BILEVEL PROGRAMMING:
THE MONTREAL SCHOOL

Luce Brotcorne* Patrice Marcotte† Gilles Savard‡

May 27, 2008

Abstract

Although leader-follower games have been around for quite some time in economics and game theory, their integration into the realm of optimization and operations research is recent. In this paper, we provide an overview of research in bilevel programming that was initiated at the University of Montreal and led to a large scale application in the field of revenue management.

Keywords. Bilevel programming. Networks. Pricing. Revenue Management. Combinatorial optimization.

1 Introduction

In the simplest version of the 'cake-cutting game'[1], Bob cuts a Sachertorte into two parts, knowing that Alice will select the larger piece. If Bob knows Alice's behaviour then, in order to maximize the size of his portion, it is self-obvious that he should cut the cake into two equal parts. This trivial game, which can be extended to several sequential players, can be modelled as a bilevel program, i.e., an optimization problem where the leader (Bob) integrates within his decision process the mathematical program of the follower (Alice). Depending on situations, it might or might not be advantageous to be the leader. In the cake-cutting game, one may safely argue that Alice has the edge, if one assumes that is impossible to perform a 'perfect' cut.

Bilevel programs are pervasive. For instance, politicians must anticipate the behaviour of electors in order to get (re)elected, while private firms must predict the reaction of customers and competing firms in order to maximize their revenues. Note that, in both these situations, the role of Alice is played by a group of agents (more on this latter). Actually, it is rather the] exception that a decision problem can be solved *in vitro*, with no exterior and rational agent in control of a subset of decision variables. Then, why has so little attention been paid so far to bilevel programming? Indeed, bilevel programs arise in the economics under various names, such as Stackelberg games, envy-free pricing, product line design, etc. Only fairly recently has the field found its niche within the mathematical programming community. Conferences devoted entirely to this topic might be few and far between, but one will encounter sessions on bilevel programming in most operational research, optimization and mathematical programming conferences, and even theoretical computer science colloquia, where researchers have analyzed the complexity of bilevel programs, whether they are labelled as such or not.

In short, Bilevel Programming is a fairly recent branch of optimization that deals with programs whose constraints embed an auxiliary mathematical program.

*LAMIH/ROI, Université de Valenciennes et du Hainaut-Cambrésis (France)

†DIRO, Université de Montréal

‡École Polytechnique (Montréal)

[1] We thank Gilbert Laporte for reminding us this paradigm.

The objective of this paper is not to review the vast body of literature devoted to bilevel programming and the related MPECs (to be introduced in the next section). For this, the reader is best referred to [7, 24, 45], a list that is not exhaustive. Rather, it is to provide an overview of the research on this topic conducted by a team mainly based in Montreal in the past thirty years, mainly focusing on (but not limited to) problems involving a network structure.

The paper is organized as follows. Following a formal description of the bilevel framework, we initially focus on algorithms. Next, we consider applications in network design, energy modelling and network pricing, the latter begin at the core of a real life model of revenue management for network-based industries.

2 Problem definition

A bilevel program is expressed mathematically as

$$\begin{aligned} \mathsf{BP}: \qquad & \max_{x,y} && f(x,y) \\ & \text{subject to} && (x,y) \in X \\ & && y \in S(x), \end{aligned}$$

where $S(x)$ denotes the set of *global* solutions of a mathematical program parameterized in x, i.e.,

$$S(x) = \arg\min_{y' \in Y(x)} g(x,y'). \tag{1}$$

In this hierarchical setting, the upper level player selects first its decision vector x, anticipating the reaction of the lower level player. This framework subsumes convex, nonconvex and mixed integer programming, and allows for the natural modelling of situations involving agents that are not under the full control of the optimizer. Obviously, addressing bilevel programs in their generality is tantamount to solving a global optimization problem. For this reason, some structure needs to be imposed in order that they be amenable to numerical algorithms. The mildest condition might be that, for given x, the lower level program is easy to solve, e.g., involves a linear, quadratic, convex or network structure. Indeed, much research has been devoted to the linear case, where all functions involved are affine, and that can be reformulated as a mixed integer program. Another class of interest is that of bilevel programs involving a convex lower level problem. If the latter satisfies some constraint qualification condition, it can be characterized by its Kuhn-Tucker conditions, thus yielding a 'standard' single-level program involving complementarity constraints, and examplifying the dual nature, continuous and combinatorial, of bilevel programs.

Bilevel programs are closely related to Stackelberg (leader-follower) games [52], to the principal-agent paradigm [53] in economics, as well as to equilibrium constrained mathematical programs (MPECs), where the lower level problem characterizes the equilibrium state of some physical or economical system, and is frequently modelled as a variational inequality. In this context, $y \in Y(x)$ belongs to $S(x)$ if and only if it solves the parametric variational inequality

$$\mathsf{PVI}: \qquad \langle F(x,y), y - y' \rangle \leq 0 \qquad \forall y' \in Y(x). \tag{2}$$

If the mapping F is the gradient of a convex function, then an MPEC reduces to a bilevel program. Actually, this result holds even if F is not a potential, since a feasible vector y is a solution of the variational inequality PVI if and only if it is a global minimizer of the (usually nonconvex) gap function defined as

$$g(x,y) = \max_{y' \in Y(x)} \langle F(x,y), y - y' \rangle.$$

3 Numerical approaches for the linear, quadratic and convex cases

Although few situations can be naturally modelled as linear bilevel programs (LBPs), the latter has nevertheless been investigated by several researchers. The term 'bilevel program' has been coined by Candler and

Norton [20] to denote such problems, that are formulated as

$$\begin{array}{lrl} \mathsf{LBP}: & \displaystyle\min_{x,y} & c_1x + d_1y \\ & \text{subject to} & A_1x + B_1y = b_1 \\ & & x \geq 0 \\ & & \\ & \displaystyle\min_{y} & c_2x + d_2y \\ & \text{subject to} & A_2x + B_2y = b_2 \\ & & y \geq 0, \end{array}$$

where, for the sake of notational clarity, the 'arg min' operator has been left out of the lower level formulation. Note that, without loss of generality, the term c_2x can be ignored, as it influences the value of the lower level objective, but not its actual solution.

To characterize the solution of LBP the following definitions, that also apply in a wider context, will be useful.

The **feasible set** of LBP is defined as

$$\Omega = \{(x, y) : x \geq 0, y \geq 0, A_1x + B_1y = b_1,\ A_2x + B_2y = b_2\}.$$

For every $x \geq 0$, the **feasible set of the lower level problem** is

$$\Omega_y(x) = \{y : y \geq 0,\ B_2y = b - A_2x\}.$$

The **trace** of the lower level problem with respect to the upper level variables is

$$\Omega_x^2 = \{x : x \geq 0,\ \Omega_y(x) \neq \emptyset\}.$$

For a given vector $x \in \Omega_x^2$, the **set of optimal solutions of the lower problem** is

$$S(x) = \arg \min_{y \in \Omega_y(x)} d_2y.$$

A point (x, y) is said to be **rational** if $x \in \Omega_x^2$ and $y \in S(x)$. The **optimal value function** is defined over Ω_x^2 as

$$v(x) = d_2y, \quad y \in \mathcal{S}(x),$$

where $x \in \Omega_x^2$ The **admissible set** (also called **induced region**) is

$$\Upsilon = \{(x, y) : x \geq 0,\ A_1x + B_1y = b_1,\ y \in S(x)\}.$$

Finally, a point (x, y) is **admissible** if it is feasible and lies in $S(x)$.

Whenever the upper level constraints involve no lower level variables, rational points are also admissible. Note that the converse statement may fail to hold in the presence of joint upper level constraints. Based on the above definitions, an admissible point (x^*, y^*) is **optimal** for LBP if, for every other admissible point (x, y), there holds $c_1x^* + d_1y^* \leq c_1x + d_1y$.

LBPs, despite their apparent simplicity, are computationally challenging. After Jeroslow [39] initially showed that LBP is $\mathcal{NP}$-hard, Hansen et al. [34] proved strongly $\mathcal{NP}$-hardness, using a reduction from KERNEL (see [31]). Vicente et al. [54] strengthened these results and proved that merely checking strict or local optimality is also strongly $\mathcal{NP}$-hard using a reduction from 3-SAT. While complexity results may be obtained from the connection with bilinear programs, it is instructive to perform reductions directly from standard combinatorial problems.

As a simple example of the reduction technique, consider the 0-1 linear program

$$\begin{array}{rl} \min\limits_{x} & cx \\ \text{subject to} & Ax = b \\ \text{for all } j & x_j \in \{0,1\}. \end{array}$$

Let us introduce an auxiliary vector u and form the LBP

$$\begin{array}{rl} \min\limits_{x,u} & cx \\ \text{subject to} & \sum\limits_j u_j = 0 \\ \max\limits_{u} & \sum\limits_j u_j \\ \text{subject to} & u_j \leq x_j \\ \text{for all } j & u_j \leq 1 - x_j \\ & u_j \in [0,1]. \end{array}$$

In the above, the upper level constraint $\sum_j u_j = 0$ can only be satisfied by setting x_j to either 0 or 1. Alternatively, to avoid dealing with disconnected sets that arise when upper level constraints are present in the formulation, one might simply append to the objective a term $K \sum_j u_j$, where this penalty scheme is exact in the sense that there exists a finite value K^* such that the solutions of the penalized and original problems coincide whenever K exceeds K^*. For classical combinatorial problems (knapsack, traveling salesman, maximal clique, kernel, etc.), it can be shown that the size of K, i.e., its logarithm, can be polynomially related to the size of the combinatorial problem being considered.

Reversely, an LBP can be formulated as a mixed binary problem. This is achieved by replacing the lower level problem by its optimality conditions, and linearizing the complementarity terms in the standard fashion (wlog, c_2 is set to zero). This yields:

$$\begin{array}{rl} \min\limits_{x,y} & c_1 x + d_1 y \\ \text{subject to} & A_1 x + B_1 y = b_1 \\ & A_2 x + B_2 y = b_2 \\ & x, y \geq 0 \\ & \lambda B_2 \leq d_2 \\ \text{for all } j & y_j \leq M u_j \\ & (d_2 - \lambda B_2)_j \leq M(1 - u_j) \\ & u_j \in \{0,1\}, \end{array}$$

where the next-to-last inequalities force the complementarity $(\lambda B_2 - d_2)y = 0$ to be satisfied, and where the 'big-M' constant M must be sufficiently large in order not to restrict the feasible domain of LBP.

The interest in these reformulations goes beyond the complexity issue. Indeed, Audet et al. [5] have uncovered equivalences between algorithms designed to solve mixed integer programs and LBP. More precisely, they have shown that the HJS algorithm (Hansen et al. [34]) designed for solving the LBP can be mapped onto a standard branch-and-bound method for addressing an equivalent mixed 0-1 program, provided that mutually consistent branching rules are implemented (e.g. [9]). One may therefore claim that the mixed 0-1 algorithm is subsumed (the authors use the term *embedded*) by the bilevel algorithm. This result shows that the structures of both problems are virtually indistinguishable, and that any algorithmic improvement on one problem can readily be adapted to the other (Audet et al. [5]). Solution techniques developed for solving mixed 0-1 programs may thus be tailored to the LBP, and vice versa.

Over the past 15 years, we have studied many such reformulations and developed efficient approaches for related combinatorial problems. Figure 1 (Audet [4]) illustrates the interrelations between many combinatorial and bilevel problems. For instance, the figure indicates that the undefined quadratic problem with quadratic constraints (QQP) and the general bilinear problem (BIL) are more general then the mixed-integer bilevel program (MILBP_{0-1}), and that LBP is more general then the bilinear disjoint problem (BILD), linear max-min (LMM) problem and quadratic concave problem (QP_{+}). The figure also indicates that the generalized linear complementarity problem (GLCP) can be transformed into a LBP without the use of a large finite constant (indicated by the single arrow), while the LBP can only be transformed into a mixed integer program (MIP_{0-1}) with the use of a large constant (indicated by the double arrow). Based on these equivalences, new algorithmic approaches have been obtained for the bilinear disjoint problems [1], the max-min problems [3] and the traveling salesman problem [48].

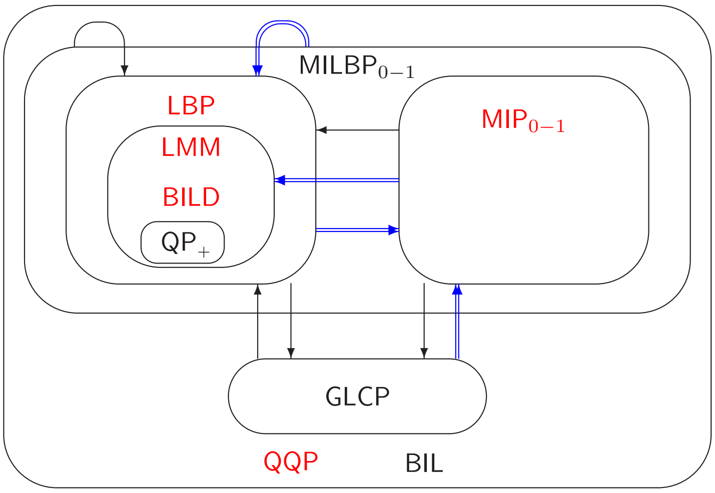

Figure 1: Some reformulations

The difficulty in solving the bilevel problem is mainly due to the nonconvexity of the solution map that corresponds to the solution set of the lower level problem. Although the global optimization community has been very prolific over the past twenty years or so (see e.g. [50]), nonconvexity is still frequently synonym of intractibility, and the success in solving a nonconvex problem depends on our ability to exploit its specific structure. For bilevel problems, the nonconvexity arises mainly from the complementarity constraints associated with the optimality conditions of the lower level problem[2]. When the optimality conditions lead to linear complementarity constraints, such for linear or quadratic bilevel programs, the nonconvexity comes solely from the disjunctive nature of the constraints and combinatorial approaches such as branch-and-bound, branch-and-cut or branch-and-price techniques are suitable. Efficient approaches to linear and quadratic bilevel programming are, expectedly, based on these techniques.

We now proceed to describe an exact branch-and-bound algorithm for solving LBPs proposed by Hansen et al. [34]. Given an LBP, one can replace the lower level problem by its optimality conditions. This yields

[2] We assume that regularity conditions are satisfied for the lower level problem, and that it be easily solvable, for instance convex.

the equivalent program

$$\begin{aligned} \min_{x,y} \quad & c_1x + d_1y \\ \text{subject to} \quad & A_1x + B_1y = b_1 \\ & A_2x + B_2y = b_2 \\ & \lambda B_2 \leq d_2 \\ & (\lambda B_2 - d_2)y = 0 \\ & x \geq 0 \\ & y \geq 0. \end{aligned}$$

A direct approach to solve this single-level problem is to relax the complementarity constraints and to branch with respect to the disjunctive constraints, as proposed by Bard and Moore [8]. Note that, were it not for complementarity constraints, the problem would separate into an optimization problem in (x, y) and a feasibility problem in λ. The HJS algorithm takes advantage of this, working in parallel on both problems. The (x, y)-subproblem, denoted LRLBP, is a relaxation of LBP. Next, instead of addressing head-on the feasibility problem in λ, which is independent of x and provides little relevant information, we consider the follower's relaxation $\mathsf{FRBLP}(x)$, together with its dual. The latter corresponds, for given x, to the lower level program. Clearly, the feasibility problem in λ is satisfied if $\mathsf{FRBLP}(x)$ is feasible and bounded. Hence, at a given node of the branch-and-bound tree, a subset of the complementarity constraints is fixed, that is, either the corresponding variables y are fixed to 0 in LRLBP and $\mathsf{FRBLP}(x)$, or the corresponding dual constraint of the dual of $\mathsf{FRBLP}(x)$ are assumed to be tight.

Algorithm HJS makes heavy use of the monotonicity analysis. Let assume, without loss of generality, that the constraints are given in an inegality form. In this context, we associate to the i^{th} second level constraint (including the non-negativity ones) the boolean variable α_i equal to 1 if the constraint is tight and equal to 0 otherwise (note that the variables α are not explicitly part of the optimisation subproblems). Next, based on monotonicity analysis, we construct a system of relations R that must be satisfied by all rational solutions. A monotonicity relation, denoted by r_k, assumes the form $\sum_{i \in I_k} \alpha_i \geq 1$. The system is updated at each node of the tree, allowing to eventually deduce the activity status (0 or free) of other constraints or variables or to conclude that no rational solution exists for the incumbent subproblem. The relations can also be used to select a branching variable y, based on criteria such as: minimum fill rate of the system, maximum number of deductions, etc. Together with monotonicity principles, HJS makes used of penalties, a concept widely used in the field of mixed integer programming. The algorithm is outlined below.

ALGORITHM HJS

1. INITIALIZATION

   Set z_{opt} to $+\infty$. Set all variables α_j free. Set $R = \emptyset$

2. COMPUTING AN LOWER BOUND

   Let $(\bar{x}, \bar{y})$ be an optimal solution of LRLBP and $\bar{z}$ its corresponding value. If $\bar{z} \geq z_{opt}$, go to 12.

3. FEASIBILITY TEST

   Solve the dual of FRBLP$(\bar{x})$. If it is infeasible, go to 12.

4. FIRST RESOLUTION TEST

   Check if $(\bar{x}, \bar{y})$ is rational for the current subproblem. Solve FRBLP$(\bar{x})$ and let y_{FS} be an optimal solution. If $d_2\bar{y} = d_2y_{FS}$ then $(\bar{x}, \bar{y})$ is rational for the current subproblem; otherwise go to 6.

5. SECOND RESOLUTION TEST

   Check if $(\bar{x}, \bar{y})$ is rational for the initial problem. Solve the second level problem for $(\bar{x})$ fixed and let y_{FS} be an optimal solution. If $d_2\bar{y} = d_2 y_{FS}$ then $(\bar{x}, \bar{y})$ is rational; otherwise go to 6. If $\bar{z} < z_{opt}$, update z_{opt} and (x_{opt}, y_{opt}) and go to 12.

6. SECOND OPTIMALITY TEST

   Compute all penalties p_i. If penalties make the bounds of all subproblems larger or equal than the incumbent objective value, go to 12.

7. FIRST CONDITIONAL OPTIMALITY TEST

   Consider the penalties p_i. For each free variable α_i, if $z_{opt} \leq \bar{z} + p_i$, fix α_i at 0 and update R.

8. THIRD OPTIMALITY TEST

   If R contains a relation r_k such that $\alpha_i = 0$ for all $i \in I_k$, go to 12.

9. RELATIONAL OPTIMALITY TEST

   For all remaining y_j appearing in the current subproblem, append to R the nonredundant logical relations on the α_j. Eliminate from R those relations which have become redundant

10. SECOND (CONDITIONAL) OPTIMALITY TEST

    If R contains a relation r_k such that $\alpha_i = 0$ for all $i \in I_k$ except for one index i, set the corresponding α_i to 1 and return to 2.

11. BRANCHING

    Select a complementarity constraint α_i by applying a specified branching rule. Fix $\alpha_j = 1$, thus creating a new subproblem and go to 2.

    Once the subtree rooted at $\alpha_i = 1$ has been completely explored, free any variable (or constraint) fixed within that subtree. Fix $\alpha_i = 0$, hence creating a new subproblem.

    Once the subtree rooted at $\alpha_i = 0$ is completely explored, free any variable fixed within that subtree.

12. BACKTRACKING

    Fathom the current node. If it is the root node, stop; (x_{opt}, y_{opt}) is an optimal solution with optimal value z_{opt}. Otherwise, pursue branching and go to 11, at the parent node in the enumeration tree.

The above algorithmic framework was initially developed for the LBP. However, it can be generalized in various ways to other classes of bilevel programs. For instance, without any modification, it can address convex/linear problems to global optimality, given the availability of a nonlinear solver at step 2. The only restrictive requirement of the algorithm about the nature of the lower level problem concerns the linearity of the complementarity constraints and functions involved in the monotonicity analysis. For a second level quadratic problem, the nonlinearity of the objective function is not satisfied and the algorithm requires an adjustment. Indeed, Jaumard et al. [38] generalized the monotonicity analysis for the quadratic instance by adding separation based on the sign values of the elements of the gradient. Hence, they showed that the convex/quadratic bilevel program can be solved to optimality in finite time. Moreover, the algorithm can be improved to take into account the structure of specific instances. For example, the max-min and the disjoint bilinear problems can be reformulated into two different but symmetric linear bilevel programs. This has allowed to improve the branching rules and generalize the concavity cuts to the bilevel instances (see [1] and [3]). Also, the link between mixed integer programming and linear bilevel programming has been exploited to derive a class of new efficient Gomory-like cuts [6].

To conclude this section, let us mention that the convex/convex bilevel program does not have yet its 'global algorithm', as it is not known how to deal efficiently with nonlinear complementarity constraints in the realm of branch-and-bound. In contrast, Colson et al. [21] have proposed an efficient algorithm based on the trust region paradigm, whereby the 'model', although nonconvex, can be solved to global optimality. In the convex/convex case, the trust region model is a linear/quadratic bilevel program that partially captures the nonconvex features of the original model. While it cannot be assured of converging to a global solution from any starting point, it can nevertheless bypass local solutions, whenever the objective functions of the leader and the follower are mildly nonlinear and non quadratic, respectively.

4 Equilibrium Network Design

Consider the problem of designing an urban road network subject to congestion, through the choice of its link capacities, the aim being to strike the right balance between expansion costs and travel delays. Due to congestion effects, one must carefully select the capacities, in order to avoid perverse effects examplified by the Braess paradox [12], where increasing the capacity of a link (or building a new link) may result in a delay increase for *every* user of the network. Hence the natural formulation as an MPEC involving a designer whose decisions impact the path choices of the users and, in turn, the performance of the network.

Throughout this section, and elsewhere in the paper, we depart from the generic notation introduced earlier, and adopt one that keeps closer to the given application and corresponding literature. In the context of our network design problem (NDP in short), z represents the vector of link capacities, v the vector of link flows, $S(v,z)$ the vector of flow-dependent link delays, and $\phi(z)$ the cost of implementing the design z. We associate with each pair of nodes k a constant demand d_k and denote by x the vector of path flows. A demand-feasible path flow vector x is in equilibrium (in the Wardrop sense) if and only, for any node pair k and paths p and p' linking the pair k, there holds, for a given design vector z,

$$x_p > 0 \Rightarrow \sum_{a\in p} S_a(v,z) \leq \sum_{a \in p'} S_a(v,z), \tag{3}$$

where v is the arc flow vector compatible with x. Upon introduction of the set V of link flows for which there exists at least one compatible feasible path flow vector x, one may rewrite (3) as the variational inequality

$$\begin{aligned} & v \in V \\ \langle S(v,z), v - v'\rangle \quad & \leq \quad 0 \qquad \forall v' \in V, \end{aligned}$$

which merely states that, at equilibrium, the current flow assignment cannot be improved upon, i.e., all flow is directed to shortest paths with respect to the *currently observed* delays $S(v,z)$. While the set of equilibrium path flow vectors is in general not unique, it can be shown, under mild conditions, that equilibrium link flows are uniquely determined. For further details and related notions of equilibrium, the reader is referred to Marcotte and Patriksson [49].

The framework we adopt for the NDP was initially proposed by Abdulaal and Leblanc [2]. It involves continuous design variables that do not correspond to a number of lanes, but rather to the flow rate (number of vehicles per time unit) that a given link can support. At the lower level, user behaviour is assumed to be consistent with Wardrop's first principle, according to which only shortest paths may carry positive flow, as described in the preceding paragraph.

This yields naturally the MPEC

$$\begin{aligned} \mathsf{NDP}: \qquad \min_{v\in V, z\in Z} \qquad & \langle v, S(v,z)\rangle + \phi(z) \\ \text{subject to} \qquad & \langle S(v,z), v-v'\rangle \leq 0 \qquad \forall v' \in V, \end{aligned}$$

where Z denotes the set of admissible designs. In the NDP literature, it is frequently assumed that S is a gradient mapping. In particular, Marcotte [46] considered the link-separable functional forms $S_a(v,z) = S_a(v_a/z_a)$ for delay, with S_a a strictly increasing function of its argument, $\phi_a(z) = \phi_a(z_a)$ for investment, and $Z = R_n^+$. Under the first of these assumptions, the variational inequality characterizing the equilibrium reduces to a convex optimization problem, and the NDP can be rewritten as the bilevel program

$$\min_{v\in V, z\geq 0} \qquad \sum_a v_a S_a(v_a/z_a) + \phi_a(z_a)$$

$$\min_{v\in V} \qquad \sum_a \int_0^{v_a} S_a(t/z_a)\, dt.$$

Note that this framework does not take into account equity considerations. Indeed, in order to improve the overall network performance, the model might penalize users that were naturally traveling on lightly congested routes, in order to re-route users whose choices were detrimental to the system.

Even in this simplified form, the NDP is a nonconvex optimization problem for which no theoretically efficient ('polynomial') algorithm has been devised yet. Nevertheless, the fact that both players wish to minimize some sort of delay (marginal delay for the leader, actual delay for the follower) suggests a number of heuristic procedures that perform well in practice. Marcotte [46] considered four such schemes. The simplest one consists in relaxing the equilibrium constraint. This yields the 'system-optimal' problem

$$\mathsf{H1}: \qquad \min_{v\in V, z\geq 0} \qquad \sum_a v_a S_a(v_a/z_a) + \phi_a(z_a) \tag{4}$$

where the designer assigns flows according to its own desire. Let $(\bar{c}, \bar{z})$ be its solution. A feasible solution to the NDP can then be obtained by computing the equilibrium flow vector $v(\bar{z})$ corresponding to $\bar{z}$. Note that, for fixed link flow v_a, one can set to zero the derivative with respect to z_a. This yields the system-optimal relationship

$$(v_a/z_a)^2 S'_a(v_a/z_a) = \phi'_a(z_a), \tag{5}$$

whose solution, unique if ϕ_a is strictly increasing, we denote by $z_a(v_a)$. Substituting this value in $H1$ yields a standard convex flow problem

$$\min_{v\in V} \qquad \sum_a v_a S_a(v_a/z_a(v_a)) + \phi_a(z_a(v_a))$$

that can be tackled by any suitable algorithm.

Another natural approach (H2) consists in iterating between 'capacity optimization' and 'flow assignment' (equilibrium computation), hoping that this block Gauss-Seidel scheme eventually converges to a suitable solution. This parallels the classical 'cobweb' procedure for solving fixed point problems. In our context, it can be considered as an approximation of a Stackelberg game by a Nash game. Alternatively, it can be viewed as a procedure that looks for an equilibrium flow that is consistent with the system-optimal relationship (5), and can be implemented by solving the variational inequality

$$\langle S(v, z(v)), v - v' \rangle \leq 0 \qquad \forall v' \in V$$

which, under our functional assumptions, reduces again to a convex flow problem.

A third approach pursues the analysis further, and introduces a program whose solution yields automatically an equilibrium flow vector, for instance

$$\mathsf{H3}: \qquad \min_{v\in V} \qquad \sum_a \int_o^{v_a} S_a(t/z_a)\, dt + \xi_a \phi_a(z_a),$$

for some choice of the positive constants ξ_a. Unfortunately, the choice of an optimal vector ξ is of the same theoretical complexity as the solution of the NDP in the first place. For this reason, it makes sense to set all ξ_a's to some constant, predetermined value.

It is worth investigating the particular case where the delay and investment functions are both convex and increasing polynomials. Indeed, the Bureau of Public Roads function assumes this form and, in certain practical situations, it is natural to consider investment costs that are linear or convex increasing. We therefore set $S_a(u) = \alpha_a + \beta_a u_a^p$ and $\phi_a(z_a) = l_a z_a^m$. Under these assumptions, the system-optimal relationship can be solved in closed form, yielding:

$$z_a(v_a) = \left(\frac{p\beta_a}{ml_a}\right)^{1/(p+m)} v_a^{(p+1)/(p+m)}.$$

Furthermore, it is interesting that, by setting $\xi = 1/(p+1)$, one realizes that Heuristic H3 subsumes H2. Since H3 can easily be solved even for large scale networks, it makes sense to do so for several values of the parameter ξ.

Our fourth and final heuristic (H4) approach consists in determining a capacity vector that makes the system-optimal flow derived from $H1$ an equilibrium flow. In the polynomial case, this can be achieved by establishing a suitable, closed form relationship between flows and capacities.

As is the case in several combinatorial optimization problems, a worst-case analysis of the various heuristics has been conducted in Marcotte [46]. More precisely, let us define

$$R_m^p(\mathsf{H}) = \sup_{\alpha,\beta,d} \frac{\text{cost solution provided by Heuristic H}}{\text{cost of the (unknown) optimal solution}} \quad .$$

In the polynomial case, the worst-case ratio has the following properties:

- $\lim_{p\to\infty} R_1^p(\mathsf{H1}) \geq 2$
- $R_1^p(\mathsf{H2}) = p + 1$
- $R_m^p(\mathsf{H4}) = m(p+1)(p+m) + p(p+m)(p+1)^{-m/p}$
- $\lim_{m\to 0} R_m^p(\mathsf{H4}) = 1$ and $\lim_{m\to\infty} R_m^p(\mathsf{H4}) = 2$
- $1 + p/\xi(p+1) \leq R_1^p(\mathsf{H3}) \leq \xi^p/(p+1)/(p+1)^{1/(p+1)}[1 + (p/\xi)(p+1)]^2$
- $2 \leq \lim_{p\to\infty} R_1^p(\mathsf{H4}) = 4$

To our knowledge, this was the first time that such bounds were derived for a bilevel problem. Recently, much attention has been directed towards deriving worst-case bound for the so-called 'price of anarchy', i.e., the ratio between the delays associated with equilibrium and system-optimal flow patterns, respectively. For polynomial delay functions, this ratio tends to infinity as the exponent p grows. In the light of these results, it is interesting to observe that our ratio stays bounded in certain situations. Also, note that the above results hold even for values of p less than 1, which corresponds to concave investment functions, i.e., cost functions that exhibit economies of scale. In a certain way, this should not be surprising. Indeed, if this is the case, the various heuristics solve concave flow problems whose optimal solutions are extremal. It follows that capacities will be concentrated on a small number of links on which both the equilibrium and system-optimal flows will also be concentrated. For instance, if only one origin-destination pair has positive demand, most heuristics will assign the entire capacity to the links of a single path, resulting in the coincidence of equilibrium and system-optimal flow patterns.

To conclude this section we mention that a suitable definition of the investment cost function allows to model the situation where an actual network must be improved. In this case, one has simply to replace $\phi_a(z_a)$ by $\max\{0, \phi_a(z_a - c_a)\}$, where c_a represents the initial capacity of arc a. In this formulation, we do *not* enforce the constraint $z_a \geq c_a$, since it might prove to *reduce* the capacity of a certain link. See Marcotte and Marquis [47] for further details.

5 Energy Markets

The energy sector is a fertile ground for the analysis and computation of economic equilibria involving suppliers and buyers. At GERAD, a university research center based in Montreal, a research team led by Alain Haurie and Richard Loulou has made important contributions to this field. In particular, it has maintained, enhanced and extended the techno-economic model MARKAL (Fishbone and Abilock [29]). Initially developed by the International Energy Agency (IEA) and based on *Activity Analysis*, MARKAL assesses the timely investment in proven or novel technologies over a finite horizon. Decision variables are related to each other by a large number of equations that ensure the conservation of investments and energy flows, and account for a variety of constraints: resource limitations, demand satisfaction, limits on rates of technological penetrations, etc. The model is driven by the optimization of a user-supplied objective, which may take different forms: minimizing the total cost, minimizing the emission of CO_2, etc. Among the various extensions implemented by the team, let us mention: (i) the inclusion of nonlinear dependencies, such as demand elasticities, (ii) the consideration of environmental factors, (iii) the partition of a 'country' into semi-autonomous 'regions', and (iv) the inclusion of discrete decision variables. Next, the team developed new models that could fit evolving and deregulated environments. Based on game-theoretic concepts, the resulting *implementable* algorithms where put to evaluate a number of situations. One important area, which led to the adoption of the bilevel framework, was the electric power sector, which was experiencing various degrees of deregulation, ranging from mild to complete. A first application aims at describing the interaction between a typical (American) power utility and 'qualifying' small power producers (QFs), such as industrial cogeneration units, which simultaneously produce thermal and electric energy from a single primary source (Haurie et al. [35]). This case study was motivated by the Public Utility Regulatory Policies Act(PURPA), an American legislation enforced during the deregulation period, which forces the utilities to buy any excess electricity from QFs at the avoided cost (a concept closed to the marginal cost). Under PURPA, the QFs react to the average production cost of utilities by acquiring cogeneration units and selling the excess electricity (at marginal value) to utilities, which are legally forced to buy it. The marginal vs average cost rationale does not, in general, yield a socially optimal solution, i.e., a solution that maximizes the sum of the consumer surplus, the utility profit, and the profits of the QFs. Consequently, when a utility invests in production capacity, its strategy must take into account the supply quantities from QFs, which will be endogenously defined as a rational reaction of cost minimizing industrial producers. Hence the bilevel structure of the model.

More specifically, let us consider a model involving two players: the utility and the aggregated industrial QFs, each represented by its own large scale techno-economic model. Each player has to satisfy inelastic demands for electricity and steam, the latter demand begin null for a utility that does not manage a district heat system. For each demand pattern (e.g. summer, winter, day and night, peak period) and each time period, electricity demand is described by a staircase load curve. Demand for steam for various pressure or temperature levels is assumed to be known over the time horizon of the study. In order to satisfy demand, each player may increase its generating capacity by investing in available technologies, while keeping a sufficient reserve margin. Investment costs are represented by the annuities corresponding to the life cycle of the equipment, and may involve different interest rates for the utility and the QFs. Players are concerned with the minimization of their net total cost (investment + production + electricity purchases - electricity sales).

The efficiency of the regulation can then be assessed by contrasting the efficient (cooperative) equilibrium resulting from a joint optimization vs the noncooperative equilibrium compatible with PURPA. The latter was computed by solving a bilevel program by a sensitivity-based algorithm. Various scenarios were run, and it was found that cogenerators would benefit most from PURPA legislation, while end-users of the energy provided by utilities would experience rate increases.

A second application was concerned with the integrated electricity sector of the Province of Québec[3], mainly operated by a single provider (Hydro-Québec) (Lavigne et al. [43]). Its objective was to develop a mathematical program of evaluating three scenarios:

- marginal cost pricing, i.e., the optimization of the entire system, notwithstanding environmental or other externalities;
- a regulated equilibrium, where prices are provided by some function of marginal costs;

[3] At this time already a state, Québec was not yet a nation officially recognized by the federal government.

- monopolistic pricing, where prices are dictated by the single supplier/distributor , which acts as a price-setter. In this context, all consumers are represented by a single large scale model, where each consumer class sets the quantity it purchases. Consumers may switch freely from electricity to substitutable energy forms.

In the model, electricity is partitioned into several ‘commodities’ according to different periods of the year, and to ‘peak’ production. This results in a number of commodities equal to 63 corresponding to seven times the length of the horizon (set to nine). An important feature of the model is that the supply and demand curves are *not* assumed to be endogenously described by simplified, aggregated, closed-form functions. Quite the opposite, supplier and consumer choices are modelled in fine detail, via highly disaggregated dynamic process models which include technology and fuel choices as explicit determinants. There follows a very realistic description of the techno-economic structure of the market. The price to pay for this realism is of course the complex, implicit nature of the resulting supply and demand curves, together with the difficulty of computing the associated equilibria.

We now proceed with the description of a model where each agent’s behaviour is explicitly described by a mathematical program of its investment process, including peak load reserve requirements. For the purpose of this section, and considering that the amounts of exchanged commodities are fixed, each model takes the form of a large scale linear program. First, we consider the production model, based on the following notation:

x : decision variables of the producer (investments, activity levels, etc)
c_1 : vector of unit costs
s : vector of commodity exchanges between supplier and consumers
p : price vector
A_1 : techno-economic matrix relating the utility’s decision variables to production levels
P_1 : polyhedron defined by any remaining constraints.

The producer model is then stated as the linear program

$$\begin{array}{lrl} \mathsf{P}: & \min\limits_{x,p} & c_1x - ps \\ & \text{subject to} & A_1x - s = b_1 \\ & & x \in P_1 \\ & & s \geq 0, \end{array}$$

where the first constraint accounts for electricity production. The production of a marginal unit induces a cost increase equal to the shadow price λ associated with this constraint. For a given s, the vector of shadow prices λ provides the value of the (implicit) inverse supply curve at the current point.

The consumer sector is modelled as the linear program

$$\begin{array}{lrl} \mathsf{C}: & \min\limits_{y,s} & c_2y + ps \\ & \text{subject to} & A_2y + s = b_2 \\ & & y \in P_2 \\ & & s \geq 0, \end{array}$$

where we have

y: consumer decision variables (investments, capacities, etc)
c_2: costs associated to the consumer variables
A_2: techno-economic matrix relating consumer decision variables to consumption
P_2: polyhedron defined by any remaining constraints.

As before, shadow prices of the first constraint define the implicit inverse demand function, with the vector s making the connection between supply and demand. Both models being linear, the inverse supply and inverse demand curves are step-functions.

An equilibrium (p^*, s^*) is achieved when, for given s^*, the producer sets its activity vector to x^* and the price vector to p^* while, for given p^*, consumers set their activity and exchange vectors to y^* and s^* respectively. Clearly, this frameworks encompasses a large class of equilibria. For instance, if the vector p is constrained to belong to a given set $S \subset R^e$, we derive a so-called *S-equilibrium*, which is a solution of the bilevel programming problem

$$\begin{aligned}
\mathsf{S-EQ:} \qquad & \min_{x,p} && c_1x - ps \\
& \text{subject to} && A_1x - s = b_1 \\
& && x \in P_1 \\
& && p \in S \\
\\
& \min_{y,s} && c_2y + ps \\
& \text{subject to} && A_2y + s = 0 \\
& && y \in P_2 \\
& && s \geq 0.
\end{aligned}$$

Note that, if p is unrestricted, i.e., the constraint $p \in S$ does not appear, then we obtain a static Stackelberg solution that maximizes the utility's profit. For the above bilevel program, we implemented an iterative algorithm based on sensitivity analysis, which we applied to three situations:

- Pure Competition;
- Regulated Equilibrium;
- Tempered Monopoly Equilibrium.

The algorithmic approach consists in performing, locally, a piecewise linear approximation of the consumer sector's reaction curve, and feeding it to the producer model. While it is well known that the net social surplus resulting from pure competition can easily be obtained through the minimization of a convex program, we showed that the regulated one can also be obtained via a modified convex program. However, the modified producer model obtained for the third equilibrium (producer surplus) leads to an integer mathematical program. Indeed, the producer now optimizes his own objective function (cost), and not the net social surplus as he did in the pure competition case. Based on a piecewise linear three-step approximation of each of the 63 inverse demand functions, it was shown that the minimal supplier's cost may be reached for each of them only at three points, each located at a discontinuity either of the current inverse demand or inverse supply functions. This observation leads to a particularly simple formulation of the supplier's problem as an integer program.

The model can be used to assess the relative benefits of partial or total deregulation. For instance, one of its output showed how the residential market reacted to various pricing mechanisms. Interestingly, it was observed that a tempered monopoly equilibrium outperformed, from the social point of view, a heavily regulated one. In this framework, the tempered monopoly may be interpreted as the limiting case on an deregulated oligopolistic market.

6 Network Pricing

The work on network pricing of Marcotte and Savard is based on discussions with Martine Labbé that took part at the Université Libre de Bruxelles in the early 1990's, and pursued with Luce Brotcorne (Université de

Valenciennes) and graduate students located either in Brussels or Montreal. The original model of Labbé et al. [41] was concerned with the setting of tolls on a subset of arcs of an uncongested transportation network, with the aim of maximizing revenue, and taking into account that users are assigned to shortest paths whose respective costs (or lengths) are the sum of the original cost and toll. The basic model was then extended to consider congestion, joint design and pricing, population segmentation with respect to the perception of travel time, and recently the incorporation of discrete choice models. This approach led to a model of revenue management that is currently applied in the airline and rail industries, and will be discussed in the next section of the paper. But, first things first, let us introduce the basic toll setting problem.

6.1 Formulation and properties

The toll setting problem TOLL is defined over a network whose arcs are partitioned into two subsets, those of toll (A_1) and toll-free arcs (A_2) respectively. The toll subset is endowed with a toll (resp. cost) vector t (resp. c) and carries a flow vector x. The respective quantities for the toll-free subset are d and y. The TOLL problem can then be expressed as the bilinear bilevel program

$$\begin{array}{lrl} \mathsf{TOLL}: & \max\limits_{t,x,y} & t\sum\limits_{k} x^k \\ & \min\limits_{x,y} & \sum\limits_{k}(c+t)x^k + dy^k \\ & \text{subject to for all } k & Ax^k + By^k = b^k \\ & & x^k, y^k \geq 0, \end{array}$$

where $[A|B]$ corresponds to the partition of the node-arc incidence matrix of the network, and b^k is the demand vector associated with the origin-destination pair indexed by k. In the remainder, we assume that there cannot exist a toll scheme that generates revenues and creates negative cost cycles in the network, and that there exists at least one path composed of toll-free arcs for each origin-destination pair. These assumptions imply that the lower level's optimal solution corresponds to a set of shortest paths. An optimal toll scheme is thus sufficiently high to generate revenue but not too high to deter the users from taking toll arcs rather than alternative routes.

As discussed by Labbé et al. [41] the leader's revenue is neither a continuous nor a convex function of t, although a semi-continuity property ensures that the solution set is nonempty. The above formulation implies that, whenever the solution set of the lower level is not a singleton, ties are broken in the leader's favour. Since a toll schedule that induces uniqueness at the lower level and yields a revenue arbitrarily close to the optimal value can be achieved through a suitable perturbation scheme, this assumption makes sense.

As shown on the example of Figure 2, an optimal solution may involve negative tolls. Let the demand be equal to one on origin-destination pairs 1-2 and 3-4, and arcs (5,6) and (6,4) be subject to tolls. In this case compensating interactions between tolls play an active role and the optimal solution, corresponding to a revenue of 8 monetary units is reached for $T_{56} = 5$ and $T_{64} = -2$. This kind of interactions occurs for example for airline pricing problems where arcs correspond to legs in the network.

Let us now return to the formulation of TOLL. Replacing the lower level linear program by its optimality

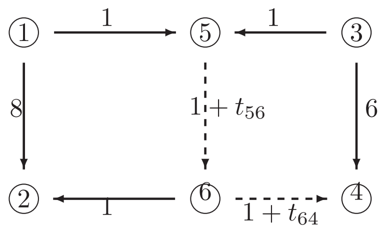

Figure 2: Negative tolls

conditions, one obtains the equivalent single level formulation

$$\begin{aligned}
\mathsf{SLP}: \qquad & \max_{t,x,y} && t\sum_k x^k && (6)\\
& \text{subject to for all } k && Ax^k + By^k = b^k\\
& && x^k, y^k \geq 0\\
& && \lambda^k A \leq c + t\\
& && \lambda^k B \leq d\\
& && (c + t - \lambda^k A)x^k = 0 && (7)\\
& && (d - \lambda^k B)y^k = 0. && (8)
\end{aligned}$$

The difficulty in solving SLP is entirely concentrated in the complementarity constraints (7)–(8). In addition to being nonlinear, these constraints may fail to satisfy any constraint qualification, indicating that the problem is numerically ill-posed.

There is a close relationship between TOLL and LBP. Indeed, through LP duality arguments, it is possible to reformulate TOLL as an LBP for which an economic interpretation has been provided by Labbé et al. [41], who also showed that there always exist extremal solutions to TOLL , and the latter is strongly NP-hard. Actually merely checking local optimality of a feasible solution is NP-hard either with (Labbé et al.) or without (Roch et al. [51]) non negativity constraints on t whenever the lower level program is degenerate. Further complexity results have been obtained by Bouhtou et al. [11, 32].

6.2 Inverse Optimization

The difficulty of the problem warrants the development of solution methods that exploit its network structure and the relationship between tolls and flows. More precisely for fixed t one may obtain the follower's optimal flows by solving shortest path problems. Conversely, for fixed x^k and y^k, tolls that maximize the leader's revenue may be obtained trough the inverse optimization program described in this section. For ease of presentation the technique will be illustrated on the single commodity problem, hence the absence of the index k in the following:

$$\begin{aligned}
\max_{t,x,y} \quad & tx\\
\text{subject to} \quad & Ax + By = b\\
& x, y \geq 0\\
& \lambda A \leq c + t\\
& \lambda B \leq d\\
& (c + t - \lambda A)x = 0\\
& (d - \lambda B)y = 0.
\end{aligned} \tag{9}$$

It is clear that, without loss of generality, one may set $t = \lambda A - c$. Actually the equality only needs to hold for the components of the flow vector x that are strictly positive. For those components whose values are zero, one is free to select any value of t that is sufficiently large with respect to (9). Some elementary algebra then yields

$$\begin{aligned} \max_{t,x,y,\lambda} \quad & \lambda b - (cx + dy) \\ \text{subject to} \quad & Ax + By = b \\ & x, y \geq 0 \\ & \lambda B \leq d \\ & (d - \lambda B)y = 0. \end{aligned}$$

Let us now consider an extremal flow solution (x, y) (this corresponds to a path in the graph), and partition y as $y = (y_0, y_+)$, where $y_0 = 0$ and $y_+ > 0$, componentwise. This leads to the linear program:

$$\begin{aligned} \max_{x,y,\lambda} \quad & \lambda b - (cx + d_+ y_+) \\ \text{subject to} \quad & Ax + B_+ y_+ = b \\ & x, y_+ \geq 0, y_0 = 0 \\ & \lambda B_0 \leq d_0 \\ & \lambda B_+ = d_+ \end{aligned}$$

which decomposes into a 'primal' (in x, y) and a 'dual' (in λ) problem. The dual problem is tantamount to solving an inverse optimization program, whereby one looks for a revenue maximizing toll vector $t = \lambda A - c$ that is consistent with a lower level solution. This development illustrates the 'pure' combinatorial nature of the problem, which is concentrated in the knowledge of nonzero toll-free flows. This feature of the problem has been exploited to design exact or heuristic approaches to the problem by Brotcorne et al. [15, 16], among others.

The dual of the inverse optimization program is defined as:

$$\begin{aligned} \min_{u} \quad & d_0 u_0 + d_+ u_+ \\ \text{subject to} \quad & B_0 u_0 + B_+ u_+ = b \\ & u_0 \geq 0 \\ & u_+ \text{ free.} \end{aligned}$$

This corresponds to a minimum cost flow problem where toll arcs have been removed, and toll free arcs that carry positive flow are bi-directional. In the multicommodity case a similar analysis yields a multicommodity network flow problem involving capacities on toll arcs.

6.3 Algorithmic approaches

Algorithmic approaches proposed in the litterature are mostly based on reformulations of TOLL as the single level optimization program SLP.

A mixed integer programming (MIP) formulation of the problem may be obtained through the introduction of binary variables z^k to lift the nonlinearities arising in constraint (7) and (8) and the objective function. To this aim, we set $\tau^k = tx^k/n^k$ and replace (7) and (8) for all $k \in K$ by the equivalent constraints

$$\begin{aligned} -Mz^k \leq \tau^k \quad & \leq \quad Mz^k & (10) \\ -M(1 - z^k) \quad & \leq \quad \tau^k - t \leq M(1 - z^k) \ , & (11) \end{aligned}$$

where M is a data-dependent, suitably large constant. The resulting formulation can be solved by an off-the-shelf solver, such as CPLEX [23]. Unfortunately, due to the large number of binary variables ($|K| \times |A|$) and the poor quality of the linear programming relaxation, the resulting MIP does not allow to solve instances of realistic sizes.

In the case where tolls must satisfy non-negativity constraints, Dewez et al. [26] have strengthened constraints (10) and (11), by tailoring the constant M to each OD pair. They also introduced two sets of cuts. The first set, derived from the complementarity constraints, improves the linear relaxation of the problem, while the second set (*galley* cuts) forces OD pairs going through nodes s and t to use a common sub-path. These cuts allow to reduce by half the integrability gap, and to sharply improve the performance of the associated branch-and-cut algorithm.

Another approach consists in penalizing the constraints (7) and (8) to derive a separable bilinear program (BILIN):

$$\begin{array}{lll} \mathsf{BILIN}: & \displaystyle\max_{t,x,y,\lambda} & \displaystyle t\sum_k x^k - M\sum_k ((c+t-\lambda^k A)x^k + (d-\lambda^k B)y^k) \\ & \text{subject to, for all } k & Ax^k + By^k = b^k \\ & & x^k, y^k \geq 0 \\ & & \lambda^k A \leq c+t \\ & & \lambda^k B \leq d. \end{array}$$

Labbé et al. [41] have shown that the penality is exact, in the sense that there exists a threshold value $\overline{M}$ such that, whenever M[4] exceeds $\overline{M}$, the *global* solutions of TOLL and BILIN coincide. It follows from known results on bilinear programming that TOLL admits optima (x, y) that are extremal points of the original primal polyhedron as claimed earlier.

Brotcorne et al. [14, 15] proposed primal-dual heuristics for solving TOLL. More precisely BILIN is sequentially solved with respect to (t, λ^k) and (x^k, y^k). For fixed (t, λ^k) each problem in $(x^k,\ y^k)$ consists in computing shortest origin-destination paths. For fixed x^k and y^k one obtains a linear program in t and λ which corresponds to the inverse optimization procedure introduced in the previous subsection. Note that alternatively the vector λ could have been set to the optimal multiplier associated with a lower level solution. In the single OD pair case as mentioned previously we can define $t = \lambda A - c$. In the general case this is not the case any more except when distinct OD pair tolls are specified.

To deal with this issue, Brotcorne et al. [15] assumed distinct tolls for every OD pair, together with a constraint specifying the equality of all OD-specific tolls. Next, they introduced a quadratic penalty on this constraint. The resulting non linear program was solved using Frank and Wolfe's linearization scheme. In the course of the algorithm, lower level solutions are generated and probed using the inverse optimization procedure. The aim of the procedure is actually not so much that of uncovering a local solution of the nonconvex master problem, as to generate a set of 'quality' lower level solutions, in the spirit of the strategy underlying metaheuristics. Corresponding tolls are then recovered through inverse optimization using a generic LP solver.

Recently, Brotcorne et al. [18] developed a tabu search heuristic based on the efficient exploration of the lower level extremal solutions (trees). The optimal toll vector corresponding to a tree solution is provided by the inverse optimization procedure introduced earlier, which is solved by an adaptation of Dantzig-Wolfe decomposition.

Within this framework, neighbours of a shortest path tree associated with an OD pair k are obtained from the pivot operation that consists in replacing one link of the tree by an out-of-tree link. At a given iteration, among all feasible moves, the one leading to the highest revenue is performed. Whenever the selected solution does not improve with respect to the best known solution , the OD pair trees update results from simplex pivots. If the revenue of the neighbour solution is larger than the previous one, the tree update results from the computation of the OD pair optimal paths associated to the tolls corresponding to the improving solution. To the best of our knowledge, the resulting scheme proved to be the most efficient procedure for addressing large scale instances.

In contrast with the arc formulation of TOLL, one may investigate formulations that integrate path variables, in conjunction or not with arc variables. One such formulation, analyzed in Didi et al. [27], is constructed around the binary variables z_p's, set to one if and only if the flow n_k corresponding to OD pair

[4] The 'large' number M does not always assume the same value. It is a generic constant used throughout the paper.

k is assigned to path $p \in P_k$, where P_k is the set of loopless paths linking the origin and destination of k. We also introduce L_k, the cost of the shortest path linking the origin and destination nodes of k. TOLL can then be expressed as the mixed integer program:

$$
\begin{aligned}
\max_{t,\tau,z,L} \quad & \sum_k n_k \tau_k \\
\text{subject to, for all } k \quad & \sum_{p \in P_k} z_p = 1 \\
\text{for all } p \in P_k \quad & z_p \in \{0,1\} \\
\text{for all } k \text{ and for all } p \quad & \tau_k \leq \sum_{a \in p \cap A_1} t_a + M(1 - z_p) \\
& L_k = \tau_k + \sum_{p \in P_k} z_p \Big(\sum_{a \in A_1} c_a + \sum_{a \in A_2} d_a \Big) \\
& L_k \geq \sum_{a \in p \cap A_1} t_a + \sum_{a \in p \cap A_1} c_a + \sum_{a \in p \cap A_2} d_a - M(1 - z_p) \\
& L_k \leq \sum_{a \in p \cap A_1} t_a + \Big(\sum_{a \in A_1} c_a + \sum_{a \in A_2} d_a \Big),
\end{aligned}
$$

where M is an arbitrarily large (finite) number. In Didi et al. [27] and Brotcorne et al. [16], it is shown how a path generation scheme, combined with the inverse optimization procedure, provides an efficient framework to solve TOLL.

More precisely, Didi et al. [27] proposed an algorithm based on a clever enumeration of path sets (so-called K-paths) whose elements are paths associated with the OD pairs, corresponding to extremal optimal solutions of the lower level problem. Given a K-path P, we denote by $v(P)$ the revenue provided by the inverse optimization procedure. Also, we define $UB(P)$ as the difference between toll free path costs and path costs obtained by setting tolls to 0 on arcs of P, and to ∞ elsewhere. It is clear that for each K-path P, $v(P)$ (respectively $UB(P)$) defines a lower bound (respectively an upper bound) on the optimal revenue. The algorithm then proceeds by generating a sequence of K-paths in decreasing order of the upper bounds and stops at the first K-path P of the sequence for which $v(P^*) \geq UB(P)$, where P^* corresponds to the best K-path found so far. In practice, the size of the sequence is limited by memory space or computing time considerations. Unfortunately, the algorithm fails to solve large instances, mainly due to the degeneracy occurring in the K-path enumeration process. Recently Brotcorne et al. [16] have decreased the upper bound $UB(P)$ and devised a generation process that obviates redundancy. This modification allowed to solve problems involving twice as many OD pairs and twice as many toll arcs.

6.4 Variants

Some particular cases dealing with special structure of the problem have been studied by Dewez et al. [25] and Heilporn et al. [37]. The Highway Pricing Problem (HPP) in which all toll arcs must be connected can represent features specific to a real highway topology. In order to allow scale economies they consider a complete toll subgraph so that every single feasible path from any origin to any destination in the network contains exactly one toll arc. For each OD pair $k \in K$ and each toll arc $a \in A_1$, let c_a^k denote the fixed cost of the unique path going from the origin to destination through toll arc a. The fixed cost on the toll free path $o^k \rightarrow d^k$ is denoted by c_{od}^k. As before, $\tau_a^k = t_a$ if OD pair k uses arc $a \in A_1$ and 0 otherwise.

$$
\begin{array}{lrl}
\mathsf{HPP}: & \max\limits_{\tau,x,t} & \sum\limits_k \sum\limits_a n^k \tau_a^k \\
& \text{subject to for all } k & \sum\limits_a x_a^k \leq 1 \\
& \text{for all } a \in A_1 & \sum\limits_b (\tau_b^k + c_b^k x_b^k) + c_{od}^k (1 - \sum\limits_b x_b^k) \leq c_a^k + t_a \\
& & t_a - M(1 - x_a^k) \leq \tau_a^k \leq M x_a^k \\
& & \tau_a^k \geq 0 \\
& & x_a^k \in \{0,1\} \qquad .
\end{array}
$$

A complete description of the convex hull has been given by Heilporn et al. [37] in the single OD pair case. The constrained highway pricing problem (CHPP) comes from HPP by adding triangle and monotonicity constraints. Dewez proposes several heuristics for the CHPP as well as an exact resolution method. She also shows that when it reduces to a single OD pair or a single toll arc, those special instances are polynomially solvable. Heilporn et al. prove the NP-Hardness of CHPP and propose some valid inequalities.

Recently a link has been established between network pricing problems and problems of designing and pricing set of products in a given economic market (Heilporn et al. [36]). More precisely, the profit problem PP (Dobson and Kalish[28]) consists in determining which subsets of products should be introduced in the market and at what price, with the aim of maximizing the seller's income. On the demand side, it is assumed that each purchaser selects the product that maximizes its utility, provided it is positive. In this setting, let I be denote the set of products and K the set of purchaser segments, with respective demand n^k. With each product $i \in I$ in the market is associated a price p_i and a fixed cost f_i for the seller. With each product-segment couple (i,k) is associated the value r_i^k of product i to segment k. The utility is then defined as the difference between the product value r_i^k and its price p_i. The profit problem is defined as follows:

$$
\begin{array}{lrl}
\mathsf{PP}: & \max\limits_{p,x,y} & \sum\limits_k \sum\limits_i n^k p_i x_i^k - \sum\limits_i f_i y_i \\
& \text{subject to} & \sum\limits_i y_i \leq Y \\
& \text{for all } k & \sum\limits_i x_i^k \leq 1 \\
& \text{for all } k \text{ and all } i & \sum\limits_j (r_j^k - p_j) x_i^k \geq (r_i^k - p_i) y_i \\
& & x_i^k \leq y_i \\
& & x_i^k, y_i \in \{0,1\}.
\end{array}
$$

By setting $Y = \infty$, $S = I$, and $y_i = 1 \ \forall i \in I$, one realizes that HPP subsumes the price variant of PP where the set of products is fixed a priori. In that context, products correspond to paths of the network, and we believe that one can exploit this relationship to devise new algorithms based on the research pursued on both these problems.

6.5 Extensions

In order to better fit real applications, the basic framework may be enhanced in several ways. We sketch some of them below.

Random utility. User behaviour can be made more realistic by incorporating randomness into the choice model of the users. One popular approach leads to discrete choice models based on random utility theory, as described in the classic book of Ben-Akiva and Lerman [10].

Design and pricing. In the telecommunication industry, it might be advantageous to determine jointly an investment and pricing policy. This problem is addressed in Brotcorne et al. [19] in the uncapacitated case and by Brotcorne et al. [17] in the capacitated case. The resulting bilevel models involve binary variables, and are solved using a Lagrangean approach, each evaluation of the dual function being performed by one of the methods described in Brotcorne et al. [15]. This problem is closely related to the a variant of the product line design problem [28]. One major difference is that in the toll problems, the set of products (paths) is not available explicitly and that tolls are set on arcs rather than paths (products).

Congestion pricing. In a congested environment, tolls can be used both for alleviating congestion or for generating revenues. In the former case, one looks for a toll vector t such that the traffic equilibrium $X(t)$ that corresponds to the modified delay function $S(v)+t$ is as efficient as possible. If tolls can be imposed on all arcs of the network, it can be shown that a delay-minimizing flow pattern can be induced, and that the associated toll vector can be easily computed by solving a linear network flow problem. This corresponds to the situation referred to as 'first best' in the economic literature. On the other hand, if tolls can only be imposed on a subset of arcs, one faces a true bilevel problem (see Fortin et al. [30]).

On the other hand, if tolls are used to generate revenues, it is possible, as advocated in [30], to reduce the problem to the standard pricing problems addressed in this section, by linearizing the delay functions S_a. Note that this approach only applies when the latter assume the separable for F_a, Furthermore, the introduction of breakpoints required in the linearization process greatly reduces the size of problems that can be addressed.

Value of time (VOT). In models involving autonomous agents at the lower level, out-of-pocket cost is not the sole determinant of disutility. For instance, users of a toll highway might ponder both cost and time before selecting a path. In VOT models, users are distributed into classes characterized by a VOT parameter α that translates delays into time units. A class-α user will then be associated the generalized cost (disutility)

$$\alpha S_a(\bar{v}) + \alpha t_a$$

where $\bar{v}$ denotes the total flow traveling on arc a of the network. If $h(\alpha)$ represents the proportion of α-users we have that $\bar{v} = \sum_\alpha v(\alpha)$ in the discrete case and $\bar{v} = \int_\alpha v(\alpha)$ in the continuous case, where $v(\alpha)$ denotes the vector of α-flows. Let us introduce the proportion $h(\alpha)$ of α-users. For fixed toll vector t, the equilibrium satisfies the variational inequality

$$\langle S(\bar{v}) + \alpha t, v(\alpha) - v'(\alpha)\rangle \leq 0 \qquad \forall v' \in h(\alpha)V, \qquad \forall \alpha.$$

Whenever S is constant (uncongested case), the resulting revenue maximization problem is akin to the basic problem, and merely involves a larger set of flow variables. When the parameter α is continuously distributed according to some density function h, the infinite-dimensional toll optimization problem can be solved by first approximating h by a mass function, and then by performing a local optimization procedure based on sensitivity results that reflect the local behaviours of equilibrium flows with respect to the toll vector. It has been shown by Marcotte et al. that a very coarse approximation of h was sufficient to initiate a search phase leading to optimal or near-optimal solutions.

7 Revenue Management

Revenue Management is the branch of operations research that is concerned with the revenue optimization of firms characterized by high investment and low operating costs. It involves issues such as capacity management, pricing, and should ideally take into account demand forecast and competition reaction. Initiated in the airline industry, where it was tradionally partitioned into four domains: pricing, seat allocation, demand forecasting and overbooking. The synonym term 'Yield Management' is frequently used to refer to the setting of dynamic rules that determine what 'products' (airline tickets) should be offered at any given instant previous to flight departure. In the recent years, the techniques of Revenue Management have been extended to fields that share with the airline industry features such as high investment and low operating costs, mentioned earlier, as well as perishable inventories, a deregulated environment, etc. One may think of the rail, telecommunication or hospitality industry, for instance. Revenue Management is, and will stay for some time, one of the fast growing areas of management science.

Bilevel programming lends itself naturally to Revenue Management. For example, Côté et al. [22] considered an airline application where an airline sets fares and capacity allocation policies, taking into account that travellers are assigned to flights that maximize their individual utilities. For each market, demand is segmented into groups, each group being characterized by its valuation of different product attributes: price, duration, class of service, etc. For instance, if 3 attributes, say price, duration and QOS, are associated with a given product p on flight g then the disutility of a customer of group g is expressed as

$$c^i_{p,f,g}(\alpha,\beta) = t^i_{p,f} + \alpha_{g,k} \times DUR_f \ + \ \beta_{g,k} \times QOS_{p,f}$$

where the fare $t^i_{p,f}$ serves as numéraire and the attributes DUR_f and $QOS_{p,f}$, with respective valuations $\alpha_{g,k}$ and $\beta_{g,k}$, correspond to flight duration and quality of service, respectively.

It is worth having a closer look at the model. Let us first introduce the notation. For additional details concerning the model, the reader is referred to Cté et al. [22].

i:	$i = 1$ refers to the leader airline and $i = 2$ to the aggregated competition
k:	market index
f:	flight index
s:	leg index
p:	fare product index
b:	booking class index
g:	user group index
K:	set of all markets (origin-destination pairs)
F^1:	set of flights supplied by the leader
F^2:	set of flights supplied by the competition
$F^i(k)$:	set of flights supplied by agent i on market k: $F^i = \cup_k F^i(k)$
S:	set of flight legs operated by the leader airline
$S(f) \subset S$:	set of flight legs making flight $f \in F^1$
$P(f)$:	set of fare products offered on flight f
$B(f)$:	set of booking classes open on flight f
G:	set of user groups
$b(p)$:	booking class of the product $p \in P(f)$.

Fares and flight attributes are:

$t^1_{p,f}$:	fare of the leader airline for product p on flight f (decision variable)
$t^2_{p,f}$:	fare of the competition for product p on flight f (exogenous data)
t^i:	fare vector
DUR_f:	duration of flight f
$QOS_{p,f}$:	quality of service associated with product p on flight f.

At the lower level, decision variables are passenger flows:

$v^i_{p,f,g}$:	number of passengers of group $g \in G$ purchasing product $p \in P(f)$ on flight $f \in F^i$.

The behavioral parameters α and β denote, respectively, the valuation of one unit of duration and one unit of quality of service. We set:

$\alpha_{g,k}$: monetary equivalent of one duration unit for passengers of group g on market k
$\beta_{g,k}$: monetary equivalent of one unit of disutility for passengers of group g on market k
$c^i_{p,f,g}(\alpha,\beta)$: perceived travel disutility:

$$c^i_{p,f,g}(\alpha,\beta) = t^i_{p,f} + \alpha_{g,k} \times DUR_f + \beta_{g,k} \times QOS_{p,f}$$

$\underline{\sigma}_k$: lower bound on targeted share of market k
$\overline{\sigma}_k$: upper bound on targeted share of market k
$\underline{\rho}_k$: lower bound on targeted revenue share of market k
$\overline{\rho}_k$: upper bound on targeted revenue share of market k.

Besides the parameters listed above, the model requires the following input:

d_k: total demand on market k over the planning horizon
$h_{g,k}$: fraction of demand d_k belonging to group g
$l_{b,f}$: number of seats available in class b on leader flight f (booking limit)
u_s: aircraft capacity of flight segment s.

The bilevel model then takes the form

$$\begin{aligned}
\max_{T^1,v^1,v^2} \quad & \sum_{f\in F^1} \sum_{p\in P(f)} \sum_{g\in G} t^1_{p,f}\; v^1_{p,f,g} \\
\text{subject to for all } k \in K \quad & \sum_{f\in F^1(k)} \sum_{p\in P(f)} \sum_{g\in G} v^1_{p,f,g} \leq \overline{\sigma}_k \\
& \sum_{f\in F^1(k)} \sum_{p\in P(f)} \sum_{g\in G} v^1_{p,f,g} \geq \underline{\sigma}_k \\
& \sum_{f\in F^1(k)} \sum_{p\in P(f)} \sum_{g\in G} v^1_{p,f,g}\; t^1_{p,f} \leq \overline{\rho}_k \\
& \sum_{f\in F^1(k)} \sum_{p\in P(f)} \sum_{g\in G} v^1_{p,f,g}\; t^1_{p,f} \geq \underline{\rho}_k
\end{aligned}$$

$$\begin{aligned}
\max_{v^1,v^2} \quad & \sum_{f\in F^1} \sum_{p\in P(f)} \sum_{g\in G} c^1_{p,f,g}(\alpha,\beta)\; v^1_{p,f,g} + \sum_{f\in F^2} \sum_{p\in P(f)} \sum_{g\in G} c^2_{p,f,g}(\alpha,\beta)\; v^2_{p,f,g} \\
\text{subject to for all } b \in B(f), f \in F^1 \quad & \sum_{\substack{p\in P(f)|\\ b(p)=b}} \sum_{g\in G} v^1_{p,f,g} \leq l_{b,f} \\
\text{for all } g \in G, k \in K \quad & \sum_{f\in F^1(k)} \sum_{p\in P(f)} v^1_{p,f,g} + \sum_{f\in F^2(k)} \sum_{p\in P(f)} v^2_{p,f,g} = d_k\, h_{g,k} \\
\text{for all } s \in S \quad & \sum_{f|s\in S(f)} \sum_{p\in P(f)} \sum_{g\in G} v^1_{p,f,g} \leq u_s
\end{aligned}$$

where, for $i \in \{1,2\}$:

$$c^i_{p,f,g}(\alpha,\beta) \;=\; t^i_{p,f} \;+\; \alpha_{g,k} \times DUR_f \;+\; \beta_{g,k} \times QOS_{p,f}.$$

Note that the above framework can easily be adapted to behavioural parameters that vary continuously by replacing discrete flow variables and summations by flow densities and integrals, respectively.

It is clear that the above framework must be refined to take into account the stochasticity and dynamics associated with real-life RM applications, which has been performed for high-speed European rail operator. The bilevel approach can be viewed as a top-down approach, with its global view of the network, in contrast with more traditional approaches that focused on a detailed representation of subsystems, for instance at the market level. Interestingly, as the latter are ‘globalized’, the convergence and interface between both approaches emerges (See Miranda-Bront et al. [55]).

8 Conclusion

The aim of the present work was to provide a quick overview of a stream of research on bilevel programming that was initiated in Montreal, together with Belgian colleagues and students from Montreal, Brussels and Valenciennes. Motivated by applications involving an underlying network structure, it led to methodological and algorithmic advances that showed that the bilevel programming paradigm was not only a powerful tool for modelling situations in economics, as well as a rich source of mathematical and computational challenges, but also could lead to the understanding of situations of practical interest.

References

[1] Alarie, S., Audet, C., Jaumard, B. and Savard, G., “Concavity Cuts for Disjoint Bilinear Programming”, *Mathematical Programming*, 90, 373–398, 2001.

[2] Abdulaal, M. and LeBlanc, L. J., “Continuous Equilibrium Network Design Models”, *Transportation Research* 13B, 19–32, 1979.

[3] Audet, C., Hansen, P., Jaumard, B. and Savard, G., “A Symmetrical Linear Maxmin Approach to Disjoint Bilinear Programming”, *Mathematical Programming*, 85, 573–592, 1999.

[4] Audet, C., *Optimisation globale structurée : propriétés, équivalences et résolution*, Thèse de doctorat, École Polytechnique de Montréal, 1997.

[5] Audet, C., Hansen, P., Jaumard, B. and Savard, G., “Links between Linear Bilevel and Mixed 0-1 Programming Problems”, *Journal of Optimization Theory and Applications*, 93, 273–300, 1997.

[6] Audet, C., Savard, G. and Zghal, W., “New branch-and-cut algorithm for bilevel linear programming”, *Journal of Optimization Theory and Applications*, 134, 353-370, 2007.

[7] Bard, J. F., Practical Bilevel Optimization: Algorithms and Applications, Kluwer Academic Publishers, 1998.

[8] Bard, J.F. and Moore, J., “A Branch-and-Bound Algorithm for the Bilevel Programming Problem”, *SIAM Journal on Scientific and Statistical Computing*, 11, 281–292, 1990.

[9] Beale, E.M.L. and Small R.E., “Mixed Integer Programming by a Branch and Bound Technique”, in W.A. Kalenich (Ed.), Proceedings of the IFIP Congress 1965, 450451, MacMillan and Washington, London, 1965.

[10] Ben-Akiva, M. and Lerman, S., Discrete Choice Analysis Theory and Application to Travel Demand, The MIT Press, 1985.

[11] Bouhtou, M., van Hoesel, S., van der Kraaij, A. and Lutton, J.-L., “Optimization in networks”, Research Memorandum 041, METEOR, Maastricht Research School of Economics of Technology and Organization, 2003.

[12] Braess, D.,“Uber ein Paradoxon aus der Verkehrsplanung”, *Unternehmensforschung 12*, 258–268, 1968.

[13] Brotcorne, L., Cirinei F., Marcotte P. and Savard G., “An exact algorithm for a bilevel pricing problem on a network”, INFORMS, San Francisco, 2005.

[14] Brotcorne, L., Labbé M., Marcotte P. and Savard G., "A bilevel model and solution algorithm for a freight tariff setting problem". *Transportation Science*, 34, 289–302 2000.

[15] Brotcorne, L, Labbé, M., Marcotte, P. and Savard, G., "A bilevel model for toll optimization on a multicommodity transportation network", *Transportation Science*, 35, 1–14, 2001.

[16] Brotcorne, L., Cirinei F., Marcotte P. and Savard G., "An exact algorithm for a bilevel pricing problem on a network", INFORMS, San Francisco, 2005.

[17] Brotcorne, L., Marcotte P., Savard G. and Wiart M., "Joint Pricing and network capacity setting problem ". In *Advanced OR and AL Methods in Transportation* (Jaszkiewicz, Kaczmarek, Zak and Kubiak, eds.). Publishing House of Poznan University of Technology, 2005.

[18] Brotcorne,L., Cirinei F., Marcotte P. and Savard G.,"A tabu search algorithm for a pricing problem on a network", Tristan VI conference (2007).

[19] Brotcorne, L. , Labbé, M. , Marcotte, P. and Savard, G., "Joint design and pricing on a network", forthcoming in Operations Research 2008.

[20] Candler, W. and Norton, R., "Multilevel programming and development policy". Technical Report 258, World Bank Staff, Washington D.C., 1977.

[21] Colson, B., Marcotte, P. and Savard, G., "A Trust-Region Method for Nonlinear Bilevel Programming: Algorithm and Computational Experience", *Computational Optimization and Applications*, 30, 211–227, 2005.

[22] Côté, J.-P., Marcotte, P. and Savard, G., "A bilevel modeling approach to pricing and fare optimization in the airline industry", *Journal of Revenue and Pricing Management*, 2, 23–36 , 2003.

[23] CPLEX, ILOG CPLEX, v10.0, 2007.

[24] Dempe, S., Foundations of Bilevel Programming, Springer, 2002.

[25] Dewez, S., *On the toll setting problem.* PhD thesis, Université Libre de Bruxelles, Institut de Statistique et de Recherche Opérationnelle, 2004.

[26] Dewez, S., Labbé, M., Marcotte, P. and Savard, G., "New formulations and valid inequalities for a bilevel pricing problem", forthcoming in *Operations Research Letters.*

[27] Didi-Biha, M., Marcotte, P. and Savard, G., "Path-based Formulations of a Bilevel Toll Setting Problem", in Optimization with Multivalued Mappings Theory: Theory, Applications and Algorithms, S. Dempe and V. Kalashnikov (eds.), Applications and Algorithms Optimization and Its Applications, Vol. 2, Springler, 2006.

[28] Dobson G. and Kalish S., "Positioning and Pricing a Product Line", *Marketing Science*, 7, 107–125, 1988.

[29] Fishbone, L.G. and Abilock, H., "MARKAL: A linear programming model for energy systems analysis technical description of the BNL version", *International Journal of Energy Research* 5, 353–375, 1981.

[30] Fortin, M., Marcotte, P. and Savard, G., "Pricing a segmented market subject to congestion", Proceedings of TRISTAN V (Triennal Symposium on Transportation Analysis) Le Gosier, 2004.

[31] Garey, M.R. and Johnson, D.S., Computers and intractability: A guide to the theory of NP-completeness, W.H. Freeman, New York, 1979.

[32] Grigoriev, A., van Hoesel, S., van der Kraaij, A., Uetz, M. and Bouhtou, M., "Pricing Network Edges to Cross a River", Research Memorandum 009, METEOR, Maastricht Research School of Economics of Technology and Organization, 2004.

[33] Grötschel, M., Lovász, L. and Schrijver, A., "The ellipsoid method and its consequences in combinatorial optimization", *Combinatorica*, 1, 169–197, 1981.

[34] Hansen, P., Jaumard, B. and Savard, G., "New Branch-and-Bound rules for Linear Bilevel Programming", *SIAM Journal on scientific and Statistical Computing*, 13, 1194-1217, 1992.

[35] Haurie, A., Loulou, R. and Savard, G., "A two player game model of power cogeneration in New England", *IEEE Transactions on Automatic Control*, 37, 1992.

[36] Heilporn G., Labbé M., Marcotte P. and Savard G., "Linking Pricing Problems in Transportation Networks and Economics", manuscript, 2008.

[37] Heilporn G., Labbé M., Marcotte P. and Savard G., "A polyhedral study of the Network Pricing Problem with Connected Toll Arcs", *Optimization Online*, `http://www.optimization-online.org/DB_HTML/2007/07/1732.html`, 2007.

[38] Jaumard, B., Savard, G. and Xiong, J., "An exact algorithm for convex quadratic bilevel programming", Technical paper, Ecole Polytechnique de Montréal, 2000.

[39] Jeroslow, R., "The Polynomial Hierarchy and a Simple Model for Competitive Analysis", *Mathematical Programming*, 32, 146–164, 1985.

[40] van der Kraaij, A., *Pricing in networks*. PhD thesis, Proefschrift Universiteit Maastricht, 2004.

[41] Labbé, M., Marcotte, P. and Savard, G., "A bilevel model of taxation and its applications to optimal highway pricing", *Management Science*, 44, 1595–1607, 1998.

[42] Labbé, M., Marcotte, P. and Savard, G., "On a class of bilevel program", In: Nonlinear Optimization and Related Topics, Di Pillo and Giannessi eds., Kluwer Academic Publishers, 183-206, 1999.

[43] Lavigne, D., Loulou, R. and Savard, G., "Pure Competition, Regulated and Stackelberg Equilibria: Application to the Energy System of Quebec", *European Journal of Operational Research*, 125, 1–17, 2000.

[44] Lawler, E.L., "A procedure to compute the K best solutions to discrete optimization problems and its application to the shortest path problem", *Management Science*, 18, 401-405, 1972.

[45] Luo, Z.-Q., Pang, J.-S. and Ralph, D., Mathematical Programs with Equilibrium Constraints, Cambridge University Press, 1996.

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

[47] Marcotte, P. and Marquis, G., "Efficient implementation of heuristics for the continuous network design problem", *Annals of Operations Research* 34, 163–176, 1992.

[48] Marcotte, P., Savard, G. and Semet, F., "A bilevel programming approach to the travelling salesman problem", *Operations Research Letters*, 32, 240–248, 2003.

[49] Marcotte, P. and Patriksson, M., "Traffic Equilibrium", in: Transportation. Handbooks in Operations Research and Management Science, C. Barnhardt and G. Laporte eds., North-Holland 623-714, 2007.

[50] Pinter, J., Global Optimization in Practice: State-of-the-Art and Perspectives, in: Complementarity, Duality, and Global Optimization, H.D. Sherali and D. Gao eds., Springer Science, New York, 2006.

[51] Roch, S., Savard, G. and Marcotte, P.,"Design and analysis of an algorithm for Stackelberg network pricing", *Optimization Online* 2003, to appear in *Networks*.

[52] Fudenberg, D. and Tirole, J., Game Theory, MIT Press, 1993.

[53] Van Ackere A., “The principal/agent paradigm: Its relevance to various functional fields”, *European Journal of Operational Research* 70, 83-103, 1993.

[54] Vicente, L.N., Savard, G. and Júdice, J.J., “Descent approaches for quadratic bilevel programming”, *Journal of Optimization Theory and Applications*, 81, 379–399, 1994.

[55] Miranda Bront, J., Méndez-Díaz, I. and Vulcano, G., “A column generation algorithm for choice-based network revenue management”, forthcoming in *Operations Research*.