EURO Journal on Computational Optimization (2020) 8:61–84
https://doi.org/10.1007/s13675-019-00120-w

ORIGINAL PAPER

Joint location and pricing within a user-optimized environment

Teodora Dan[1] · Andrea Lodi[2] · Patrice Marcotte[1]

Received: 30 September 2018 / Accepted: 29 August 2019 / Published online: 24 September 2019

Abstract
In the design of service facilities, whenever the behaviour of customers is impacted by queueing or congestion, the resulting equilibrium cannot be ignored by a firm that strives to maximize revenue within a competitive environment. In the present work, we address the problem faced by a firm that makes decisions with respect to location, service levels and prices and that takes explicitly into account user behaviour. This situation is modelled as a nonlinear mathematical program with equilibrium constraints that involves both discrete and continuous variables, and for which we propose an efficient algorithm based on an approximation that can be solved for its global optimum.

Keywords Pricing · Location pricing · Bilevel programming · Mixed-integer programming · Equilibrium · Queueing · Nonconvex

Mathematics Subject Classification 90C11 · 90C26 · 90C30 · 90C33

1 Introduction

In a competitive market, service levels and pricing, along with facility locations, are critical decisions that a service provider faces in order to capture demand and maximize

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13675-019-00120-w) contains supplementary material, which is available to authorized users.

✉ Teodora Dan
teodora.dan@umontreal.ca

Andrea Lodi
andrea.lodi@polymtl.ca

Patrice Marcotte
marcotte@iro.umontreal.ca

[1] Université de Montréal, C.P. 6128, succursale Centre-ville, Montreal, QC H3C 3J7, Canada

[2] Polytechnique Montréal, P.O. Box 6079, Station Centre-Ville, Montreal, QC H3C 3A7, Canada

profit. In this context, an important trait of a user-choice market is congestion, which has been often overlooked in the pricing literature, where one routinely assumes that users patronize the closest facility, disregarding the delays that may arise at facilities in the form of queues. However, in real-life situations, customers are sensitive to service level as well as to prices. Actually, low prices that attract customers to a facility may in turn induce large waiting times that will deter customers and shift them to the competition. Alternatively, the smaller number of clients buying high-priced items might be offset by the better experience associated with lower waiting times. In such an environment, the firm that makes location and pricing decisions must take into account not only the price and location attributes of its competitors, but also the user-optimized behaviour of its potential customers, who patronize the facility that maximizes their individual utility. This situation fits the framework of a Stackelberg game and is best formulated as a bilevel program or, more generally, a mathematical program with equilibrium constraints (MPEC). At the upper level, the firm makes revenue-maximizing location and pricing decisions, taking into account the user equilibrium resulting from those decisions.

The resulting MPEC, which involves highly nonlinear queueing terms, as well as continuous (user flows) and discrete (location decision) variables, looks formidable. The aim of this paper is to show that it is yet amenable to a strategy that involves approximation by a tractable mixed-integer linear program. The paper's contributions are fourfold:

- The integration of location, service rates and prices as decision variables within a user-choice process based on service level, queueing and pricing considerations.
- The integration of congestion and competition in the context of mill pricing, i.e., prices that can vary between facilities.
- The explicit modelling of the queueing process that takes place at the facilities.
- The design of an efficient heuristic algorithm based on mixed discrete, continuous linear approximations and reformulations.

The remainder of this paper is organized as follows. In Sect. 1.1, we provide an overview of the existing facility location and pricing literature. Section 2 is devoted to the model, while, in Sect. 3, we describe the algorithmic framework. Numerical experiments and a discussion of our results are reported in Sect. 4. Finally, in Sect. 5, we draw conclusions and mention possible extensions of the current work.

1.1 Literature review

In this section, we outline works that are relevant to ours, either from the modelling (facility location, pricing, user equilibrium) or from the computational (bilevel programming, MPECs) points of view. For a more complete overview on facility location and pricing, one may refer to Eiselt et al. (2015).

Although the facility location problem (FLP) has a rich history, most works disregard user behaviour related to congestion and competition; i.e., similar users are assigned to a single path leading to the facility they patronize. While some models incorporate congestion in the form of capacity limits, more elaborate ones capture congestion through nonlinear functions that can be derived from queueing theory.

With respect to congestion, an early model can be found in Desrochers et al. (1995), who studies a centralized facility location problem where travel time increases with traffic, and users are assigned in a way that minimizes the total delay and costs. Towards the end, the authors mention a bilevel user-choice version of their model, but do not provide a solution algorithm. Within the same centralized framework, Fischetti et al. (2016) propose a Benders decomposition method for a capacitated FLP. Similarly, Marianov (2003) formulates a model for locating facilities in a centralized system subject to congestion, and where demand is elastic with respect to travel time and queue length. Users are assigned to centres that maximize total demand. In Castillo et al. (2009), users are assigned to facilities so as to minimize the sum of the number of waiting customers and the total opening and service costs. Similar to Marianov (2003), Berman and Drezner (2006), Aboolian et al. (2008) and Aboolian et al. (2012) consider models characterized by elastic demand, subject to constraints on the waiting time at facilities. Moreover, in Zhang et al. (2010) a model maximizing the participation rate is considered, in a preventive healthcare setting, when demand is elastic and users choose the facilities to patronize based on the waiting and travel time. Note that neither of the above papers consider competition or pricing.

With respect to competitive congested facility location problems (CC–FLP), we mention the work of Marianov et al. (2008), who were the first to address congestion within a competitive user-choice environment. Similarly, Sun et al. (2008) consider a generic bilevel facility location model, in which the upper level selects locations with the aim of minimizing the sum of total cost and a congestion function, while the lower level (users) minimizes a nonlinear cost. Both papers employ heuristics for solving their model. A more recent development is that of Dan and Marcotte (2019), who solve the competitive congested FLP using matheuristics and approximation algorithms. The present work can be considered a pricing extension of Dan and Marcotte (2019). Moreover, Ljubić and Moreno (2018) address a market share-maximization competitive FLP, where captured customer demand is represented by a multinomial logit model. The authors solve this problem using two branch-and-cut techniques, namely outer approximation cuts and submodular cuts.

The pricing literature is vast. Actually, many authors have addressed joint location and pricing problems, the common practice being to operate in a hierarchical manner: locations are specified first, and then price competition is defined according to the Bertrand model (Pérez et al. 2004; Panin et al. 2014). This approach can be justified by the fact that location decisions are frequently made for the long term, while prices may fluctuate in the short term. However, determining the pricing strategy after the locations have been set limits the price choices and can yield suboptimal locations, as argued in Hwang and Mai (1990), Cheung and Wang (1995) and Aboolian et al. (2008). A joint decision is more suited in some practical applications and can provide valuable insights into whether or not entry into a market is profitable.

To the best of our knowledge, the first paper to consider simultaneous decisions on location, price and capacity is Dobson and Stavrulaki (2007), who investigate a monopolistic market where a firm sells a product to customers located on the Hotelling line (Hotelling 1929). In his PhD thesis, Tong (2011) considers two profit maximizing models in a network, single facility and multifacility, respectively. Competition is not present, and demand is elastic with respect to travel distance, waiting time and price.

The author analyses both a centralized system and a user-choice system. Within the same framework, Abouee-Mehrizi et al. (2011) consider a model in which demand is elastic with respect to price only, and clients spread among facilities based on proximity, according to a multinomial logit random utility model. Congestion, which arises at facilities, is characterized by queueing equations, and customers might balk upon arrival. Furthermore, Pahlavani and Saidi-Mehrabad (2011) address a pricing problem within a user-choice competitive network. Locations are fixed, and users select the facility to patronize based on price and proximity. Also, they might balk and veer, upon observation of the queue length. The authors propose two metaheuristics for solving their model. More recent contributions are given by Hajipour et al. (2016) and Tavakkoli-Moghaddam et al. (2017), who investigate multiobjective models for the centralized facility location problem with congestion and pricing policies. Demand is elastic with respect to price and distance, while profit and congestion (waiting time, and idle probability) are decision variables. We also mention the work of Lüer-Villagra and Marianov (2013), who formulate and solve a hub location and pricing problem in a hub and spoke competitive network. An extensive review of the literature concerning competition in queueing systems is provided in Hassin (2016).

From the algorithmic point of view, our approach borrows ideas from the bilevel pricing literature, which was initiated by Labbé et al. (1998) and extended along several directions to include population heterogeneity, congestion or design, as exemplified in the papers by Meng et al. (2012) or Brotcorne et al. (2008), to name only two representative publications. We will in particular adapt a linearization technique introduced in Julsain (1999) for coping with pricing of the arcs of a packet-switched communication network.

2 Model formulation

The problem under consideration involves a firm that enters a market that is already served by competitors that can accommodate the total demand. At the upper level of the hierarchical model, a firm must make decisions pertaining to location, prices and quality of service, anticipating that users will reach an equilibrium where individual utilities are maximized. Note that, when it comes to pricing, three strategies are typically considered (Hanjoul et al. 1990):

- mill pricing: prices can vary between facilities;
- uniform pricing: all facilities charge the same price;
- discriminative pricing: customers patronizing the same facility can be charged different prices.

In the present work, we consider mill pricing, which might be the most challenging from the computational point of view. Indeed, uniform pricing involves a much smaller number of decision variables, while discriminative pricing allows for more flexibility. In the latter case, the problem would actually separate into distinct problems for each commodity, were it not for queueing at facilities.

At the lower level, customers purchase an item (this could be a service as well) at the facility where their disutility, expressed as the weighted sum of (constant) travel time, queueing delay and price, is minimized. For the sake of simplicity, facilities are modelled as $M/M/1$ queues, endowed with only one server. Nevertheless, any $M/M/s$ queues can be considered, provided that the number of servers s is fixed, and the decision variable is the service rate μ.

Our decision to adopt service rate as decision variable is motivated by the argument that it 'leads to cleaner analytical results' (Berman and Krass 2015) and that this framework makes sense in a variety of applications. A medical clinic, for instance, requires different types of personnel (doctors, nurses, machines, etc.), and it might be easier to model the number of people served per hour rather than to model each server separately. Alternatively, queues with continuous service rate provide a reasonable approximation to multiserver queues and are more tractable from computational point of view.

The assignment of users to facilities thus follows Wardrop's user equilibrium principle, i.e., disutility is minimized with respect to current flows.

We now introduce the parameters and variables of the model.

Sets

I: set of demand nodes;
J: set of candidate facility locations (leader and competitors); $J = J_1 \bigcup J_c$
J_1: set of leader's candidate sites;
J_c: set of competitors facilities;

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 price in the disutility formula;
f_c: fixed cost associated with opening a new facility;
v_c : cost per unit of service.

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 a facility $j \in J$; typically 0 if the facility is closed;
p_j: price at an open facility $j \in J$.

Additional variables

x_{ij}: arrival rate at facility $j \in J$ originating from demand node $i \in I$;
$\lambda_j = \sum_{i \in I} x_{ij}$: arrival rate at node $j \in J$;
w_j: mean queueing time at facility j.

At an open facility j, the mean waiting time in the system, w_j, is a bivariate function depending on both the arrival rate and the service rate, namely

$$w_j(\lambda_j, \mu_j) = \frac{1}{\mu_j - \lambda_j}. \qquad (1)$$

In the above, the waiting time w_j is only defined for open facilities, i.e., those for which μ_j is positive. However, one can generalize Eq. (1) to all facilities, open or not, through multiplication by $\mu_j - \lambda_j$:

$$w_j\mu_j - w_j\lambda_j = y_j. \tag{2}$$

Indeed, when facility j is closed, $y_j = \mu_j = \lambda_j = 0$, and w_j can assume any value. On the other hand, Eqs. (1) and (2) are equivalent when $y_j = 1$. Nevertheless, for simplicity and without loss of generality, we keep the original form (1) in the model and will specify in Sect. 3.3 how we deal with null service rates.

At the lower level, let γ_i denote the minimum disutility for users originating from node i. The Wardrop conditions are expressed as the set of logical constraints

$$t_{ij} + \alpha w(\lambda_j, \mu_j) + \beta p_j \begin{cases} = \gamma_i, & \text{if } x_{ij} > 0 \\ \geq \gamma_i, & \text{if } x_{ij} = 0 \end{cases} \quad i \in I;\ j \in J$$

In other words, the disutility of the paths having positive flow must be lower or equal than the utility of paths carrying no flow. These conditions can alternatively be formulated as the complementarity system

$$\begin{aligned} & t_{ij} + \alpha w(\lambda_j, \mu_j) + \beta p_j - \gamma_i \geq 0 && i \in I;\ j \in J \\ & x_{ij} \cdot \big(t_{ij} + \alpha w(\lambda_j, \mu_j) + \beta p_j - \gamma_i\big) = 0 && i \in I;\ j \in J \\ & x_{ij} \geq 0 && i \in I;\ j \in J. \end{aligned}$$

Typically, the equilibrium equations should only be enforced for open facilities. However, in our case, they are automatically satisfied for closed facilities, for the following reason: if a facility j is closed, the service rate μ_j and implicitly λ_j and x_{ij} will be null, and w_j can be set to any large value. Additionally, in our model, p_j can take any value for a closed facility (although this is suboptimal from an economic standpoint), as its contribution to the objective value is cancelled by the null terms x_{ij}. It follows that the equilibrium constraints are satisfied even for closed facilities.

Our model is as follows:

P: (Leader)

$$\max_{y,\mu,x,p,\gamma} \quad z = \sum_{i\in I}\sum_{j\in J_1} x_{ij}p_j - \sum_{j\in J_1}\big(f_c \cdot y_j + v_c \cdot \mu_j\big) \tag{3}$$

$$\text{s.t.} \quad \mu_j \leq M_1 \cdot y_j \qquad j \in J_1 \tag{4}$$

$$y_j \in \{0, 1\} \qquad j \in J_1 \tag{5}$$

$$\mu_j \geq 0 \qquad j \in J_1 \tag{6}$$

(Users)

$$t_{ij} + \alpha w(\lambda_j, \mu_j) + \beta p_j - \gamma_i \geq 0 \qquad i \in I;\ j \in J \tag{7}$$

$$x_{ij} \cdot \big(t_{ij} + \alpha w(\lambda_j, \mu_j) + \beta p_j - \gamma_i\big) = 0 \qquad i \in I;\ j \in J \tag{8}$$

$$w_j\mu_j - w_j\lambda_j = y_j \qquad j \in J \tag{9}$$

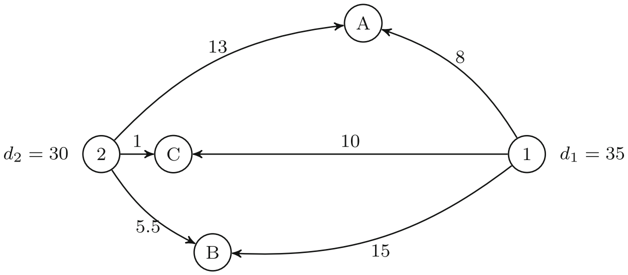


Fig. 1 Example of a two-demand node network, two location candidate sites

$$\lambda_j = \sum_{i \in I} x_{ij} \qquad j \in J \tag{10}$$

$$\sum_{j \in J} x_{ij} = d_i \qquad i \in I \tag{11}$$

$$\lambda_j \leq \mu_j \qquad j \in J \tag{12}$$

$$x_{ij} \geq 0 \qquad i \in I; j \in J. \tag{13}$$

The decision variables are the location vectors y (binary) and service rate μ (continuous).

The user assignment x is the solution of an equilibrium problem that can be reduced to a convex optimization problem. The leader's objective in Eq. (3) is to maximize the difference between the total profit and the opening and service costs. Constraint (4) ensures that the service rate is strictly positive only at open facilities. When $y = 1$, it also helps strengthen the formulation by computing a tight value for M_1 such that μ values yielding solely negative profit are eliminated.

Constraints (7), (8) and (13) characterize the user equilibrium problem, where γ_i is the optimal disutility that users originating from node i are willing to incur. Typically, the equilibrium equations should only be enforced for open facilities. However, we can extend these equations to all facilities, as previously explained. Finally, constraint (11) ensure that the total number of users originating from a demand point amounts to the demand associated with this node, and Eq. (12) guarantees that the arrival rate does not exceed the service rate at facility j.

For the sake of illustration, let us consider the example corresponding to the graph and data of Fig. 1, where nodes 1 and 2 are endowed with a demand of 35 and 30, respectively. The competitor's facility situated at node C operates at a service rate of 70.5 and charges a price of 12. The fixed and variable costs are set to 50 and 1, respectively, $\alpha = 20$ and $\beta = 10$. The values on the arcs represent the travel time between demand nodes and facilities. In this example, the leader opens facilities at both available sites. The profit is shown as a function of prices charged at the two facilities, for service rates set to 37.3 for A and 32.5 for B.

The associated profit curve is illustrated in Fig. 2. While it lacks the discontinuities associated with the basic network pricing problem (see Labbé et al. 1998), due to

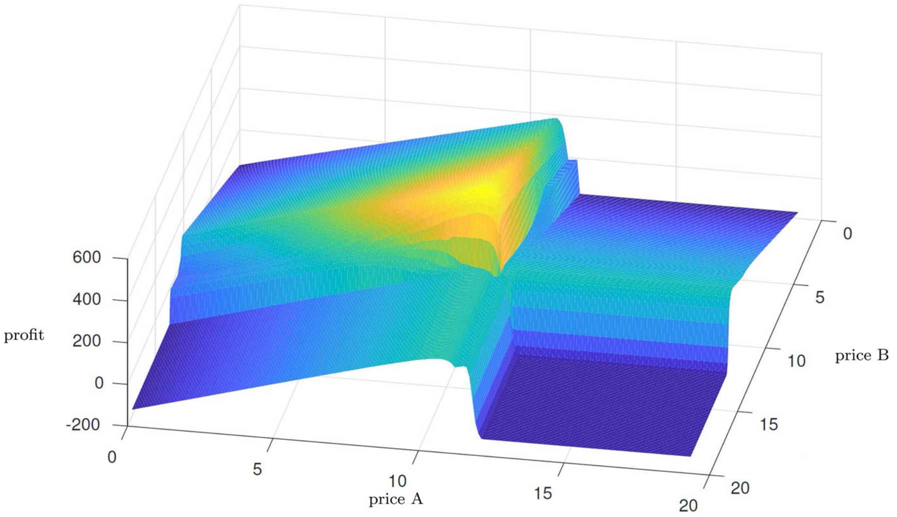


Fig. 2 Profit associated with open facilities A and B, for the network displayed in Fig. 1

the smoothing effect of the nonlinear queueing terms, it is still highly nonlinear and nonconvex.

Observation 1 *The waiting time* w_j *is jointly convex in* μ_j *and* λ_j*, for all* $\mu_j > 0, \lambda_j < \mu_j$.

3 A mixed-integer linear approximation

The general idea that underlies the algorithmic approach is to replace the original problem by a more manageable mixed-integer linear program (MILP) that we can further solve using an off-the-shelf software. This idea is not entirely novel, as it has been exploited before with different variants. For instance, in Dan and Marcotte (2019), the lower-level problem is linearized using tangent planes, before the optimality conditions are written. This yields a model containing bilinear and other nonlinear terms, which are further linearized, for instance, by using the triangle method of D'Ambrosio et al. (2010). Our approach is related to that of Julsain (1999), where univariate congestion functions are linearized in the context of a network pricing problem. In our case, concepts from network pricing and CC–FLP are merged into a single model, which makes the problem much more challenging by the presence of facility location and service level decision variables, as well as bivariate queueing delays.

The main steps of our resolution method are:

- Replace the bilinear terms in the objective with functions derived from the equilibrium constraints.
- Perform linear approximations of the complementarity constraints and the remaining nonlinear terms through the introduction of binary variables and 'big-M' constants.

- Use off-the-shelf MILP technology to solve the resulting MILP, or a carefully designed sequence of MILPs.

3.1 Reformulation of the objective function

The key issue is to eliminate the bilinear terms $x_{ij}p_j,\ j \in J_1$, in the objective, through substitution and other algebraic manipulations of the model's constraints. From Eq. (8), we have

$$x_{ij}p_j = -\frac{1}{\beta}\left(t_{ij}x_{ij} + \alpha x_{ij}w_j - x_{ij}\gamma_i\right), \quad j \in J_1, \tag{14}$$

whose summation over $i \in I$ and $j \in J_1$ leads to

$$\sum_{i\in I}\sum_{j\in J_1} x_{ij}p_j = -\frac{1}{\beta}\left(\sum_{i\in I}\sum_{j\in J_1} t_{ij}x_{ij} + \alpha\sum_{i\in I}\sum_{j\in J_1} x_{ij}w_j - \sum_{i\in I}\sum_{j\in J_1} x_{ij}\gamma_i\right), \quad j \in J_1. \tag{15}$$

The RHS of Eq. (15) now contains linear and nonlinear terms. We can simplify some of the most 'complicating' ones, namely the bilinear $x_{ij}\gamma_i$, as follows.

$$\sum_{i\in I}\sum_{j\in J_1} x_{ij}\gamma_i = \sum_{i\in I}\left(\sum_{j\in J} x_{ij}\gamma_i - \sum_{j\in J_c} x_{ij}\gamma_i\right), \tag{16}$$

and since $J = J_1 \cup J_c$ and $J_1 \cap J_c = \emptyset$,

$$\sum_{i\in I}\sum_{j\in J_1} x_{ij}\gamma_i = \sum_{i\in I} d_i\gamma_i - \sum_{i\in I}\sum_{j\in J_c} x_{ij}\gamma_i. \tag{17}$$

For the bilinear terms $x_{ij}\gamma_i$ in the RHS of Eq. (17), we write the following equations, derived from Eq. (8):

$$x_{ij}\gamma_i = t_{ij}x_{ij} + \alpha x_{ij}w_j + \beta x_{ij}p_j, \quad i \in I, j \in J_c \tag{18}$$

or, equivalently,

$$\sum_{i\in I}\sum_{j\in J_c} x_{ij}\gamma_i = \sum_{i\in I}\sum_{j\in J_c}\left(t_{ij}x_{ij} + \alpha x_{ij}w_j + \beta x_{ij}p_j\right). \tag{19}$$

Recall that the price is fixed at competitors' facilities (i.e., $j \in J_c$), so $x_{ij}p_j$ is not a bilinear term when $j \in J_c$. Then, the only nonlinear terms in the RHS of Eq. (19) are $x_{ij}w_j$. Putting together Eqs. (15), (17) and (19) yields:

$$\sum_{i\in I}\sum_{j\in J_1} x_{ij}p_j = -\frac{1}{\beta}\left(\sum_{i\in I}\sum_{j\in J} t_{ij}x_{ij} + \alpha\sum_{j\in J}\frac{\sum_{i\in I}x_{ij}}{\mu_j - \sum_{i\in I}x_{ij}} \right.$$
$$\left. - \sum_{i\in I} d_i\gamma_i + \beta\sum_{i\in I}\sum_{j\in J_c} p_j x_{ij}\right)$$

and, since $\lambda_j = \sum_{i\in I} x_{ij}$, the objective function can be written as

$$z = -\frac{1}{\beta}\sum_{i\in I}\sum_{j\in J} t_{ij}x_{ij} - \frac{\alpha}{\beta}\sum_{j\in J}\frac{\lambda_j}{\mu_j - \lambda_j} + \sum_{i\in I}\frac{d_i}{\beta}\gamma_i$$
$$-\sum_{i\in I}\sum_{j\in J_c} p_j x_{ij} - \sum_{j\in J_1}\left(f_c\cdot y_j + v_c\cdot\mu_j\right). \tag{20}$$

All terms in Eq. (20) are linear, with the exception of $\lambda_j/(\mu_j - \lambda_j)$. Additionally, these terms are undefined for $\mu_j = 0$. We overcome these issues during the linearization process, as mentioned in Sect. 3.3. We now discuss some of their properties.

Proposition 3.1 *Each term* $\frac{\alpha}{\beta}\frac{\lambda_j}{\mu_j - \lambda_j}$ *is*

- *convex in* λ_j*, and convex in* μ_j
- *neither convex, nor concave jointly in* λ_j *and* μ_j *(see Fig. 3).*
- *pseudolinear jointly in* λ_j *and* μ_j*.*

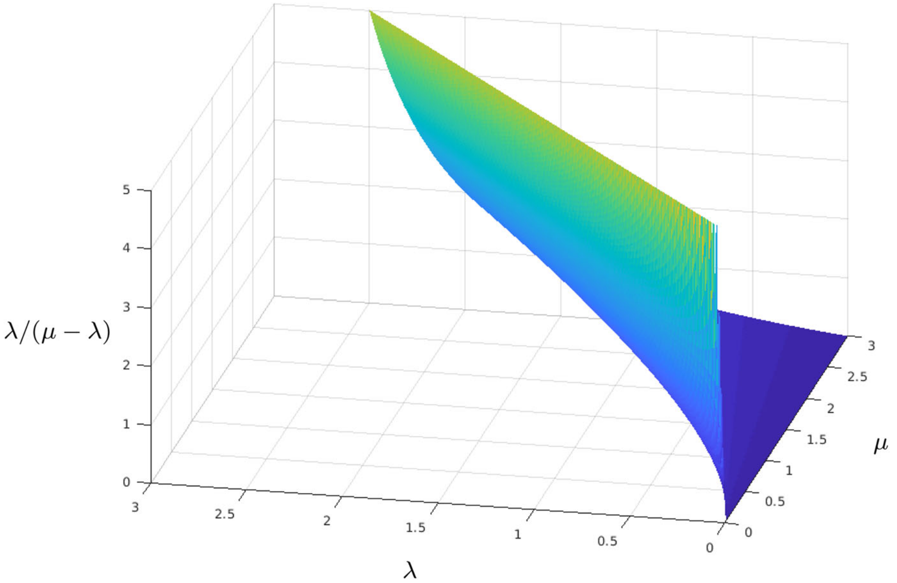


Fig. 3 Function $\lambda/(\mu - \lambda)$. Although neither convex nor concave, it is pseudolinear (pseudoconvex, and pseudoconcave). The nonconvexity is more accentuated in the vicinity of the origin

Proof The first statement is obvious. The proof of the second rests on the fact that the Hessian of the function $f(x, y) = x/(y - x)$ is indefinite. As for the pseudolinearity claim, let us consider pseudoconcavity first. The gradient of f is

$$\nabla f(x, y) = \left(\frac{y}{(y-x)^2}, \ \frac{-x}{(y-x)^2} \right)$$

Let $a = (x_a, y_a)$ and $b = (x_b, y_b)$, such that $\nabla f(a) \cdot (b - a) \geq 0$. We have that

$$\nabla f(a) \cdot (b-a) = \left(\frac{y_a}{(y_a - x_a)^2}, \ \frac{-x_a}{(y_a - x_a)^2} \right) \cdot (x_b - x_a, y_b - y_a) = \frac{y_a x_b - x_a y_b}{(y_a - x_a)^2} \tag{21}$$

and

$$\frac{y_a x_b - x_a y_b}{(y_a - x_a)^2} \geq 0 \Rightarrow y_a x_b - x_a y_b \geq 0. \tag{22}$$

We now proceed by contradiction. Let us assume that $f(b) < f(a)$. Then, $x_b/(y_b - x_b) < x_a/(y_a - x_a)$. This means that $x_b y_a - x_a y_b < 0$ and $x_b y_a - x_a y_b \geq 0$ by Eq. (22), a contradiction. This implies that

$$\nabla f(a) \cdot (b-a) \geq 0 \Rightarrow f(a) \leq f(b), \tag{23}$$

as required.

Using the same arguments, we can prove the pseudoconvexity of $-f$ and the pseudolinearity of $(\alpha/\beta)(\lambda_j)/(\mu_j - \lambda_j)$ follows. □

3.2 Bounds on *w*, *p* and *μ*

Special attention is paid to tight bounds on the variables, since these will improve the numerical efficiency of the resolution algorithm. Based on the parameters of the network, we can derive upper and lower bounds for the waiting time at facilities, the prices set by the emerging firm and the service rate profitable for the leader. It is obvious that in order to make nonnegative profit, the minimum price that the leader can set must exceed the variable cost v_c associated with the service rate

$$p_{\min} = v_c.$$

Let $(x', \lambda', w', \gamma')$ be the solution of the lower level problem under a competing oligopoly. Then, the maximum disutility that users originating from node i are willing to incur in order to access service is

$$\gamma'_i = \max_{j \in J_c} \left\{ t_{i,j} + \alpha w'_j + \beta p_j \right\}.$$

The equilibrium constraints guarantee that the above equation is satisfied even when the new firm enters the market. Then, for all couples (i, j) that have positive flows, the associated utility cannot exceed γ'_i

$$t_{i,j} + \alpha w_j + \beta p_j \leq \gamma_i',$$

and the bounds on p and w follow directly

$$\begin{aligned} w_j &\leq (u_{\max} - \beta p_{\min})/\alpha, & p_j &\leq u_{\max}/\beta \\ w_{\max} &= (u_{\max} - \beta p_{\min})/\alpha, \text{ and} & p_{\max} &= u_{\max}/\beta, \end{aligned} \tag{24}$$

where $u_{\max} = \max_{i\in I}\{\gamma_i'\}$.

The service rate at any given facility is limited by the service cost, the maximum price, fixed cost and total demand. The maximum possible profit of the firm is obtained when all the demand is attracted, the maximum price is charged and only one facility is open (fixed cost is minimal). Since the profit (objective function) must be nonnegative, we must have that

$$p_{\max} \sum_{i\in I} d_i - f_c - \mu_{\max} v_c \geq 0,$$

and the upper bound on μ follows directly:

$$\mu_{\max} = \frac{p_{\max}}{v_c} \sum_{i\in I} d_i - \frac{f_c}{v_c}.$$

3.3 Linear approximation

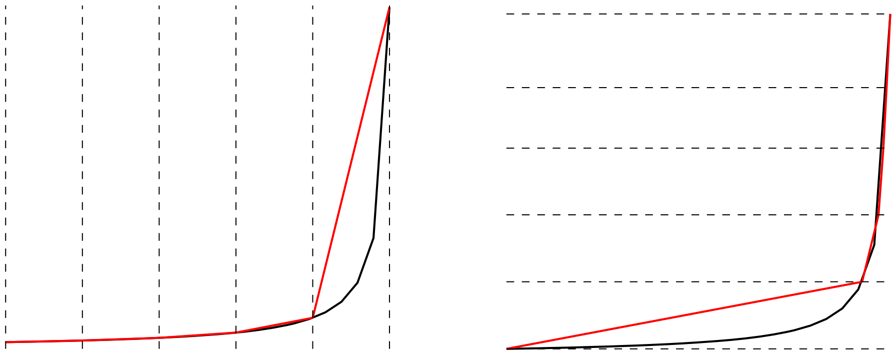
This section is devoted to a detailed description of the techniques that allow to transform the original problem into a mixed-integer linear program.

Sampling We performed piecewise linear interpolations of the nonlinear functions involved in our model, namely $\lambda_j/(\mu_j - \lambda_j)$ and $1/(\mu_j - \lambda_j)$. These functions are bivariate for the leader (μ is a decision variable) and univariate for the competitors.

For the leader, we consider $N_\mu + 1$ equidistant sampling points on the x axis, within the interval $[0, \mu_{\max}]$: $\{\tilde{\mu}^n\}$, $n \in \{1, \ldots, N_\mu\}$ such that $\tilde{\mu}^i < \tilde{\mu}^j$ for all $1 \leq i < j \leq N_\mu$. Next, for each sample $\tilde{\mu}^n$, we define $\lambda^n_{\max} = \tilde{\mu}^n - 1/w_{\max}$, and we sample each interval $[0, \lambda^n_{\max}]$ using N_λ points $\{\tilde{\lambda}^{nk}\}$, $k \in \{1, \ldots, N_\lambda\}$, such that $\tilde{\lambda}^{ni} < \tilde{\lambda}^{nj}$ for all $1 \leq i < j \leq N_\lambda$. A similar sampling is performed for every facility of the competitor, where the sampling interval for λ is $[0, \mu_j]$, $\forall j \in J_c$.

Special attention is paid to the type of sampling we use for λ. The sampling can be equidistant either 'horizontally' or 'vertically'. In the 'horizontal' case, for a given $\tilde{\mu}^n$ the difference between two consecutive values, $\tilde{\lambda}^{ni} - \tilde{\lambda}^{ni+1}$, remains constant. In contrast, in the vertical case, the samples are computed such that, for a given $\tilde{\mu}^n$, and for any two consecutive λ samples, $\tilde{\lambda}^{ni}$ and $\tilde{\lambda}^{ni+1}$, the difference between their respective waiting time, $1/(\tilde{\mu}^n - \tilde{\lambda}^{ni}) - 1/(\tilde{\mu}^n - \tilde{\lambda}^{ni+1})$, is constant. Both cases are illustrated in Fig. 4.

When using samples that are equidistant on the x axis, the approximation of waiting times is best on the region where the slope is small. It is important that this function be well approximated in this area, as a small change in the waiting time value would cause a significant change in the x-variable and thus approximate badly the resulting objective function. On the other hand, a rougher approximation of the congested part



(a) sampling equidistant on x axis **(b)** sampling equidistant on y axis

Fig. 4 Illustration of the impact of sampling type on the approximation

would not yield a large error in the x-values, which justifies performing the sampling equidistant on y axis.

Piecewise linearization We now detail the linear approximation of the terms $\frac{\lambda_j}{\mu_j - \lambda_j}$ in the reformulated objective function, and $\frac{1}{\mu_j - \lambda_j}$ in constraints (9). To this end, we use the sampling described above in a triangle piecewise linearization technique from D'Ambrosio et al. (2010). At a given point $(\tilde{\lambda}, \tilde{\mu})$, the function of interest is approximated by a convex combination of the function values at the vertices of the triangle containing the point $(\tilde{\lambda}, \tilde{\mu})$.

First, we approximate $\frac{\lambda_j}{\mu_j - \lambda_j}$ and $\frac{1}{\mu_j - \lambda_j}$ for the leader, using the following sets of variables:

- $\underline{l}_{j,n,k}$ and $\bar{l}_{j,n,k}$ are binary variables denoting the lower and upper triangles, respectively, used for evaluating the convex combinations for $n \in \{1, \ldots, N_\mu\}$, $k \in \{1, \ldots, N_\lambda\}$, $j \in J$. In a feasible solution, these variables equal 1 if the point of interest falls inside their associated triangle, and 0 otherwise.
- $\bar{s}_{j,n,k}$ represents the weight of the convex combination associated with the vertices of the triangle containing the point of interest.
- $\bar{u}$ and $\bar{w}$ hold the approximated values of $\frac{\lambda_j}{\mu_j - \lambda_j}$ and $\frac{1}{\mu_j - \lambda_j}$, respectively.

The following constraints allow to linearize $\frac{\lambda_j}{\mu_j - \lambda_j}$ and $\frac{1}{\mu_j - \lambda_j}$ in the original model. Since they are not defined for $\mu_j = 0$, by convention, we set them to 0, whenever $\mu_j = 0$. The motivation is that users cannot patronize a facility offering no service, yielding a null waiting time at facilities. To accommodate this, the summation starts at $n = 2$ in constraints (30) and (31).

$$\sum_{n=1}^{N_\mu} \sum_{k=1}^{N_\lambda} \left(\bar{l}_{j,n,k} + \underline{l}_{j,n,k} \right) = 1 \qquad j \in J_1 \tag{25}$$

$$\bar{s}_{j,n,k} \le \bar{l}_{j,n-1,k} + \underline{l}_{j,n-1,k-1} + \bar{l}_{j,n,k} + \underline{l}_{j,n,k} + \bar{l}_{j,n-1,k-1} + \underline{l}_{j,n,k-1} \qquad j \in J_1;\, n \in \{1, \dots, N_\mu\};\, k \in \{1, \dots, N_\lambda\} \tag{26}$$

$$\sum_{n=1}^{N_\mu} \sum_{k=1}^{N_\lambda} \bar{s}_{j,n,k} = 1 \qquad j \in J_1 \tag{27}$$

$$\lambda_j = \sum_{n=1}^{N_\mu} \sum_{k=1}^{N_\lambda} \bar{s}_{j,n,k} \tilde{\lambda}^{nk} \qquad j \in J_1 \tag{28}$$

$$\mu_j = \sum_{n=1}^{N_\mu} \sum_{k=1}^{N_\lambda} \bar{s}_{j,n,k} \tilde{\mu}^{n} \qquad j \in J_1 \tag{29}$$

$$\overline{w}_j = \sum_{n=2}^{N_\mu} \sum_{k=1}^{N_\lambda} \frac{1}{\tilde{\mu}^n - \tilde{\lambda}^{nk}} \cdot \bar{s}_{j,n,k} \qquad j \in J_1 \tag{30}$$

$$\overline{u}_j = \sum_{n=2}^{N_\mu} \sum_{k=1}^{N_\lambda} \frac{\tilde{\lambda}^{nk}}{\tilde{\mu}^n - \tilde{\lambda}^{nk}} \cdot \bar{s}_{j,n,k} \qquad j \in J_1 \tag{31}$$

$$\bar{l}_{j,n,k}, \underline{l}_{j,n,k} \in \{0, 1\} \qquad j \in J_1;\, n \in \{1, \dots, N_\mu\};\, k \in \{1, \dots, N_\lambda\} \tag{32}$$

$$0 \le \bar{s}_{j,n,k} \le 1 \qquad j \in J_1;\, n \in \{1, \dots, N_\mu\};\, k \in \{1, \dots, N_\lambda\} \tag{33}$$

$$\bar{l}_{j,n,0} = 0, \quad \underline{l}_{j,n,0} = 0 \qquad j \in J_1;\, n \in \{0, \dots, N_\mu\}. \tag{34}$$

$$\bar{l}_{j,n,N_\lambda} = 0, \quad \underline{l}_{j,n,N_\lambda} = 0 \qquad j \in J_1;\, n \in \{0, \dots, N_\mu\} \tag{35}$$

$$\bar{l}_{j,0,k} = 0, \quad \underline{l}_{j,0k} = 0 \qquad j \in J_1;\, k \in \{0, \dots, N_\lambda\} \tag{36}$$

$$\bar{l}_{j,N_\mu,k} = 0, \quad \underline{l}_{j,N_\mu,k} = 0 \qquad j \in J_1;\, k \in \{0, \dots, N_\lambda\}. \tag{37}$$

We perform a similar linearization for the competitor. Recall that, in this case, the service rate, μ_j, is constant. We introduce variables, $\hat{l}$, $\hat{s}$ $\hat{w}$ and $\hat{u}$, having similar meaning to their leader counterparts. Given $w_{\max}$, we compute $\hat{\lambda}^j_{\max} = \mu_j - 1/w_{\max}$, and we sample the interval $[0, \hat{\lambda}^j_{\max}]$ using N_c points $\hat{\lambda}^{jn}$, $n \in \{1, \dots, N_c\}$ such that $\hat{\lambda}^{jn} < \hat{\lambda}^{jm}$ for all $1 \le n < m \le N_c$, and obtain the linearization

$$\sum_{n=1}^{N_c} \hat{s}_{j,n} = 1 \qquad j \in J_c \tag{38}$$

$$\lambda_j = \sum_{n=1}^{N_c} \hat{s}_{j,n} \hat{\lambda}^{jn} \qquad j \in J_c \tag{39}$$

$$\hat{w}_j = \sum_{n=1}^{N_c} \frac{1}{\mu_j - \hat{\lambda}^{jn}} \cdot \hat{s}_{j,n} \qquad j \in J_c \tag{40}$$

$$\hat{u}_j = \sum_{n=1}^{N_c} \frac{\hat{\lambda}^{j,n}}{\mu_j - \hat{\lambda}^{jn}} \cdot \hat{s}_{j,n} \qquad j \in J_c \tag{41}$$

$$\sum_{n=1}^{N_c} \hat{l}_{j,n} = 1 \qquad j \in J_c \tag{42}$$

$$\hat{s}_{j,n} \leq \hat{l}_{j,n} + \hat{l}_{j,n-1} \qquad j \in J_c;\ n \in \{1, \ldots, N_c\} \tag{43}$$

$$\hat{l}_{j,n} \in \{0, 1\} \qquad j \in J_c;\ n \in \{1, \ldots, N_c\} \tag{44}$$

$$0 \leq \hat{s}_{j,n} \leq 1 \qquad j \in J_c;\ n \in \{1, \ldots, N_c\} \tag{45}$$

$$\hat{l}_{j,0} = 0, \hat{l}_{j,N_c} = 0 \qquad j \in J_c. \tag{46}$$

At last, the complementarity constraints Eq. (8) are linearized through the introduction of binary variables and big-M constants as follows:

$$t_{ij} + \alpha \overline{w}_j + \beta p_j - \gamma_i \leq M_{2,i} s_{ij} \qquad i \in I;\ j \in J_1 \tag{47}$$

$$t_{ij} + \alpha \hat{w}_j + \beta p_j - \gamma_i \leq M_{2,i} s_{ij} \qquad i \in I;\ j \in J_c \tag{48}$$

$$x_{ij} \leq M_{3,i}(1 - s_{ij}) \qquad i \in I;\ j \in J \tag{49}$$

$$s_{ij} \in \{0, 1\} \qquad i \in I;\ j \in J. \tag{50}$$

The values of $M_{2,i}$ and $M_{3,i}$ must be sufficiently large not to forbid feasible solutions, but not too large that they slow down the enumeration algorithm, due to a weak continuous relaxation. Based on the network's parameters, the following 'tight' values for $M_{2,i}$ and $M_{3,i}$ hold:

$$\begin{aligned} M_{2,i} &= \max_{j \in J}\{t_{ij}\} + \alpha w_{\max} + \beta p_{\max} \\ M_{3,i} &= d_i. \end{aligned}$$

Putting together all linear terms yields the following MILP approximation of P:

PL:

$$\begin{aligned} \max_{y,c,x,\gamma} \quad & z = -\frac{1}{\beta} \sum_{i \in I} \sum_{j \in J} t_{ij} x_{ij} - \frac{\alpha}{\beta} \sum_{j \in J_1} \overline{u}_j - \frac{\alpha}{\beta} \sum_{j \in J_c} \hat{u}_j + \sum_{i \in I} \frac{d_i}{\beta} \gamma_i - \sum_{i \in I} \sum_{j \in J_c} p_j x_{ij} - \sum_{j \in J_1} \left(f_c \cdot y_j + v_c \cdot \mu_j \right) \\ \text{s.t.} \quad & t_{ij} + \alpha \overline{w} + \beta p_j - \gamma_i \geq 0 \qquad i \in I;\ j \in J_1 \\ & t_{ij} + \alpha \hat{w} + \beta p_j - \gamma_i \geq 0 \qquad i \in I;\ j \in J_c \\ & \text{constraints (4)}-(6), (10)-(13), (25)-(50). \end{aligned} \tag{51}$$

An interesting feature of this reformulation–linearization is that, since we use the same set of variables and constraints to approximate two different functions simultaneously, the number of variables is greatly reduced. This would not be the case if we were to linearize separately the waiting time and the bilinear terms $x_{ij} p_j$ present in the original formulation.

Table 1 Instances characteristics for each problem size

	Problem size		
	15 nodes	20 nodes	25 nodes
No of demand nodes	15	20	25
No of location nodes	15	20	25
No of competitor nodes	4–5	5–8	6–8
Travel time	0–150	0–150	0–150
Demand rate	1–50	1–70	1–50
Competitor service rate	1–120	25–210	20–110
Competitors prices	8–20	9–25	6–20

The last four rows display the range of uniform random variables

Another interesting feature of this reformulation is the pseudolinearity of the functions replacing the bilinear terms in the objective. Although we do not exploit this property directly, we expect the linearization to be well behaved.

Finally, an alternative algorithmic approach based on the power-based transformation originally proposed in Teles et al. (2011) was initially implemented but did not perform satisfactorily. The main idea is to transform nonlinear polynomial problems into an MILP, by applying a term-wise disaggregation scheme, notwithstanding, with additional discrete and continuous variables. Kolodziej et. al incorporate this technique into a global optimization algorithm for bilinear programs (Kolodziej et al. 2013). The authors argue that this technique scales better than the piecewise McCormick envelopes and is comparable with global optimization solvers.

For the sake of completeness, and to warn other researchers tempted by that path, we thought it is useful to mention it. The interested reader can find it in the appendix of this Ph.D. thesis (Dan 2018).

4 Experimental setup and results

The algorithm has been tested on randomly generated data. We focused on challenging instances, in which, at optimality, the number of open facilities represents more than one-fifth of the nodes. Our experiments have been conducted on synthetic data, where ten instances (numbered 0–9) were generated for each problem size, the latter defined as the total number of nodes (location and demand). Table 1 displays the network features for 15-, 20- and 25-sized problems. The competitors' service rates have been adjusted such that the entire demand could be satisfied before the entrance of the new firm. In order to generate challenging instances, the combinations of fixed and variable costs were chosen such that there exist feasible solutions yielding nonnegative profit involving a large (more than half) number of open facilities.

Travel times were generated as follows. First, the networks were split into 4 to 7 components, and the distances between nodes were set to random values in the interval (0, 50). Next, we ensured that the graph was connected by setting the distance between random pair of nodes to some large number (100). This operation was performed such

Table 2 CPU time (seconds) on 15-node networks for different number of samples

Test #	# of samples (λ and μ)					
	5	10	20	30	40	60 (gap%)
1	4	9	25	9473	1363	14,205
2	14	20	110	398	3883	86,409 (10.95)
3	9	26	30	361	19,837	86,404 (9.97)
4	4	32	172	13,814	21,694	34,066
5	6	18	149	11,025	52,951	73,124
6	2	5	54	5982	18,408	86,402 (1.03)
7	5	15	92	18,006	8831	86,402 (3.94)
8	3	11	51	3535	9160	86,402 (1.86)
9	2	10	88	30,486	24,153	86,402 (8.22)
10	1	9	52	8010	9830	1406
Average	5	14	82	10,109	17,011	64,122 (3.60)

as to generate challenging instances where the optimal solution would involve more than one open facility.

All procedures were implemented in Java, and the MILP formulations were solved by IBM CPLEX Optimizer version 12.6. The tests were performed on a computer equipped with 96 GB of RAM, and two 6-core Intel(R) Xeon(R) X5675 processors running at 3.07 GHz. The default values of the parameters α and β were set to 20 and 10, respectively, unless specified otherwise. In all tests, the maximum tree size was set to 30 GB. Throughout this section, the *estimated objective* refers to the MILP objective value as returned by CPLEX, whereas the *recovered objective* is computed as follows: decision variables (locations, service levels and prices) were set to the optimal values found by CPLEX, and the convex lower-level optimum was solved to optimality, using Frank–Wolfe algorithm. The obtained solution (arrival rates at facilities) were then plugged into the objective function of the original formulation in order to obtain its associated objective value.

4.1 Solving the MILP with different number of samples

An initial set of experiments was intended to assess the performance of the linear approximation method. The algorithm was stopped as soon as the optimality gap dropped below CPLEX's default value (10^{-4}), the 86,400 s (24 h) limit was reached, or the tree size exceeded 30GB. Tables 2, 3 and 4 report the CPU needed, for various number of approximating samples or linear segments. The relative MILP gap is shown in percentage, next to the CPU. The gap is omitted if the algorithm terminated at optimality (i.e., gap $< 10^{-4}$).

For five and ten samples, the algorithm needs less than 100 s, and on average less than 35 s, to reach optimality. The CPU increases abruptly with the number of samples, which is to be expected. For 15-node networks, all tests finished at optimality when the number of samples is lower than 60. However, six over ten instances exceeded

Table 3 CPU time (seconds) on 20-node networks for different number of samples

Test #	# of samples (λ and μ)				
	5	10	20 (gap%)	30 (gap%)	40 (gap%)
1	22	94	1459	64,348 (0.30)	86,402 (5.17)
2	6	15	1297	59,626	77,542
3	12	52	86,401 (3.60)	86,403 (0.95)	86,402 (2.04)
4	7	24	1035	1853	86,401 (0.24)
5	13	20	86,402 (0.27)	86,402 (6.12)	86,401 (4.76)
6	7	13	782	86,402 (0.13)	52,097 (0.75)
7	6	27	228	30,892	86,401 (1.73)
8	7	20	305	2462	28,330
9	20	78	86,401 (0.07)	86,401 (0.04)	86,402 (6.71)
10	3	9	146	86,401 (0.56)	18,096
Average	10	35	26,446 (0.39)	59,119 (0.81)	69,447 (2.14)

Table 4 CPU time (seconds) on 25-node networks for different number of samples

Test #	# of samples (λ and μ)				
	5	10	20 (gap%)	30 (gap%)	40 (gap%)
1	3	5	143	86,402 (0.59)	22,702 (0.48)
2	9	23	259	5891 (2.25)	86,403 (3.97)
3	2	11	233	78,143 (0.50)	37,895 (1.15)
4	8	32	86,401 (0.73)	25,010 (0.84)	16,177 (2.51)
5	8	24	86,413 (0.76)	86,401 (4.18)	86,403 (5.14)
6	4	12	58,331	68,406 (2.43)	86,403 (2.27)
7	3	24	86,402 (2.40)	15,545 (3.08)	7,864 (3.88)
8	5	30	9650	86,405 (3.12)	71,371 (2.50)
9	2	16	170	6633 (0.69)	68,635 (0.54)
10	3	17	9127 (0.36)	86,402 (1.57)	8,789 (4.10)
Average	4	19	33,713 (0.43)	54,524 (1.93)	49,264 (2.65)

the allotted time or memory when using 60 samples. For larger, 20-node networks, the algorithm terminated at optimality on very few instances, when using more than 30 samples, and ran out of time on all 25-node network instances. Figure 5 displays the algorithm's average behaviour over all 20- and 25-node instances, respectively. Both charts suggest that the good solutions are found in the early stages, while the remaining steps are used to close the gap and prove optimality.

The increase in the running time is compensated by an improvement in the approximation quality, as illustrated in Fig. 6. These charts show averages over all instances that were able to find feasible solutions on all tests, within 24 h. For this reason, instances 9 and 5 were removed from the 20- and 25-node tests, respectively. The difference between the estimated (MILP) optimal objective value and the recovered

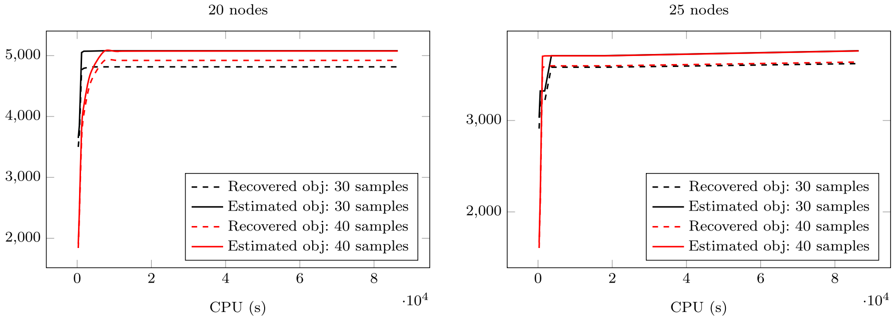


Fig. 5 Variation of estimated (MILP) and recovered objective value with respect to time

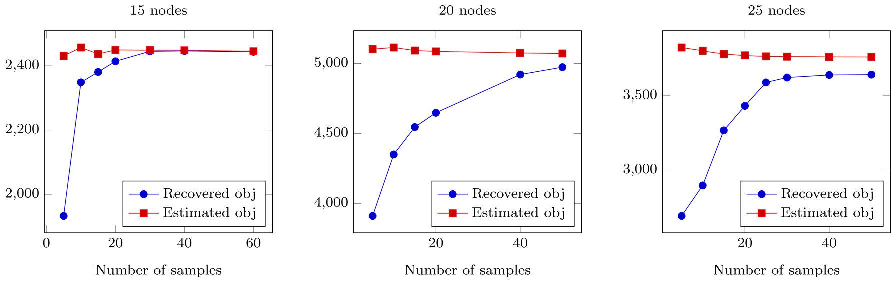


Fig. 6 Evolution of the MILP objective value ('Estimated') and the true objective value ('Recovered'), as the number of samples increases

one is decreasing with the increase in the number of samples, suggesting a solid improvement in the quality of the approximation. Note that not all instances finished to optimality, but they were within 6 relative gap.

4.2 A matheuristic approach

After careful inspection of the solutions, we have noticed that the number of facilities opened at optimality does not vary significantly with the number of samples, nor with the allotted execution time (on average around 5–7 are opened for the 20 and 25-node instances). This suggests that quasi-optimal locations are found on the early stages of the algorithm, or for coarse approximations.

Next, we assessed the quality of these opened facilities, restricting the problem to the determination of price and service levels, which remains a difficult nonlinear bilevel problem. We now solve the linearized problem PL using the following algorithm whose main steps are:

I. Solve PL for a small number of samples and a limited time.
II. Retrieve the locations associated with the incumbent.
III. Solve PL, where locations are fixed from step II, using a more fine-grained sampling, for a limited time.
IV. Retrieve the associated solution (μ and p) and compute the lower-level equilibrium required to obtain the true objective. This last operation can be achieved

Table 5 Objective value comparison on 20-node networks, when 40 samples are used for linearization, locations are fixed and the CPU is limited to 1 h in total (including the warm start)

Test #	40 samples, 1 h in total			40 samples, 1 h	50 samples, 1 h	50 samples, 24 h
	From 5 samples	From 10 samples	From 30 samples, 30 min			
1	3454.01	3454.01	3454.01	345.14	–	3455.85
2	4931.14	4931.14	4931.14	4931.14	4933.98	4933.98
3	10,083.30	10,083.30	10,091.46	10,091.46	–	10,145.76
4	4892.30	4892.30	4892.30	4892.30	4887.66	4887.66
5	5106.06	5862.84	5788.60	5757.25	6219.17	6201.88
6	4200.60	4200.60	4200.60	4200.60	4227.83	4227.83
7	4398.22	4398.22	4201.22	4345.16	4401.96	4401.96
8	3141.79	3141.79	3141.79	3141.79	3154.11	3154.11
9	3318.63	3318.63	3318.63	3291.84	3325.85	3354.89
10	–	–	–	–	–	–

by solving a convex program. To this purpose, we implemented the classical Frank–Wolfe algorithm.

This matheuristic version of our algorithmic approach has been tested on instances involving 5, 10 and 30 samples, and a time limit of 30 min, at step I, and 40 samples and a time limit of 1 h in total, for all three steps. Tables 5 and 6 show the comparison between the values obtained in this way, and the objective values yielded by the original algorithm for 40- and 50-sample approximations, with running time limited to 1 h, and a 50-sample approximation running for 24 h, for 20- and 25-node networks, respectively.

For the 20-node networks, the best performance corresponds to the ten-sample starting point. On one instance, it outperformed the 50-sample approximation, and on eight instances, it falls, on average, within 2.4% of the optimum found by the latter, at a much smaller computational cost (1 h for the ten-sample starting point as opposed to 24 h for the 50 samples). On most tests, the deviation is less than 1%, but the average is increased by an outlier (instance # 5) that has an error of 11%. The five- and 30-sample starting point yields similar results. In almost all cases in which the 40- and 50-sample algorithm finds an initial solution in 1 h, such a solution is as good, or even better than the 40-sample boosted by the ten-sample locations. However, the boosted version looks more robust.

Table 6 tells a similar story about the 25-node networks. On almost half of the instances, the 30-sample starting point outperforms the 50-sample approximation, and on the other half of instances, it falls, on average, within 0.3% of the optimum, and at a much smaller computational cost (1 h for the 40-sample starting point as opposed to 24 h for the 50 samples). When the 40- and 50-sample algorithm finds an initial solution in 1 h, such a solution is equally good, or even better than the 40 samples boosted by the 30 samples locations.

Table 6 Objective value comparison on 25-node networks, when 40 samples are used for linearization, locations are fixed and the CPU is limited to 1 h

Test #	40 samples, 1 h in total			40 samples, 1 h	50 samples, 1 h	50 samples, 24 h
	From 5 samples	From 10 samples	From 40 samples, 30 min			
1	2783.93	2783.93	2820.28	2820.28	2840.15	2840.15
2	3653.74	3751.08	3775.86	3775.86	–	3775.44
3	3531.34	3531.34	3549.39	3549.39	3550.34	3550.34
4	3477.32	3477.32	3482.76	3482.76	–	3482.60
5	3793.96	3841.20	3849.36	3849.36	3793.38	3849.02
6	3211.12	3211.12	3223.18	3223.18	–	–
7	3401.53	3441.98	3450.50	3450.50	3427.26	3452.59
8	2881.09	2881.09	2881.09	2881.07	–	2883.04
9	4590.49	4590.49	4590.49	4590.49	4590.80	4592.41
10	4277.79	4277.79	4304.77	4353.92	4347.62	4347.62

These results demonstrate that ‘good’ locations are found in the initial stages of the algorithm. From an execution time point of view, it is advantageous to stop the algorithm early on, retrieve the locations and then solve for optimal service levels and prices, using a limited number of samples, for a small running time.

4.3 Comparison with general-purpose solvers

Finally, we compare our linear approximation algorithm with a general-purpose solvers for mixed-integer nonlinear optimization problems, such as BARON. We have measured the objective values yielded by BARON, and we compare them with the results of our reformulation technique run for 1000s.

Next, we attempted to improve the solutions found by our algorithm, using IPOPT, an open-source software for large-scale nonlinear optimization based on a primal–dual interior-point algorithm (Wächter and Biegler 2006). For this experiment, we fixed the locations given by a 30-sample approximation within 1 h, yielding a fully continuous restricted problem. We have warm-started IPOPT with the respective 30-sample price, service levels and user flows. The results are shown in Table 7.

All BARON and IPOPT tests were run for 1000 s on the NEOS server, on computers equipped with 64 GB of RAM, and processors running at a frequency between 2.2 and 2.8 GHz.[1]

Our reformulation technique clearly outperforms BARON on all instances. IPOPT is capable of improving the initial solution only in three instances while, on the others, the solution worsens significantly. On one instance, marked with * in the table, the

[1] A detailed description of the NEOS server computers’ specifications can be found here https://neos-guide.org/content/FAQ.

Table 7 Objective value comparison with BARON and IPOPT on 20-node networks

	30 samples (1000 s)	BARON (1000 s)	y from 30 samples, 1 h, IPOPT (1000 s)
1	3454.28	3330.10	3139.12
2	4932.44	4444.79	3625.29
3	9926.58	9385.23	6147.16
4	4891.93	4323.95	3053.51
5	5336.68	4446.12	5195.17
6	4105.17	3901.16	3965.02
7	4426.14	3789.63	* −6419.41
8	3093.31	2550.13	2852.44
9	3215.63	2666.95	2374.10
10	4053.45	2689.02	667.08

objective value is negative, despite being warm started with a good (positive) solution, likely indication of numerical difficulties.

5 Conclusions

In this paper, we addressed a highly nonlinear bilevel pricing location model involving both combinatorial and continuous elements and proposed for its solution an algorithm based on reformulation and piecewise linear approximations.

Our results are encouraging, but our algorithms have some limitations. For instance, one of the remaining challenges is to design algorithms that scale well and can be applied successfully on large networks.

Future work could integrate other realistic features, such as variable demand. On the algorithmic side, an interesting development could be a method that exploits the pseudolinearity property of the nonlinear terms present in the reformulated objective function.

References

Aboolian R, Berman O, Krass D (2008) Optimizing pricing and location decisions for competitive service facilities charging uniform price. J Oper Res Soc 59(11):1506–1519. https://doi.org/10.1057/palgrave.jors.2602493

Aboolian R, Berman O, Krass D (2012) Profit maximizing distributed service system design with congestion and elastic demand. Transp Sci 46(2):247–261. https://doi.org/10.1287/trsc.1110.0392

Abouee-Mehrizi H, Babri S, Berman O, Shavandi H (2011) Optimizing capacity, pricing and location decisions on a congested network with balking. Math Methods Oper Res 74(2):233–255

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

Berman O, Krass D (2015) Stochastic location models with congestion. Springer, Cham, pp 443–486. https://doi.org/10.1007/978-3-319-13111-5_17

Brotcorne L, Labbé M, Marcotte P, Savard G (2008) Joint design and pricing on a network. Oper Res 56:1104–1115. https://hal.archives-ouvertes.fr/hal-01255555. Language of publication: en

Castillo I, Ingolfsson A, Sim T (2009) Socially optimal location of facilities with fixed servers, stochastic demand and congestion. Prod Oper Manag 18(6):721–736

Cheung FK, Wang X (1995) Spatial price discrimination and location choice with non-uniform demands. Reg Sci Urb Econ 25(1):59–73

D'Ambrosio C, Lodi A, Martello S (2010) Piecewise linear approximation of functions of two variables in MILP models. Oper Res Lett 38(1):39–46

Dan T (2018) Algorithmic contributions to bilevel location problems with queueing and user equilibrium: exact and semi-exact approaches. Ph.D. thesis, University of Montreal

Dan T, Marcotte P (2019) Competitive facility location with selfish users and queues. Oper Res 67(2):479–497. https://doi.org/10.1287/opre.2018.1781

Desrochers M, Marcotte P, Stan M (1995) The congested facility location problem. Locat Sci 3(1):9–23

Dobson G, Stavrulaki E (2007) Simultaneous price, location, and capacity decisions on a line of time-sensitive customers. NRL 54(1):1–10. https://doi.org/10.1002/nav.20169

Eiselt HA, Marianov V, Drezner T (2015) Competitive location models. Springer, Cham, pp 365–398. https://doi.org/10.1007/978-3-319-13111-5_14

Fischetti M, Ljubić I, Sinnl M (2016) Benders decomposition without separability: a computational study for capacitated facility location problems. Eur J Oper Res 253(3):557–569. https://doi.org/10.1016/j.ejor.2016.03.002

Hajipour V, Farahani RZ, Fattahi P (2016) Bi-objective vibration damping optimization for congested location-pricing problem. Comput Oper Res 70(C):87–100. https://doi.org/10.1016/j.cor.2016.01.001

Hanjoul P, Hansen P, Peeters D, Thisse JF (1990) Uncapacitated plant location under alternative spatial price policies. Manag Sci 36(1):41–57. https://doi.org/10.1287/mnsc.36.1.41

Hassin R (2016) Rational queueing. CRC Press, Boca Raton

Hotelling H (1929) Stability in competition. Econ J 39(153):41–57

Hwang H, Mai CC (1990) Effects of spatial price discrimination on output, welfare, and location. Am Econ Rev 80(3):567–575

Julsain H (1999) Tarification dans les réseaux de télécommunications [microforme] : une approche par programmation mathématique à deux niveaux. Canadian theses. Thèse (M.Sc.A.)–École polytechnique de Montréal. https://books.google.ca/books?id=5uiKtgAACAAJ

Kolodziej S, Castro PM, Grossmann IE (2013) Global optimization of bilinear programs with a multiparametric disaggregation technique. J Glob Optim 57(4):1039–1063. https://doi.org/10.1007/s10898-012-0022-1

Labbé M, Marcotte P, Savard G (1998) A bilevel model of taxation and its application to optimal highway pricing. Manag Sci 44(12):1608–1622. https://doi.org/10.1287/mnsc.44.12.1608

Ljubić I, Moreno E (2018) Outer approximation and submodular cuts for maximum capture facility location problems with random utilities. Eur J Oper Res 266(1):46–56

Lüer-Villagra A, Marianov V (2013) A competitive hub location and pricing problem. Eur J Oper Res 231(3):734–744. https://doi.org/10.1016/j.ejor.2013.06.006

Marianov V (2003) Location of multiple-server congestible facilities for maximizing expected demand, when services are non-essential. Ann Oper Res 123(1–4):125–141. https://doi.org/10.1023/A:1026171212594

Marianov V, Ríos M, Icaza MJ (2008) Facility location for market capture when users rank facilities by shorter travel and waiting times. Eur J Oper Res 191(1):32–44

Meng Q, Liu Z, Wang S (2012) Optimal distance tolls under congestion pricing and continuously distributed value of time. Transp Res Part E Logist Transp Rev 48(5):937–957. https://doi.org/10.1016/j.tre.2012.04.004 **(Selected papers from the 14th ATRS and the 12th WCTR Conferences, 2010)**

Pahlavani A, Saidi-Mehrabad M (2011) Optimal pricing for competitive service facilities with balking and veering customers. Int J Innov Comput Inf Control 7:3171–3191

Panin AA, Pashchenko M, Plyasunov AV (2014) Bilevel competitive facility location and pricing problems. Autom Remote Control 75(4):715–727

Pérez MDG, Hernández PF, Pelegrín BP (2004) On price competition in location-price models with spatially separated markets. Top 12(2):351–374. https://doi.org/10.1007/BF02578966

Sun H, Gao Z, Wu J (2008) A bi-level programming model and solution algorithm for the location of logistics distribution centers. Appl Math Model 32(4):610–616

Tavakkoli-Moghaddam R, Vazifeh-Noshafagh S, Taleizadeh AA, Hajipour V, Mahmoudi A (2017) Pricing and location decisions in multi-objective facility location problem with m/m/m/k queuing systems. Eng Optim 49(1):136–160. https://doi.org/10.1080/0305215X.2016.1163630

Teles JP, Castro PM, Matos HA (2011) Multi-parametric disaggregation technique for global optimization of polynomial programming problems. J Glob Optim 55(2):227–251. https://doi.org/10.1007/s10898-011-9809-8

Tong D (2011) Optimal pricing and capacity planning in operations management. Ph.D. thesis

Wächter A, Biegler LT (2006) On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Mathe Program 106(1):25–57. https://doi.org/10.1007/s10107-004-0559-y

Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Trans 42(12):865–880

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.