CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Competitive Facility Location with Selfish Users and Queues

Teodora Dan
Patrice Marcotte

July 2017

CIRRELT-2017-46

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

Competitive Facility Location with Selfish Users and Queues

Teodora Dan*, Patrice Marcotte

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Computer Science and Operations Research, Université de Montréal, P.O. Box 6128, Station Centre-ville, Montréal, Canada H3C 3J7

Abstract. In a competitive environment, we consider the problem faced by a service firm that makes decisions with respect to both the location and service levels of its facilities, taking into account that users patronize the facility that maximizes their individual utility, expressed as the sum of travel time, queueing delay, and a random term. This situation can be modeled as a bilevel program that involves discrete and continuous variables, as well as linear and nonlinear functions. We design for its solution an approximation algorithm that provides "quasi-optimal" solutions, as well as heuristics that exploit the very structure of the problem.

Keywords: Location, bilevel programming, equilibrium, queueing, non-convex.

Acknowledgement. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 14997.

---

* Corresponding author: Teodora.Dan@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2017

1 Introduction

1.1 Contribution of this paper

While the literature concerning discrete facility location is vast, few studies have focused on user choice, where the latter frequently involves congestion, either along the paths leading to a facility, or at the facility itself. The aim of this paper is to provide a model that captures the key features of congestion and competition within a user choice environment, yielding a bilevel program where the leader firm's objective function integrates the stochastic equilibrium resulting from the choice of locations and the associated service levels. Beyond the analysis of the model's theoretical properties, the main part of the paper is devoted to the design and analysis of efficient algorithms, whose nature is either based on approximations ('semi-exact') or heuristic.

Our work is closely related to that of [Marianov et al., 2008], who analyze a location model where queueing (and balking) is explicitly taken into account, while users are assigned to facilities according to a logit discrete choice model, yielding a mathematical program involving user-equilibrium constraints. The model is well suited to a variety of applications, such as location of shops, restaurants, walk-in clinics, etc., where user flows are not in direct control of the optimizer, but are dictated by utility maximization principles. One aim of this paper is to extend and improve the model, both from the modelling and algorithmic standpoints. Its main contributions are the following:

- The introduction of service rate as endogenous variables, as well as the correct modelling of the balking process, by integrating within a user's utility the probability of service denial.
- The explicit consideration of competition.
- The embedding of a discrete choice model of user behaviour, as well as the study of the deterministic (Wardrop) limiting case.
- The reformulation of the model as a standard bilevel model, thus allowing an approximate reformulation as a mixed integer linear program.
- The design of a heuristic algorithm and its validation against the MILP solution.

The remainder of this paper is organized as follows. Section 1.2 is devoted to the literature review, and Section 2 to a description of the model, together with a study of its theoretical properties. Section 3 is dedicated to algorithms: a linear approximation algorithm in Subsection 3.1, and a user-driven heuristic in Subsection 3.2. Numerical experiments, discussion of our results, as well as an illustrative case are detailed in Section 4. Extensions of the current framework are mentioned in the concluding Section 5.

1.2 Literature review

Location problems have been widely studied, due to their simple structure and numerous real-life applications. Most literature is concerned with versions of the problem where users

are simply assigned to shortest paths, and thus sidesteps the nonlinearities associated with the important issue of user behaviour, including congestion. In our model, customers select their own path and whenever congestion occurs, customers leaving from the same origin may travel along different paths or patronize different facilities. This user behavior principle fits the framework of a Wardrop equilibrium in the deterministic case, and of stochastic user equilibrium when a random utility model of delay is assumed. The overallbilevel model belongs to the class of mathematical programs with equilibrium constraints (MPEC), where the equilibrium can be expressed as a variational inequality. It can be reformulated as an NP-hard discrete nonlinear bilevel program which, it goes without saying, poses formidable challenges from the computational point of view.

Competitive location models were introduced by [Hotelling, 1929]. In his seminal paper, the author addresses the simple situation where two firms engage in spatial competition, with the purpose of maximizing individual profit through the location of a point along a segment located at respective distances a and b from the endpoints. It is assumed that demand is uniformly distributed along the line segment, and customers patronize the closest facility. This work represents the cornerstone for a plethora of articles concerned with the topic of competitive facility location. The environment considered therein was generalized to a network by [Hakimi, 1983], who studied variants of the weighted p-median problem involving competition. [Labbé and Hakimi, 1991] address a two-stage location-allocation game, where location is decided at the first stage while, at the second stage, two firms engage in a Cournot game with respect to quantities. An interesting development is considered by [Küçükaydin et al., 2011], where one firm decides the sites and attractiveness for new facilities in order to maximize its profit. In this Stackelberg (leader-follower) setting, the competitor responds to the leader's action and adjusts its attractiveness level to maximize its profit, while user behavior is characterized by Huff's gravity law. In the work of [Beresnev, 2013], two competing firms strive to maximize profit as well, but user preferences are provided by a linear order relation. The model is then solved by branch-and-bound techniques. [Drezner et al., 2015] address a leader-follower competitive coverage model, where the attractiveness of a facility is related to an attraction radius, and customers are spread evenly among facilities that fall within this radius. The leader can open new facilities or adjust the attractiveness of existing ones, while the competitor responds accordingly. Both firms compete for market share within budget limits.

Congestion occurs naturally in an environment with limited resources. It can arise either at facilities, or along the road. Although basic models are content to incorporate congestion in the form of maximum capacity, more elaborate models capture congestion through functional forms derived or not from queueing theory. Within this framework we note the work of [Desrochers et al., 1995] who consider an extension of a deterministic facility location problem, where individual delays (travel time) increase with traffic. The model is centralized, namely, users are assigned as to minimize the sum of opening cost, waiting delays, and travel times experienced by the users. Although the authors mention a user-choice version of their model that fits the bilevel programming paradigm, they do not suggest solution algorithms for its solution. A related formulation, where service rates are endogenous, is considered by [Castillo et al., 2009]. Users are assigned to facilities as to minimize the sum of the number of waiting customers and the total opening and service costs. Within the framework

of centralized systems, [Marianov, 2003] formulates a model for locating facilities subject to congestion where demand is elastic with respect to travel time and queue length. Customers are assigned to centers in order to maximize total demand. Location of congested facilities when demand is elastic has also been investigated by [Berman and Drezner, 2006]. Similar to [Marianov, 2003], the objective is to maximize total demand, subject to constraints on the waiting time at facilities. Heuristic procedures are proposed for its solution.

Another work worth mentioning is that of [Zhang et al., 2010] who propose a methodology for addressing a congested facility network design, with the aim of improving healthcare accessibility, i.e., maximize the participation rate. The environment is user-choice, and users patronize the facility minimizing the sum of waiting and travel time, while demand is elastic with respect to total expected time experienced by clients. The authors illustrate the performance of a metaheuristic procedure on data issued from a network of mammography centers in Montreal, Canada. Congestion has also been considered by [Abouee-Mehrizi et al., 2011] in the context of simultaneous decision-making over the location, service rate and price, for facilities located on vertices of a network. They assumed that demand be elastic with respect to price, and clients spread among facilities based on proximity only, according to a multinomial logit random utility model. Congestion, which arises at facilities, is characterized by queueing equations. For a more elaborate review of congestion models in the context of facility location, the reader is referred to [Boffey et al., 2007].

Although congestion and competition have been previously combined, few papers have tackled both within a user-choice environment. Actually, most papers that incorporate congestion do not account for competition. On the other hand, when competition is present, users select facilities based on congestion-free traits such as distance or attractiveness. To the best of our knowledge, the only paper to address congestion in a competitive user-choice environment is that of [Marianov et al., 2008]. A taxonomy of the models most relevant to our research is provided in the e-companion to this article.

2 The model

2.1 Preliminaries

Let us consider the problem faced by a firm (a service center, for instance) that makes location and service level decisions, with the aim of maximizing the number of customers to attract with respect to its competitors, under a budget constraint. A salient feature of the model is that user behavior is explicitly taken into account. Precisely, users patronize the facility that maximizes their individual utility, i.e., minimizes their disutility. The latter is estimated as the sum of travel time to the facility, queueing at the facility, plus the actual probability of accessing a server (facilities are modeled as finite-length queues).

Since our model is closely related to that of [Marianov et al., 2008], we provide a detailed description of the latter. In that work, the authors consider an oligopoly scenario in which firm A locates p new facilities in a market where competitors already operate. The 'game' takes place over a bipartite graph $V = I \times J$, where a vertex v may correspond to either a location ($v \in J$) or a demand node ($v \in I$), the latter endowed with demand d_v. We denote by $J_1 \subseteq J$ the set of candidate locations for firm A, and by J_c the set of locations of its

competitors. A customer leaving vertex $i \in I$ for facility $j \in J$ incurs a fixed travel time t_{ij}. At facility j, this customer enters an $M/M/s/K$ queue that involves s servers with identical mean service time μ, and an associated waiting time w_j. Whenever the queue reaches length $K - s$ (which corresponds to K customers in the system), any arriving customer is denied access and leaves the system as a lost customer. The disutility u_{ij} of a customer is defined as a linear (convex) combination of travel time t_{ij} and queueing delay w_j, and ignores the actual constant service time, i.e.,

$$u_{ij} = \alpha t_{ij} + (1 - \alpha) w_j, \tag{1}$$

for some scalar α between 0 and 1.

The arrival and service processes are governed by Poisson (memoryless) processes. If the arrival rate at facility j is λ_j, the probability that n customers are in the queue (or are served) is

$$p_{nj} = \begin{cases} (\rho_j^n / n!) p_{0j} & \text{if } n \leq s, \\ (\rho_j^n / (s! s^n - s)) p_{0j} & \text{if } s < n \leq K, \\ 0 & \text{if } n > K, \end{cases} \tag{2}$$

where $\rho_j = \lambda_j / \mu$ is the intensity of the queueing process and

$$p_{0j} = \left[1 + \sum_{n=1}^{s} \frac{\rho_j^n}{n!} + \frac{\rho_j^s}{s!} \sum_{n=s+1}^{K} \left(\frac{\rho_j}{s} \right)^{n-s} \right]^{-1}. \tag{3}$$

The demand side is cast within the framework of a random utility model, where flows between vertices i and j are determined according to the logit formula

$$x_{ij} = \frac{y_j e^{-\theta u_{ij}}}{\sum_{k \in J_1} y_k e^{-\theta u_{ik}} + \sum_{k \in J_c} e^{-\theta u_{ik}}}, \tag{4}$$

where y_j is a binary variable set to 1 if a facility is open at vertex $j \in J_1$, and to 0 otherwise. Competitor's facilities are already open, thus the absence of term y_k in Eq. (4). Parameter θ is set to $\pi / (\sigma \sqrt{6})$, where σ is the standard deviation of the Gumbel randon variable yielding the probabilities (or proportions) x_{ij}. The y variables are needed solely for the leader, as the competitors facilities are already open. If one denotes by λ_j the arrival rate at node j, and by $\overline{\lambda}_j$ the throughput rate, the model of [Marianov et al., 2008] takes the form of the

mathematical program

$$
\begin{aligned}
\max_{y} \quad & \sum_{j\in J_1} \lambda_j \\
\text{s.t.} \quad & \lambda_j = \sum_{i\in I} d_i x_{ij}, && \forall j \in J_1 \cup J_c \\
& x_{ij} = \frac{y_j e^{-\theta u_{ij}}}{\sum_{k\in J_1} y_k e^{-\theta u_{ik}} + \sum_{k\in J_c} e^{-\theta u_{ik}}}, && \forall i \in I, \forall j \in J_1 \cup J_c \\
& u_{ij} = \alpha t_{ij} + (1-\alpha) w_j, && \forall i \in I, \forall j \in J \\
& w_j = L_j / \overline{\lambda}_j, && \forall j \in J \\
& L_j = \sum_{n=s}^{K} (n-s) p_{nj}, && \forall j \in J \\
& \overline{\lambda}_j = \lambda_j (1 - p_{Kj}), && \forall j \in J \\
& x_{ij} \leq y_j, && \forall i \in I, \forall j \in J_1 \\
& \sum_{j\in J} x_{ij} = 1, && \forall i \in I \\
& \sum_{j\in J_1} y_j = p, \\
& 0 \leq x_{ij} \leq 1, && \forall i \in I, \forall j \in J \\
& \lambda_j \geq 0, && \forall j \in J \\
& y_j \in \{0,1\}, && \forall j \in J_1,
\end{aligned}
$$

where the only decision variables are the binary location variables y_j. Once these are set, the remaining quantities are determined through the solution of a nonlinear fixed point problem, where the probabilities x_{ij} of choosing a facility j depend on waiting times, which are themselves functions of the demand rate vector λ, while demand rates depend on the probabilities x_{ij}. This yields a mathematical program with equilibrium constraints that can be formulated in the compact form

$$
\begin{aligned}
\max_{y} \quad & \sum_{j\in J_1} f_j(\lambda, y) \\
\text{s.t.} \quad & \sum_{j\in J_1} y_j = p, \\
& y_j \in \{0,1\}, && \forall j \in J_1,
\end{aligned}
$$

where the arrival rate vector λ satisfies the fixed point equation $f_j(\lambda, y) = \lambda_j, \forall j \in J$. The authors show that this equation admits a unique solution, and propose a variant of Newton-Raphson algorithm for its determination. The model is then addressed by a two-phase metaheuristic procedure that combines GRASP (Greedy Randomized Adaptive Search Procedure) and Tabu Search. In the initial phase, facility locations are selected and a nonlinear assign-

ment problem is solved. In the second phase, Tabu Search is used to improve upon the initial location decisions.

A key feature of the model is the possible occurrence of balking, due to a fixed buffer of size $K - s$. Besides its practical important, balking allows the arrival rate at a facility to actually exceed the service rate, without the queues growing unbounded. However, this has two important consequences. First, note that the objective is to maximize the number of clients $\sum_{j \in J_1} \lambda_j$ *showing up* at the facilities and not the number of clients $\sum_{j \in J_1} \bar{\lambda}_j$ that *actually* access service. It follows that a solution with a low rate of served clients might be preferred to one with a high rate, if both its arrival and rejection rates are very high. This situation is illustrated in Figure 1. In this example, facilities can be set up at three sites (A,B and D), coinciding with two demand vertices. The competitor's facility is located at C. Demand d_1 is 200 at vertex 1 and $d_2 = 10$ at vertex 2, while distances between vertices are shown next to the edges of the network. On the supply side, the common service rate at all facilities is equal to 100. Facilities are modelled as $M/M/1/99$ queues. For simplicity, we assume $\theta = \infty$, the limiting case of the random utility model. Accordingly, at equilibrium, clients issued from a common origin will experience identical delays (travel time plus queueing delay).

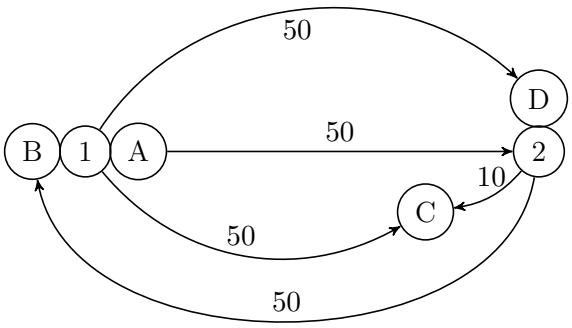

Figure 1: Paradox when maximizing λ instead of $\bar{\lambda}$.

Assuming that the leader's budget only allows two facilities to be opened, the options are to open sites A and B, or sites A and D (B–D is equivalent to A–D). In the first case, demand d_1 is assigned to sites A and B, while d_2 patronize the competitor's facility. Basic arithmetic shows that the total arrival rate at the leader's facilities is $\lambda = \lambda_1 + \lambda_2 = 200$, and that the number being serviced is $\bar{\lambda} = \bar{\lambda}_1 + \bar{\lambda}_2 = 198$. If facilities are opened at sites A and D, d_1 is assigned to site A, and d_2 to site D, with no client assigned to the competitor. The total arrival rate at the leader's facilities is $\lambda = \lambda_1 + \lambda_2 = 210$ and the amount of customers receiving service is $\bar{\lambda} = \bar{\lambda}_1 + \bar{\lambda}_2 = 101$. In either case, the maximum λ corresponds to a much smaller value of $\bar{\lambda}$. In other words, the solution that attracts more customers is less profitable, as roughly half of the clients will balk, due to no vacancies in the queue, and thus experience low delays at the facility.

The issue is also due to the definition of customer utility, which embeds travel and queueing delays, but ignores balking. Returning to the example of Figure 1, when sites A and

D are open, demand d_1 originating in 1 patronizes site A, notwithstanding a probability of balking close to 50%. This situation is not realistic, given that facilities located at site D and C are relatively close and have low waiting times and probability of rejection. Since the queueing delay is directly related to the buffer capacity $K-1$, facilities with small buffers (or none at all!) will turn down most arriving customers, in contrast with facilities equipped with large buffer zones. This leads to the paradoxical situation where customers will favour facilities where the probability of balking is high, since it will minimize the overall time spent in the system! This effect is exacerbated by the maximization of the arrival rate (rather than the throughput rate) and will only disappear if buffers have infinite capacities.

2.2 A new model

We now consider a variant of the model of [Marianov et al., 2008] that differs in three significant ways: the objective is the throughput rate (rather than the arrival rate), service rates are decision variables, and users integrate within their utility function the probability of accessing the service. Additionally, the leader has a limited budget B that can be spent on building facilities or improving service rate. The fixed cost of locating a new facility f is set to c_f, while the cost of improving the service rate of an $M/M/1/K$ queue ($K-1$ available places in the queue, and 1 place at the server) by one unit is c_μ. A customer observes the queue upon arrival, and opts for balking if there are more than $K-1$ customers already waiting. In this context, the probability p_{nj} of having n customers in the queue (or being served) at facility j is given by

$$p_{nj} = \begin{cases} \rho_j^n \dfrac{1-\rho_j}{1-\rho_j^{K+1}}, & n \leq K,\ \rho_j \neq 1 \\ \dfrac{\rho_j}{K+1}, & n \leq K,\ \rho_j = 1 \\ 0, & n > K, \end{cases} \tag{5}$$

where $\rho_j = \lambda_j/\mu_j$ is the intensity of the process. At facility j, the expected number L_j of customers in the system is

$$L_j = \sum_{n=0}^{K} n p_{nj}. \tag{6}$$

The effective arrival rate, i.e., the number of customers that access the service, is denoted by $\overline{\lambda}$, i.e.,

$$\overline{\lambda}_j = \lambda_j(1 - p_{Kj}), \quad \forall j \in J. \tag{7}$$

The average waiting time w_j in the system (including service time) is a function of the service and arrival rates. According to Little's formula, we have that

$$w_j = \frac{L_j}{\overline{\lambda}_j}, \quad \forall j \in J. \tag{8}$$

Basic algebra yields the expression of the waiting time at open facilities:

$$w_j(\lambda_j, \mu_j) = \begin{cases} \dfrac{1}{\mu_j}\left(K + \dfrac{K}{\rho_j^K - 1} - \dfrac{1}{\rho_j - 1}\right), & \rho_j \neq 1 \\ \dfrac{K+1}{2\mu_j}, & \rho_j = 1. \end{cases} \tag{9}$$

2.2.1 Stochastic assignment

In a random utility model, clients patronize the facility that minimizes their individual disutility, expressed as a linear combination of travel time, queueing, and probability of accessing service. In this framework, the disutility of facility j for a customer issued from demand node i is given by

$$\begin{aligned} \tilde{u}_{ij} &= -u_{ij} + \varepsilon_{ij} \\ &= -(t_{ij} + \alpha w_j + \beta p_{Kj}) + \varepsilon_{ij}, \end{aligned}$$

where ε_{ij} are independant Gumbel variates with comon scale parameter θ and variance $\pi^2/(6{\cdot}\theta^2)$. In this multinomial logit framework (see [McFadden, 1974]), the demand generated at node i that patronizse an open facility j is given by the expression

$$x_{ij} = d_i \frac{e^{-\theta\,(t_{ij} + \alpha w_j + \beta p_{Kj})}}{\sum_{l \in J^*} e^{-\theta\,(t_{il} + \alpha w_l + \beta p_{Kl})}}, \tag{10}$$

where J^* represents the set of open facilities. For small values of θ, users are spread more or less evenly between facilities while, when θ is large, the assignment approaches that of a Wardrop equilibrium (see [Fisk, 1980]). According to our assumptions, the problem can be formulated as the equilibrium-constrained nonlinear mixed integer program involving a

leader and a follower (users):

$$\text{(P) LEADER:} \quad \max_{y,\mu} \quad \sum_{j\in J_1} \overline{\lambda}_j \tag{11}$$

$$\sum_{j\in J_1} c_f y_j + \sum_{j\in J_1} c_\mu \mu_j \le B, \tag{12}$$

$$\mu_j \le M y_j, \qquad \forall j \in J_1 \tag{13}$$

$$\overline{\lambda}_j = \lambda_j (1 - p_{Kj}), \qquad \forall j \in J \tag{14}$$

$$y_j \in \{0,1\}, \mu_j \ge 0, \qquad \forall j \in J_1 \tag{15}$$

$$\text{USERS:} \quad x_{ij} = d_i \frac{y_j \cdot e^{-\theta\,(t_{ij} + \alpha w_j + \beta p_{Kj})}}{\sum_{l\in J^*} e^{-\theta\,(t_{il} + \alpha w_l + \beta p_{Kl})}}, \qquad \forall i \in I;\ \forall j \in J \tag{16}$$

$$\lambda_j = \sum_{i\in I} x_{ij}, \qquad \forall j \in J \tag{17}$$

$$w_j = \frac{1}{\mu_j}\left(K + \frac{K}{\rho_j^K - 1} - \frac{1}{\rho_j - 1}\right), \qquad \forall j \in J \tag{18}$$

$$p_{Kj} = \rho_j^K \frac{1-\rho_j}{1-\rho_j^K + 1}, \qquad \forall j \in J. \tag{19}$$

The decision variables are the vectors μ and y, while the user assignment x is the solution of a fixed point problem. In Eq. (13), M is a sufficiently large constant that can be set to $M = (B - c_f)/c_\mu$.

The limiting case $\theta = \infty$ yields a deterministic version of (P) where customers are assigned to facilities according to Wardrop's equilibrium principle. If $c_i(\mu)$ denotes the minimum disutility (travel + waiting time and probability of balking) for users originating from node i, the optimal solution x^* isthen characterized by the complementarity system

$$t_{ij} + \alpha w_j(x^*,\mu) + \beta p_{Kj}(x^*,\mu) \begin{cases} = c_i(\mu), & \text{if } x^*_{ij} > 0 \\ \ge c_i(\mu), & \text{if } x^*_{ij} = 0, \end{cases} \tag{20}$$

and the deterministic version of (P) takes the form

$$\text{(P*) LEADER:} \quad \max_{y,\mu} \quad \sum_{j\in J_1} \overline{\lambda}_j \tag{21}$$

$$\text{s.t.} \quad \text{constraints (12), (13), (14) and (15)} \tag{22}$$

$$\text{USERS:} \quad t_{ij} + \alpha w_j(x^*,\mu) + \beta p_{Kj}(x^*,\mu) - c_i(\mu) \ge 0, \qquad \forall i \in I;\ \forall j \in J \tag{23}$$

$$x_{ij}\left(t_{ij} + \alpha w_j(x^*,\mu) + \beta p_{Kj}(x^*,\mu) - c_i(\mu)\right) = 0, \qquad \forall i \in I;\ \forall j \in J \tag{24}$$

$$x_{ij} \ge 0, \qquad \forall i \in I;\ \forall j \in J \tag{25}$$

$$\text{constraints (17), (18), (19).} \tag{26}$$

In (P), the solution of the lower level equilibrium problem can be obtained by solving a convex optimization problem akin to [Fisk, 1980]. In our framework, this program takes the

form

$$\text{(P2)} \quad \min_{x} \quad \sum_{i\in I}\sum_{j\in J^*}\left[\frac{1}{\theta}x_{ij}\ln x_{ij} + x_{ij}t_{ij}\right] + \alpha\sum_{j\in J^*}\int_0^{\lambda_j} w_j(q,\mu_j)dq + \beta\sum_{j\in J^*}\int_0^{\lambda_j} p_{Kj}(q,\mu_j)dq \tag{27}$$

$$\text{s.t.} \quad \sum_{j\in J^*} x_{ij} = d_i, \qquad \forall i \in I \tag{28}$$

$$x_{ij} \geq 0, \qquad \forall i \in I; \forall j \in J^* \tag{29}$$

$$\lambda_j = \sum_{i\in I} x_{ij}, \qquad \forall j \in J^*. \tag{30}$$

Indeed, it is easy to check that, if θ is finite, x_{ij} cannot be zero at the solution, which implies that the Lagrange multiplier associated with Eq. (29) is 0, thus useless. If we let a_i, and c_j be the Lagrange multipliers associated with Equations (28) and (30), respectively, the first-order necessary and sufficient optimality conditions are given by

$$\frac{\partial L}{\partial x_{ij}} = 0 \;\Rightarrow\; \frac{1}{\theta}\left(\ln x_{ij} + 1\ \right) + t_{ij} - a_i + c_j = 0 \tag{31}$$

$$\frac{\partial L}{\partial \lambda_j} = 0 \;\Rightarrow\; \alpha w_j(\lambda_j,\mu_j) + \beta p_{Kj}(\lambda_j,\mu_j) - c_j = 0. \tag{32}$$

It follows that $c_j = \alpha w_j(\lambda_j,\mu_j) + \beta p_{Kj}(\lambda_j,\mu_j)$, and Equation (31) yields

$$x_{ij} = \frac{e^{-\theta u_{ij}}}{e^{-\theta a_i}+1}.$$

By substituting x_{ij} into (28) we obtain

$$x_{ij} = d_i\frac{e^{-\theta u_{ij}}}{\sum_{l\in J^*} e^{-\theta u_{il}}}.$$

Now, replacing the fixed point problem by its optimization counterpart, the original model can be formulated as a bilevel program. At the upper level, the firm maximizes total market capture, subject to some budget constraints, while, at the lower level, the follower solves Problem (P2). The main advantage of this reformulation is that we can adapt for its solution methods and algorithms from convex bilevel programming, as we will detail further in Section 3.

2.3 Properties of the model

This subsection is devoted to the properties and features of our model. First, let us consider the indefinite integrals of the waiting time and probability of balking, $W_j(q,\mu_j)$ and

$P_{Kj}(q,\mu_j)$ respectively, that enter the lower level's objective function. We have

$$W_j(q,\mu_j) = \int w_j(q,\mu_j)dq = \begin{cases} \int \frac{1}{\mu_j}\left(K + \frac{K}{\rho_j^K - 1} - \frac{1}{\rho_j - 1}\right) d\lambda, & \text{if } q \neq \mu_j \\ \int \frac{K+1}{2\mu_j} dq, & \text{if } q = \mu_j. \end{cases}$$

$$P_{Kj}(q,\mu_j) = \int p_{Kj}(q,\mu_j)dq = \begin{cases} \int \frac{\rho^K - \rho^{K+1}}{1-\rho^{K+1}} dq, & \text{if } q \neq \mu_j \\ \int \frac{1}{K+1} dq, & \text{if } q = \mu_j. \end{cases}$$

where $\rho_j = q/\mu_j$. Let $l_w = \frac{1}{\mu_j}\int \frac{-1}{\rho_j - 1}dq$, and $l_p = \int \frac{\rho_j^K}{1-\rho_j^{K+1}}dq$. Then

$$l_w = \begin{cases} -\ln(\rho_j - 1), & \text{if } q > \mu_j \\ -\ln(1-\rho_j), & \text{if } q < \mu_j \end{cases} \quad \text{and} \quad l_p = \begin{cases} -\frac{\ln(\rho^{K+1} - 1)}{K+1}\mu_j, & \text{if } q > \mu_j \\ -\frac{\ln(1-\rho^{K+1})}{K+1}\mu_j, & \text{if } q < \mu_j, \end{cases} \tag{33}$$

which yields the following expression for the integral of the waiting time:

$$W_j(q,\mu_j) = \begin{cases} K\rho + l_w + K\int \frac{1}{\rho^K - 1}d\rho, & \text{if } q \neq \mu_j \\ \frac{K+1}{2}\rho, & \text{if } q = \mu_j \end{cases} \tag{34}$$

and for the integral of the balking probability:

$$P_{Kj}(q,\mu_j) = \begin{cases} q + l_p + \mu_j \int \frac{1}{\rho^K - 1}d\rho, & \text{if } q \neq \mu_j \\ \frac{q}{K+1}, & \text{if } q = \mu_j. \end{cases} \tag{35}$$

Note that $\int \frac{1}{\rho^K - 1}d\rho = -\rho F_1^2(1, 1/K; 1 + 1/K; \rho^K)$, where F_1^2 stands for the hypergeometric function, and does not have a closed-form expression for general K, although it can be evaluated for any fixed value of K. We have that

$$\int_0^{\lambda_j} w_j(q,\mu_j)dq = W_j(\lambda_j,\mu_j) - W_j(0,\mu_j).$$

Since $W_j(0,\mu_j)$ is constant at the lower level, it can be removed from the objective function. Applying a similar operation to P_{Kj}, the lower level objective takes the form

$$\sum_{i\in I}\sum_{j\in J^*}\left[\frac{1}{\theta}x_{ij}\ln x_{ij} + x_{ij}t_{ij}\right] + \alpha\sum_{j\in J^*} W_j(\lambda_j,\mu_j) + \beta\sum_{j\in J^*} P_{Kj}(\lambda_j,\mu_j). \tag{36}$$

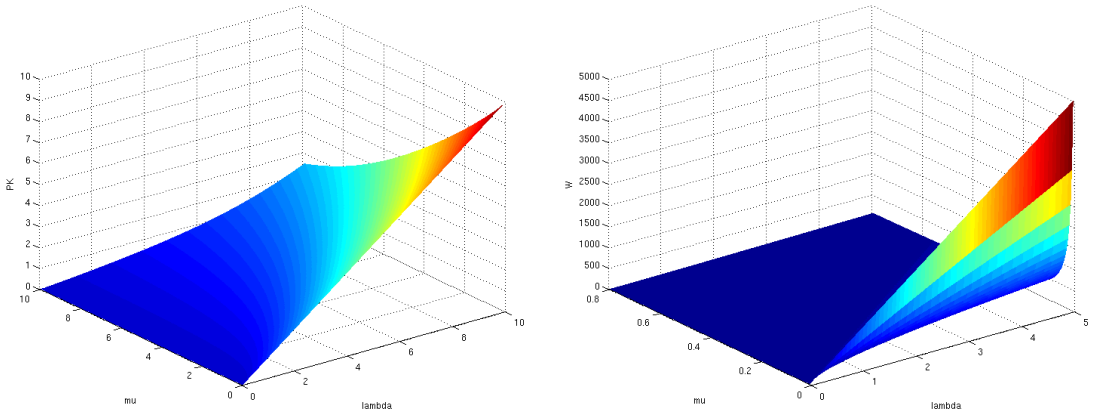

Figure 2: Integrals of probability of balking (P_{Kj}) and waiting time (W_j) for $K = 10$. Although convex in λ_j, neither of them are convex overall, especially in the vicinity of the origin.

Proposition 1. *The waiting time w_j is increasing in λ_j.*

Proposition 2. *The probability of balking p_{Kj} is increasing in λ_j.*

From the convexity of the function $x_{ij} \ln x_{ij}$, and Propositions 1 and 2 it follows that:

Proposition 3. *The lower level objective function (36) is convex in x, hence Problem (P2) is convex.*

Proposition 4. *When $K = \infty$, i.e., balking does not occur (in this case, the model admits a solution only if the total service rate exceeds the total demand rate), the lower level objective function is convex jointly in λ and μ.*

Although the integral of the waiting time and probability of balking are convex in x_{ij} and λ_j, they are not *j*ointly convex in λ_j and μ_j. Figure 2 illustrates the situation.

Proposition 5. *The integral of the waiting time, $W_j(\lambda_j, \mu_j)$ is pseudoconvex.*

The proofs of Propositions 1, 2, 4 and 5 are provided in the e-companion to this paper.

3 Algorithms

This section is concerned with the design of algorithms, both 'semi-exact' and heuristic, for addressing the bilevel location problem. Our 'semi-exact' approach is related to that of [Gilbert et al., 2015] for solving a bilevel toll problem involving logit user assignment. It is based on mixed integer linear programming (MILP) approximations of the original problem. Whenever the approximation is fine-grained, we expect its solution to be nearly optimal,

hence the term 'semi-exact'. In contrast, the heuristic algorithm is based on a surrogate problem, and is akin to the approach of [Marcotte, 1986] for addressing a network design problem involving user-optimized (Wardrop) flows, where the issue of enforcing equilibrium constraints is sidestepped.

3.1 A semi-exact method

By linearizing the upper level nonlinear terms $\bar{\lambda}_j$ and the lower-level objective of the bilevel program, it is possible to reformulate (P) as a mixed integer linear bilevel program, which can be further reduced to a MILP. This is achived through the following five operations:

1. Approximate the lower-level objective function by a piecewise linear approximation.
2. Write the KKT optimality conditions of the lower-level linear program to obtain a single-level mathematical program involving complementarity constraints (MPEC).
3. Formulate the MPEC as an MILP, through the introduction of binary variables and 'big-M' constants.
4. Solve the resulting MILP for optimum values of μ and y.
5. Solve the original nonlinear lower-level problem to recover the true values of the assignment vector x associated with μ and y.

We now provide a detailed description of the linear approximation used at the first step of the algorithm. We let

$$\tilde{d} = \max_{i \in I}\{d_i\}, \quad \bar{\mu} = (B - c_f)/c_\mu, \quad \text{and} \quad \tilde{\mu} = \max\left\{\bar{\mu}, \max_{j \in J_c}\{\mu_j\}\right\},$$

and sample the interval $(0, \tilde{d}]$ using N points $x^n, n = 1, .., N$ such that $x^i < x^j$ for all $i < j$, and consider the linearization

$$f^n(x) = x(\ln x^n + 1) - x^n = a_f^n x + b_f^n. \tag{37}$$

Similarly, let $\tilde{\lambda} = \sum_{i \in I} d_i$ be the maximum arrival rate. We sample the interval $(0, \tilde{\lambda}]$ using R points $\lambda^r, r = 1, \ldots, R$ such that $\lambda^i < \lambda^j$ for $i < j$. We also generate P samples of μ over $(0, \tilde{\mu}]$ over the same interval. Let λ^r and μ^p be the samples hence obtained. We linearize $W_j(\lambda_j, \mu_j)$ and $P_{Kj}(\lambda_j, \mu_j)$ using tangent plane at points (λ^r, μ^p) for $r = 1, \ldots, M$, $p = 1, \ldots, P$ such that $\lambda^r \neq \mu^p$. Based on the gradients

$$\nabla W_j(\lambda_j, \mu_j) = (w_j(\lambda_j, \mu_j), -w_j(\lambda_j, \mu_j)\rho_j) \tag{38}$$

$$\nabla P_{Kj}(\lambda_j, \mu_j) = \left(p_{kj}(\lambda_j, \mu_j), \frac{1}{\mu_j} P_{Kj}(\lambda_j, \mu_j) - p_{kj}(\lambda_j, \mu_j)\rho_j\right), \tag{39}$$

we write the first-order Taylor approximations of $W_j(\lambda_j, \mu_j)$ and $P_{Kj}(\lambda_j, \mu_j)$, respectively:

$$g^{rp}(\lambda, \mu) = W_j(\lambda^r, \mu^p) + \nabla W_j(\lambda^r, \mu^p) \begin{pmatrix} \lambda - \lambda^r \\ \mu - \mu^p \end{pmatrix} = a_g^{rp}\lambda + b_g^{rp}\mu + c_g^{rp},$$

$$h^{rp}(\lambda, \mu) = P_{Kj}(\lambda^r, \mu^p) + \nabla P_{Kj}(\lambda^r, \mu^p) \begin{pmatrix} \lambda - \lambda^r \\ \mu - \mu^p \end{pmatrix} = a_h^{rp}\lambda + b_h^{rp}\mu + c_h^{rp}.$$

Next, we convexify W_j, P_{Kj} and $x \ln x$ by setting them to the maximum of their linear approximations:

$$x_{ij} \ln x_{ij} \approx \max_{n \in N} \{f^n(x_{ij})\} \tag{40}$$

$$W_j(\lambda_j, \mu_j) \approx \max_{r \in R, p \in P} \{g^{rp}(\lambda_j, \mu_j)\} \tag{41}$$

$$P_{Kj}(\lambda_j, \mu_j) \approx \max_{r \in R, p \in P} \{h^{rp}(\lambda_j, \mu_j)\} \tag{42}$$

Upon the introduction of additional variables, the linear approximation of (P2) takes the form

$$\text{(P2-lin)} \quad \min_{x, v, u, z} \quad \sum_{i \in I} \sum_{j \in J^*} \left[\frac{1}{\theta} v_{ij} + x_{ij} t_{ij} \right] + \alpha \sum_{j \in J^*} u_j + \beta \sum_{j \in J^*} z_j \tag{43}$$

$$\text{s.t.} \quad \sum_{j \in J^*} x_{ij} = d_i, \qquad \forall i \in I \tag{44}$$

$$\lambda_j = \sum_{i \in I} x_{ij}, \qquad \forall j \in J^* \tag{45}$$

$$v_{ij} - a_f^n x_{ij} \geq b_f^n, \qquad \forall i \in I; \forall j \in J^*; \forall n \in N \tag{46}$$

$$u_j - a_g^{rp}\lambda_j - b_g^{rp}\mu_j \geq c_g^{rp}, \qquad \forall j \in J^*; \forall r \in R; \forall p \in P \tag{47}$$

$$z_j - a_h^{rp}\lambda_j - b_h^{rp}\mu_j \geq c_h^{rp}, \qquad \forall j \in J^*; \forall r \in R; \forall p \in P \tag{48}$$

$$x_{ij} \geq 0, \qquad \forall i \in I; \forall j \in J^*. \tag{49}$$

We close this section by noting that (P2-lin) is an entirely linear formulation, and thus the variables x_{ij} could assume the value 0, although this cannot occur in the initial formulation (P2), due to the presence of the logarithmic barrier term $\ln x_{ij}$.

To achieve a MILP formulation, we first perform a linear approximation of constraint (14)

using the triangle technique described in [D'Ambrosio et al., 2010]. This yields the equalities

$$\sum_{r=1}^{R-1}\sum_{p=1}^{P-1}\left(\bar{l}_{jrp}+\underline{l}_{jrp}\right)=1, \qquad \forall j\in J_1 \tag{50}$$

$$s_{jrp}\leq \bar{l}_{jrp}+\underline{l}_{jrp}+\bar{l}_{jrp-1}+\underline{l}_{jr-1p-1}+\bar{l}_{jr-1p-1}+\underline{l}_{jr-1p}, \qquad \forall j\in J_1;\forall r\in R;\forall p\in P \tag{51}$$

$$\sum_{r=1}^{R}\sum_{p=1}^{P}s_{jrp}=1, \qquad \forall j\in J_1 \tag{52}$$

$$\lambda_j=\sum_{r=1}^{R}\sum_{p=1}^{P}s_{jrp}\lambda^r, \qquad \forall j\in J_1 \tag{53}$$

$$\mu_j=\sum_{r=1}^{R}\sum_{p=1}^{P}s_{jrp}\mu^p, \qquad \forall j\in J_1 \tag{54}$$

$$e_j=\sum_{r=1}^{R}\sum_{p=1}^{P}s_{jrp}\left(\lambda^r(1-p_{Kj}(\lambda^r,\mu^p))\right), \qquad \forall j\in J_1. \tag{55}$$

Next, we write the optimality conditions of (P2-lin). Let γ_i, δ_j, ν_{ij}^n, π_j^{rp} and η_j^{rp} denote the dual variables associated with constraints (44), (45), (46), (47) and (48), respectively. We replace constraints (16), (17), (18) and (19) in (P) with the optimality conditions of (P2-lin), which yields a nonlinear program involving complementarity constraints. The standard method of dealing with this nonlinearity is to linearize these constraints through the introduction of binary variables and 'big-M' constants. Alternatively, one can substitute to the complementarity constraints the equality of the lower level primal and dual objectives. The latter involves bilinear terms that can be further linearized. Technical details, together with the corresponding MILP formulation, can be found in the e-companion. The MILP can be solved by an off-the-shelf software such as CPLEX. For given location variables y and service rates μ, a feasible assignment matrix x is then recovered by solving a convex assignment program that involves a simple structure. The corresponding running time is negligible. Note that, due to approximation errors in the MILP, the recovered solution is not necessarily identical to the one yielded by the MILP.

Bound on the linearization error for the $M/M/1/\infty$ case

If facilities are modeled as $M/M/1/\infty$ (infinite capacity) queues, the waiting time at a facility j is $w_j(\lambda_j,\mu_j)=1/(\mu_j-\lambda_j)$, and its indefinite integral $W_j(\lambda_j,\mu_j)=-\log(\mu_j-\lambda_j)$, which is convex. We have the following underlying hypotheses:

i. The total service rate in the network can satisfy the entire demand.

ii. In all open facilities, $\mu_j\geq\psi+\lambda_j$, where $\psi>0$.

The latter condition ensures that waiting time at facilities is finite. In practice, ψ can be as as small as desired, and we have that $w_j\leq 1/\psi=w_{MAX}$. Let t_{MIN} and t_{MAX} represent the minimum and maximum travel time in the network, respectively. Furthermore,

$w_{\text{MIN}} = 1/\mu_{\text{MAX}}$, $\text{diam}(t) = t_{\text{MAX}} - t_{\text{MIN}}$ and $\text{diam}(w) = w_{\text{MAX}} - w_{\text{MIN}}$. We define μ_{MAX} as the maximum service rate possible in the network, either allowed by the budget at leader's facilities, or at competitor's facilities.

If both conditions are satisfied, we obtain:

$$x_{ij} = \frac{e^{-\theta\,(t_{ij}+\alpha w_j)}}{\sum\limits_{k\in J^*} e^{-\theta\,(t_{i,k}+\alpha w_k)}} \geq \frac{e^{-\theta\,(t_{\text{MAX}}+\alpha w_{\text{MAX}})}}{\sum\limits_{k\in J^*} e^{-\theta\,(t_{\text{MIN}}+\alpha w_{\text{MIN}})}} = \frac{e^{-\theta\,(\text{diam}(t)+\alpha\ \text{diam}(w))}}{|J^*|} = r_{min}, \tag{56}$$

Now, let $g(\mu, x)$ be the lower-level objective function, i.e.

$$g(\mu,x) = \underbrace{\sum_{i\in I}\sum_{j\in J^*}\left[\frac{1}{\theta}x_{ij}\log x_{ij} + t_{ij}x_{ij}\right]}_{g_1(\mu,x)} + \alpha \underbrace{\sum_{j\in J^*} W_j(\mu_j, x)}_{g_2(\mu,x)}. \tag{57}$$

The lower-level problem can be written as:

$$(\text{P}^\infty)\ \min_x\ g(\mu,x) = g_1(\mu,x) + \alpha g_2(\mu,x) \tag{58}$$

$$\text{s.t.}\ \sum_{j\in J^*} x_{ij} = d_i \qquad \forall i \in I \tag{59}$$

$$x_{ij} \geq 0 \qquad \forall i \in I, \forall j \in J^*. \tag{60}$$

Next, we define the compact set $D = \left\{x \in \mathbb{R}^{|I|\cdot|J|} \mid \sum\limits_{j\in J^*} x_{ij} = d_i, \forall i \in I;\ x_{ij} \geq 0, \forall i \in I, \forall j \in J^*\right\}$, and the function $G(\mu, x) : D \to \mathbb{R}$, $G(\mu,x) = \nabla_x g(\mu,x) = \nabla_x g_1(\mu,x) + \alpha\nabla_x g_2(\mu,x) = G_1(\mu,x) + \alpha G_2(\mu_x)$. Note that D is a compact set. Note that (P^∞) can be written simply as $\min\limits_{x\,\in\, D} g(\mu,x)$, and we have the following results, whose proofs are provided in the e-companion to this paper.

Proposition 6. *G_1 is strongly monotone in x of modulus $\theta \cdot d_{MAX}$.*

Proposition 7. *G_2 is monotone in x.*

It follows directly that

Proposition 8. *G is strongly monotone in x, with modulus $\theta \cdot d_{MAX}$.*

Theorem 1. *The approximation error of the upper-level objective function is at most $O(1/N_1 + 1/N_2)$, where N_1 and N_2 are the number of samples for the linearization of g_1 and g_2, respectively.*

We now illustrate Theorem 1 for the instance based on the network illustrated in Figure 3. It involves two demand nodes, which are potential locations as well. Demand rates in 1 and 2 are set to 5.5 and 15.0, respectively. The fixed cost of opening a facility is set to 5 and the

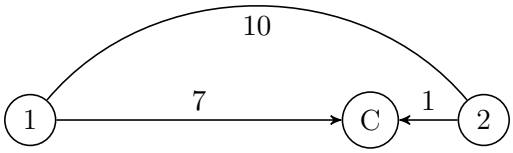

Figure 3: A three-node network.

unit service cost to 1, for a total budget of 25. The competition owns a facility with service rate 25.1. On the demand side, parameter α is set to 10 and parameter θ to 0.2.

For each set of open locations, the problem can be approximately solved by sampling a very large number of values of the parameter μ. This yields an optimal solution with objective 10.197, where both facilities are open, with respective service rates 5.325 and 9.675.

The semi-exact algorithm was then run for different sample sizes, and we report the optimal of the approximation MILP, as well as the true objective values corresponding to these solutions. The results are displayed in Figure 4, where we observe that

- The approximated objective mostly overestimates the true objective.
- The true objective obtained by solving for the actual equilibrium with respect to the service levels quickly reaches a near-optimal solution, and actually does so for a sample size as small as 4.
- The true objective does not increase in a monotone fashion, but stabilizes fairly quickly close to the optimum.

3.2 A surrogate-based heuristic

In this section we present a parameterized heuristic based on replacing the original model by a single-level model involving a surrogate objective, whose optimal solution automatically satisfies the fixed point constraint. This strategy is akin to that proposed by [Marcotte, 1986] for addressing a bilevel network design problem involving user-optimized flow patterns.

The rationale behind this strategy is that both the leader and the users have a shared interest in minimizing delays. We therefore expect that, if the lower level is given full control, the resulting design should favor access to the leader's facilities, and therefore yield a high throughput. Incorporating the budget constraint to ensure feasibility, we obtain the single-

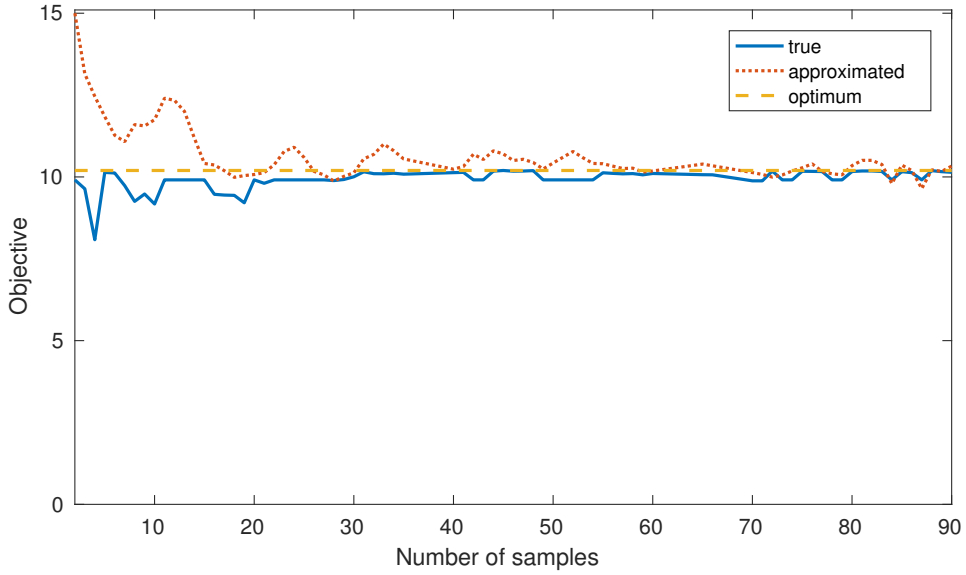

Figure 4: Evolution of the semi-exact MILP objective value with respect to sample size. We use the same number of samples on x, and λ. Legend: the 'approximated' line corresponds to the optimal objective of the approximate MILP. The 'true' line is the true objective value corresponding to the MILP solution.

level mixed nonlinear program

$$\text{(PH)} \quad \min_{y,\mu,x} \sum_{i\in I}\sum_{j\in J^*}\left[\frac{1}{\theta}x_{ij}\ln x_{ij} + x_{ij}t_{ij}\right] + \alpha\sum_{j\in J^*}\int_0^{\lambda_j} w_j(q,\mu_j)dq + \beta\sum_{j\in J^*}\int_0^{\lambda_j} p_{Kj}(q,\mu_j)dq \tag{61}$$

$$\text{s.t.} \sum_{j\in J} x_{ij} = d_i, \qquad \forall i \in I \tag{62}$$

$$\sum_{j\in J_1} y_j c_f + \sum_{j\in J_1} c_\mu \mu_j \le B, \tag{63}$$

$$\lambda_j = \sum_{i\in I} x_{ij}, \qquad \forall j \in J \tag{64}$$

$$y_j \in \{0,1\}, \qquad \forall j \in J \tag{65}$$

$$x_{ij} \ge 0, \qquad \forall i \in I; \forall j \in J, \tag{66}$$

whose x-solution is a logit flow assignment with respect to the design variables y and μ. For $\theta = \infty$, the limiting case (PH*) is a mathematical program involving user-equilibrium

(Wardropian with respect to queueing delays) flows, and is expressed as

$$\text{(PH*)} \quad \min_{y,\mu,x} \quad \sum_{i\in I}\sum_{j\in J^*} x_{ij}t_{ij} + \alpha \sum_{j\in J^*}\int_0^{\lambda_j} w_j(q,\mu_j)dq + \beta \sum_{j\in J^*}\int_0^{\lambda_j} p_{Kj}(q,\mu_j)dq$$
$$\text{s.t.} \quad \text{constraints (62) –(66).}$$

Properties of the surrogate model

The surrogate model always yields feasible solutions for the original model, and inherits some of its properties, such as nonconvexity of its objective. However, some properties may help to understand why it is computationally tractable, as will be confirmed in Section 4. Proofs are provided in the e-companion.

Proposition 9. *If $K = \infty$ and there are no fixed costs, the surrogate model is convex.*

Proposition 10. *At the optimum of (PH*), if $K = \infty$, queue waiting times are equal for all leader's facilities.*

We close this section with an example that shows that, in the worst case, the difference between the heuristic optimum and the true optimum can be arbitrarily large. Let us consider the network shown in Figure 5, with sites A and B being potential opening nodes for the leader, with null fixed cost. Let $D_1 > 1$ and nD_1 be the demand in nodes 1 and 2, respectively. The total service rate available to the leader is $\bar{\mu} = (2n+4)D_1$. The service rate at the competitor's facility is set to $\mu_c = 2nD_1$. From proposition 10, waiting times at facilities A and B must be equal. Since $t_{2,2} = t_{2,c}$, half of population issued from 2 chooses facility C, while the other half chooses facility B. It is easy to check that the solution of (PH*) is $\mu_1 = \mu_2 = (n+2)D_1$, with each leader's facility capturing D_1 customers, for a total of $2D_1$ number of served customers. On the other hand, if we set $\mu_1 = 2D_1$ and $\mu_2 = (2n+2)D_1$, the leader captures D_1 customers at facility A, and $D_1(n+2)/2$ at facility B, for a total number of customers of $D_1(n+4)/2$. Since K is infinite, no customers are lost. The ratio between the better option (described above) and the one found by the heuristic is $(n+4)/4$, which can be arbitrarily large.

A parameterized surrogate heuristic

One drawback of the heuristic solution presented in the previous section is that, for $K = \infty$ and $\theta = \infty$, queueing delays are equal, a property that might not hold at the true optimum. Actually, in order to maximize efficiency, one expects the leader to adapt its service rates to arrival rates. This can be achieved by incorporating a service-dependent linear term into the objective. This term depends on a set of positive parameters ξ_j, to be tuned, one for each facility. The resulting mathematical program is

$$\text{(PH}(\xi)\text{)} \quad \min_{y,\mu,x} \quad \sum_{i\in I}\sum_{j\in J^*}\left[\frac{1}{\theta}x_{ij}\ln x_{ij} + x_{ij}t_{ij}\right] + \alpha\sum_{j\in J^*} W_j(x,\mu_j) + \beta\sum_{j\in J^*} P_{Kj}(x,\mu_j) + \sum_{j\in J_1}\xi_j\mu_j$$
$$\text{s.t.} \quad \text{constraints (62), (63), (65), (66).}$$

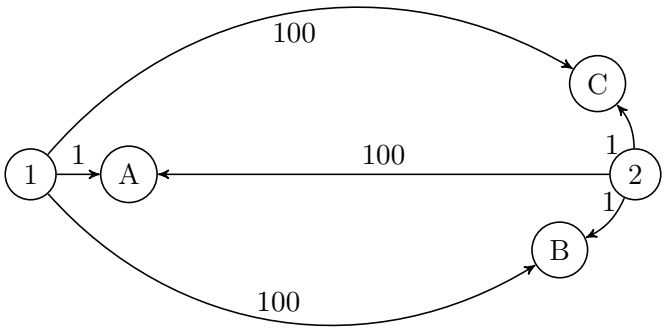

Figure 5: An instance where the gap between the heuristic and optimal value of the objective function can be as large as desired.

This program is transformed and solved as a MILP where the linearization is based on the techniques presented in Section 3.1. As before, a feasible flow assignment x compatible with the location vector y and the service rate vector μ is retrieved by solving a convex program. We now focus on the case $K = \infty$ and $\theta = \infty$, when there are no fixed costs:

$$\begin{aligned} (\text{PHY}^*(\xi)) \quad \min_{y,\mu,x} \quad & \sum_{i\in I}\sum_{j\in J^*} x_{ij}t_{ij} - \alpha \sum_{j\in J^*} \ln(\mu_j - (\sum_{i\in I} x_{ij})) + \sum_{j\in J_1} \xi_j \mu_j \\ \text{s.t.} \quad & \text{constraints (62), (65), (66), (70),} \end{aligned}$$

for which we provide a theoretical result, whose proof is provided in the e-companion.

Proposition 11. *There exists a value of ξ^* for which (PHY*(ξ^*)) yields an optimal solution for (P*).*

While the complexity of determining an optimal ξ vector is equivalent to that of solving the initial problem, educated guesses may yield good values, as will be observed later.

4 Experimental setup and results

The MILP formulation was solved by IBM ILOG CPLEX Optimizer version 12.5. All tests, either using the semi-exact method or heuristics, were performed on a 16 core Xeon(R) Intel(R) processor running at 2.4GHz frequency. For the semi-exact method, we opted for the MILP formulation based on the equality between the primal and dual lower level objectives. Surprisingly, while approximate, this formulation outperformed that based on complementarity constraints.

An initial set of experiments was intended to compare the linear approximation-based method and the parameterized heuristic described in Sections 3.1 and 3.2, respectively, that involve the parameterized model (PH(ξ)). The latter is solved for different values of the parameter ξ. For each facility j, ξ_j is set to the negative of a scalar that increases with

demand and decreases with distance:

$$\xi_j = -c \sum_{i \in I} \frac{d_i}{t_{ij} + 1}, \tag{67}$$

for some nonnegative parameter c. This is motivated by the fact that it makes sense, from the leader's perspective, to assign high service rates to facilities located close to high demand nodes: the lower ξ_j, the larger μ_j in the optimal solution. The term 1 in the denominator was added to t_{ij} to avoid dividing by a small number. The linear approximations involve 7, 5 and 5 uniformly distributed samples for x, λ and μ, respectively. The parameter α was set to 10, while the algorithms were run for different combinations of parameters θ and β. Travel times were varied between 0 and 100 for nodes belonging to a common cluster. Two sensible choices for the parameter β are 50 or 100, as previously explained in Section 2.2.1.

In CPLEX branching rules, priority was given to the strategic location variables over the binary variables required in the linearization process. The algorithm was stopped as soon as the optimality gap dropped below 1%, CPLEX ran out of memory (4GB), or that CPU exceeded 2,000 seconds.

		heuristic over semi-exact ratio			semi-exact relative gap(%)			CPU(seconds)			semi-exact gap ≤0.1%	
θ	β	$c=0$	$c=1$	best	min	average	max	semi-exact	$c=0$	$c=1$	PI(%)	CPU(s)
0.2	50	0.99	0.93	1.01	0.99	11.3	25.4	1,778	110	11	−0.66	31,239
0.5	50	1.00	0.95	1.01	0.98	12.1	25.6	1,834	17	8	0.08	14,375
2.0	50	0.99	0.93	1.00	0.88	11.5	25.6	1,833	9	7	1.20	44,832
0.2	100	0.99	0.98	1.00	0.98	11.8	26.0	1,930	101	10	1.14	13,852
0.5	100	0.98	0.97	0.99	0.98	11.1	26.1	1,836	18	9	0.00	13,888
2.0	100	1.03	1.01	1.04	0.99	11.9	26.0	1,929	9	8	3.46	13,874

Table 1: Comparison between the semi-exact method and two heuristics. Budget set to 500. Averages taken over 10 instances.

Tables 1 and 2 report mean CPU times (in seconds), the optimality gap when the stopping criteria is met, and the average ratio between the objective value found by the heuristics and by the semi-exact method (as described in Section 3.1), for two values of the available budget. Heuristics are run for different values of parameter c, as in Eq. (67). We also report the best solutions found across these runs in the *best* column. Additionally, we let CPLEX run to optimality (gap<0.1%), regardless of the execution time, comparing the objective value obtained within 2,000 seconds and the one obtained with no time limit; we report the percentage increase (the PI column).

In most cases, CPLEX could not reach a gap less than 1% in the allotted CPU. As shown in Tables 1 and 2, the average optimality gap lies in the [11,14] interval, when time is limited. However, as illustrated in Figure 6, the optimal solutions are frequently found in the early stages of the Branch-and-Bound process, while the remaining iterations are merely used to prove optimality. The above observation is supported by the numbers in the PI column. The percentage increase in objective value when running to optimality is not significant (less

		heuristic over semi-exact ratio				semi-exact relative gap(%)			CPU(seconds)				semi-exact gap ≤0.1%	
θ	β	$c=0$	$c=1$	$c=10$	best	min	average	max	semi-exact	$c=0$	$c=1$	$c=10$	PI(%)	CPU(s)
0.2	50	0.86	0.86	0.62	0.94	0.93	12.0	24.2	1,862	20	12	5	0.14	56,616
0.5	50	0.83	0.86	0.62	0.93	2.22	13.7	21.6	2,011	10	9	5	1.40	22,871
2.0	50	0.83	0.86	0.63	0.94	2.25	12.8	21.3	2,010	9	8	5	0.10	39,029
0.2	100	0.84	0.86	0.58	0.88	0.99	11.3	20.7	1,826	15	10	6	−0.60	23,990
0.5	100	0.83	0.84	0.62	0.90	1.92	12.4	21.7	2,009	9	9	6	0.30	22,850
2.0	100	0.82	0.84	0.59	0.87	0.99	10.9	19.3	1,903	8	8	6	0.25	11,089

Table 2: Comparison between the semi-exact method and three heuristics. Budget set to 250. Averages taken over 10 instances.

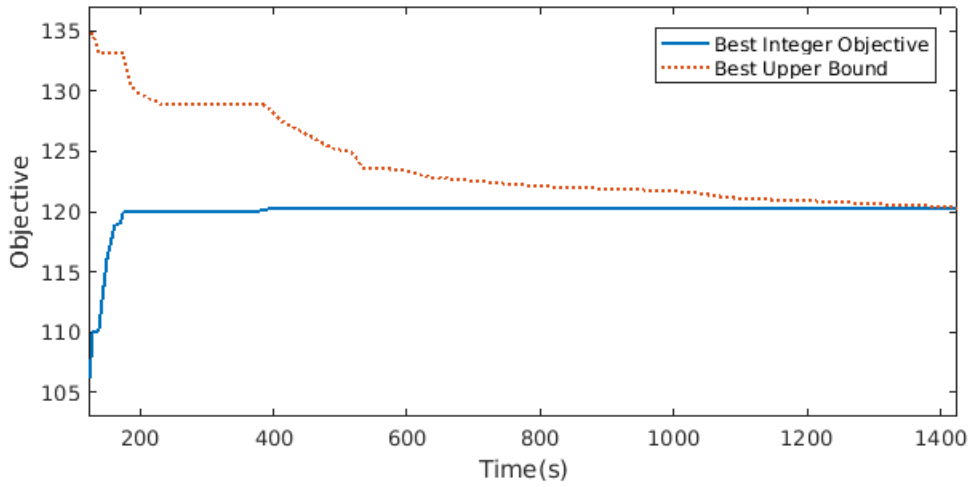

Figure 6: Lower and upper bounds throughout the branch-and-bound process for an instance of (P-lin).

than 1.5%, in most cases, and 3.5% when the budget is 500, $\theta = 2.0$ and $\beta = 100$), despite a large increase in CPU. In some cases we observe a small decrease in the objective value, which is explained as follows: when running to optimality, there can be a small increase in the approximate objective value (the one found by solving the MILP), however the optimal solution corresponds to a slightly small true objective. We remind you that the MILP is only an approximated version of a highly nonlinear program.

Table 1 shows that, for a high budget, heuristics perform well, managing to attract and serve, on average, the same number of customers as the semi-exact method, and in some cases, outperforms it. This inconsistency is made possible due to approximation errors in the various linearizations performed at both the lower level and in the objective function of the semi-exact method.

Table 2 tells a similar story. In this case (budget = 250), taken individually, heuristics for $c = 0$, $c = 1$ and $c = 10$ do not perform very well, capturing as little as 58% of the semi-exact value in one case. However, when retaining the best out of the three, the objective value is around $87 - 94\%$ of the semi-exact objective, at a much lower computational cost. For instance, for budget = 250, the CPU required by the semi-exact method exceeds by a factor of 50 ($\theta = 0.2, \beta = 50$) up to 91 ($\theta = 2.0, \beta = 50$) the combined CPU of the three heuristics. This illustrates the limitations of the semi-exact method, which, although superior in terms of solution quality, does not scale well. We also observed that, in the heuristic case, for identical values of the parameter β, CPU is a decreasing function of θ. We recall that this parameter is inversely proportional to the standard deviation of the Gumbel random variable embedded into the logit process. When θ is small, users are spread over the facilities, regardless of their disutility, making for highly nonlinear instances that are difficult to linearize. In contrast, when θ is large, variance is small, and users focus on a limited number of destination facilities.

Within the same experimental setup, it is interesting to compare the number of facilities opened by the various algorithms. As displayed in Table 3 and Table 4, the semi-exact method opens between 4 and 6 facilities, and on average $5.6 - 5.8$ for a budget of 500. When the budget is set to 250, the number of facilities opened by the semi-exact method is reduced by more than one, on average. For both values of the budget, the leader opens less facilities for $\beta = 100$ than for $\beta = 50$. Indeed, as β increases, users require a higher service rate to make for the higher probability of balking. The budget is thus spent more on service rate and a little less on opening new facilities.

		number of open facilities			fraction of common facilities	
θ	β	exact	$c = 0$	$c = 1$	$c = 0$	$c = 1$
0.2	50	5.9	8.3	6.8	0.52	0.54
0.5	50	6.0	8.4	6.8	0.56	0.45
2.0	50	5.9	8.4	6.8	0.66	0.55
0.2	100	5.7	8.2	7.4	0.62	0.55
0.5	100	5.7	8.2	7.6	0.54	0.48
2.0	100	5.6	8.2	7.3	0.48	0.51

Table 3: Number of open facilities. Budget set to 500. Averages over 10 different runs.

For the high budget and low values of c (0 or 1) the heuristics open on average $6.8 - 8.4$ facilities. Only half of the facilities opened by the semi-exact method are among them. Nevertheless, the heuristic facilities yield large values of the objective function. For low budget, a similar situation occurs although all methods open, on average, less facilities. Overall, we notice a trend among heuristics: the average number of open facilities decreases with c. The larger values of c yield smaller values of ξ, therefore, larger values of μ, and the heuristics put more emphasis on providing high service rates, versus opening several facilities. These results highlight the fact that determining the optimal facility locations is hard, and that solutions of similar values can vastly differ in their topologies.

Although Table 2 suggests that heuristics do not perform very well when budget is small, a closer inspection reveals that for some values of c, they yield results close to those of the

		number of open facilities				fraction of common facilities		
θ	β	exact	$c=0$	$c=1$	$c=10$	$c=0$	$c=1$	$c=10$
0.2	50	3.8	5.7	5.7	2.7	0.54	0.54	0.28
0.5	50	4.1	5.9	5.7	2.8	0.45	0.40	0.21
2.0	50	3.8	5.9	5.6	2.7	0.51	0.50	0.26
0.2	100	3.5	5.7	5.5	3.0	0.50	0.50	0.40
0.5	100	3.7	5.7	5.6	2.8	0.54	0.55	0.35
2.0	100	3.7	5.7	5.5	3.0	0.54	0.54	0.33

Table 4: Number of open facilities. Budget set to 250. Averages over 10 different runs.

semi-exact method, as shown in Table 5, where the best results among those run for values of c ranging from 0 to 10 are displayed. The best results were usually related to low values of c. In this setting, heuristics manage to capture from 90% up to 95% of the number of customers obtained by the semi-exact method, at a much lower computational cost.

θ	β	heuristic over semi-exact ratio	total CPU (sec.)
0.2	50	0.95	133
0.5	50	0.96	86
2.0	50	0.95	69
0.2	100	0.90	132
0.5	100	0.92	95
2.0	100	0.88	75

Table 5: Parameter c runs from 0 to 10. Budget set to 250.

In Table 6, we report the impact of c on the number of facilities opened, as well as on the number of served customers, for 3 randomly chosen tests in our dataset. We vary the c from 0 to 10, and report the best solution found for each test. We then compute the average ratio between the latter and the optimum found by the semi-exact method. As c increases, more importance is given to μ, and less budget is available for opening facilities. A second trend is the concave-like behaviour (increasing, levelling, decreasing) of the number of served customers with respect to c, shown in Table 6.

Finally, we decided to assess the performance of the heuristics, given an optimal set of open facilities provided by the semi-exact method. Restricted to the determination of service levels, the problem remains a hard nonlinear bilevel program. All tests have been performed on the same aforementioned dataset, using 10 samples for x and 9 for λ and μ. The results are displayed in Table 7, where we observe a sharp improvement. Actually, due to the approximation errors in the semi-exact method, the latter was outperformed by the theoretically suboptimal heuristics.

	# of open facilities			served customers		
c	test 1	test 2	test 3	test 1	test 2	test 3
0	6	6	6	113.92	124.70	122.57
1	6	6	6	113.92	124.88	122.64
2	6	6	6	114.82	124.95	123.29
3	6	6	6	115.94	122.55	123.33
4	4	5	7	99.39	119.54	123.75
5	3	4	6	84.42	98.82	124.07
6	3	1	6	84.42	45.70	123.99
7	2	1	5	74.21	45.70	116.87
8	1	1	4	53.64	45.70	106.71
9	1	1	3	53.64	45.70	90.48
10	1	1	1	53.64	45.70	62.30

Table 6: Sensitivity of analysis with respect to c in formula (67).

		heuristic over semi-exact ratio			
θ	β	$c = 0$	$c = 1$	$c = 10$	best
0.2	50	1.02 (5)	1.02 (4)	0.84 (1)	1.02
0.5	50	1.02 (6)	1.01 (2)	0.86 (2)	1.02
2.0	50	1.02 (6)	1.00 (4)	0.83 (1)	1.02
0.2	100	1.01 (7)	1.00 (3)	0.88 (1)	1.02
0.5	100	1.02 (5)	1.00 (5)	0.89 (0)	1.02
2.0	100	1.02 (8)	1.00 (3)	0.89 (3)	1.02

Table 7: Heuristics run from facility locations provided by the semi-exact method. Budget set to 250. Within parentheses: number of instances for which the corresponding value of c yielded the best result. The sum of values exceed in some cases the total number of tests, as sometimes, different heuristics yield the same optimum.

Accuracy of linearization

In order to measure the impact of the number of sample points involved in the approximation of the nonlinear functions $\tilde{W}$ and $\tilde{P}_K$ (K was set to 10), we vary λ and μ for values ranging from 1 to 10, for a step of 0.1. We then compute the absolute difference between W and P_K, and their linearized counterparts across this fine-grained domain. Note that, due to nonconvexity in the vicinity of the origin (see Figure 2), the tangents in this area can be very steep and thus wildly overestimate the true value of the function. For this reason, linearization sample points were not selected close to 0. As observed in Table 8, increasing the number of sample points can actually worsen the approximation, due to non-convexity of the original functions. The way around this issue would be to make *nonconvex* piecewise linear approximations, the drawback being the addition of a significant number of binary variables, and thus a sharp increase in the running time of the algorithm. When selecting

a number of samples, one has to achieve a trade-off between the error on W, on P_K, the running time and the quality of the solution.

# of samples		Error (average)	
R (on λ)	P (on μ)	W	P_K
3	3	1.34	0.29
3	5	1.33	0.36
3	7	1.77	0.38
3	10	2.94	0.41
5	3	1.13	0.38
5	5	1.24	0.41
5	7	2.67	0.41
5	10	4.37	0.42

# of samples		Error (average)	
R (on λ)	P (on μ)	W	P_K
7	3	2.10	0.20
7	5	2.00	0.42
7	7	2.17	0.42
7	10	5.51	0.42
10	3	2.05	0.26
10	5	2.00	0.43
10	7	3.36	0.43
10	10	3.18	0.43

Table 8: Linearization error for the waiting time and probability of balking, for different number of samples, when $K = 10$.

# of samples			CPLEX			true no of	estimated no of
N (on x)	R (on λ)	P (on μ)	CPU limit(s)	CPU(s)	gap(%)	served customers	served customers
2	2	2	1,000	562	9.71	88.15	80.35
5	3	3	2,000	829	0.92	97.89	100.65
7	3	3	5,000	1,057	0.97	98.33	100.65
7	5	5	7,000	5,752	0.94	102.24	103.42
10	5	5	10,000	8,856	7.78	100.66	103.81
10	7	7	15,000	12,478	1.14	104.91	106.80
12	7	7	20,000	16,921	16.56	94.13	93.69

Table 9: Number of attracted and served customers for different number of samples ($K = 10$), $\theta = 0.2$, $\beta = 50$.

Finally, we investigate the impact of sample size on the quality of the optimal solution of the generated MILP. Surprisingly, as observed in Table 9 this impact is almost negligible, and the objective can actually decrease when the sample size increases. A similar behaviour has been observed in [Marcotte, 1986] for a bilevel pricing model where a probability density function was approximated by a coarse-grained histogram. This behaviour can also be explained by factors such as travel time. For instance, if a facility is located far from a demand point, a small error in the waiting time will not significantly impact the number of arriving customers.

According to the results displayed in Table 9, we observe that the value of the objective function estimated by the approximate model does not correlate well with the actual optimal value obtained by performing an assignment of users with respect to the service rate vector μ. Note that for $10, 7, 7$ and $12, 7, 7$ CPLEX was not able to find a feasible solution in the alloted time, for 3 out of 10 tests. Since the true number of attracted and served customers

is quite insensitive to the number of samples, it is clearly advantageous to set those number to values as small as possible, but yet not too small.

An illustrative case

In this section we illustrate our methods on a fictitious case study that fits well the model described in Section 2.2. We consider the construction of walk-in clinics in the Mont-Tremblant, Canada area. Walk-in clinics provide professional assessment and treatment for minor illnesses or injuries, for people who do not have a family doctor, and often function without an appointment, on a first-come first-served basis. According to Statistics Canada (2017) Health Fact Sheets, in 2014 25.2% of Quebec residents were without a regular doctor. Having a regular doctor plays a key role in the early screening and treatment of various diseases. The problem is to decide the location and service rate of new facilities as to maximize the number of patients served by the clinics.

	Number of open facilities								
	Budget=15			Budget=20			Budget=25		
θ	$\beta = 10$	$\beta = 50$	$\beta = 100$	$\beta = 10$	$\beta = 50$	$\beta = 100$	$\beta = 10$	$\beta = 50$	$\beta = 100$
0.01	2	2	2	2	3	3	3	3	3
0.1	2	2	2	2	2	2	3	3	2
0.2	2	2	2	2	2	2	2	2	2
0.5	2	2	2	2	2	2	2	2	2

Table 10: Parametric analysis on θ, β and the budget.

Mont-Tremblant has 17 population zones, to which we assign demand nodes, which we assume to be spatially located in the center of each zone. The population count per demand node is generated as follows. The initial population data is taken from Statistics Canada [Census, 2016], out of which only 25.2% would be interested to visit a walk-in clinic. Considering 250 days a year, 8 hours a day, and an average of 4 doctor visits per year, per person, the hourly demand count represents only 0.05% of the initial population.

There are already 4 medical clinics in Mont-Tremblant that we consider serving on average between 1 and 3 clients per hour. Assuming the balking threshold at 10 (people balk if there are 10 or more people waiting in line), and a fixed cost/variable cost ratio of 5:1, we perform a parametric analysis on β, θ and the budget. We show the results in Table 10.

Note that for small values of θ, the number of open facilities increases with the budget, which is expected. For higher values of θ, only two facilities are open, regardless of the increase in the budget. When θ is close to 0, clients choose facilities with almost no respect to their disutility. When θ is higher, the clinic must ensure low waiting time and probability of balking. For instance, when $\theta = 0.1$ it opens 3 facilities for $\beta = 10$ and only 2 when $\beta = 100$, for a budget of 25. This happens due to β (the balking coefficient in the disutility formula). When clients place a low importance on the probability of balking (e.g. $\beta = 10$), more money can be spent in opening new clinics. On the other hand, when $\theta = 100$, we open only two facilities and we invest more in a higher service rate.

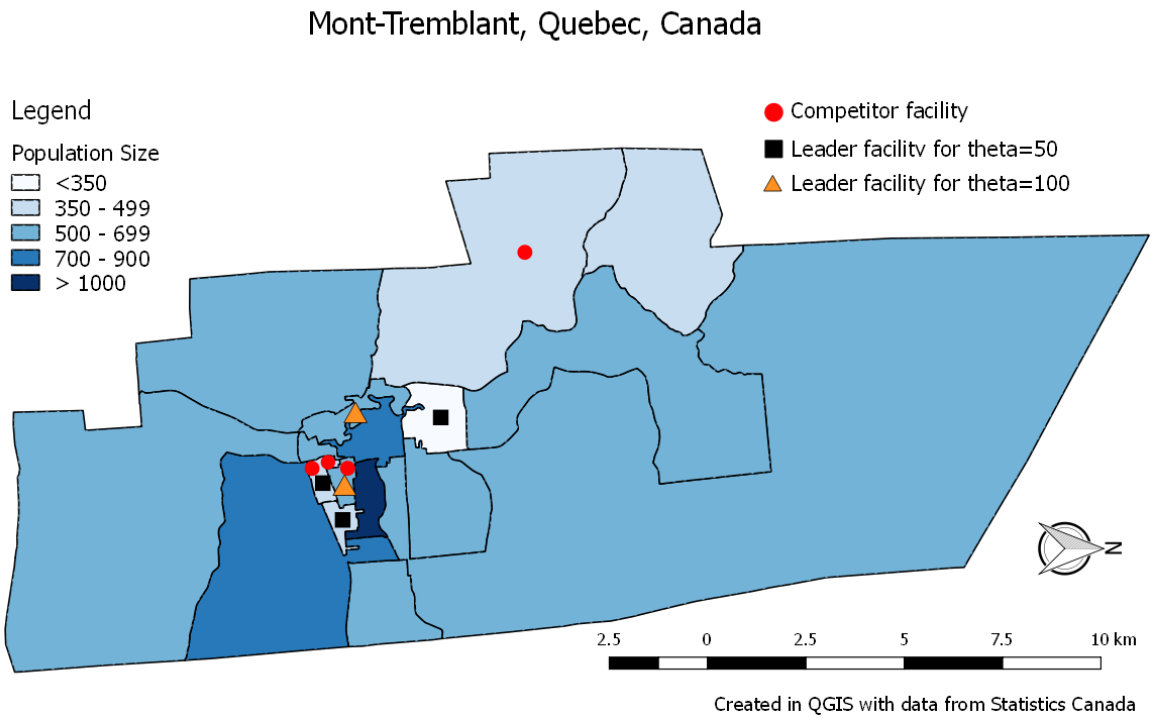

Figure 7: Population Map of Mont-Tremblant, Qc, Canada

In Figure 7 we illustrate the spatial repartition of the facilities, for the cases mentioned above. The main observation is that the facilities are opened adjacent to the highly populated areas, but not within them. This demonstrates the complexity of the problem, showing that the most populated areas are not always the best choice for an optimal location. We also note that the emerging facilities seem to be close to the competitor's facilities.

5 Conclusion and extensions

In this paper, we have addressed a complex location problem that, beyond the combinatorial nature of location decisions, involves two sources of nonlinearity, one related to queueing at the facilities, and the second to the random utility model that characterizes user behaviour. Cast within a bilevel setting, we proposed for its solution a semi-exact algorithm, as well as a parameterized heuristic. We also provided an illustrative case of a real-life application.

While the results are more than encouraging, our findings raise a number of issues, from either the modelling, theoretical or algorithmic viewpoints. For instance, the surprising result that the standard linearization of the lower level complementarity constraints proved less efficient, numerically, than an approach based on a triangular approximation involving a larger number of binary variables, is certainly worth investigating.

On the modelling side, future work will integrate features such as variable demand and

the possibility of either increasing or decreasing the service rates of existing facilities. This will involve a piecewise affine investment function whose two slopes reflect the fact that economies resulting from lowering service are less than those of increasing it. More realistic models where the price of service depends on location should also be considered.

On the algorithmic side, three avenues can be pursued: (i) the design of improved approximations for the nonlinear terms involved in the semi-exact method, and (ii) the design of fast heuristics for determining good sets of facility locations, from which efficient methods for determining optimal service rates can be initiated and, finally (iii) the investigation of approximations based on the exact mixed integer formulation of the logit-based location models proposed by [Haase, 2009], [Benati and Hansen, 2002], [Zhang et al., 2012], and numerically analyzed by [Haase and Müller, 2014].

References

[hea, 2017] (2017). Health Fact Sheets. Statistics Canada Online Catalogue no. 82-625-X.

[Abouee-Mehrizi et al., 2011] Abouee-Mehrizi, H., Babri, S., Berman, O., and Shavandi, H. (2011). Optimizing capacity, pricing and location decisions on a congested network with balking. *Mathematical Methods of Operations Research*, 74(2):233–255.

[Averbakh et al., 2007] Averbakh, I., Berman, O., Drezner, Z., and Wesolowsky, G. O. (2007). The uncapacitated facility location problem with demand-dependent setup and service costs and customer-choice allocation. *European Journal of Operational Research*, 179(3):956–967.

[Benati and Hansen, 2002] Benati, S. and Hansen, P. (2002). The maximum capture problem with random utilities: Problem formulation and algorithms. *European Journal of Operational Research*, 143(3):518–530.

[Beresnev, 2013] Beresnev, V. (2013). Branch-and-bound algorithm for a competitive facility location problem. *Computers & Operations Research*, 40(8):2062–2070.

[Berman and Drezner, 2006] Berman, O. and Drezner, Z. (2006). Location of congested capacitated facilities with distance-sensitive demand. *IIE Transactions*, 38(3):213–221.

[Boffey et al., 2007] Boffey, B., Galvão, R., and Espejo, L. (2007). A review of congestion models in the location of facilities with immobile servers. *European Journal of Operational Research*, 178(3):643–662.

[Camacho-Vallejo et al., 2014] Camacho-Vallejo, J.-F., Cordero-Franco, A. E., and González-Ramírez, R. G. (2014). Solving the bilevel facility location problem under preferences by a stackelberg-evolutionary algorithm. *Mathematical Problems in Engineering*, 2014:14.

[Castillo et al., 2009] Castillo, I., Ingolfsson, A., and Sim, T. (2009). Socially optimal location of facilities with fixed servers, stochastic demand and congestion. *Production & Operations Management*, 18(6):721–736.

[Census, 2016] Census (2016). Dissemination Area Boundary File, 2016 Census. Statistics Canada Catalogue no. 92-169-X.

[D'Ambrosio et al., 2010] D'Ambrosio, C., Lodi, A., and Martello, S. (2010). Piecewise linear approximation of functions of two variables in MILP models. *Operations Research Letters*, 38(1):39–46.

[Desrochers et al., 1995] Desrochers, M., Marcotte, P., and Stan, M. (1995). The congested facility location problem. *Location Science*, 3(1):9–23.

[Drezner et al., 2015] Drezner, T., Drezner, Z., and Kalczynski, P. (2015). A leader-follower model for discrete competitive facility location. *Computers & Operations Research*, 64:51–59.

[Fisk, 1980] Fisk, C. (1980). Some developments in equilibrium traffic assignment methodology. *Transportation Research B*, 14(3):243–256.

[Gilbert et al., 2015] Gilbert, F., Marcotte, P., and Savard, G. (2015). A numerical study of the logit network pricing problem. *Transportation Science*, 49(3):706–719.

[Haase, 2009] Haase, K. (2009). Discrete location planning. Technical report, Institute of Transport and Logistics Studies, University of Sydney.

[Haase and Müller, 2014] Haase, K. and Müller, S. (2014). A comparison of linear reformulations for multinomial logit choice probabilities in facility location models. *European Journal of Operational Research*, 232(3):689–691.

[Hakimi, 1983] Hakimi, S. (1983). On locating new facilities in a competitive environment. *European Journal of Operational Research*, 12(1):29–35.

[Hotelling, 1929] Hotelling, H. (1929). Stability in competition. *The Economic Journal*, 39(153).

[Kim, 2013] Kim, S. (2013). Heuristics for congested facility location problem with clearing functions. *Journal of the Operational Research Society*, 64(12):1780–1789.

[Küçükaydin et al., 2011] Küçükaydin, H., Aras, N., and Altınel, I. K. (2011). Competitive facility location problem with attractiveness adjustment of the follower: A bilevel programming model and its solution. *European Journal of Operational Research*, 208(3):206–220.

[Labbé and Hakimi, 1991] Labbé, M. and Hakimi, S. L. (1991). Market and locational equilibrium for two competitors. *Operations Research*, 39(5):749–756.

[Marcotte, 1986] Marcotte, P. (1986). Network design problem with congestion effects: A case of bilevel programming. *Mathematical Programming*, 34(2):142–162.

[Marianov, 2003] Marianov, V. (2003). Location of multiple-server congestible facilities for maximizing expected demand, when services are non-essential. *Annals of Operations Research*, 123(1-4):125–141.

[Marianov et al., 2008] Marianov, V., Ríos, M., and Icaza, M. J. (2008). Facility location for market capture when users rank facilities by shorter travel and waiting times. *European Journal of Operational Research*, 191(1):32–44.

[Marianov and Serra, 2001] Marianov, V. and Serra, D. (2001). Hierarchical location-allocation models for congested systems. *European Journal of Operational Research*, 135(1):195–208.

[Marić et al., 2012] Marić, M., Stanimirović, Z., and Milenković, N. (2012). Metaheuristic methods for solving the bilevel uncapacitated facility location problem with clients' preferences. *Electronic Notes in Discrete Mathematics*, 39(0):43 – 50. EURO Mini Conference.

[McFadden, 1974] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *FRONTIERS IN ECONOMETRICS*, pages pp. 105–142.

[Rahmati et al., 2014] Rahmati, S. H. A., Ahmadi, A., Sharifi, M., and Chambari, A. (2014). A multi-objective model for facility location-allocation problem with immobile servers within queuing framework. *Computers & Industrial Engineering*, 74(0):1–10.

[Vidyarthi and Jayaswal, 2014] Vidyarthi, N. and Jayaswal, S. (2014). Efficient solution of a class of location-allocation problems with stochastic demand and congestion. *Computers & Operations Research*, 48(0):20–30.

[Zhang et al., 2010] Zhang, Y., Berman, O., Marcotte, P., and Verter, V. (2010). A bilevel model for preventive healthcare facility network design with congestion. *IIE Transactions*, 42(12):865–880.

[Zhang et al., 2012] Zhang, Y., Berman, O., and Verter, V. (2012). The impact of client choice on preventive healthcare facility network design. *OR Spectrum*, 34(2):349–370.

A Notation and proofs

In this e-companion we present the notation used throughout this paper, and we complete the proofs of some propositions.

B Notation

Sets

I:	set of demand nodes;
J:	set of candidate facility locations (leader and competition);
J_c:	set of competition's facilities;
J_1:	set of leader's candidate sites;
$J_1^* \subseteq J_1$:	set of leader's open facilities
$J^* \subseteq J$:	set of open facilities (leader and competitor).

Parameters

d_i: demand originating from node $i \in I$;
t_{ij}: travel time between nodes $i \in I$ and $j \in J$;
α: coefficient of the waiting time in the disutility formula;
β: coefficient of the balking probability in the disutility formula;
B: available budget (for opening new facilities and associated service rates);
c_f: fixed cost associated with opening a new facility;
c_μ : cost per unit of service;
$\bar{\mu}$: maximum service rate allowed by the budget;
p: number of facilities to open.

Basic decision variables

y_j: binary variable set to 1 if a facility is open at site j, and to 0 otherwise;
μ_j: service rate at open facilities.

Additional variables

x_{ij}: arrival rate at at facility $j \in J$ originating from demand node $i \in I$;
λ_j: arrival rate at node $j \in J$;
ρ_j: utilization rate of facility $j \in J$;
$\bar{\lambda}_j$: throughput rate (customers accessing service) at node $j \in J$;
w_j: mean queueing time at facility j.

C Proofs of Propositions 1, 2, 4, 5, 6, 7, 9, 10, 11 and Theorem 1

Proposition 1. *The waiting time w_j is increasing in λ_j.*

Proof. Proof. The derivative of w_j with respect to λ_j (see Equation (9)) is

$$\frac{\partial w_j}{\partial \lambda_j} = \frac{\partial w_j}{\partial \rho_j}\frac{\partial \rho_j}{\partial \lambda_j} = \frac{\partial w_j}{\partial \rho_j}\frac{1}{\mu_j}.$$

To show that $\partial w_j/\partial \rho_j$ is nonnegative for all $\rho_j \neq 1$, let us consider

$$\frac{\partial w_j}{\partial \rho_j} = \frac{1}{\mu_j}\left(-\frac{K^2\rho_j^K - 1}{\left(\rho_j^K - 1\right)^2} + \frac{1}{(\rho_j - 1)^2}\right).$$

Basic algebraic manipulation yields

$$\frac{1}{(\rho_j - 1)^2} \geq \frac{K^2\rho_j^K - 1}{\left(\rho_j^K - 1\right)^2} \iff \sum_{i=0}^{K-1} \rho_j^i \geq K\rho_j^{(K-1)/2}. \tag{68}$$

To prove that the right-hand inequality holds true, we consider two cases.

If K is odd:

$$\begin{aligned}\sum_{i=0}^{K-1} \rho_j^i &= \sum_{i=0}^{(K-1)/2-1} \left(\rho_j^i + \rho_j^{K-1-i}\right) + \rho_j^{(K-1)/2} \\ &\geq 2 \sum_{i=0}^{(K-1)/2-1} \rho_j^{(K-1)/2} + \rho_j^{(K-1)/2} = K\rho_j^{(K-1)/2}.\end{aligned}$$

If K is even:

$$\sum_{i=0}^{K-1} \rho_j^i = \sum_{i=0}^{(K-2)/2} \left(\rho_j^i + \rho_j^{K-1-i}\right) \geq 2 \sum_{i=0}^{(K-2)/2} \rho_j^{(K-1)/2} = K\rho_j^{(K-1)/2}.$$

It follows that w_j is an increasing function of λ_j. □

Proposition 2.

Proposition 2. *The probability of balking p_{Kj} is increasing in λ_j.*

Proof. Proof. The derivative of p_{Kj} with respect to λ_j is

$$\begin{aligned}p'_{Kj} &= \frac{\lambda_j^{K-1}\mu_j}{\left(\lambda_j^{K+1} - \mu_j^{K+1}\right)^2}\left[\lambda_j^{K+1} - (K+1)\lambda_j\mu_j^K + K\mu_j^{K+1}\right] \\ &= \sigma[x^{K+1} - (K+1)\,x + K],\end{aligned}$$

where σ is a positive number and $x = \lambda_j/\mu_j$. By differentiating with respect to x, we find that the right-hand-side achieves its minimum value 0 at $x = 1$, which concludes the proof. □

Proposition 4.

Proposition 4. *When $K = \infty$, i.e., balking does not occur (in this case, the model admits a solution only if the total service rate exceeds the total demand rate), the lower level objective function is convex jointly in λ and μ.*

Proof. Proof. If $K = \infty$, the probability of balking can be removed from the objective, since it is equal to 0. Moreover, $w_j = 1/(\mu_j - \lambda_j)$, and the lower level objective takes the form

$$\sum_{i \in I} \sum_{j \in J^*} \left[\frac{1}{\theta} x_{ij} \ln x_{ij} + x_{ij} t_{ij}\right] - \alpha \sum_{j \in J^*} \ln(\mu_j - \lambda_j).$$

Basic algebra shows that its Hessian is positive semidefinite, hence the function is convex. □

Proposition 5.

Proposition 5. *The integral of the waiting time, $W_j(\lambda_j, \mu_j)$ is pseudoconvex.*

Proof. Proof. Let $x = (\lambda_x, \mu_x)$ and $y = (\lambda_y, \mu_y)$. Assume that $\nabla W(x)(y - x) \geq 0$. Then we have:

$$\begin{aligned} & \left(w_j(x), -\frac{\lambda_x}{\mu_x} w_j(x) \right) (\lambda_y - \lambda_x, \mu_y - \mu_x) \geq 0 \\ \Rightarrow \quad & (\rho_y - \rho_x) w_j(x) \geq 0 \\ \Rightarrow \quad & \rho_y \geq \rho_x, \end{aligned} \tag{69}$$

since w_j is nonnegative. On the other hand, $\partial W_j / \partial \rho = \mu_j w_j$ is nonnegative, we have that W_j is increasing in ρ, so $\rho_y \geq \rho_x \Rightarrow W_j(y) \geq W_j(x)$. From Eq (69) it follows that if $\nabla W(x)(y - x) \geq 0$ then $W_j(y) \geq W_j(x)$, hence W_j is pseudoconvex. □

Proposition 6.

Proposition 6. *G_1 is strongly monotone in x of modulus $\theta \cdot d_{MAX}$.*

Proof. Proof. [Gilbert et al., 2015] have already argued that G_1 is strongly monotone. Indeed, the associated Jacobian is a positive definite diagonal matrix over D, with the smallest possible eigenvalue $1/(\theta \cdot d_{\text{MAX}})$. It follows that G_1 is strongly monotone with modulus $\theta \cdot d_{\text{MAX}}$. □

Proposition 7.

Proposition 7. *G_2 is monotone in x.*

Proof. Proof.

$$\begin{aligned} \langle G_2(\mu, x) - G_2(\mu, y), x - y \rangle &= \sum_{i \in I} \sum_{j \in J^*} \left(\frac{1}{\mu_j - \sum_{l \in I} x_{l,j}} - \frac{1}{\mu_j - \sum_{l \in I} y_{l,j}} \right) \cdot (x_{ij} - y_{ij}) \\ &= \sum_{j \in J^*} \left[\frac{\mu_j - \sum_{l \in I} y_{l,j} - \mu_j + \sum_{l \in I} x_{l,j}}{\left(\mu_j - \sum_{l \in I} x_{l,j} \right) \cdot \left(\mu_j - \sum_{l \in I} y_{l,j} \right)} \cdot \sum_{i \in I} (x_{ij} - y_{ij}) \right] \\ &= \sum_{j \in J^*} \left[\frac{\sum_{l \in I} (x_{l,j} - y_{l,j}) \sum_{l \in I} (x_{l,j} - y_{l,j})}{\left(\mu_j - \sum_{l \in I} x_{l,j} \right) \cdot \left(\mu_j - \sum_{l \in I} y_{l,j} \right)} \right] \\ &\geq 0 \end{aligned}$$

□

Proposition 9.

Proposition 9. *If $K = \infty$ and there are no fixed costs, the surrogate model is convex.*

Proof. Proof. According to Proposition 4, the objective is jointly convex in μ and λ. Moreover one can, without loss of generality, open all facilities and hence dispense with the binary vector y. Notwithstanding, a facility can be closed by setting its service level to zero. □

Proposition 10.

Proposition 10. *At the optimum of (PH*), if $K = \infty$, queue waiting times are equal for all leader's facilities.*

Proof. Proof. For fixed y variables, Equation (63) can be rewritten as

$$\sum_{j \in J^*} \mu_j \leq \bar{\mu}, \tag{70}$$

where $\bar{\mu}$ is the maximum possible total service rate allowed by the budget. But $K = \infty$, so $w_j(\lambda_j, \mu_j) = 1/(\mu_j - \lambda_j)$ and $p_{Kj}(\lambda_j, \mu_j) = 0$, which yields the mathematical program

$$\begin{aligned} \text{(PHY*)} \qquad \min_{\mu, x} \quad & \sum_{i \in I} \sum_{j \in J^*} x_{ij} t_{ij} - \alpha \sum_{j \in J^*} \ln(\mu_j - (\sum_{i \in I} x_{ij})) \\ \text{s.t.} \quad & \text{constraints (62), (65), (66), (70)} \end{aligned}$$

Let δ_i, π_{ij} and γ be the Lagrange multipliers associated with Equations (62), (66) and (70), respectively. Variables δ_i are free, while γ and π_{ij} are restricted to be nonnegative. The stationarity conditions of the above program are:

$$\frac{\partial L}{\partial x_{ij}} = 0 \;\Rightarrow\; t_{ij} + \alpha w_j(\lambda_j, \mu_j) - \delta_i - \pi_{ij} = 0, \;\forall i \in I, \forall j \in J^* \tag{71}$$

$$\frac{\partial L}{\partial \mu_j} = 0 \;\Rightarrow\; -\alpha w_j(\lambda_j, \mu_j) + \gamma = 0, \;\forall j \in J^* \cap J_1, \tag{72}$$

and the conclusion follows from Equation (72). □

We observe, after plugging $\alpha w_j(\lambda_j, \mu_j)$ from Equation (72) into Equation (71) for a given demand node i, that only one flow x_{ij} is nonzero, provided that transportation times to the leader's facilities are distinct.

Proposition 11.

Proposition 11. *There exists a value of ξ^* for which (PHY*(ξ^*)) yields an optimal solution for (P*).*

Proof. Proof. Let y^* and μ^* be optimal for (P*). Without loss of generality (there are no fixed costs) we assume that all facilities are open. At equilibrium, let c_i^* be the cost associated with demand node i and optimal service rate μ^*. Let x^*, $w_j(x^*, \mu_j^*)$ and c_i^* satisfy Equation (23) and (24). If x_{ij} is positive, we have:

$$t_{ij} + \alpha w_j(x^*, \mu^*) = c_i^*, \quad \forall j \in J, \forall i \in I. \tag{73}$$

Let $C = \max_{i \in I} \{c_i^*\}$ in the initial formulation. For $j \in J$, we let $\xi_j = c_i^* - t_{ij} - C$ and select and index i corresponding to a positive flow x_{ij}^*. If no such i exists, then $\mu_j^* = 0$, otherwise the leader would waste monetary resources. We then set $\xi_j = -C$.

Now, let δ_i, π_{ij} and γ be the Lagrange multipliers associated with Equations (62), (66) and (70), respectively. Variables δ_i and γ are free, while π_{ij} are restricted to be nonnegative. The stationarity conditions of the program above take the form

$$\frac{\partial \mathcal{L}}{\partial x_{ij}} = 0 \;\Rightarrow \qquad t_{ij} + \alpha w_j(x, \mu_j) - \delta_i = 0, \quad \text{if } x_{ij} > 0, \;\forall i \in I, \forall j \in J \tag{74}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_j} = 0 \;\Rightarrow \qquad -\alpha w_j(x, \mu_j) + \gamma + \xi_j = 0, \;\forall j \in J_1. \tag{75}$$

Note that the derivative of the Lagrangian with respect to x_{ij} is left unchanged, i.e., Equation (74) is equivalent to Equation (71). If $\gamma = C$, we derive from Equation (75) that $\alpha w_j(x, \mu_j) = c_i^* - t_{ij}$, which is equivalent to Equation (74). This completes the proof, since for the given values of ξ, variables x and μ match the optimal solution of (P*). □

Theorem 1.

Theorem 1. *The approximation error of the upper-level objective function is at most $O(1/N_1 + 1/N_2)$, where N_1 and N_2 are the number of samples for the linearization of g_1 and g_2, respectively.*

Proof. Proof. Let $\bar{G}$ be an approximation of G. We note $\bar{x}$ = solution of IV$(\bar{G}(\mu, \cdot), D)$, and x = solution of IV$(G(\mu, \cdot), D)$. Then the following inequalities hold:

$$\begin{aligned} \langle G(\mu, x), \bar{x} - x \rangle &\geq 0 \\ \langle \bar{G}(\mu, \bar{x}), x - \bar{x} \rangle &\geq 0 \\ \Rightarrow \langle \bar{G}(\mu, \bar{x}) - G(\mu, x), x - \bar{x} \rangle &\geq 0 \end{aligned} \tag{76}$$

From the strong monotonicity of G and Eq. (76) it follows that

$$\langle \bar{G}(\mu, \bar{x}) - G(\mu, \bar{x}), x - \bar{x} \rangle \geq \frac{1}{\theta \cdot d_{\text{MAX}}} ||x - \bar{x}||^2. \tag{77}$$

We write the norm, and we obtain

$$\theta \cdot d_{\text{MAX}} \cdot ||\bar{G}(\mu, \bar{x}) - G(\mu, \bar{x})|| \geq ||x - \bar{x}||. \tag{78}$$

It follows that

$$\begin{aligned} |f(x) - f(\bar{x})| = |\sum_{i \in I} \sum_{j \in J_1^*} (x_{ij} - \bar{x}_{ij})| &\leq \sqrt{|I| \cdot |J|} ||x - \bar{x}|| \quad \text{(Cauchy-Schwarz inequality)} \\ &\leq \sqrt{|I| \cdot |J|}\; \theta \cdot d_{\text{MAX}} \cdot ||\bar{G}(\mu, \bar{x}) - G(\mu, \bar{x})||. \end{aligned} \tag{79}$$

We perform two separate linear approximations on g_1 and g_2, respectively. Then the vector functions $\bar{G}_1(x)$ and $\bar{G}_2(x)$ are piecewise constant approximations, that we detail separately.

A. $\bar{G}_1$: Each component (i,j) of this vector is a constant approximation of $\log(x_{ij})$, satisfying:

i) there are N_1 total samples on x_{ij}, starting from r_{min} to d_{MAX}.

ii) the sampling points are chosen so that the slopes are equidistant

iii) the slopes are tangents of $x\log(x)$, evaluated in the sampling points.

We note with Δ_1 the difference between two consecutive slopes: $\Delta_1 = \dfrac{\log(d_{\text{MAX}}) - \log(r_{min})}{N_1 - 1}$. Then $|\bar{G}_{1(i,j)} - G_{1(i,j)}| \leq \Delta_1$, which yields:

$$||\bar{G}_1(x) - G_1(x)|| = \sqrt{\sum_{i \in I} \sum_{j \in J^*} |\bar{G}_{1(i,j)} - G_{1(i,j)}|^2} \leq \frac{(\log(d_{\text{MAX}}) - \log(r_{min}))\sqrt{|I| \cdot |J|}}{N_1 - 1} \tag{80}$$

B. $\bar{G}_2$: This is a vector function whose component corresponding to a pair (i,j) is a constant piecewise approximation of $1/q_j$, where $= \mu_j - \sum_{i \in I} x_{ij}$. Similar to $\bar{G}_1$, this linearization satisfies the following:

i) there are N_2 total samples, starting from ψ to μ_{MAX}.

ii) the sampling points are chosen so that the slopes are equidistant

iii) the slopes are tangents of $-\log(q)$, evaluated in the sampling points.

We note with Δ_2 the difference between two consecutive slopes: $\Delta_2 = \dfrac{\frac{1}{\psi} - \frac{1}{\mu_{\text{MAX}}}}{N_2 - 1}$. Then $|\bar{G}_{2(i,j)} - G_{2(i,j)}| \leq \Delta_2$, which yields

$$||\bar{G}_2(x) - G_2(x)|| = \sqrt{\sum_{i \in I} \sum_{j \in J^*} |\bar{G}_2(x_{ij}) - G_2(x_{ij})|^2} \leq \frac{\left(\frac{1}{\psi} - \frac{1}{\mu_{\text{MAX}}}\right)\sqrt{|I| \cdot |J|}}{N_2 - 1} \tag{81}$$

From Eq. (79) it follows that, given y and μ:

$$|f(x) - f(\bar{x})| \leq \theta \cdot d_{\text{MAX}}|I| \cdot |J| \left[\frac{(\log(d_{\text{MAX}}) - \log(r_{min}))}{N_1 - 1} + \alpha \frac{\frac{1}{\psi} - \frac{1}{\mu_{\text{MAX}}}}{N_2 - 1} \right] \in O(\frac{1}{N_1} + \frac{1}{N_2}). \tag{82}$$

□

Theorem 1 has several implications.

- For a given set of open facilities, the absolute difference between the optimal and the approximated objective value is bounded by the right-hand-side of inequality (82). For large values of N_1 and N_2, the two values are very close.
- If the optimal solution is unique in terms of the location vector y, and the absolute difference between the objective and other solutions objectives are lower than the right hand side of inequality (82), the approximation algorithm will find the optimum locations.

D Linearization of optimality conditions

D.1 Complementarity constraints for Program (P2-lin)

Let γ_i, δ_j, ν_{ij}^n, π_j^{rp}, η_j^{rp}, and ϕ_{ij} be the dual variables associated with constraints (44), (45), (46), (47), (48) and (49), respectively. Then the complementarity constraints for program (P2-lin) can be written as:

$$\gamma_i \left(\sum_{j \in J^*} x_{ij} - d_i \right) = 0 \qquad \forall i \in I \tag{83}$$

$$\delta_j \left(\lambda_j - \sum_{i \in I} x_{ij} \right) = 0 \qquad \forall j \in J^* \tag{84}$$

$$\nu_{ij}^n \left(v_{ij} - a_f^n x_{ij} - b_f^n \right) = 0 \qquad \forall i \in I; \;\; \forall j \in J^*; \;\; \forall n \in N \tag{85}$$

$$\pi_j^{rp} \left(u_j - a_g^{rp} \lambda_j - b_g^{rp} \mu_j - c_g^{rp} \right) = 0 \qquad \forall j \in J^*; \;\; \forall r \in R; \;\; \forall p \in P \tag{86}$$

$$\eta_j^{rp} \left(z_j - a_h^{rp} \lambda_j - b_h^{rp} \mu_j - c_h^{rp} \right) = 0 \qquad \forall j \in J^*; \;\; \forall r \in R; \;\; \forall p \in P \tag{87}$$

$$\phi_{ij} x_{ij} = 0 \qquad \forall i \in I; \;\; \forall j \in J^*, \tag{88}$$

and can be linearized in the standard fashion, through the introduction of binary variables and big-M constants. For instance, the last constraint is replaced by the inequalities

$$\begin{aligned} \phi_{ij} &\leq M u_{ij} \\ x_{ij} &\leq M(1 - u_{ij}), \end{aligned}$$

where $u_{ij} \in \{0, 1\}$. It is possible to find a valid upper bound for the variable ϕ_{ij} however, a large value of M is required, which leads to a poor relaxation and consequently a ill-behaved branch-and-bound algorithm.

D.2 Equality between primal and dual objectives

Alternatively, constraints (83) – (88) can be replaced with constraint (89), which represents the equality between between the primal and dual objective of (P2-lin). Then the optimality

constraints of (P2-lin) are

$$\sum_{i \in I} \gamma_i d_i + \sum_{n \in N} \sum_{i \in I} \sum_{j \in J} \nu_{ij}^n b_f^n + \sum_{r \in R} \sum_{p \in P} \sum_{j \in J} \left(b_g^{rp} \mu_j \pi_j^{rp} + b_h^{rp} \mu_j \eta_j^{rp} + c_g^{rp} \pi_j^{rp} + c_h^{rp} \eta_j^{rp} \right)$$
$$= \sum_{i \in I} \sum_{j \in J} \left[\frac{1}{\theta} v_{ij} + x_{ij} t_{ij} \right] + \alpha \sum_{j \in J} u_j + \beta \sum_{j \in J} z_j, \tag{89}$$

$$\sum_{j \in J} x_{ij} = d_i, \qquad \forall i \in I$$

$$\lambda_j = \sum_{i \in I} x_{ij}, \qquad \forall j \in J$$

$$v_{ij} - a_f^n x_{ij} \geq b_f^n, \qquad \forall i \in I; \forall j \in J; \forall n \in N$$

$$u_j - a_g^{rp} \lambda_j - b_g^{rp} \mu_j \geq c_g^{rp}, \qquad \forall j \in J; \forall r \in R; \forall p \in P$$

$$z_j - a_h^{rp} \lambda_j - b_h^{rp} \mu_j \geq c_h^{rp}, \qquad \forall j \in J; \forall r \in R; \forall p \in P$$

$$\gamma_i + \delta_j - \sum_{n \in N} a_f^n \nu_{ij}^n \leq t_{ij}, \qquad \forall i \in I;\ \forall j \in J$$

$$-\delta_j - \sum_{r \in R} \sum_{p \in P} \left(a_g^{rp} \pi_j^{rp} + a_h^{rp} \eta_j^{rp} \right) = 0, \qquad \forall j \in J$$

$$\sum_{n \in N} \nu_{ij}^n = \frac{1}{\theta}, \qquad \forall i \in I;\ \forall j \in J$$

$$\sum_{r \in R} \sum_{p \in P} \pi_j^{rp} = \alpha, \qquad \forall j \in J$$

$$\sum_{r \in R} \sum_{p \in P} \eta_j^{rp} = \beta, \qquad \forall j \in J$$

$$\pi_j^{rp}, \eta_j^{rp} \geq 0, \qquad \forall j \in J; \forall r \in R; \forall p \in P$$

$$x_{ij} \geq 0,\ \nu_{ij}^n \geq 0, \qquad \forall i \in I; \forall j \in J; \forall n \in N.$$

To obtain a MILP formulation, we linearize the nonlinear terms $\mu_j \pi_j^{rp}$ and $\mu_j \eta_j^{rp}$ via the triangle method described in [D'Ambrosio et al., 2010]. For each term $\mu_j \pi_j^{kq}$ we introduce $2(R-1)(P-1)$ binary variables $\overline{l}_{jrpkq}^{\pi}$ and $\underline{l}_{jrpkq}^{\pi}$ associated with the upper and lower triangles, respectively, of the rectangle defined by the intervals $[\pi^r, \pi^{r+1})$ and $[\mu^p, \mu^{p+1})$. Note that the values of π and η are upper bounded by α and β, respectively. Additionally, μ is bounded by the maximum value allowed by the leader's budget, $\bar{\mu}$. Next, we introduce $J_1 RP$ continuous variables $s_{jrpkq} \in [0,1]$ which will be used to express the couple (π_j^{kq}, μ_j) as a convex combination of triangle vertices. We introduce a similar linearization for the term

$\mu_j \eta_j^{kh}$. The approximation of $\mu_j \pi_j^{kq}$ and $\mu_j \eta_j^{kq}$ is then

$$\sum_{r=1}^{R-1} \sum_{p=1}^{P-1} \left(\overline{l}^{\pi}_{jrpkq} + \underline{l}^{\pi}_{jrpkq} \right) = 1, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{90}$$

$$s^{\pi}_{jrpkq} \leq \overline{l}^{\pi}_{jrpkq} + \underline{l}^{\pi}_{jrpkq} + \overline{l}^{\pi}_{jrp-1kq} + \underline{l}^{\pi}_{jr-1p-1kq} + \overline{l}^{\pi}_{jr-1p-1kq} + \underline{l}^{\pi}_{jr-1pkq}, \quad \forall j \in J_1; \forall r \in R; \forall p \in P; \forall k \in R; \forall q \in P \tag{91}$$

$$\sum_{r=1}^{R} \sum_{p=1}^{P} s^{\pi}_{jrpkq} = 1, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{92}$$

$$\pi_j^{kq} = \sum_{r=1}^{R} \sum_{p=1}^{P} s^{\pi}_{jrpkq} \pi^r, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{93}$$

$$\mu_j = \sum_{r=1}^{R} \sum_{p=1}^{P} s^{\pi}_{jrpkq} \mu^p, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{94}$$

$$e^{\pi}_{jkq} = \sum_{r=1}^{R} \sum_{p=1}^{P} s^{\pi}_{jrpkq} \pi^r \mu^p, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{95}$$

$$\sum_{r=1}^{R-1} \sum_{p=1}^{P-1} \left(\overline{l}^{\eta}_{jrpkq} + \underline{l}^{\eta}_{jrpkq} \right) = 1, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{96}$$

$$s^{\eta}_{jrpkq} \leq \overline{l}^{\eta}_{jrpkq} + \underline{l}^{\eta}_{jrpkq} + \overline{l}^{\eta}_{jrp-1kq} + \underline{l}^{\eta}_{jr-1p-1kq} + \overline{l}^{\eta}_{jr-1p-1kq} + \underline{l}^{\eta}_{jr-1pkq}, \quad \forall j \in J_1; \forall r \in R; \forall p \in P; \forall k \in R; \forall h \in P \tag{97}$$

$$\sum_{r=1}^{R} \sum_{p=1}^{P} s^{\eta}_{jrpkq} = 1, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{98}$$

$$\eta_j^{kq} = \sum_{r=1}^{R} \sum_{p=1}^{P} s^{\eta}_{jrpkq} \eta^r, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{99}$$

$$\mu_j = \sum_{r=1}^{R} \sum_{p=1}^{P} s^{\eta}_{jrpkq} \mu^p, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{100}$$

$$e^{\eta}_{jkq} = \sum_{r=1}^{R} \sum_{p=1}^{P} s^{\eta}_{jrpkq} \eta^r \mu^p, \qquad \forall j \in J_1; \forall k \in R; \forall q \in P \tag{101}$$

The complete MILP formulation is presented below. It involves variables associated with the original fixed point (or bilevel) formulation (y, μ, x), together with variables issued from the

linearizations and primal-dual optimality conditions.

$$
\begin{aligned}
\text{(P-lin)} \qquad \max \quad & \sum_{j \in J_1} e_j \\
& \sum_{j \in J_1} y_j c_f + \sum_{j \in J_1} c_\mu \mu_j \leq B, \\
& \mu_j \leq \bar{\mu} y_j, && \forall j \in J_1 \\
& \sum_{r \in R} \sum_{p \in P} \sum_{j \in J_c} \left(b_g^{rp} \pi_j^{rp} \mu_j + b_h^{rp} \eta_j^{rp} \mu_j + c_g^{rp} \pi_j^{rp} + c_h^{rp} \eta_j^{rp} \right) + \sum_{n \in N} \sum_{i \in I} \sum_{j \in J} \nu_{ij}^n b_f^n \\
& + \sum_{r \in R} \sum_{p \in P} \sum_{j \in J_1} \left(b_g^{rp} e_{jrp}^{\pi} + b_h^{rp} e_{jrp}^{\eta} + c_g^{mp} \pi_j^{rp} + c_h^{rp} \eta_j^{rp} \right) + \sum_{i \in I} \gamma_i d_i \\
& \quad = \sum_{i \in I} \sum_{j \in J} \left[\frac{1}{\theta} v_{ij} + x_{ij} t_{ij} \right] + \alpha \sum_{j \in J} u_j + \beta \sum_{j \in J} z_j, \\
& \sum_{j \in J} x_{ij} = d_i, && \forall i \in I \\
& \lambda_j = \sum_{i \in I} x_{ij}, && \forall j \in J \\
& v_{ij} - a_f^n x_{ij} \geq b_f^n, && \forall i \in I; \forall j \in J; \forall n \in N \\
& u_j - a_g^{rp} \lambda_j - b_g^{rp} \mu_j \geq c_g^{rp}, && \forall j \in J; \forall r \in R; \forall p \in P \\
& z_j - a_h^{rp} \lambda_j - b_h^{rp} \mu_j \geq c_h^{rp}, && \forall j \in J; \forall r \in R; \forall p \in P \\
& \gamma_i + \delta_j - \sum_{n \in N} a_f^n \nu_{ij}^n \leq t_{ij}, && \forall i \in I; \ \forall j \in J \\
& -\delta_j - \sum_{r \in R} \sum_{p \in P} \left(a_g^{rp} \pi_j^{rp} + a_h^{rp} \eta_j^{rp} \right) = 0, && \forall j \in J \\
& \sum_{n \in N} \nu_{ij}^n = \frac{1}{\theta}, && \forall i \in I; \ \forall j \in J \\
& \sum_{r \in R} \sum_{p \in P} \pi_j^{rp} = \alpha, && \forall j \in J \\
& \sum_{r \in R} \sum_{p \in P} \eta_j^{rp} = \beta, && \forall j \in J \\
& \text{constraints (90) – (101) and (50) – (55),} \\
& y_j \in \{0,1\}, \ \mu_j, \pi_j^{rp}, \eta_j^{rp} \geq 0, && \forall j \in J; \forall r \in R; \forall p \in P \\
& x_{ij} \geq 0, \ \nu_{ij}^n \geq 0, && \forall i \in I; \forall j \in J; \forall n \in N.
\end{aligned}
$$

D.3 Example of lower level linearization when $K = \infty$

Recall that, according to Proposition 4, the function is convex if the buffer zone is infinite (no balking). In that situation, the maximum of the linear approximations is consistent with the original function, give or take the approximation error.

Proceeding as before, we obtain

$$g^{rp}(\lambda, \mu) = a_g^{rp}\lambda + b_g^{rp}\mu + c_g^{rp} = \frac{\alpha}{\mu^p - \lambda^r}\lambda - \frac{\alpha}{\mu^p - \lambda^r}\mu - \alpha(\ln(\mu^p - \lambda^r) - 1). \tag{102}$$

This yields the linearized lower level program

$$\begin{aligned}
(\mathrm{P2}^\infty) \quad \min_{x,\, v,\, u} \quad & \sum_{i \in I} \sum_{j \in J^*} \left[\frac{1}{\theta} v_{ij} + x_{ij} t_{ij}\right] + \alpha \sum_{j \in J^*} u_j & & (103)\\
\text{s.t.} \quad & \sum_{j \in J^*} x_{ij} = d_i, & \forall i \in I & (104)\\
& \lambda_j = \sum_{i \in I} x_{ij}, & \forall j \in J^* & (105)\\
& v_{ij} - a_f^n x_{ij} \geq b_f^n, & \forall i \in I; \forall j \in J^*; \forall n \in N & (106)\\
& u_j - a_g^{rp}\lambda_j - b_g^{rp}\mu_j \geq c_g^{rp}, & \forall j \in J^*; \forall r \in R; \forall p \in P & (107)\\
& x_{ij} \geq 0, & \forall i \in I; \forall j \in J^*. & (108)
\end{aligned}$$

D.4 Taxonomy

We now provide a taxonomy of the models most relevant to our research, with respect to four features: (i) user choice environment (yes or no), (ii) stochastic (or not), (iii) inclusion of congestion (or not) at facilities, (iv) inclusion (or not) of competition. The relevant information is displayed in Table 11.

Authors	user choice	stochastic	congestion	competition
[Abouee-Mehrizi et al., 2011]	×	×	×	
[Averbakh et al., 2007]	×			
[Berman and Drezner, 2006]	×		×	
[Camacho-Vallejo et al., 2014]	×			
[Castillo et al., 2009]		×	×	
[Desrochers et al., 1995]			×	
[Kim, 2013]			×	
[Küçükaydin et al., 2011]	×	×		×
[Labbé and Hakimi, 1991]				×
[Marianov and Serra, 2001]			×	
[Marianov, 2003]		×	×	
[Marianov et al., 2008]	×	×	×	×
[Marić et al., 2012]	×			
[Rahmati et al., 2014]		×	×	
[Vidyarthi and Jayaswal, 2014]		×	×	
[Zhang et al., 2010]	×		×	

Table 11: Taxonomy of congested facility location models