ & C_\mu \mu^{2,2}[\lambda - \lambda_1]). \end{aligned} \tag{43}$$

Note that $f(\lambda_1, \mu^{2,1}[\lambda_1], \mu^{2,2}[\lambda - \lambda_1])$ is simply the blocking and service capacity costs of $Cost^2(p)$.

Let $(\mu^{2,1}[\lambda_1])^*$ and $(\mu^{2,2}[\lambda - \lambda_1])^*$ be the optimal capacities. Then, with the optimal capacities, we can denote $f(\lambda_1, (\mu^{2,1}[\lambda_1])^*, (\mu^{2,2}[\lambda - \lambda_1])^*)$ as a function of λ_1 only, and according to the Envelope Theorem [13], we can take the first and second derivatives of f with respect to λ_1:

$$\frac{\partial f(\lambda_1, (\mu^{2,1}[\lambda_1])^*, (\mu^{2,2}[\lambda - \lambda_1])^*)}{\partial \lambda_1} = \frac{\partial f(\lambda_1, \mu^{2,1}[\lambda_1], \mu^{2,2}[\lambda - \lambda_1])}{\partial \lambda_1},$$

and:

$$\frac{\partial^2 f(\lambda_1, (\mu^{2,1}[\lambda_1])^*, (\mu^{2,2}[\lambda - \lambda_1])^*)}{\partial \lambda_1^2} \leq \frac{\partial^2 f(\lambda_1, \mu^{2,1}[\lambda_1], \mu^{2,2}[\lambda - \lambda_1])}{\partial (\lambda_1)^2}.$$

So,

$$\frac{df(\lambda_1)}{d\lambda_1} = C_B(1-p)^2 (\frac{2\lambda_1}{\lambda_1 + \mu^{2,1}} - \frac{(\lambda_1)^2}{(\lambda_1 + \mu^{2,1})^2} - \frac{2(\lambda - \lambda_1)}{\lambda - \lambda_1 + \mu^{2,2}} + \frac{(\lambda - \lambda_1)^2}{(\lambda - \lambda_1 + \mu^{2,1})^2}). \tag{44}$$

Note that when $\lambda_1 = \frac{\lambda}{2}$, the capacities are equal to each other, i.e., $\mu^{2,1} = \mu^{2,2}$. Let $\mu \equiv \mu^{2,1} = \mu^{2,2}$. So, we get:

$$\frac{df(\lambda_1)}{d\lambda_1}_{\lambda_1 = \frac{\lambda}{2}} = 0. \tag{45}$$

Further,

$$\begin{aligned} \frac{d^2 f(\lambda_1)}{d\lambda_1^2} \quad = \quad & C_B(1-p)^2 (\frac{2(\lambda_1)^2}{(\lambda_1 + \mu^{2,1})^3} - \frac{4\lambda_1}{(\lambda_1 + \mu^{2,1})^2} + \frac{2}{\lambda_1 + \mu^{2,1}} \\ + \quad & \frac{2(\lambda - \lambda_1)^2}{(\lambda - \lambda_1 + \mu^{2,2})^3} - \frac{4(\lambda - \lambda_1)}{(\lambda - \lambda_1 + \mu^{2,2})^2} + \frac{2}{\lambda - \lambda_1 + \mu^{2,2}}) \\ = \quad & \frac{2(\mu^{2,1})^2}{(\lambda_1 + \mu^{2,1})^3} + \frac{2(\mu^{2,2})^2}{(\lambda_2 + \mu^{2,2})^3} > 0. \end{aligned} \tag{46}$$

So, $0 < \frac{d^2 f(\lambda_1)}{d\lambda_1^2} \leq \frac{\partial^2 f(\lambda_1, \mu^{2,1}[\lambda_1], \mu^{2,2}[\lambda - \lambda_1])}{\partial (\lambda_1)^2}$, implying that $f(\lambda_1)$ is convex in λ_1 and has a minimum point at $\lambda_1 = 0.5\lambda$.

Proof of Observation 2:

Using the method of Lagrange multipliers: $\Lambda(\{\lambda_j\}_{j\in\{1,\ldots,m\}},\alpha) = \sum_{j=1}^{m} f(\lambda_j)+\alpha(\sum_{j=1}^{m}\lambda_j-\lambda)$. Thus, we require that $\frac{\partial\Lambda(\{\lambda_j\}_{j\in J},\alpha)}{\partial\lambda_j} = f'(\lambda_j)+\alpha = 0\ \forall j\in J$, implying that the λ_j are equal. As $\sum_{j=1}^{m}\lambda_j = \lambda$, we conclude that $\lambda_j = \frac{\lambda}{m}\ \forall j\in\{1,\ldots,m\}$.

Proof of Lemma 1:

Given $m \geq 2$ $M/M/1/1$ facilities, and the approximations (38) for their capacities; we have m functions $f[\lambda_j]$. Each one is the sum of facility j's blocking and service capacity costs, and its form is:

$$f(\lambda_j) \equiv \sum_{l\in L} \frac{P_l C_B(\lambda_j(p)^l)^2}{\lambda_j(p)^l + \overline{\mu}^{m,j}(p)} + C_\mu \overline{\mu}^{m,j}(p),$$

with P_l the probability of failure event $l\in L$ and $\lambda_j(p)^l$ the demand rate faced by facility j at that event. Note that both $\lambda_j(p)^l$ and $\overline{\mu}^{m,j}(p)$ are positive and linear increasing in λ_j. Thus,

$$\frac{\partial^2 f[\lambda_j]}{\partial(\lambda_j)^2} \quad = \quad \sum_{l\in L} \frac{2C_B P_l(\overline{\mu}^{m,j}(p) - \lambda_j(p)^l \frac{\partial\overline{\mu}^{m,j}(p)}{\partial\lambda_j})^2}{(\lambda_j(p)^l + \overline{\mu}^{m,j}(p))^3} > 0,$$

and so $f[\lambda_j]$ are convex in λ_j, $j\in J$.

B Appendix

Algorithm 1

Input and Initialization: $G, \lambda_i\ \forall i \in I, C_B, C_\mu, C_K, c, K$. Calculate the shortest path matrix of G.
Procedure 0**: Initial guess**:
$m \leftarrow 0$, $Partial \leftarrow M$, where M is a very large number.
for $m = 1$ to $|I|$ **do** State Solve the relaxed-m-median problem.
  $Travel(m) \leftarrow$ traveling cost of the solution.
  if $Travel(m) + mC_K < Partial$ **then** $\hat{m} \leftarrow m$, $Partial \leftarrow Travel(m) + mC_K$.
  else stop Procedure 0.
  end if
end for
$m \leftarrow \hat{m}$.
Phase I**: Location problem**:
Solve the m-median problem.
$Travel(m) \leftarrow$ traveling cost of the solution.
$Partial(m) \leftarrow Travel(m) + mC_K$
Phase II**: Find** m^*:
(IIa) Repeat Phase I with $(m-1)$ replacing m.
if $Partial(m-1) < Partial(m)$ **then** $m \leftarrow (m-1)$, go to (IIa).
else(IIb) Repeat Phase I with $(m+1)$ replacing m.
  if $Partial(m+1) < Partial(m)$ **then** $m \leftarrow (m+1)$, go to (IIb).
  else stop Phase II.
  end if
end if
$m^* \leftarrow m$.
Phase III**: Optimal capacities**: Load $Partial(m^*)$ from Phase I.
for $j \in J$ **do** Evaluate $\lambda^{m^*,j}$.
end for
for $j \in J$ **do**
  $(\mu^{m^*,j})^* \leftarrow$ the real positive solution of $\frac{\partial(\sum_{j\in J} C_\mu \mu^{m^*,j} + \sum_{j\in J} C_B \lambda^{m^*,j} P_B(\lambda^{m^*,j}, \mu^{m^*,j}))}{\partial \mu^{m^*,j}} = 0$.
end for
for $j \in J$ **do**
  $Capacity(m^*) \leftarrow \sum_{j\in J} C_\mu (\mu^{m^*,j})^*$
  $Block(m^*) \leftarrow \sum_{j\in J} C_B \lambda^{m^*,j} P_B(\lambda^{m^*,j}, (\mu^{m^*,j})^*)$
end for
$Total(m^*) \leftarrow Partial(m^*) + Block(m^*) + Capacity(m^*)$.
Output $m^*, Total(m^*)$, and $(\mu^{m^*,j})^*\ \forall j \in J$.

Algorithm 2

Input and Initialization: $G, \lambda_i\ \forall i \in I, C_B, C_\mu, C_K, c, K, p$. Calculate the shortest path matrix of G.

Procedure 0**: Initial guess**:

for $m = 1$ to $|I|$ **do**

Solve the relaxed m-median problem.

$Travel(m) \Leftarrow$ traveling cost of the solution.

$Travel(m)_{LOW} \Leftarrow \sum_{h=0}^{m-1} Travel(h)\binom{m}{h}p^h(1-p)^{m-h}$.

$Block(m)_{LOW}(m, \mu^m) \Leftarrow mC_B \sum_{h=0}^{m-1} \binom{m}{h}p^h(1-p)^{m-h}\frac{\lambda}{m-h}P_B(\frac{\lambda}{m-h}, \mu^{m-h})$.

$Capacity(m, \mu^m) \Leftarrow mC_\mu\mu^m$.

$(\mu^m)^* \leftarrow$ the real positive solution of $\frac{\partial(Block(m)_{LOW}+Capacity(m,\mu^m))}{\partial\mu^m} = 0$.

$Setup(m) \Leftarrow mC_K$.

$Failure(m) \Leftarrow C_B\lambda p^m$.

$Total(m) \Leftarrow Travel(m) + Block_{LOW}(m, (\mu^m)^*) + Capacity(m, (\mu^m)^*) + Setup(m) + Failure(m)$.

end for

$\hat{m} \Leftarrow \arg\min_{m\in I} Total(m)$.

$m \leftarrow \hat{m}$.

Procedure I**:** m**-median**

(a) Solve the m-median problem.

for $i = 1$ to $|I|$ **do**

Generate sorted list L_i, of all m facilities by non decreasing distance from i

end for

$Travel(m) \Leftarrow \sum_{i\in I} \lambda_i d_{iL_i^h} p^{h-1}(1-p)$

$Block_{PARTIAL}(m, \mu^{m,j}) \Leftarrow \underline{\Pi^3}$

$Capacity(m, \mu^{m,j}) \Leftarrow mC_\mu\mu^{m,j}$.

for $i = 1$ to m **do**

$(\mu^{m,j})^* \leftarrow$ the real positive solution of $\frac{\partial(Block_{PARTIAL}(m,\mu^{m,j})+Capacity(m,\mu^m))}{\partial\mu^{m,j}} = 0$.

end for

Evaluate $Block_{PARTIAL}(m, (\mu^{m,j})^*)), Capacity(m, (\mu^{m,j})^*)$

$\mu^{\min} \leftarrow \arg\min_{i=1}^{m}(\mu^{m,j})^*$

$Block_{ADD} \Leftarrow$ blocking cost of a facility with a demand rate λ and a service rate $\mu^{\min}$.

$Probability_{ADD} \Leftarrow$ probability of $4, 5, \ldots, m-1$ facilities failure events.

$Block(m, (\mu^{m,j})^*) \Leftarrow Block_{PARTIAL}(m, \mu^{m,j})^* Probability_{ADD} Block_{ADD}$

(b) $Total(m) \Leftarrow Travel(m) + Block(m, (\mu^{m,j})^*) + Capacity(m, (\mu^{m,j})^*) + mC_K + C_B\lambda p^m$.

$h \Leftarrow m$. Repeat (a)-(b) with $h-1$ replacing m.

if $Total(h-1) < Total(m)$ **then** let $h \Leftarrow h-1$. Repeat (a)-(b) with $h-1$ replacing m.

else let $h \equiv m$. Repeat (a)-(b) with $h+1$ replacing m.

if $Total(h+1) < Total(m)$ **then** let $h \equiv h+1$. Repeat (a)-(b) with $h+1$ replacing m.

else Stop.

end if

end if

Output $m \Leftarrow \arg\min Total(m)$.

Procedure II**: Fair-**m**-median**

Repeat Procedure I, but use the fair-m-median solution at (a).

C Appendix

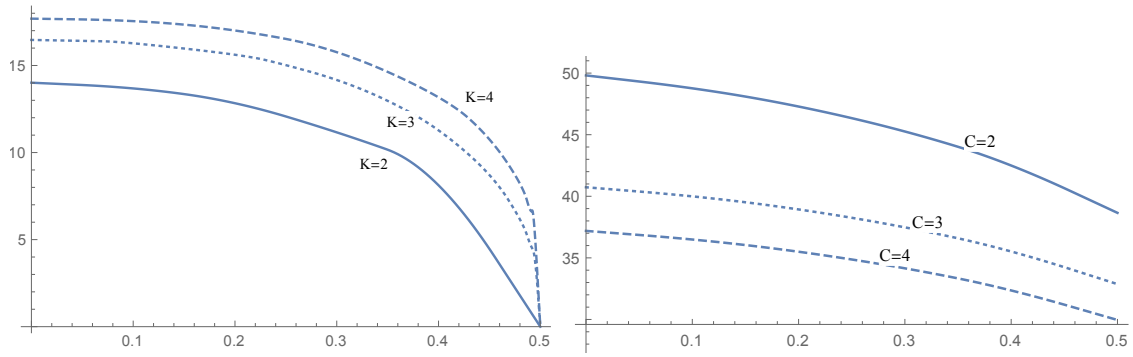
$M/M/c/K$ facilities

We numerically consider facilities that are modeled as $M/M/c/K$ queues with $1 \leq c \leq K$ (because the blocking probability formula, given by (1-2), is less elegant, and so are the analytical results). We provide some concrete examples that show similarities and differences from the results obtained for the $M/M/1/1$ facilities.

The following example shows that the optimal capacities increase with K, and decrease with c.

Example 5. ***Optimal capacities increase with*** K ***and decrease with*** c***:*** *Figure 7a presents* $(\mu^{2,1})^*$ *as a function of* p*, for* $c = 1$ *and* $K = 2, 3, 4$*, and Figure 7b presents* $(\mu^{2,1})^*$ *for* $K = 4$*, and* $c = 2, 3, 4$ *as a function of* p*.*

It is clear that the optimal capacities increase with K*, and decrease with* c*. This is also very intuitive: when* K *increases, more customers are allowed at the system. So, the (single) server may need to serve more customers, and thus its optimal service capacity should increase. On the other hand, when* c *decreases, more servers serve the same number of customers. Then, each server needs to serve less customers, and thus its optimal service capacity should decrease.*



(a) $(\mu^{2,1})^*$ for $c = 1$ and $K = 2, 3, 4$ (b) $(\mu^{2,1})^*$ for $K = 4$ and $c = 2, 3, 4$

Figure 7: $(\mu^{2,1})^*$ as a function of c and K

The following example is Example 2 for $M/M/c/K$ facilities.

Example 6. ***Example 2 revisited*** *Let* $C_K = 0, c = 2$*, and* $K = 3$*. Figure 8 presents the blocking and total costs of* $Cost^1(p)$ *and* $Cost^2(p)$*. The costs elements of* $Cost^1(p)$ *are drawn by a solid line, and the ones of* $Cost^2(p)$ *are drawn by a dashed line.*

At Example 2, the blocking cost of $Cost^1(p)$ *is decreasing as a function of* p*, the blocking cost of* $Cost^2(p)$ *is concave as a function of* p*, and it is higher than the blocking cost of* $Cost^1(p)$*. Here, the*

blocking cost of $Cost^1(p)$ *is close to a straight line, almost parallel to the p-axis, the blocking cost of* $Cost^2(p)$ *is concave as a function of p, and it is still higher than the blocking cost of* $Cost^1(p)$*, as proved in Theorem 6 for the* $M/M/1/1$ *case.*

The differences between the two upper curves are the C_K^**. The highest* C_K^* *is around* 25 *for* $p \in [0.2, 0.3]$*, and the lowest* C_K^* *is around* 12 *for* $p \approx 0.5$*. That is,* C_K^**s are not necessarily higher than* λ_1*.*

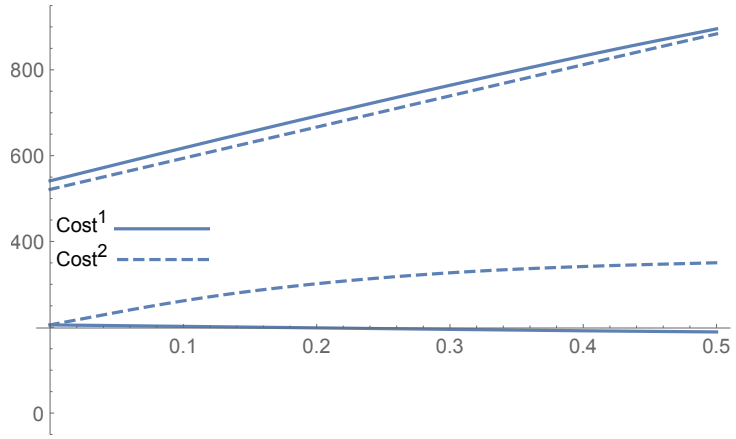


Figure 8: Blocking (the lower curves) and total costs of $Cost^1(p)$ and $Cost^2(p)$