A bilevel optimization model for load balancing in mobile networks through price incentives

Jean Bernard Eytard*, Marianne Akian*, Mustapha Bouhtou†, and Stéphane Gaubert *
*INRIA, CMAP, Ecole Polytechnique, CNRS
Route de Saclay, 91128 Palaiseau, France
Email: {jean-bernard.eytard, marianne.akian, stephane.gaubert}@inria.fr
†Orange Labs
44 avenue de la République, 92320 Chatillon, France
Email: mustapha.bouhtou@orange.com

***Abstract*—We propose a model of incentives for data pricing in large mobile networks, in which an operator wishes to balance the number of connexions (active users) of different classes of users in the different cells and at different time instants, in order to ensure them a sufficient quality of service. We assume that each user has a given total demand per day for different types of applications, which he may assign to different time slots and locations, depending on his own mobility, on his preferences and on price discounts proposed by the operator. We show that this can be cast as a bilevel programming problem with a special structure allowing us to develop a polynomial time decomposition algorithm suitable for large networks. First, we determine the optimal number of connexions (which maximizes a measure of balance); next, we solve an inverse problem and determine the prices generating this traffic. Our results exploit a recently developed application of tropical geometry methods to mixed auction problems, as well as algorithms in discrete convexity (minimization of discrete convex functions in the sense of Murota). We finally present an application on real data provided by Orange and we show the efficiency of the model to reduce the peaks of congestion.**

I. Introduction

With the development of new mobile data technologies (3G, 4G), the demand for using the Internet with mobile phones has increased rapidly. Mobile service providers (MSP) have to confront congestion problems in order to guarantee a sufficient quality of service (QoS).

Several approaches have been developed to improve the quality of service, coming from different fields of the telecommunication engineering and economics. For instance, one can refer to Bonald and Feuillet [1] for some models of performance analysis to optimize the network in order to improve the QoS. One of the promising alternatives to solve such problems consists in using efficient pricing schemes in order to encourage customers to shift their mobile data consumption. In [2], Maillé and Tuffin describe a mechanism of auctions based on game-theoretic methods for pricing an Internet network, see also [3]. In [4], Altman et al. study how to price different services by using a noncooperative game. These different approaches are based on congestion games. In the present work, we are interested in how a MSP can improve the QoS by balancing the traffic in the network. We wish to determine in which locations, and at which time instants, it is relevant to propose price incentives, and to evaluate the influence of these incentives on the quality of service.

This kind of problem belongs to smart data pricing. We refer the reader to the survey of Sen et al. [5] and also to the collection of articles [6]. Finding efficient pricing schemes is a revenue management issue. The first approach consists in usage-based pricing; the prices are fixed monthly by analysing the use of the former months. It is possible to improve this scheme by identifying peak hours and non-peak hours and proposing incentives in non-peak hours in order to decrease the demand at peak hours and to better use the network capacity at non-peak hours. This leads to time-dependent pricing. Such a scheme for mobile data is developed by Ha et al. in [7]. The prices are determined at different time slots and based on the usage of the previous day in order to maximize the utility of the customers and the revenue of the MSP. This pricing scheme was concretely implemented by AT&T, showing the relevance of such a model. In another approach, Tadrous et al. propose a model in which the MSP anticipates peak hours and determines incentives for proactive downloads [8].

The latter models concern only the time aspects. One must also take into account the spatial aspect in order to optimize the demand between the different locations. In [9], Ma, Liu and Huang present a model depending on time and location of the customers where the MSP proposes prices and optimizes his profit taking into account the utility of the customers.

Here, we assume (as in [9]) that the MSP proposes incentives at different time and places. Then, customers optimize their data consumption by knowing these incentives and the MSP optimizes a measure of the QoS. In this way, we introduce a bilevel model in which the provider proposes incentives in order to balance the traffic in the network and to avoid as much as possible the congestion (high level problem), and customers optimize their own consumption for the given incentives (low level problem).

Bilevel programs have been widely studied, see the surveys of Colson, Marcotte and Savard [10] and of Dempe [11]. They represent an important class of pricing problems in sense that they model a leader wanting to maximize his profit and

This work is supported in part by Orange Labs through INRIA Contract CRE 8306.

978-3-901882-90-6/17 ©2017 IFIP

proposing prices to some followers who maximize themselves their own utility. Most classes of bilevel programs are known to be NP-hard. Several methods have been introduced to solve such problems. For instance, if the low level program is convex, it can be replaced by its Karush-Kuhn-Tucker optimality conditions and the bilevel problem becomes a classical one-stage optimization problem, which is however generally non convex. If some variables are binary or discrete, and the objective function is linear, the global bilevel problem can be rewritten as a mixed integer program, as in Brotcorne et al. [12].

In the present work, we optimize the consumption of each customer in a large area (large urban agglomerations) during typically one day divided in time slots of one hour, taking into account the different types of customers and of applications that they use. Therefore, we have to confront both with the difficulties inherent to bilevel programming and with the large number of variables (around 10^7). Hence, we need to find polynomial time algorithms, or fast approximate methods, for classes of problems of a very large scale, which, if treated directly, would lead to mixed integer linear or nonlinear programming formulations beyond the capacities of current off-the-shelve solvers.

This motivated us to introduce a different approach, based on tropical geometry. Tropical geometry methods have been recently applied by Baldwin and Klemperer in [13] to an auction problem. This has been further developed by Yu and Tran [14]. In these approaches, the response of an agent to a price is represented by a certain polyhedral complex (arrangement of tropical hypersurfaces). This approach is intuitive since it allows one to vizualize geometrically the behavior of the agents: each cell of the complex corresponds to the set of incentives leading to a given response. Then, we vizualize the collective response of a group of customers by "superposing" (refining) the polyhedral complexes attached to every customer in this group. We apply here this idea to represent the response of the low-level optimizers in a bilevel problem. This leads to the following decomposition method: first we compute, among all the admissible consumptions of the customers, the one which maximizes a measure of balance of the network; then, we determine the price incentive which achieves this consumption. In this way, a bilevel problem is reduced to the minimization of a convex function over a certain Minkowski sum of sets. We identify situations in which the latter problem can be solved in polynomial time, by exploiting the discrete convexity results developed by Murota [15]. In this approach, a critical step is to check the membership of a vector to a certain Minkowski sum of sets of integer points of polytopes. In our present model, these polytopes, which represent the possible consumptions of one customer, have a remarkable combinatorial structure (they are hypersimplices). Exploiting this combinatorial structure, we show that this critical step can be performed quickly, by reduction to a shortest path problem in a graph. This leads to an exact solution method when there is only one type of contract and one type of application sensitive to price incentive, and to a fast approximate method in the general case.

We finally present the application of this model on real data from Orange and show how price incentives can improve the QoS by balancing the number of active customers in an urban agglomeration during one day. These results indicate that a price incentive mechanism can effectively improve the satisfaction of the users by displacing their consumption from the most loaded regions of the space-time domain to less loaded regions.

The paper is organized as follows. In Section II, we present the bilevel model. In Section III, we explain how a certain polyhedral complex can be used to represent the user's responses. In Section IV, we describe the decomposition method. In Section V, we deal with the high level problem and identify special cases which are solvable in polynomial time. In Section VI, we propose a general relaxation method. The application to the instance provided by Orange is presented in Section VII.

II. A BILEVEL MODEL

We consider a time horizon of one day, divided in T time slots numbered $t \in [T] = \{1, \dots T\}$, and a network divided in L different cells numbered $l \in [L]$. We assume K customers, numbered $k \in [K]$, are in the network. The customers have different types of contracts $b \in [B]$ and they make requests for different types of applications $a \in [A]$ (web/mail, streaming, download, ...). We denote by $\mathcal{K}^b$ the set of customers with the contract b. A given customer $k \in \mathcal{K}^b$ is characterized by the following data. We denote by $L_t^k \in [L]$ the position of the customer k at each time $t \in [T]$, so that the sequence $(L_1^k, \dots, L_T^k)$ represents the trajectory of this customer. We assume that this trajectory is deterministic, so we consider customers with a regular daily mobility (for example, the trip between home and work). We denote by $\rho_k^a(t)$ the inclination of a customer k to make a request for an application of type a at time $t \in [T]$. We suppose that customer k wishes to make a fixed number of requests $R_k^a \leqslant T$ using the application a during the day. We consider a set of time slots $\mathcal{I}_k^a \subset [T]$ in which the customer k decides not to consume the application a.

We denote by $u_k^a(t)$ the consumption of the customer k for the application a at time t, setting $u_k^a(t) = 1$ if k is active at time t and makes a request of type a and $u_k^a(t) = 0$ otherwise. Therefore, the number $N^{a,b}(t,l)$ of active customers with contract b for the application a at time t and location l is given by $N^{a,b}(t,l) = \sum_{k \in \mathcal{K}^b} u_k^a(t) \mathbb{1}(L_t^k = l)$, where $\mathbb{1}$ denotes the indicator function, and the total number of active customers $N(t,l)$ at time t and location l is given by $N(t,l) = \sum_a \sum_b N^{a,b}(t,l)$.

We consider the following two-stage model of price incentives. The first stage consists for the operator in announcing a discount $y^{a,b}(t,l)$ at time t and location l for the customers of contract b making requests of type a. We consider only nonnegative discounts, so $y^{a,b}(t,l) \geqslant 0$. The second stage models the behavior of customers who modify their consumption by taking the discounts into account. We will

assume the preference of a customer k for consuming at time t becomes $\rho_k^a(t) + \alpha_k^a y^{a,b}(t, L_t^k)$, where α_k^a denotes the sensitivity of customer k to price incentives for the application a. It corresponds to classical linear utility functions, see e.g. [13]. We also assume that the customers cannot make more than one request at each time, that is $\forall t \in [T], \sum_a u_k^a(t) \leqslant 1$. Therefore, each customer k determines his consumptions $u_k^a = (u_k^a(t))_{t\in[T]} \in \{0,1\}^T$ for the applications, as an optimal solution of the linear program:

Problem II.1 (Low-level, customers)**.**

$$\max_{u_k^a \in \{0,1\}^T} \sum_{a\in[A]} \sum_{t=1}^{T} \left[\rho_k^a(t) + \alpha_k^a y^{a,b}(t, L_t^k)\right] u_k^a(t) \quad (1)$$

$$s.t. \;\; \forall a \in [A], \sum_{t=1}^{T} u_k^a(t) = R_k^a, \quad \forall t \in [T], \sum_{a\in[A]} u_k^a(t) \leqslant 1$$
$$\forall t \in \mathcal{I}_k^a, \forall a \in [A], u_k^a(t) = 0$$

Consequently, each price $y^{a,b} = (y^{a,b}(t,l))_{t\in[T],\, l\in[L]}$ determines the possible individual consumptions u_k^a for the users with contract b, and so the possible cumulated traffic vectors $N^{a,b} = (N^{a,b}(t,l))_{t\in[T],\, l\in[L]}$ and $N = \sum_a \sum_b N^{a,b}$. The aim of the operator is, through price incentives, to balance the load in the network into the different locations and time slots to improve the quality of service perceived by each customer. We introduce a coefficient γ_b relative to the kind of contracts of the different customers in order to favor some classes of premium customers. In [16], Lee et al. suppose that the satisfaction of a customer depends on his perceived throughput, which can be considered as inversely proportional to the number of customers in the cell. Here, we assume that the satisfaction of each customer k in the cell $l \in [L]$ is a decreasing function $s_l^{a,b}$ of the total number of active customers in the cell $N(t,l)$, depending on the characteristics of the cell, of the type of application the user wants to do (some applications like streaming need a higher rate than other) and on the type of contract. We also assume the satisfaction of all the customers with contract b using a given application a in a given cell is maximal until the number of active customers reaches a certain threshold $N_l^{a,b}$, then $s_l^{a,b}(N(t,l)) = 1$ for $N(t,l) \leqslant N_l^{a,b}$. After this threshold, the satisfaction decreases until a critical value N_l^C. We add the constraint $\forall t \in [T], \forall l \in [L], N(t,l) \leqslant N_l^C$ to prevent the congestion. For non-real time services like web, mail, download, the satisfaction function can be viewed as a concave function of the throughput, like $1 - e^{-\delta/\delta_c}$ where δ denotes the throughput, see Moety et al. [17]. Hence, we will consider that for contents like web, mail and download, $N_l^{a,b} = N_l^1$, $s_l^{a,b}(n) = 1$ for $n \leqslant N_l^1$ and $s_l^{a,b}(n) = 1 - \lambda_b \exp\left(-\frac{2N_l^C}{n - N_l^1}\right)$ for $N_l^1 \leqslant n \leqslant N_l^C$ where λ_b is a positive parameter depending on the kind of contract of the customer. The more expensive the contract of the customer is, the larger is λ_b. We can prove that this function is concave for $0 \leqslant n \leqslant N_l^C$. For real time services like video streaming, the customers need a more important throughput to ensure a good QoS [16]. We will here consider the same type of functions $s_l^{a,b}$ but with N_l^1 replaced by $N_l^{a,b} = 0$, that is $s_l^{a,b}(n) = 1 - \lambda_b \exp\left(-\frac{2N_l^C}{n}\right)$ for $0 < n \leqslant N_l^C$.

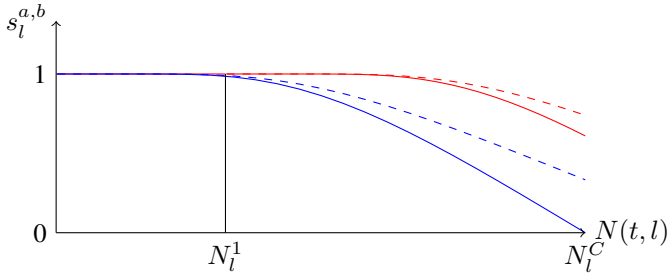

Fig. 1. Different kind of satisfaction functions of the number of active customers in a cell. The blue ones are those for streaming contents whereas the red ones are those for web, mail and download contents. The dashed ones corresponds to the satisfaction of standard customers, the continuous ones to the satisfaction of premium customers.

So, the first stage consists in maximizing the global satisfaction function s which depends on the vectors $N^{a,b} \in \mathbb{N}^{T\times L}$ and is defined by:

$$\begin{aligned} s(N^{a,b}) &= \sum_{t=1}^{T} \sum_{a\in[A]} \sum_{b\in[B]} \sum_{k\in\mathcal{K}^b} \gamma_b s_{L_t^k}^{a,b}(N(t, L_t^k)) u_k^a(t) \\ &= \sum_{t=1}^{T} \sum_{a\in[A]} \sum_{b\in[B]} \sum_{k\in\mathcal{K}^b} \sum_{l=1}^{L} \gamma_b s_l^{a,b}(N(t,l)) \mathbb{1}(L_t^k = l) u_k^a(t) \\ &= \sum_{t=1}^{T} \sum_{l=1}^{L} \sum_{a\in[A]} \sum_{b\in[B]} \gamma_b N^{a,b}(t,l) s_l^{a,b}(N(t,l)) \end{aligned}$$

with $\forall b \in [B], \gamma_b > 0$. Our final model consists in solving the following bilevel program:

Problem II.2 (High-level, provider)**.**

$$\max_{y^{a,b} \in \mathbb{R}_+^{T\times L}} \sum_{t=1}^{T} \sum_{l=1}^{L} \sum_{a\in[A]} \sum_{b\in[B]} \gamma_b N^{a,b}(t,l) s_l^{a,b}(N(t,l)) \quad (2)$$

where $\forall t \in [T], \; l \in [L], \; N(t,l) = \sum_{a=1}^{A} \sum_{b=1}^{B} N^{a,b}(t,l)$, *and* $N(t,l) \leqslant N_l^C$, $\forall t \in [T], \; l \in [L], \; a \in [A], \; b \in [B], \; N^{a,b}(t,l) = \sum_{k\in\mathcal{K}^b} u_k^a(t) \mathbb{1}(L_t^k = l)$, *and* $\forall k\, [K]$, *the vectors* u_k^a *are solutions of the problem II.1.*

III. A TROPICAL APPROACH FOR THE BILEVEL PROBLEM

We will present a decomposition method for solving the previous bilevel problem. In this section, and in the two next ones, we suppose that there is only one kind of application and one kind of contract. This special case is already relevant in applications: it covers the case when, for instance, only the download requests are influenced by price incentives, whereas other requests like streaming or web are fixed. Whereas the analytical results of the present section carry over to the general model, the results of the next two sections (polynomial time solvability) are only valid under these restrictive assumptions.

We shall return to the general case in Section VI, developing a fast approximate algorithm for the general model based on the present principles.

In this special case, the bilevel model can be rewritten:

$$\max_{y\in\mathbb{R}_+^{T\times L}} \sum_{t=1}^{T}\sum_{l=1}^{L} N(t,l)s_l(N(t,l))$$

where $\forall t,l\ N(t,l) = \sum_{k\in[K]} u_k(t)\mathbb{1}(L_t^k = l)$, and $\forall k$, the vectors u_k are solutions of the problem:

$$\max_{u_k\in\{0,1\}^T} \sum_{t=1}^{T} \left[\rho_k(t) + \alpha_k y(t, L_t^k)\right] u_k(t)$$

$$\text{s.t. } \sum_{t=1}^{T} u_k(t) = R_k, \qquad \forall t\in\mathcal{I}_k, u_k(t) = 0,$$

In order to deal more abstractly with the bilevel model, we introduce the notation $u_k(t,l) = u_k(t)\mathbb{1}(L_t^k = l)$. Hence, we have $u_k(t,l) = 0$ if $L_t^k \neq l$. By defining the set $\mathcal{J}_k = \{(t,l) \mid t\in\mathcal{I}_k \text{ or } L_t^k \neq l\}$ and $\rho_k(t,l) = \rho_k(t)\mathbb{1}(L_t^k = l)/\alpha_k$, we can rewrite each low-level problem:

Problem III.1 (Abstract low-level problem)**.**

$$\max_{u_k\in F_k} \sum_{t,l} \left[\rho_k(t,l) + y(t,l)\right] u_k(t,l) \tag{3}$$

where $F_k = \{x\in\{0,1\}^{T\times L} | \sum_{t,l} x(t,l) = R_k$ *and* $\forall (t,l)\in \mathcal{J}_k, x(t,l) = 0\}$.

Because the functions s_l are concave and decreasing, we can prove that the functions $f_l : x\mapsto xs_l(x)$ defined for $x\geqslant 0$ are also concave. The global bilevel problem is:

Problem III.2 (Bilevel problem)**.**

$$\max_{y\in\mathbb{R}_+^{T\times L}} \sum_{t,l} f_l(N(t,l)) \quad \textit{s.t. } \forall (t,l), N(t,l) = \sum_{k=1}^{K} u_k(t,l) \tag{4}$$

with u_k *solutions of the problem III.1.*

The lower-level component of our bilevel problem can be studied thanks to tropical techniques. Tropical mathematics refers to the study of the max-plus semifield $\mathbb{R}_{\max}$, that is the set $\mathbb{R}\cup\{-\infty\}$ endowed with two laws $\oplus$ and $\odot$ defined by $a\oplus b = \max(a,b)$ and $a\odot b = a+b$, see [18], [19], [20], [21] for background. We first consider the relaxation in which the price vector y can take any real value, i.e. $y\in\mathbb{R}^{T\times L}$. Each customer k defines his consumption u_k by solving the problem:

$$\max_{u_k\in F_k} \sum_{t,l} \left[\rho_k(t,l) + y(t,l)\right] u_k(t,l) = \max_{u_k\in F_k} \langle \rho_k + y, u_k\rangle \ , \tag{5}$$

The map $P_k : y\mapsto \max\langle \rho_k + y, u_k\rangle$ is convex, piecewise affine, and the gradients of its linear parts are integer valued. It can be thought of as a tropical polynomial function in the variable y. Indeed, with the tropical notation, we have

$$P_k(y) = \bigoplus_{u_k\in F_k} (\rho_k(1,1)\odot y(1,1))^{\odot u_k(1,1)}\odot \cdots \odot (\rho_k(T,L)\odot y(T,L))^{\odot u_k(T,L)} \ ,$$

where $z^{\odot p} := z\odot\cdots\odot z = p\times z$ denotes the pth tropical power. In this way, we see that all the monomials of P_k have degree $\sum_{t,l} u_k(t,l) = R_k$, so that P_k is homogeneous of degree R_k, in the tropical sense. If we denote by $e = (1\dots 1)\in\mathbb{R}^{T\times L}$, it means that $\forall y\in\mathbb{R}^{T\times L}, \forall\beta\in\mathbb{R},\ P_k(y+\beta e) = P_k(y)$. An important corollary is:

Lemma III.3. *The value of the bilevel problem coincides with the value of the relaxed problem with* $y\in\mathbb{R}^{T\times L}$.

By definition, the *tropical hypersurface* associated to a tropical polynomial function is the nondifferentiability locus of this function. Since the monomial P_k is homogeneous, its associated tropical hypersurface is invariant by the translation by a constant vector. Therefore, it can be represented as a subset of the tropical projective space $\mathbb{TP}^{T\times L-1}$. The latter is defined as the quotient of $\mathbb{R}^{T\times L}$ by the equivalence relation which identifies two vectors which differ by a constant vector, and it can be identified to $\mathbb{R}^{T\times L-1}$ by the map $\mathbb{TP}^{T\times L-1}\to\mathbb{R}^{T\times L-1}$, $y\mapsto (y(t,l) - y(T,L))_{(t,l)\in[T]\times[L]\setminus\{(T,L)\}}$.

Example III.4. Consider a simple example with $T = 3$ time steps (for instance morning, afternoon and evening), $L = 1$, $K = 5$ and $\mathcal{J}_k = \emptyset$ for each k. For brevity, we will write y_t instead of $y(t,l)$. The parameters of the customers are

$$\rho_1 = [0,0,0]\,,\ R_1 = 1, \quad \rho_2 = [0,-1,0]\,,\ R_2 = 2\ ,$$
$$\rho_3 = [-1,1,0]\,,\ R_3 = 1\ \ \rho_4 = [1/2,1/2,0]\,,\ R_4 = 2,$$
$$\rho_5 = [1/2,2,0]\,,\ R_5 = 1\ .$$

The tropical polynomial of the first customer is $P_1(y) = \max(y_1, y_2, y_3)$, meaning that this customer has no preference and consumes when the incentive is the best. Its associated tropical hypersurface is a tropical line (since P_1 has degree 1), so it splits $\mathbb{TP}^2$ in three different regions corresponding to a choice of the vector u_1 among $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, see Figure 2. E.g., the cell labeled by $(1,0,0)$ represents a consumption concentrated the morning, induced by a price $y_1 > y_2$ and $y_1 > y_3$.

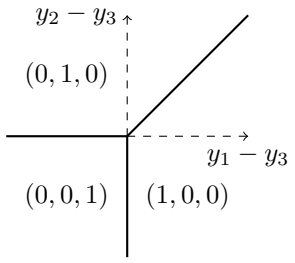

Fig. 2. A customer response: a tropical line splits the projective space into three cells. Each cell corresponds to a possible customer response

To study jointly the responses of the five customers, we represent the arrangement of the tropical hypersurfaces associated to the $P_k,\ k \in [5]$, with

$$P_2(y) = \max(y_1 + y_2 - 1, y_1 + y_3, y_2 + y_3 - 1),$$
$$P_3(y) = \max(y_1 - 1, y_2 + 1, y_3),$$
$$P_4(y) = \max(y_1 + y_2 + 1, y_1 + y_3 + 1/2, y_2 + y_3 + 1/2),$$
$$P_5(y) = \max(y_1 + 1/2, y_2 + 2, y_3) \quad .$$

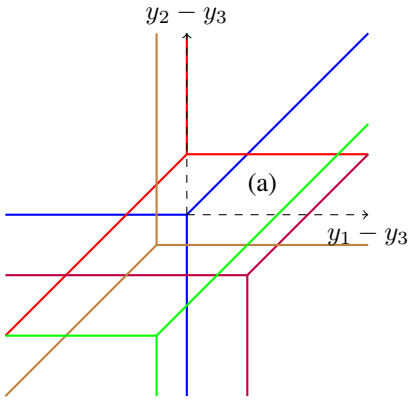

Fig. 3. Arrangement of tropical hypersurfaces: each tropical hypersurface corresponds to a customer response. For example, the cell (a) corresponds to discounts y with responses (1,0,0) for customer 1, (1,0,1) for customer 2, (0,1,0) for customer 3, (1,1,0) for customer 4 and (0,1,0) for customer 5 . Hence, the total number of customers in the network with these discounts is (3,3,1).

Lemma III.5 (Corollary of [14, §4, Lemma 3])**.** *Each cell of the arrangement of tropical hypersurfaces corresponds to a collection of customers responses* $(u_1, ..., u_K)$ *and to an unique traffic vector* N*, defined by* $N = \sum_k u_k$.

IV. Decomposition Theorem

We next show that the present bilevel problem can be solved by decomposition. We note that the function to optimize for the higher level problem, i.e. the optimization problem of the provider, depends only on N. The variables $y(t,l)$ allow one to generate the different possible vectors N. So we will characterize the feasible vectors N in order to optimize directly the satisfaction function on the set of feasible N. This idea is motivated by the tropical approach thanks to Lemma III.5.

Most of the following results are applications of classical notions of convex analysis which can be found in [22]. It is convenient to define for every k the polytope Δ_k as the convex hull of F_k, together with the convex function φ_k defined by $\varphi_k(u) = -\langle \rho_k, u \rangle$ if $u \in \Delta_k$, and $\varphi_k(u) = +\infty$ otherwise. The value of each low level problem (5) can therefore be viewed as the value of the Legendre-Fenchel transform of φ_k at point y, i.e., $\varphi_k^*(y) = \sup_{u_k \in \Delta_k} [\langle y, u_k \rangle - \varphi_k(u_k)]$.

Lemma IV.1. $\Delta_k = \{x \in [0,1]^{T \times L} \,|\, \sum_{t,l} x(t,l) = R_k$ *and* $\forall (t,l) \in \mathcal{J}_k, x(t,l) = 0\}$.

So, a vector $u_k \in \Delta_k$ is a solution of the low-level problem iff $u_k \in \partial \varphi_k^*(y)$ where $\partial \varphi_k^*$ denotes the subdifferential of the convex function φ_k^*. A feasible N is a sum of such vectors u_k. We have, by [22, Th. 23.8], that N is feasible iff $\exists y \in \mathbb{R}^{T \times L}$, $N \in \sum_k \partial \varphi_k^*(y) = \partial \left(\sum_k \varphi_k^* \right)(y)$, i.e., $\exists y, N \in \partial \psi^*(y)$, or equivalently $\exists y, y \in \partial \psi(N)$, where $\psi = \square_k \varphi_k$ is the inf-convolution of the functions φ_k. The function ψ is polyhedral (as the inf-convolution of polyhedral convex functions) and it is finite at every point $N \in \sum_k \Delta_k$. So, $\forall N \in \sum_k \Delta_k, \partial \psi(N)$ is a non-empty polyhedral convex set [22, Th. 23.10] and N is feasible. Moreover, we have the following lemma:

Lemma IV.2. *Let* $N = \sum_k u_k$ *with* $u_k \in \Delta_k \ \forall k$*. The following assertions are equivalent:*

1) *There exists* $y \in \mathbb{R}_+^{T \times L}$ *such that each* u_k *is a solution of the low-level problem of customer* k *with discount vector* y*;*
2) *The vectors* $u_1, \ldots, u_K$ *realize the minimum in the inf-convolution* ψ*, i.e.* $\psi(N) = -\sum_k \langle \rho_k, u_k \rangle$.

In our problem, we are not interested in the vectors N which are sums of optimal solutions of each low level problem, but in the ones which are sums of integer optimal solutions of each low level problem. These vectors belong to $\sum_k F_k$. Let $N \in \sum_k F_k$. We have $N \in \sum_k \Delta_k$, so it can be written as the sum of optimal solutions $u_k \in \Delta_k$ of each low level problem. According to the previous lemma, we have $(u_1, \ldots, u_K)$ optimal solution of:

$$\max \sum_k \langle \rho_k, v_k \rangle$$
$$\text{s.t.} \begin{cases} \forall k, t, l,\ 0 \leqslant v_k(t,l) \leqslant 1 \ , \\ \forall k,\ \sum_{t,l} v_k(t,l) = R_k \ , \\ \forall k,\ \forall (t,l) \in \mathcal{J}_k,\ v_k(t,l) = 0 \ , \\ \forall t, l,\ \sum_k v_k(t,l) = N(t,l) \ . \end{cases}$$

The polytope defined by the constraint can be written $Av \leqslant b$ where A is a totally unimodular matrix and b is an integer vector. So, the optimal solutions of this problem are integer vectors and each vector u_k belongs to F_k. Hence, the feasible set of the high-level problem is exactly $\sum_k F_k$.

We arrive at the following method.

Theorem IV.3. (Decomposition) *The bilevel program can be solved as follows:*

1) *Find an optimal solution* N^* *to the high level problem with unknown* N*:*
$$\max_{N \in \sum_k F_k} \sum_{t,l} f_l(N(t,l)) \quad s.t.\ N(t,l) \leqslant N_l^C \ \forall t, l \ . \quad (6)$$
2) *Find optimal requests vectors* u_k^* *by solving the inf-convolution problem:*
$$\max_{\substack{u_1 \in F_1, \ldots, u_K \in F_K \\ \sum_k u_k = N^*}} \sum_k \langle \rho_k, u_k \rangle \ .$$
3) *Find a vector* y^* *such that* $\forall k, u_k^*$ *is a solution of the low level problem.*

The second step of this theorem consists in solving a linear program. We next show that the third step reduces to a linear feasibility problem.

Lemma IV.4. *u_k^* is an optimal solution of the low level problem iff the set of indices (t,l) such that $u_k^*(t,l) = 1$ coincides with the set of indices of the R_k highest coordinates not included in $\mathcal{J}_k$ of the vector $\rho_k + y$, i.e. $\forall (t,l), (t',l') \notin \mathcal{J}_k$ such that $u_k^*(t,l) = 1$, $u_k^*(t',l') = 0$, we have $\rho_k(t,l) + y(t,l) \geqslant \rho_k(t',l') + y(t',l')$.*

For every k, the latter inequalities define a polytope, and we have to find y^* in the intersection of all these polytopes.

V. Algorithm for solving the bilevel problem

A. Solving the high-level problem

We next explain how to solve Problem (III.2). We will use some elements of discrete convexity developed by Murota [15]. An integer set $B \subset \mathbb{Z}^n$ is *M-convex* [15, Ch. 4, p.101] if $\forall x, y \in B, \forall i \in [n]$ such that $x_i > y_i, \exists j \in [n]$ such that $x_j < y_j$, $x - e_i + e_j \in B$ and $y + e_i - e_j \in B$, where e_i is the i-th vector of the canonical basis in $\mathbb{R}^n$.

Lemma V.1. *The feasible domain of the high-level program $B = \{N \in \sum_k F_k | \forall t,l \; N(t,l) \leqslant N_l^C\}$ is a M-convex set of $\mathbb{Z}^{T \times L}$.*

We have to maximize a separable concave function on a M-convex set. This is easy, because local optimality is equivalent to global optimality, as shown by the following result:

Theorem V.2 ([15, Th. 6.26, p.148])**.** *Let f be a separable concave function on $\mathbb{Z}^n$, B a M-convex set, and $N^* \in B$. Then, N^* is a maximum point of f over B iff $\forall i,j \in [n]$ such that $N^* - e_i + e_j \in B, f(N^* - e_i + e_j) \leqslant f(N^*)$.*

Moreover, Murota ([15], ch.10, p.281) gives an algorithm which runs in pseudo-polynomial time to maximize separable concave functions on M-convex sets.

Algorithm 1 Murota's algorithm to minimize a M-convex function f on a M-convex set B.

1) Find $N \in B$;
2) Find $i,j \in \underset{k,l \in [n] \text{ s.t. } N - e_k + e_l \in B}{\arg\max} f(N - e_k + e_l)$;
3) If $f(N - e_i + e_j) \leqslant f(N)$ then $N^* = N$ is a global minimizer of f over B;
4) Else $N := N - e_i + e_j$ and go back to Step 2;

B. A polynomial time algorithm for the bilevel problem

Algorithm V-A can be applied to the high-level problem (6) of Theorem IV.3, with $f(N) = \sum_{t,l} f_l(N(t,l))$ and $B = \sum_k F_k$. The most critical part of this algorithm is to check for a given $N \in \sum_k F_k$ whether $N - e_i + e_j$ for $i,j \in [T] \times [L]$ belongs to $\sum_k F_k$. However, it can be easily done.

Lemma V.3. *Let $u_k \in F_k$ for each $k \in [K]$ such that $\psi(N) = -\sum_k \langle \rho_k, u_k \rangle$. Consider the quantity $w_{\alpha\beta}^k$ for $k \in [K]$ and $\alpha, \beta \in [T] \times [L]$ defined by $w_{\alpha\beta}^k = \rho_k(\alpha) - \rho_k(\beta)$ if $u_k(\alpha) = 1$ and $u_k(\beta) = 0$ and $w_{\alpha\beta}^k = +\infty$ otherwise. The optimal $v_k \in F_k$ such that $\psi(N - e_i + e_j) = -\sum_k \langle \rho_k, v_k \rangle$ can be obtained by solving the shortest path problem between i and j in a graph of $T \times L$ nodes with edges weighted by $w_{\alpha\beta} = \min_k w_{\alpha\beta}^k$.*

This leads to the following algorithm. Note that the pseudo-polynomial time bound for Murota algorithm leads in this special case to a polynomial time bound.

Algorithm 2 Solving the bi-level problem, for one application and one type of contract

1) Find $N \in \sum_k F_k$ with this optimal decomposition $N = \sum_k u_k^*$;
2) For each $i,j \in [T] \times [L]$, calculate the shortest path between i and j in the graph of weights $w_{\alpha\beta}$ defined in the former lemma and deduce if $N - e_i + e_j \in \sum_k F_k$ and the optimal decomposition $N - e_i + e_j = \sum_k v_k^*$;
3) Find $i,j \in \underset{(k,l) \text{ s.t. } N - e_k + e_l \in \sum_k F_k}{\arg\max} f(N - e_k + e_l)$;
4) If $f(N - e_i + e_j) \leqslant f(N)$ then $N^* = N$ and go to Step 6;
5) Else $N := N - e_i + e_j$ and go back to Step 2;
6) Find $y^* \in \mathbb{R}^{T \times L}$ verifying the property of Lemma IV.4 and return y^*.

Theorem V.4. *Algorithm 2 returns a global optimizer in polynomial time.*

Example V.5. Consider again Example III.4 together with the concave function $f : N \mapsto -\sum_{t,l} N(t,l)^2$. We suppose that $\forall k, \mathcal{J}_k = \emptyset$. Hence, we can prove that $\sum_k F_k = \{N \in \mathbb{N}^3 | \sum_{i=1}^3 N_i = 7 \text{ and } \max(N_i) \leqslant 5\}$. First, we want to solve $\max_{N \in \sum_k F_k} -(N_1^2 + N_2^2 + N_3^2)$. We start from $N^{(0)} = (5,2,0)$, a feasible point. Following Algorithm V-A, we compute $N^{(1)} = (4,2,1)$ and $N^{(2)} = (3,2,2)$ which is a minimizer. We take $N^* = (3,2,2)$. Now, we solve $\max_{u_1 \in F_1, \ldots, u_5 \in F_5, \sum_{k=1}^5 u_k = N^*} \sum_k \langle \rho_k, u_k \rangle$. We obtain $u_1^* = [1,0,0]$, $u_2^* = [1,0,1]$, $u_3^* = [0,1,0]$, $u_4^* = [1,0,1]$, $u_5^* = [0,1,0]$. Applying Lemma IV.4, we obtain the linear inequalities $y_1^* - y_2^* \leqslant 3/2$, $0 \leqslant y_1^* - y_3^*$ and $-1 \leqslant y_2^* - y_3^* \leqslant -1/2$. In particular, $y^* = (3/4, 0, 3/4)$ is an optimal solution.

VI. The general algorithm

In this section, we come back to the general bilevel problem II.2 proposed in Section II, and extend the Algorithm of Section V to it. In the low level problem of each customer, the consumptions for different contents verify the constraints $\forall a \in [A], \sum_{t=1}^T u_k^a(t) = R_k^a, \forall t \in \mathcal{I}_k^a, a \in [A], u_k^a(t) = 0$ and $\forall t \in [T], \sum_{a \in [A]} u_k^a(t) \leqslant 1$. We make the assumption that for each customer k, the sets of possible instants at which this customer makes a request for the different applications are disjoint, meaning that for any two applications $a \neq a'$, the complements of $\mathcal{I}_k^a$ and $\mathcal{I}_k^{a'}$ in $[T]$ have an empty intersection. Then the constraint $\forall t \in [T], \sum_{a \in [A]} u_k^a(t) \leqslant 1$ is automatically verified and the low-level problem of each customer can be separated into different optimization problems corresponding to the consumption vector u_k^a of each customer k for each application a. Each of these problems takes the following form:

Problem VI.1.

$$\max_{u_k^a \in \{0,1\}^T} \sum_{t=1}^{T} \left[\rho_k^a(t) + \alpha_k^a y^{a,b}(t, L_t^k)\right] u_k^a(t) \qquad (7)$$

$$s.t. \sum_{t=1}^{T} u_k^a(t) = R_k^a, \qquad \forall t \in \mathcal{I}_k^a, a \in [A], u_k^a(t) = 0 \ .$$

We denote by F_k^a the feasible set of this problem. The above assumption (that the complements of $\mathcal{I}_k^a$ and $\mathcal{I}_k^{a'}$ have an empty intersection) is relevant in particular if only one kind of application is sensitive to price incentives. For instance, requests for downloading data can be anticipated (see [8]) and it makes sense to assume that customers are only sensitive to incentives for this kind of contents. In this case, the assumption means that customers wanting to download data can shift their consumption only at instants when they do not request another kind of content. Under this assumption, the decomposition theorem is still valid.

The high-level problem consists in maximizing the separable function $\sum_{t,l}\left(\sum_{a\in[A]}\sum_{b\in[B]}\gamma_b N^{a,b}(t,l)s_l^{a,b}(N(t,l))\right)$ where each vector $N^{a,b}$ belongs to a M-convex set $\sum_{k\in\mathcal{K}^b} F_k^a$ according to Theorem IV.3 and Lemma V.1. Because each function $s_l^{a,b}$ is concave decreasing and each $N^{a,b}(t,l)$ is positive, we notice that $\forall a' \in [A], b' \in [B]$, the function wich sends $N^{a',b'}(t,l)$ to $\sum_{a\in[A]}\sum_{b\in[B]}\gamma_b N^{a,b}(t,l)s_l^{a,b}(N(t,l))$ is still concave. Consequently, the function to optimize in the high level problem is M-concave in each vector $N^{a,b} \in \mathbb{Z}^{T\times L}$ considered separately. This leads to a block descent method (Algorithm 3), in which we maximize the objective function, successively, over every vector $N^{a,b}$. We denote by $f(N^{1,1}, \ldots, N^{A,B})$ the objective function of the high-level problem. Step 2 of this algorithm can be implemented by solving the shortest path problem of certain graphs as in Lemma V.3. Unlike Algorithm 2, Algorithm 3 is not guaranteed to give a globally optimal solution.

Algorithm 3 Solving the bilevel problem for an arbitrary number of types of contracts.

1) Find $\forall a, b, N^{a,b} \in \sum_{k\in\mathcal{K}^b} F_k^a$
2) Find, for each $a \in [A], b \in [B]$, $(i^{a,b}, j^{a,b})$ belonging to
$$\underset{(k,l) \text{ s.t. } N^{a,b}-e_k+e_l\in\sum_{k\in\mathcal{K}^b} F_k^a}{\arg\max} f(N^{1,1} - e_{i^{1,1}} + e_{j^{1,1}}, \ldots, N^{a,b} - e_k + e_l, \ldots, N^{A,B});$$
3) If $f(N^{1,1} - e_{i^{1,1}} + e_{j^{1,1}}, \ldots, N^{A,B} - e_{i^{A,B}} + e_{j^{A,B}}) \leqslant f(N^{1,1}, \ldots, N^{A,B})$ then $\forall a, b$, return the optimal solution $N^{*,a,b} = N^{a,b}$.
4) Else for each a, b, $N^{a,b} := N^{a,b} - e_{i^{a,b}} + e_{j^{a,b}}$ and go back to Step 2;

VII. Experimental results

We consider an application based on real data provided by Orange. It involves the data consumptions in an area of $L = 43$ cells, during one day divided in time slots of one hour, that is $T = 24$ time slots. We will focus here our study on price incentives only for download contents. During this day, a number K of more than 2500 customers make some requests for downloading data in this area and we are interested in balancing the number of active customers in the network. Even though they are insensitive to price incentives, other kind of requests (web, mail, etc.) have to be satisfied and they are taken into account in the high level optimization problem. We consider two classes of users: standard and premium customers. The premium ones demand a better quality of service. Hence, they are less satisfied than the standard customers if they share their cell with a given number of active customers. We therefore define the satisfaction function as in Section II. The provider wants to favor the premium customers. Hence, we take $\gamma_b = 2$ for the latter ones and $\gamma_b = 1$ for the standard customers, in the high-level optimization problem. We also assume that the premium customers are less sensitive to the incentives, and thus take $\alpha_k^a = 1/2$ for all standard customers and $\alpha_k^a = 1$ for all premium customers in the low-level problem II.1. We estimate very simply the parameters ρ_k. We take $\rho_k(t) = 1$ when the customer k consumes download at time t without incentives, $\rho_k(t) = 0$ when he does not make any request without incentives but makes a request for download at times $t-1$ or $t+1$ (we assume he could shift his consumption of one hour) and $\rho_k(t) = -\infty$ otherwise.

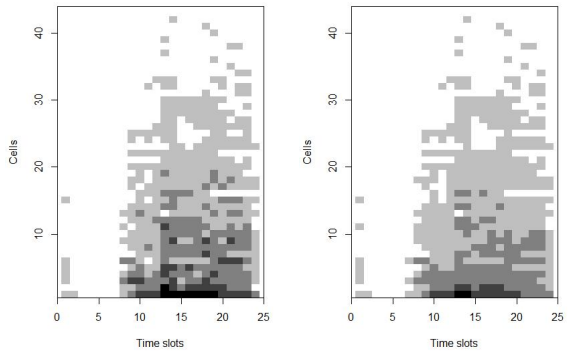

Fig. 4. Satisfaction of premium customers for streaming without (left) and with (right) incentives. The grey level indicates the satisfaction: critical unsatisfaction, $s < 0.3$ (black), $0.3 < s < 0.7$ (dark grey), $0.7 < s < 0.9$ (grey), $0.9 < s < 0.99$ (light grey) and complete satisfaction $0.99 < s$ (white).

We solve the bilevel problem using Algorithm 3, implemented in `Scilab`. The computation took 9526 seconds on a single core of an Intel i5-4690 processor @ 3.5 GHz.

On Figures 4– 7, we show the evolution of the satisfaction of different kind of customers for different kind of contents without and with incentives. These results show that price incentives have an effective influence on the load, especially in the most loaded cells (the number of black regions in the space-time coordinates, in which the unsatisfaction of the users is critical, is considerably reduced). Moreover, Figure 8 reveals that the consumption of users is not only moved in time, but also in space: not only some consumption is moved from the peak hour to the night (off peak), but the surface of the dark grey region, representing the total download consumption in

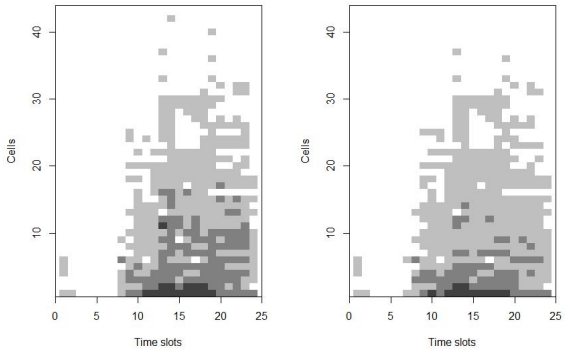

Fig. 5. Satisfaction of standard customers for streaming without (left) and with (right) incentives

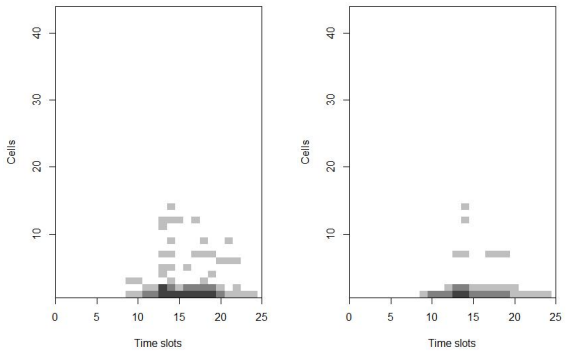

Fig. 6. Satisfaction of premium customers for web, mail or download without (left) and with (right) incentives

the cell over the whole day, is decreased, indicating that some part of the consumption has been shifted to other cells.

VIII. Conclusion

We presented here a bilevel model for price incentives in data mobile networks. We solved this problem by a decomposition method based on discrete convexity and tropical geometry. We finally applied our results to real data. In further work, we shall consider more general models: unfixed number of requests, nonlinear preferences of the customers, satisfaction functions of the provider taking into account the profit. Stochastic models shall also be considered in particular to take into account the partial information of the provider about the customers preferences and trajectories.

IX. Acknowledgements

We thank the reviewers for their remarks and comments, helping us to improve this work.

References

[1] T. Bonald and M. Feuillet, *Network performance analysis*. John Wiley & Sons, 2013.

[2] P. Maillé and B. Tuffin, "Pricing the internet with multibid auctions," *IEEE/ACM transactions on networking*, vol. 14, no. 5, pp. 992–1004, 2006.

[3] ——, *Telecommunication network economics: from theory to applications*. Cambridge University Press, 2014.

[4] E. Altman, D. Barman, R. El Azouzi, D. Ros, and B. Tuffin, "Pricing differentiated services: A game-theoretic approach," *Computer Networks*, vol. 50, no. 7, pp. 982–1002, 2006.

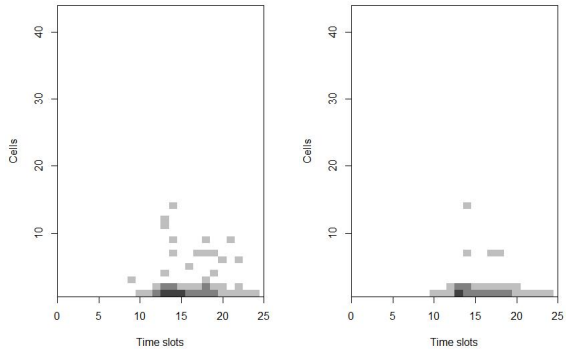

Fig. 7. Satisfaction of standard customers for web, mail or download without (left) and with (right) incentives

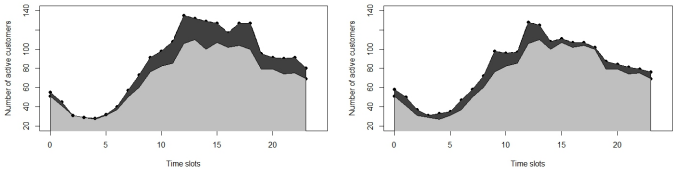

Fig. 8. Traffic in the most loaded cell. The light grey part represents the web, mail and streaming customers who have no incentives and are fixed. The dark grey part corresponds to the download customers in the cell without (left) and with (right) incentives

[5] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 15, 2013.

[6] ——, *Smart Data Pricing*. John Wiley & Sons, 2014.

[7] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: time-dependent pricing for mobile data," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 247–258, 2012.

[8] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Pricing for demand shaping and proactive download in smart data networks," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 3189–3194.

[9] Q. Ma, Y.-F. Liu, and J. Huang, "Time and location aware mobile data pricing," in *Communications (ICC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3235–3240.

[10] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, no. 1, pp. 235–256, 2007.

[11] S. Dempe, *Bilevel programming: A survey*. Dekan der Fak. für Mathematik und Informatik, 2003.

[12] L. Brotcorne, M. Labbé, P. Marcotte, and G. Savard, "A bilevel model and solution algorithm for a freight tariff-setting problem," *Transportation Science*, vol. 34, no. 3, pp. 289–302, 2000.

[13] E. Baldwin and P. Klemperer, "Tropical geometry to analyse demand," Working paper, Oxford University, Tech. Rep., 2012.

[14] N. M. Tran and J. Yu, "Product-mix auctions and tropical geometry," *arXiv preprint arXiv:1505.05737*, 2015.

[15] K. Murota, *Discrete convex analysis*. SIAM, 2003.

[16] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Non-convex optimization and rate control for multi-class services in the internet," *IEEE/ACM transactions on networking*, vol. 13, no. 4, pp. 827–840, 2005.

[17] F. Moety, M. Bouhtou, T. En-Najjary, and R. Nasri, "Joint optimization of user association and user satisfaction in heterogeneous cellular network," in *28th International Teletraffic Congress*, 2016.

[18] F. Baccelli, G. Cohen, G. Olsder, and J. Quadrat, *Synchronization and Linearity*. Wiley, 1992.

[19] I. Itenberg, G. Mikhalkin, and E. I. Shustin, *Tropical algebraic geometry*. Springer Science & Business Media, 2009, vol. 35.

[20] P. Butkovič, *Max-linear systems : theory and algorithms*, ser. Springer monographs in mathematics. Springer, 2010.

[21] D. Maclagan and B. Sturmfels, *Introduction to Tropical Geometry*, ser. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2015, vol. 161.

[22] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1970.