Bounds for global optimization of capacity expansion and flow assignment problems

H.P.L. Luna a, P. Mahey b, *

a Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
b Laboratoire d’Informatique, de Modélisation et d’Optimisation des Systèmes, Université Blaise Pascal, Campus des Cézeaux, Aubière, France

Received 1 September 1998; received in revised form 1 December 1999

Abstract

This paper provides new bounds related to the global optimization of the problem of mixed routing and bandwidth allocation in telecommunication systems. The combinatorial nature of the problem, related to arc expansion decisions, is embedded in a continuous objective function that encompasses congestion and investment line costs. It results in a non-convex multicommodity flow problem, but we explore the separability of the objective function and the fact that each associated arc cost function is piecewise-convex. Convexifying each arc cost function enables the use of efficient algorithms for convex multicommodity flow problems, and we show how to calculate sharp bounds for the approximated solutions. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Network design; Capacity expansion of telecommunication systems; Nonconvex multicommodity flow problem; Separable convexification

1. Introduction

This paper presents an alternative approach for some models and algorithms that have been addressed to problems of telecommunication and computer network expansion and operation. Many such models are special issues of the general network design problem and lead to large-scale nonconvex or combinatorial optimization problems (see [2] for example). Approximation algorithms and heuristics are frequently employed to treat these problems and the need for tight bounds of the global optimal value is a crucial issue we want to address here.

The motivation behind our modelling approach concerns the need to integrate the routing mechanisms (commodity flows) together with the control of bandwidth allocation (capacity assignment) in emerging technologies (the so-called Capacity and Flow Assignment problem or CFA). The integrated problem is classical in network planning, but it has now a greater importance because capacity assignment can
also be done in real time. The literature focusing on the capacity and flow assignment problem is very limited, as cited in [5], where lower and upper bounds are provided for a problem addressed to the selection of a primary route for each communicating origin–destination pair. Topology design and bandwidth allocation models with alternative routes have been addressed by Gerla and collaborators [6,7] and solved to local optimality. Using several initial flow assignments as starting points, the basic procedure iterates between the cost minimizing capacity assignment algorithms, and a flow assignment phase in which a measure of the average delay is minimized, until a local optimum is reached. A weakness common to all these attempts to solve the bifurcated problem is that no means, either theoretical or empirical, are provided in order to evaluate the quality of the heuristic solution generated. This may seriously interfere with the progress of real life applications, and the state of the art is that few nonlinear capacity and flow assignment problems have been solved to global optimality with exact methods in the literature.

Special hypotheses on the cost structure have been studied in the literature to identify easier-to-solve models (see, for instance, the recent book by Konno et al. [10]). The case of concave costs under networks constraints has been too widely studied in telecommunications and transportation problems, as it corresponds to the effect of economies of scale. Unfortunately, it induces NP-hard problems in most cases and only very special situations can be treated by polynomial algorithms [12]. On the other hand, the general case of indefinite quadratic programming has been shown to be equivalent to convex programming (see [1]), but only when affine equality constraints are present in the model. Indeed, it is the presence of affine inequality constraints which affects the combinatorial complexity of these models as it is well known that the minimization of a concave function on a polyhedron occurs at an extreme point and that most other vertices are local minima. Anyway, these situations do not apply to the nonconvex model we are aiming at, and even if it is observed in [2] that the (CFA) problem presents many local optima, the general complexity issues are still open questions.

Our present objective is to study some bounds for the global optimization of a network flow problem with implicit capacity assignment on the arcs, which are expanded as far as induced by congestion costs. We are not concerned here with pure topology design of the backbone network, and assume that an initial graph is given, such that there exists at least one path between each origin–destination pair and that in each arc is installed a given capacity which is supposed to be able to support a standard traffic load. The problem is formulated by minimizing the link expansion costs and the congestion costs associated with the total flow circulating in each link. The application concerns the simultaneous use of routing and bandwidth allocation in order to ensure an acceptable performance level at a minimum cost.

In the next section we formulate the model and present the main result, related to a lower bound for a network optimization problem where the variables refer to routing and bandwidth allocation. In Section 3 we explore some properties of typical congestion functions, thus giving a characterization of the separable objective function and of the convex hull of the component associated with each arc. In Section 4 we exemplify how to calculate an explicit gap value with a simple class of quadratic congestion functions. In the concluding section, we discuss some practical extensions.

2. The network expansion model

We present in this section the network expansion model. The basic component of the model is a digraph $G = (V, E)$ with $m$ nodes and $n$ arcs representing a communication network. Any kind of traffic between a given pair of nodes $(O_k, D_k)_k$ is treated as a separate commodity $k$. The kind of traffic indexed by $k$ is also related to the parameter $d_k$, which represents consumer demand for commodity $k$. Given a commodity $k$, we consider a given set of directed paths $P_{kh}, h = 1, \ldots, N_k$ joining $O_k$ and $D_k$. This set may be the set of all simple directed paths or a restricted set of feasible paths, for instance with a limited number of hops. Let $x_{kh}$ be the amount of flow of commodity $k$ through the path $P_{kh}$ and let $a_{kh}^j$ be its arc-path incidence vector defined by

$$
a_{kh}^j = \begin{cases} 1 & \text{if arc } j \in P_{kh}, \\ 0 & \text{otherwise.} \end{cases}
$$
Each component $x_j$ of the vector $x$ denotes the total flow on arc $j$. We assume that we are given a congestion plus leasing cost function $f_j(x_j)$ on each arc $j$, which is increasing on $[0, c_{ij}]$, where $c_{ij}$ is the value of the expanded capacity on arc $j$. To be more precise, each $f_j$ is assumed to be continuous, defined on $\mathbb{R}$ with values on the extended reals $\mathbb{R} \cup \{+\infty\}$ and minorized by at least one affine function.

The cost structure aims at distributing the load among all capacitated arcs to reduce the total cost of leasing (or capacity investment) lines and congestion expressed in monetary values. We remark that the capacity in any arc may be possibly increased affording a given expansion cost, in such a case reducing the marginal time delay. We will not digress here either on the question of the tradeoff between money and marginal time delay, or on the approximation in this case of any statistical congestion function like Kleinrock’s delay function (see for example [2]).

Each link $j$ has an initial capacity $c_{0j} \geq 0$ and we assume a fixed cost $\pi_j$ to expand the capacity from $c_{0j}$ to $c_{ij} = c_{0j} + \delta_j$, $\delta_j > 0$. We can now state the network expansion problem (NEP) as follows:

$$\begin{align*}
\min_{\xi \geq 0, x \geq 0} & \sum_{j=1}^{n} f_j(x_j) \\
\text{subject to} & \\
& d_k - \sum_{h=1}^{K} \xi_{kh} = 0 \quad k = 1, \ldots, K, \\
& \sum_{h=1}^{K} \sum_{k=1}^{K} a_{kh} \xi_{kh} - x_j = 0 \quad j = 1, \ldots, n.
\end{align*}$$

Constraints (2) ensure that the demand for each commodity is supplied by the path flows activated for the commodity, while constraints (3) impose that the total flow on each arc is equal to the sum of the individual path flows which use that arc. We seek for path flows, arc flows and implicit arc capacities in order to minimize a total cost that includes capacity and queuing components. We will detail the properties of the cost function later in the next section. For the moment, we remind that each function $f_j$ depends on the actual arc capacity $c_{0j}$, on the expanded arc capacity $c_{ij}$ and on the associated expansion cost $\pi_j$.

The main result on which the present note relies is based on the convexification of each arc cost function to yield a lower bound of the global minimal value. Indeed, these functions are likely to be nonconvex, making the search for a global optimum a very hard task. The motivation here is the sequential use of convex multicommodity flow algorithms.

Let $X$ be the nonempty convex set of arc flow vectors for which we have feasible multicommodity flows, i.e., there exists $x \in X$ and corresponding path flows $\xi_{kh}$ satisfying constraints (2) and (3). We will denote by $\text{conv } f$ the closed convex hull of a function $f$, i.e. the greatest closed convex function majorized by $f$ (to be consistent, that definition requires that $f$ admits at least one minorizing affine function). It is shown in [8] that $\text{conv } f$ can be defined for instance by the supremum of all affine functions minorizing $f$, i.e. $\text{conv } f(x) = \sup \{s.x - b; s.y - b \leq f(y), \forall y \in \mathbb{R}\}$, where the supremum is taken over all $(s, b)$.

The following result on convexification gaps is a special case of known results most of them investigated by Falk (see [3]). We will reproduce the formal proof below mainly to define some notations used later.

**Proposition 1.** Suppose that each function $f_j$ is bounded below and that problem (1–3) has an optimal solution $x^*$ with optimal value $z^*$. Then

$$\tilde{z} = \inf_{x \in X} \left\{ \sum_j \text{conv } f_j(x_j) \right\}$$

is a lower bound of the optimal value, i.e. $z^* \geq \tilde{z}$.

Moreover,

$$z^* - \tilde{z} \leq \sum_j \max_{x} \left[ f_j(x_j) - \text{conv } f_j(x_j) \right] = \Delta.$$

**Proof.** From the definition of $\tilde{z}$, we have

$$\tilde{z} \leq \sum_j \text{conv } f_j(x_j), \quad \forall x \in X.$$

Then, using the definition of the convex hull of each arc function, we obtain that $\tilde{z} \leq \sum_j f_j(x_j), \quad \forall x \in X$, and thus $\tilde{z} \leq z^*$.

Let $\tilde{x} \in X$ be the vector with components $\tilde{x}_j$ for each arc $j$ such that $\tilde{z} = \sum_j \text{conv } f_j(\tilde{x}_j)$.

$$z^* - \tilde{z} = \sum_j f_j(x^*_j) - \sum_j \text{conv } f_j(\tilde{x}_j)$$
\[ \leq \sum_j [f_j(\tilde{x}_j) - \text{conv } f_j(\tilde{x}_j)] \]
\[ \leq \sum_j \max_{x_j} [f_j(x_j) - \text{conv } f_j(x_j)] = \Delta. \] (4)

**Observations.** (1) The maximal gap \( \Delta \) associated with the above lower bound is, in general, greater than zero as the convex hull of the sum of a set of functions is, in general, different from the sum of the convex hulls of each function. This is true even if the functions are separable because we have coupling constraints, constraints (3).

(2) Our motivation here comes from the fact that the convex hull of certain one-dimensional functions is relatively easy to compute explicitly. In such a case, we can compute \( \tilde{z} \) and the corresponding solution \( \tilde{x} \) by solving the convex multicommodity flow problem by any efficient algorithm (see [9] or [11] for example). If then, \( f_j(\tilde{x}_j) = \text{conv } f_j(\tilde{x}_j), \forall j \), we can conclude that \( \tilde{x} \) is an optimal solution and \( \tilde{z} \) is the global optimal value of the problem. Otherwise, if the effective gap at \( \tilde{x} \) is greater than a given tolerance, we can switch to a local improvement method, taking \( \tilde{x} \) as a starting point, to reduce it gradually.

### 3. Properties of the arc cost function

The cost function is assumed to be separable with respect to arcs and is intended to integrate quality of service (on the arc) and expansion cost. Quality of service is typically an increasing function which measures congestion on the arc where commodities are considered as competitive users of a limited resource.

The choice of the cost function on the arc, denoted hereafter by \( f(x) \), omitting temporarily the arc index \( j \), does not intend to represent a specific traffic model, but should preserve the general characteristics of a large family of congestion measures and should be sufficiently simple to permit explicit computation as will be shown.

Let \( \pi \) be the fixed cost of expanding the arc, \( f^0 \) and \( f^1 \) be the congestion cost functions before and after expansion. Then, the integrated cost function is defined on \([0, c_1]\) by
\[ f(x) = \min\{ f^0(x), f^1(x) + \pi \}. \] (5)

Fig. 1 illustrates the integrated arc cost function \( f(x) \), as it is constructed on the basis of the congestion functions \( f^0 \) and \( f^1 \) and of the fixed expansion cost \( \pi \). We remark that, from properties (6–7), when \( x \) increases from 0, initially \( f(x) \) assumes the value of \( f^0(x) \). Because of (8), the difference between the congestion costs \( f^0(x) \) and \( f^1(x) \) increases until it reaches the value \( \pi \), when a breakeven point

\[ f^0(0) = f^1(0) = 0, \] (6)
\[ f^1(x) < f^0(x) \quad \forall x \in (0, c_1), \] (7)
\[ f^1(x) < f^{0'}(x) \quad \forall x \in (0, c_1), \] (8)
\[ f^{0'}(c_0) \geq M \quad \text{and} \quad f^{1'}(c_1) \geq M. \] (9)

**Observations.** (1) Relation (7) means that the congestion functions decrease when the capacity increases.

(2) Relation (8) means that derivatives induce increasing marginal costs.

(3) As the congestion increases with the flow, we force the derivative at the saturation level to be greater than a given high value \( M \) in relation (9).
1. The Kleinrock barrier function (see [4]):

$$f^n(x) = \frac{x}{c_n - x}, \quad n = 0, 1$$

Here $f^n$ and $f^{n'}$ tend to infinity when $x \uparrow c_n$. This is the kind of function which has been used to trace the integrated function in Fig. 1.

2. Quadratic function

To satisfy the above hypotheses with a quadratic function, we might use

$$f^n(x) = a_n x^2 + b_n x, \quad n = 0, 1$$

The parameter $b_n$ is the derivative at $x = 0$ and could be equal to $\pi_n / c_n$, thus representing the cost to transmit one unit of capacity across the channel. In order to satisfy (9), we might fix the derivative at $x = c_n$ to the value $M$, i.e.

$$2a_n c_n + \pi_n = M,$$

thus resulting in

$$a_n = \frac{M - \pi_n}{2c_n}.$$

As $M$ is large, we can verify that $a_n > 0$ and $a_n > a_{n+1}$.

4. An example of an explicit gap value

For the sake of simplicity, we will work with a pure quadratic function, neglecting the linear term. We will consider the case of two quadratic congestion functions and compute explicitly the convex hull of $f$. As $2a_n c_n = M$, the coefficient $a_n$ will be such that

$$a_0 = \frac{c_1}{c_0},$$

We can observe in Fig. 1 that $f$ is a piecewise convex function with a unique break point $\gamma$, representing the flow for which it is more economic to increase the arc capacity from $c_0$ to $c_1$. We begin by computing the break point $\gamma$ and the value $f(\gamma)$ by solving

$$a_0 \gamma^2 = a_1 \gamma^2 + \pi$$

which leads to:

$$\gamma = \sqrt{\frac{\pi c_1}{a_0(c_1 - c_0)}}$$

and

$$f(\gamma) = \pi \frac{c_1}{c_1 - c_0}.$$
Now, we can compute the maximal lower bound $\hat{z}$ induced by the convex hull of $f$:

$$\hat{z} = a_0 \frac{c_0}{c_1} \gamma^2 + 2a_0 \sqrt{\frac{c_0}{c_1}} (\gamma - \sqrt{\frac{c_0}{c_1} \gamma})$$

$$= a_0 \left( 2 \sqrt{\frac{c_0}{c_1}} - \frac{c_0}{c_1} \right) \gamma^2.$$ 

Thus, for $a_0 = 1$, the gap value associated with a single arc network is given by

$$\Lambda = f(\gamma) - \hat{z} = n \left( \sqrt{c_1} - \sqrt{c_0} \right)^2.$$

**Example.** Taking $c_0 = 64\text{kb/s}$ and $c_1 = 128\text{kb/s}$, we obtain a relative gap $\Lambda / f(\gamma)$ equal to $(\sqrt{c_1} - \sqrt{c_0})^2 / c_1$ which is here only 8%.

**5. Final comments**

This paper has proposed a modelling framework to integrate design and operation in modern backbone networks. In practice, the capacity of an arc may be expanded from $c_0$ to any value of an ordered set of standard capacity levels $c_1, c_2, \ldots, c_N$. Let $\pi_n$ be the fixed cost of expanding the arc from the capacity $c_0$ to the capacity $c_n$, for $n = 1, 2, \ldots, N$. Let $f^0$ be the congestion cost function before expansion, and let $f^n, n = 1, 2, \ldots, N$ be the congestion cost function after the corresponding expansion. Then, instead of (5), the integrated cost function is defined on $[0, c_N]$ by

$$f(x) = \min \{ f^0(x), f^1(x) + \pi_1, \ldots, f^N(x) + \pi_N \}$$

and a sequential reduction on the number of pieces of the convex hulls can invoke the approach studied here for this general case.

Our analysis fits in the framework referred to as *topology tuning* by Gerla et al. [7], since we address the issue of bandwidth allocation together with the conventional use of routing strategies. Indeed, our procedure may run as part of network management for large networks for which the dynamic adjustment of bandwidth is too complex to be carried out manually. We have shown how to consider discrete changes of capacity within a continuous modelling framework, and we have provided theoretical means in order to evaluate heuristic solutions with respect to global optimality in this class of problems.

**References**