This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

Multi-Period Network Rate Allocation with End-to-End Delay Constraints

Mohammad H. Hajiesmaili, M. Sadegh Talebi, and Ahmad Khonsari

***Abstract*—QoS-aware networking applications such as real-time streaming and video surveillance systems require nearly fixed average end-to-end delay over long periods to communicate efficiently, although may tolerate some delay variations in short periods. This variability exhibits complex dynamics that makes rate control of such applications a formidable task. This paper addresses rate allocation for heterogeneous QoS-aware applications that preserves the long-term end-to-end delay constraint while seeking the maximum network utility cumulated over a fixed time interval. To capture the temporal dynamics of sources, we incorporate a novel time-coupling constraint in which delay-sensitivity of sources is considered such that a certain end-to-end average delay for each source over a pre-specified time interval is satisfied. We propose an algorithm, as a dual-based solution, which allocates source rates for the next time interval in a distributed fashion, given the knowledge of network parameters in advance. Also, we extend the algorithm to the case that the problem data is not known fully in advance to capture more realistic scenarios. Through numerical experiments, we show that our proposed algorithm attains higher average link utilization and a wider range of feasible scenarios in comparison with the best, to our knowledge, rate control schemes that may guarantee such constraints on delay.**

***Index Terms*—Rate allocation, network utility maximization, end-to-end delay, convex optimization.**

I. Introduction

NOWADAYS, a plethora of emerging computer applications exhibit delay-sensitivity and may require some guarantee on delay. Some of such applications are less sensitive to the instantaneous delay, but rather, concern the end-to-end delay averaged over some specified time interval. A notable instance is media streaming where the average end-to-end delay is obliged not to exceed some threshold to ensure continuous playback. Several other examples include some applications of Wireless Sensor Networks (WSNs) and networked control systems. Due to temporal variations in both source traffic and network characteristics, in such scenarios it is crucial to accomplish rate allocation capable of capturing such dynamic behavior.

Rate allocation in networked systems has been well studied in the framework of Network Utility Maximization (NUM); see, e.g., [2]–[4]. In its simplest form, NUM concerns a network of sources connected through a set of links of fixed capacity. To each source, a utility function is associated, which maps its rate to its perceived quality and/or preference [4]. The notion of utility function indeed quantifies the satisfaction of the source for a given amount of allocated resource. Each source sends packets towards its destination through a subset of links referred to as its route. The fixed routing structure and link capacities dictate a set of linear capacity constraints. The goal of NUM is to find an allocation of source rates maximizing the network utility given capacity constraints.

M. H. Hajiesmaili is with the Whiting School of Engineering, the Johns Hopkins University, Baltimore, United States (e-mail: hajiesmaili@jhu.edu).

M. S. Talebi is with the Department of Automatic Control, School of Electrical Engineering, KTH The Royal Institute of Technology, 10044, Stockholm, Sweden (e-mail: mstms@kth.se).

A. Khonsari is with the Department of Electrical and Computer Engineering, The University of Tehran, Tehran 14176-14418, Iran, and also with the School of Computer Science, Institute for Research in Fundamental Sciences, Tehran 19538-33511, Iran (e-mail: a_khonsari@ut.ac.ir; ak@ipm.ir).

The preliminary version of this paper [1] has appeared in IEEE CDC'15.

A number of studies have thus far incorporated end-to-end delay into the NUM model [5]–[12]. In these studies, end-to-end delay either is included in the objective function of the problem (as in, e.g., [6], [11], [12]) or introduced constraints (as in, e.g., [5], [8], [9]). We further discuss these two lines of research in Section II. Amongst the latter works, Qiu et al. [9] consider a NUM formulation with some constraints on delay. Interestingly, their delay model delicately integrates with the NUM as it does not require precise knowledge of underlying packet arrival models, but rather, relies on the first order derivative of the delay function. Their formulation, however, cannot be effectively applied to the case of rate allocation with average end-to-end delay requirements. The main drawback in doing so is that [9] considers a *one-shot* or *static* allocation, namely it only concerns the instantaneous delay instead of the average delay. Therefore, the approach of [9] may result in a conservative rate allocation, in particular for applications capable of tolerating instantaneous delay in short intervals.

In order to tailor rate allocation to cases where some guarantees on average end-to-end delays are required, one should instead seek a *dynamic* allocation supporting source rate fluctuations. To this aim, in the present paper we consider a *multi-period* NUM that strives to maximize the network utility aggregated over a fixed time interval. In this multi-period formulation, utilities are coupled across time through end-to-end delay constraints whereas they are made dependent across sources through capacity constraints. By permitting rate fluctuations over time, the resulting allocation may sacrifice network utility in some periods so as to maintain delay while compensating for it in some other periods. Arguably, this property essentially enlarges the set of feasible scenarios in comparison with, e.g., [9]. In addition, it endows us the ability of maintaining several delay constraints for each source, where each delay constraint is relevant to a particular time interval.

We make the following contributions:

1. We present a multi-period NUM formulation capturing a set of general constraints on average end-to-end delay, coupled

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

over time. The form of constraints we consider is more general than those considered in [9], [13], and [14].

2. We present an algorithm (in Section IV) for the proposed multi-period NUM problem, called *Delay-Aware Dynamic Network Utility Maximization* (DA-DNUM), which is a distributed algorithm that leverages dual decomposition approach [4], [15]. We provide convergence analysis of the DA-DNUM algorithm in Theorem 1 and Proposition 2, where we derive explicit bounds on the step size of the algorithm. This is in contrast with existing Lyapunov-based convergence analysis in similar studies (as in, e.g., [9], [13]), where no bound on the step size is reported. Through numerical experiments, we examine the performance of the DA-DNUM algorithm and provide some comparison scenarios to demonstrate its superiority over the relevant existing rate allocation schemes [9], [13] that are delay-aware variants of the basic single-period NUM.

3. DA-DNUM relies on the knowledge of network parameters for the next time interval in advance. To relax this requirement, in Section V we provide an algorithm based on model predictive control [16], which approximately solves a variant of the problem that relies on the estimate of link capacities in each time interval.

The rest of this paper is organized as follows. First, in Section II we briefly review the related work. In Section III, we introduce the temporal-aware system model, characterize delay constraints, and state the rate allocation problem. In Section IV, we present an iterative solution to the rate allocation problem, present DA-DNUM, and provide theoretical guarantees on its convergence. Section V is devoted to introduce a solution when the network parameters are not known in advance. Section VI presents some experimental results. Finally, in Section VII we provide concluding remarks and outline some future directions.

A. Basic Notations and Terminologies

We use the following notations. For any vector $\mathbf{z}$ (resp. matrix Z), $\mathbf{z} \geq 0$ (resp. $Z \geq 0$) implies that all components of vector $\mathbf{z}$ (resp. matrix Z) are nonnegative. The vector $\mathbf{e}_j$ denotes the j-th unit vector. The operator $\|\cdot\|$ is the standard Euclidean norm. The domain of a function f is denoted by $\mathbf{dom}\ f$. Moreover, $\mathbb{1}\{A\}$ is the indicator of event A, i.e. it equals 1 if A occurs; it is 0 otherwise. We use $[\cdot]_{\mathcal{P}}$ to denote the projection onto the set $\mathcal{P}$. Projection onto nonnegative orthant will be shown by $[\cdot]^+$.

In what follows, we give some necessary definitions that can be found in, e.g., [17].

Definition 1. *A function $f(\cdot)$ is a G-Lipschitz function if*

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq G\|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbf{dom}\ f.$$

Definition 2. *A convex function $f(\cdot)$ is κ-strongly convex if and only if there exists a constant $\kappa > 0$ such that the function $f(\mathbf{x}) - \frac{\kappa}{2}\|\mathbf{x}\|^2$ is convex.*

Note that if $f(\cdot)$ is twice differentiable, then $f(\cdot)$ is κ-strongly convex if there exists constant κ such that $\nabla^2 f(\mathbf{x}) - \kappa I$ is positive semidefinite.

II. Related Work

In the recent years, many studies have leveraged the NUM framework to propose efficient protocols and algorithms for network applications under different types of utility models, capturing different traffic characteristics, and constraints (we refer to [2] and the references therein). We may identify two lines of work that investigated delay within the NUM framework.

A. Delay as Objective Function

Li et al. [6] consider a NUM formulation in which delay penalizes the utility function. In other words, the corresponding NUM seeks to maximize the aggregate utility of sources while reducing the end-to-end delays. Based on a delay-sensitive utility function introduced in [18], authors in [11] and [12] present application-oriented rate allocation schemes employing an alternative utility definition. Both approaches, however, fail to provide any guarantees for the delay, thus becoming improper for QoS-aware applications with hard long term average delay requirements.

B. Delay as Constraint

In another line of works as in, e.g., [5], [8], [9], [19], the delay experienced by sources is incorporated into the NUM formulation as constraints. By introducing the notion of Virtual Link Capacity Margin (VLCM) to characterize source delay as constraint of the problem, the authors of [9] and [13] propose a joint rate allocation and scheduling scheme in multi-hop wireless networks. Using a different approach than that of [9], Dogahe et al. [8] present a NUM formulation to address joint power and rate control under source delay constraints. Moreover, in [5], using an elegant fluid model of multi-class flows with different delay requirements, another distributed and stable delay-aware algorithm is proposed. It is remarkable that lack of support for temporal variation of network and source characteristics in single-period NUM may lead to a broad range of infeasible rate allocation scenarios. We further investigate this issue in our numerical evaluations in Section VI.

C. Extensions to Dynamic Network Utility Maximization

We conclude this section by mentioning some studies that consider the multi-period variant of NUM. To the best of our knowledge, such a variant is first investigated in [14], where it is referred to as *Dynamic NUM (DNUM)*. In DNUM, utilities are coupled across time through *delivery contracts*, which impose a set of linear constraints on source rates. The DNUM framework has been extended in several ways by subsequent works (e.g., [20], [21]). In [20], focusing on video streaming applications, temporal variations of source rates are incorporated into the utility functions. It is remarkable that though the notion of delivery contracts could capture some type of QoS constraints, they cannot model end-to-end delay requirements. To the best of our knowledge, none of the studies that investigate multi-period NUM problems similarly to [14] (e.g., [20], [21]) provide guarantees on delay.

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

III. Model and Problem Formulation

A. Network Model

Our model is based on that of DNUM [14], which considers rate allocation over a discrete-time interval $\mathcal{T} = \{1, \ldots, T\}$[1]. We consider a network formed by a set $\mathcal{L} = \{1, \ldots, L\}$ of links shared among a set $\mathcal{S} = \{1, \ldots, S\}$ of sources. We represent the possibly time-varying routing in the network in time period t by a routing matrix $R_t = [(R_t)_{ls}]_{L\times S}$, where

$$(R_t)_{ls} = \begin{cases} 1 & \text{if source } s \text{ passes through link } l \text{ at period } t \\ 0 & \text{otherwise.} \end{cases}$$

We let c_{tl} denote the capacity of link l at period t and $\mathbf{c}_t = [c_{tl}]_{l\in\mathcal{L}}$ be the corresponding vector of link capacities.

Moreover, we let $x_{st} \in \mathcal{X}_s$ be the transmission rate of source s at period t, where we define $\mathcal{X}_s := [w_s, W_s]$ with w_s and W_s respectively denoting the minimum and the maximum rates of source s.[2] We let $X = [x_{st}]_{S\times T}$ be the *rate matrix* and define $\mathcal{X} = \{X \in \mathbb{R}^{S\times T} : x_{st} \in \mathcal{X}_s\}$.

B. Capacity Constraints

To introduce capacity constraints, we first recall the definition of *link margin* variables from [9]. For each link l and time period t, link margin variable σ_{tl} is defined as the difference between capacity of link l and the maximum allowable flow passing through it [9]. Unlike [9], however, our setup does not admit schedulability constraints and hence we proceed to formulate link margin as follows. Consider capacity constraint for link l at period t given by

$$\sum_{s\in\mathcal{S}} (R_t)_{ls} x_{st} + \sigma_{tl} = c_{tl} \quad \text{and} \quad \sigma_{tl} \ge 0.$$

We relax the equality constraint above and establish the following constraints for link l at period t:

$$\sum_{s\in\mathcal{S}} (R_t)_{ls} x_{st} + \sigma_{tl} \le c_{tl} \quad \text{and} \quad \sigma_{tl} \ge 0.$$

Although the relaxation above constricts resource usage (i.e., capacity), it will play an important role in limiting the flow of link l and thereby proves essentially useful to control the queuing delay of link l. Introducing $\boldsymbol{\sigma}_t = [\sigma_{tl}]_{l\in\mathcal{L}}$ and $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_t]_{t\in\mathcal{T}}$, we then represent the capacity constraints compactly as

$$R_t X \mathbf{e}_t + \boldsymbol{\sigma}_t \le \mathbf{c}_t \quad \text{and} \quad \boldsymbol{\sigma}_t \ge 0, \quad \forall t \in \mathcal{T}. \tag{1}$$

These constraints constitute a set of $2T \times L$ linear inequalities.

C. Average Delay Constraints

Having defined the notion of link margin, we next introduce the link delay as a function of link margin. For simplicity of presentation, we assume that all links have the same delay function $D(\cdot)$, which maps link margin variable of a given link to its queuing delay. That is, $D(\sigma_{tl})$ equals the delay of link l at period t. Clearly, the dependence of $D(\sigma)$ on σ is determined by the packet arrival process model. For instance, for M/M/1 queuing model whose packet arrival is a Poisson process, we have

$$D(\sigma_{tl}) = \frac{1}{\sigma_{tl}}. \tag{2}$$

Another notable instance is the case of M/G/1 queuing model whose delay function is given in [8], [22].

In what follows, we make the following assumptions on the delay function $D(\cdot)$:

A1. $|D'(z)| \le \alpha_D$ for all $z \in \mathbf{dom}\ D$.
A2. $D(\cdot)$ is κ_D-strongly convex.

We remark that assumptions **A1** and **A2** are provided to facilitate presentation of the analysis in Section IV for the generic delay function $D(\cdot)$, and we may compute α_D and κ_D as soon as a problem instance is specified. These assumptions can be relaxed if lower and upper bounds on σ_{tl} for all t and l are known, which would be possible as soon as delay function $D(\cdot)$ is specified. This property is satisfied by most delay functions, e.g., those for M/M/1 and M/G/1 queues. In particular, we refer to Proposition 2, where we compute α_D and κ_D in terms of problem parameters for the case of M/M/1 arrival process, whose delay function is given in (2).

In the present study, we only consider queuing delays and hence, for each source s, we obtain the end-to-end delay by simply adding up all link delays along the path of s. Writing ϕ_{st} for the end-to-end queuing delay of source s at period t, we get

$$\phi_{st} = \sum_{l\in\mathcal{L}} (R_t)_{ls} D(\sigma_{tl}).$$

We further introduce $\boldsymbol{\phi}_s = [\phi_{st}]_{t\in\mathcal{T}}$. Next, we define the constraint on average end-to-end delay as follows: Assume that source s requires its average end-to-end queuing delay over some interval of interest $\mathcal{T}_\Delta \subseteq \mathcal{T}$ with length Δ be less than some constant d. This constraint is formally given by

$$\frac{1}{\Delta} \sum_{t\in\mathcal{T}_\Delta} \phi_{st} \le d. \tag{3}$$

To model a general scenario for the introduced delay constraint, we assume that each source s declares K_s delay constraints of the form (3), indexed by $k \in \mathcal{K}_s = \{1, \ldots, K_s\}$. Each delay constraint $k \in \mathcal{K}_s$ concerns a specific time interval, where overlap between various intervals is allowed. In order to encode delay constraints of the form (3), for each source s, we introduce the *delay indicator matrix* $M_s = [(M_s)_{kt}]_{K_s\times T}$ as follows

$$(M_s)_{kt} = \begin{cases} \frac{1}{G_k^s} & \text{if } k\text{-th delay constraint of } s \text{ concerns } t, \\ 0 & \text{otherwise,} \end{cases}$$

where $G_k^s = \sum_{t\in\mathcal{T}} \mathbb{1}\{(M_s)_{kt} \neq 0\}$. Now, we can write the k-th delay constraint of source s as

$$\sum_{t\in\mathcal{T}} (M_s)_{kt} \phi_{st} \le d_{sk},$$

[1]The duration of each period t and the whole time horizon T is an application-specific design parameter. As an example, in a previous work [20], where the underlying application is video streaming, each period is set based on the length of the video frames and the time horizon T is determined according to the length of GoPs (Group of Pictures).

[2]The results in this paper can be easily extended to the case where $\mathcal{X}_s$ is dependent on t.

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

where d_{sk} is the average delay requirement of source s for its k-th delay constraint. Note that the elements of every row of M_s add up to one and therefore, we may interpret the left-hand side of the constraint above, like that of (3), as the end-to-end queueing delay of s averaged over time instants $\{t \in \mathcal{T} : (M_s)_{kt} = 1\}$. Defining $\mathbf{d}_s = [d_{sk}]_{k\in\mathcal{K}_s}$ yields the following vector representation for delay constraints:

$$M_s \boldsymbol{\phi}_s \leq \mathbf{d}_s, \qquad \forall s \in \mathcal{S}. \tag{4}$$

These constraints constitute a set of $\sum_{s\in\mathcal{S}} K_s$ inequalities that are nonlinear in $\boldsymbol{\sigma}$.

1) An Illustrative Example in Mission-Oriented WSNs: To motivate the appropriateness of the model above, we provide a practical application of it in mission-oriented wireless sensor networks (WSN) [23]. In such WSNs, there are several coexisting applications (henceforth, *missions*). Let us look at a surveillance application that employs various types of sensors such as video, motion detector, and thermal sensors, to provide assistive ambient intelligence in, e.g., disaster recovery scenarios.

The naive approach is to require each sensor to periodically transmit its readings at specific time intervals. Albeit simple to implement, this approach is inefficient as each mission might possess particular QoS requirement in terms of end-to-end delay. For instance, a video mission may demand for a long-time delay constraint to work efficiently. In contrast, the thermal mission may report the temperature periodically on a regular basis and thereby declares a short-term delay requirement at certain periods.

In order to achieve the best efficiency, it is therefore crucial to properly schedule transmissions. Besides other parameters, one could set $\mathcal{T}_\Delta = \mathcal{T}$ for the real-time video mission, as it records and streams data to the sink continuously. The choice of $\mathcal{T}_\Delta$ for the thermal sensor is different and could be set, for example in the case of $T = 60$, as follows: $\mathcal{T}_{\Delta_1} = \{1, 2, 3\}$, $\mathcal{T}_{\Delta_2} = \{21, 22, 23\}$, and $\mathcal{T}_{\Delta_3} = \{41, 42, 43\}$. Hence, the thermal sensor reports its readings in 3 different steps as mentioned above.

In summary, one can identify several other application scenarios (such as in emerging Internet of Things (IoT) or networked control systems), where different competing goals (missions in some contexts) with diverse QoS characteristics coexist under a unified application, but with heterogeneous requirements.

D. Optimization Problem

We associate a utility function $U_{st}(\cdot)$ to each source s at period t. The value $U_{st}(x_{st})$ quantifies satisfaction of source s at period t when it sends data at rate x_{st} in that period [2], [4]. We make the following assumptions for utility functions: for all s and t,

A3. $U_{st}(\cdot)$ is continuous, monotonically increasing, and twice continuously differentiable.
A4. $U_{st}(\cdot)$ is α_U-Lipschitz.
A5. $-U_{st}(\cdot)$ is κ_U-strongly convex.

TABLE I: Key Notations

Notation	Definition
$\mathcal{T}$	The set of time slots (periods), $T := \lvert\mathcal{T}\rvert$
$\mathcal{L}$	The set of links, $L := \lvert\mathcal{L}\rvert$
$\mathcal{S}$	The set of sources, $S := \lvert\mathcal{S}\rvert$
R_t	The routing matrix at period t
c_{tl}	The capacity of link l at period t
σ_{tl}	Link margin of link l at period t
$D(\cdot)$	The delay function of any link
x_{st}	The transmission rate of source s at t
w_s	The minimum rate of source s
W_s	The maximum rate of source s
ϕ_{st}	The end-to-end queuing delay of source s at period t
$\mathcal{K}_s$	The index set of delay constraints of source s, $K_s := \lvert\mathcal{K}_s\rvert$
M_s	The delay indicator matrix of source s
d_{sk}	The average delay requirement of source s for its k's delay constraint
$U_{st}(\cdot)$	The utility function of source s at t

Similarly to [14], we define the network utility $U(\cdot)$ as the sum of all utilities over time horizon $\mathcal{T}$:

$$U(X) = \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}} U_{st}(x_{st}).$$

Now, we cast the rate allocation problem as

$$\begin{aligned} \mathsf{P1}: \quad & \max_{X\in\mathcal{X},\boldsymbol{\sigma}\geq 0} \quad U(X) \\ & \text{subject to}: \ R_t X \mathbf{e}_t + \boldsymbol{\sigma}_t \leq \mathbf{c}_t, \qquad \forall t \in \mathcal{T}, \\ & \qquad M_s \boldsymbol{\phi}_s \leq \mathbf{d}_s, \qquad \forall s \in \mathcal{S}, \\ & \qquad \phi_{st} = \sum_{l\in\mathcal{L}} (R_t)_{ls} D(\sigma_{tl}), \quad \forall s \in \mathcal{S}, \forall t \in \mathcal{T}. \end{aligned}$$

First we highlight that constraints of P1 constitute a compact set. Hence, at least one optimal solution exists. Furthermore, P1 is a strongly convex optimization problem, and hence its optimal solution is unique. We remark that P1 is nonseparable over time due to coupled delay constraints. It's worth noting that in the absence of average delay constraints, problem P1 degenerates to DNUM problem of [14] without delivery contracts. In the above formulation, we address QoS requirements mainly through end-to-end delay constraints and thus avoid augmenting delivery contracts, i.e. linear constraints on source rates over $\mathcal{T}$. We stress, however, that the solution procedure below permits having delivery contracts as well.

The main notations of the paper are summarized in Table I.

IV. Optimal Rate Allocation Algorithm

In this section, we solve P1 and develop a distributed rate allocation algorithm. First, we provide the following proposition for the strong duality of P1.

Proposition 1. *Problem* P1 *has the strong duality property.*

Proof. Problem P1 is a convex problem since (i) objective function is a nonnegative sum of concave functions (see Assumptions **A3-A5**), (ii) capacity constraints are linear, and (iii) average delay constraints are convex constraints (see Assumption **A2**). In addition, there exist strictly feasible points; for example we can set $x_{st} = w_s$, for all s. Recall that $w_s \geq 0$ is the minimum rate requirement of source s, which is the input to the problem and ensures that there exists at least one feasible solution. Hence, Slater's constraint qualification (see,

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

$$\begin{aligned} L(X,\boldsymbol{\sigma},\boldsymbol{\lambda},\boldsymbol{\mu}) &= \sum_t \sum_s U_{st}(x_{st}) - \sum_t \boldsymbol{\lambda}_t^\top \left(R_t X \mathbf{e}_t - \mathbf{c}_t + \boldsymbol{\sigma}_t\right) - \sum_s \boldsymbol{\mu}_s^\top \left(M_s \boldsymbol{\phi}_s - \mathbf{d}_s\right) \\ &= \sum_t \sum_s \left(U_{st}(x_{st}) - \lambda^{st} x_{st}\right) - \sum_t \sum_l \left(\mu^{tl} D(\sigma_{tl}) + \lambda_{tl}\sigma_{tl}\right) + \sum_t \boldsymbol{\lambda}_t^\top \mathbf{c}_t + \sum_s \boldsymbol{\mu}_s^\top \mathbf{d}_s. \end{aligned} \tag{5}$$

e.g., [15, pp. 226–227]) is satisfied and strong duality holds for problem P1. □

An immediate consequence of Proposition 1 is that we can solve P1 optimally through its dual. To this end, we introduce the following notations. We let $\boldsymbol{\lambda}_t = [\lambda_{tl}]_{l\in\mathcal{L}}$ and $\boldsymbol{\mu}_s = [\mu_{sk}]_{k\in\mathcal{K}_s}$ respectively denote the Lagrange multipliers (dual variables) associated to the capacity constraints for period t and average delay constraints for source s. Moreover, we let $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_t]_{t\in\mathcal{T}}$ and $\boldsymbol{\mu} = [\boldsymbol{\mu}_s]_{s\in\mathcal{S}}$.

We introduce the partial Lagrangian of P1 in (5), shown at the top of this page, where

$$\begin{aligned} \lambda^{st} &:= \sum_l (R_t)_{ls}\lambda_{tl}, \\ \mu^{tl} &:= \sum_s \sum_{k\in\mathcal{K}_s} (M_s)_{kt}(R_t)_{ls}\mu_{sk}. \end{aligned}$$

The dual function of P1, denoted by $g(\boldsymbol{\lambda},\boldsymbol{\mu})$, is then given by:

$$\begin{aligned} g(\boldsymbol{\lambda},\boldsymbol{\mu}) &= \max_{X\in\mathcal{X},\boldsymbol{\sigma}\geq 0} L(X,\boldsymbol{\sigma},\boldsymbol{\lambda},\boldsymbol{\mu}) \\ &= \max_{X\in\mathcal{X}} \sum_t \sum_s \left(U_{st}(x_{st}) - \lambda^{st}x_{st}\right) \\ &\quad + \max_{\boldsymbol{\sigma}\geq 0} \sum_t \sum_l \left(\mu^{tl}D(\sigma_{tl}) + \lambda_{tl}\sigma_{tl}\right). \end{aligned} \tag{6}$$

Thus, the dual problem associated to P1 is [17]:

$$\mathsf{D1}: \quad \min_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}\geq 0} g(\boldsymbol{\lambda},\boldsymbol{\mu}).$$

Given $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, let $X^\star = [x^\star_{st}]_{T\times S}$ and $\boldsymbol{\sigma}^\star_t = [\sigma^\star_{tl}]_{l\in\mathcal{L}}$ be the maximizers of maximization problems in (6). To derive these solutions, first note that partial derivatives of the Lagrangian are given by:

$$\begin{aligned} \frac{\partial L}{\partial x_{st}} &= U'_{st}(x_{st}) - \lambda^{st}, \quad \forall s, \forall t, \\ \frac{\partial L}{\partial \sigma_{tl}} &= \mu^{tl}D'(\sigma_{tl}) + \lambda_{tl}, \quad \forall t, \forall l. \end{aligned}$$

The maximizers are stationary points of the Lagrangian. We therefore get, through preliminary manipulations,

$$\begin{aligned} x^\star_{st}(\boldsymbol{\lambda}) &= \left[U'^{-1}_{st}(\lambda^{st})\right]_{\mathcal{X}_s}, \quad \forall s, \forall t, \\ \sigma^\star_{tl}(\boldsymbol{\lambda},\boldsymbol{\mu}) &= \left[D'^{-1}\left(-\frac{\lambda_{tl}}{\mu^{tl}}\right)\right]^+, \quad \forall t, \forall l. \end{aligned} \tag{7}$$

One consequence of strong convexity of P1 is that the dual function $g(\boldsymbol{\lambda},\boldsymbol{\mu})$ is differentiable in its domain (see, e.g., [17]). Hence, we can employ the *gradient projection method* [17] to solve D1. We therefore use Danskin's Theorem [17, Proposition B.25],[3] to obtain partial derivatives of dual function $g(\boldsymbol{\lambda},\boldsymbol{\mu})$:

$$\begin{aligned} \frac{\partial g}{\partial \lambda_{tl}} &= c_{tl} - \sigma_{tl} - \sum_s (R_t)_{ls}x_{st}, \quad \forall l, \forall t, \\ \frac{\partial g}{\partial \mu_{sk}} &= d_{sk} - \sum_t \sum_l (M_s)_{kt}(R_t)_{ls}D(\sigma_{tl}), \quad \forall s, \forall k \in \mathcal{K}_s. \end{aligned}$$

Hence, we obtain the following relation for the update of the dual variable for link l at time t:

$$\lambda^{(j+1)}_{tl} = \left[\lambda^{(j)}_{tl} + \gamma\left(\sum_s (R_t)_{ls}x^{(j)}_{st} + \sigma^{(j+1)}_{tl} - c_{tl}\right)\right]^+,$$

where $x^{(j)}_{st} = x^\star_{st}(\boldsymbol{\lambda}^{(j)})$, $\sigma^{(j+1)}_{tl} = \sigma^\star_{tl}(\boldsymbol{\lambda}^{(j)},\boldsymbol{\mu}^{(j)})$, and $\gamma > 0$ is a sufficiently small step size, which will be determined later. Similarly, for each source s and $k \in \mathcal{K}_s$, we update μ_{sk} as follows:

$$\mu^{(j+1)}_{sk} = \left[\mu^{(j)}_{sk} + \gamma\left(\sum_t \sum_l (M_s)_{kt}(R_t)_{ls}D(\sigma^{(j)}_{tl}) - d_{sk}\right)\right]^+.$$

Note that it is crucial to choose step size γ properly to guarantee the convergence of the iterative solution above. Given appropriate γ, update equations for dual variables converge to minimizers of D1. Strong duality then guarantees that optimal values of D1 and P1 coincide and that $X^\star$ and $\boldsymbol{\sigma}^\star$ can be obtained accordingly. The following theorem determines the values of the step size γ that guarantee the convergence of the iterative solution to the optimal solution of P1 for a generic delay function $D(\cdot)$.

Theorem 1. *Assume that γ satisfies $0 < \gamma < \frac{2}{Q}$, where*

$$\begin{aligned} Q :=& \frac{1}{\mu_{\min}\kappa_D}\left(1 + \alpha_D^2 TL\Big(\sum_s K_s\Big)^2 + 2\alpha_D\sqrt{TL\sum_s K_s}\right) \\ &+ \frac{LS}{\kappa_U} \end{aligned} \tag{8}$$

and $\mu_{\min} = \min_{t,l}\mu^{tl}$. Then, starting from any initial feasible point, the limit point $(X^\star,\sigma^\star,\mu^\star,\lambda^\star)$ of the sequence $\{X^{(j)},\sigma^{(j)},\mu^{(j)},\lambda^{(j)}\}_{j\geq 1}$ generated by the aforementioned iterative solution is primal-dual optimal and $(X^\star,\sigma^\star)$ is the unique optimal solution to P1.

Proof. See Appendix A. □

We stress that the parameter $\mu_{\min}$ is necessarily positive, since otherwise σ_{tl} for some (t,l) would tend to zero and consequently, some delay constraint in (4) would become violated.

[3] The Danskin's Theorem provides a formula for the derivative of functions of the form $f(z) = \max_{y\in Y}\phi(z,y)$, where Y is a compact set and $\phi(.,y)$ is a convex function for every $y \in Y$ [17, Proposition B.25].

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

When delay function $D(\cdot)$ is specified, the parameters $\mu_{\min}$, α_D, and κ_D can be expressed explicitly. As a consequence, an explicit bound on the step size γ can be obtained. The following proposition provides such an explicit bound for the case of M/M/1 packet arrival model. To set up the notation, let

$$d_{\min} = \min_{s,k\in\mathcal{K}_s} d_{sk}, \quad d_{\max} = \max_{s,k\in\mathcal{K}_s} d_{sk}, \quad c_{\max} = \max_{t,l} c_{tl}.$$

Proposition 2. *For the case of M/M/1 delay model, we have*

$$\mu_{\min} \geq \frac{LT^2\alpha_U}{d_{\min}^2}, \quad \alpha_D = T^2 d_{\max}^2, \quad \kappa_D = \frac{2}{c_{\max}^3}.$$

Proof. See Appendix B. □

Next we compare Theorem 1 with [9, Proposition 2]. Theorem 1 relies on Descent Lemma [17], which yields a step size bound guaranteeing convergence. In contrast, [9, Proposition 2] leverages a Lyapunov approach to guarantee convergence to the optimal solution. As such, it does not necessarily lead to an explicit bound on the step size, but rather, establishes convergence provided that the step size is sufficiently small. Arguably, the need for an explicit step size, similar to those provided in Theorem 1 and Proposition 2, is crucial from a practical standpoint. We thus believe that not only does Theorem 1 sound more interesting, but it also could be of more practical impact.

A. The DA-DNUM Algorithm

Next, we give a distributed iterative algorithm, named *Delay-Aware Dynamic Network Utility Maximization (DA-DNUM)*, which is based on a distributed implementation of the above iterative solution. As gradient-based algorithms are not finitely convergent, the DA-DNUM algorithm relies on a parameter `th` to determine the stopping criterion of the iterative procedure. The DA-DNUM algorithm relies on both the knowledge of network parameters in advance of time interval $\mathcal{T}$ and the ability of explicitly/implicitly exchanging dual variables between sources and links (more precisely, between each source s and links on the path of s). The pseudo-code of DA-DNUM is shown in Algorithm 1.

V. A Solution with Limited Future Knowledge

The solution presented in Sections IV is based on the assumption that the problem data (input parameters) for the entire time horizon is available ahead of time. Dependence of these solutions on the precise knowledge of future network parameters stimulates devising another scheme that efficiently works under uncertainty of the parameters. In this section, we extend our solution in a way such that the problem data is not fully known in advance. Without loss of generality, we assume that only the link capacities are revealed at the beginning of each period. Our approach in this section is based on a causality constraint so that the source rates at period t are a function of the link capacities up to period t. We further note that this is a convex stochastic problem, where the goal is to maximize the expected aggregated utility of all sources subject to the capacity, average delay, and causality constraints.

Algorithm 1 DA-DNUM Algorithm

Collect network parameters for the next time horizon $\mathcal{T}$.
Initialize $X^0, \boldsymbol{\sigma}^0, \boldsymbol{\lambda}^0$, and $\boldsymbol{\mu}^0$. Set $j = 0$.
while $\max_{s,l,t}\left\{|x_{st}^{(j+1)} - x_{st}^{(j)}|, |\sigma_{tl}^{(j+1)} - \sigma_{tl}^{(j)}|\right\} > \mathtt{th}$ **do**
At each link l, for each period t, obtain $\mu^{tl,(j)}$ and update:

$$\sigma_{tl}^{(j+1)} = \left[D'^{-1}\left(-\frac{\lambda_{tl}^{(j)}}{\mu^{tl,(j)}}\right)\right]^+,$$

$$\lambda_{tl}^{(j+1)} = \left[\lambda_{tl}^{(j)} + \gamma\left(\sum_s (R_t)_{ls} x_{st}^{(j)} + \sigma_{tl}^{(j+1)} - c_{tl}\right)\right]^+.$$

At each source s, for each period t, obtain $\lambda^{st,(j)}$ and compute:

$$x_{st}^{(j+1)} = \left[U_{st}'^{-1}\left(\lambda^{st,(j)}\right)\right]_{\mathcal{X}_s},$$

$$\mu_{sk}^{(j+1)} = \left[\mu_{sk}^{(j)} + \gamma\left(\sum_t\sum_l (M_s)_{kt}(R_t)_{ls} D(\sigma_{tl}^{(j)}) - d_{sk}\right)\right]^+.$$

$j \leftarrow j+1$
end while

Although, similarly to the conventional NUM problems this problem could be efficiently solved by centralized approaches, here we are interested in decentralized ones.

To this aim, we construct our solution based on Model Predictive Control (MPC) [16]. To calculate the source rates (x_{st}s) and link margin values (σ_{tl}s) for any particular period τ, instead of solving problem P1, we solve the following problem:

$$\begin{aligned}
\mathsf{P2:}\quad & \max_{x_{st}, s\in\mathcal{S}, t\in\mathcal{T}^\tau} \sum_{t\in\mathcal{T}^\tau}\sum_s U_{st}(x_{st}) \\
& \text{subject to:} \\
& \quad R_\tau X\mathbf{e}_\tau + \boldsymbol{\sigma}_\tau \leq \mathbf{c}_\tau, \\
& \quad R_t X\mathbf{e}_t + \boldsymbol{\sigma}_t \leq \hat{\mathbf{c}}(t|\tau), \qquad \forall t\in\mathcal{T}^\tau, \\
& \quad M_s\boldsymbol{\phi}_s \leq \mathbf{d}_s, \qquad \forall s\in\mathcal{S}, \\
& \quad \phi_{st} = \sum_l (R_t)_{ls} D(\sigma_{tl}), \quad \forall s\in\mathcal{S}, \forall t\in\mathcal{T},
\end{aligned}$$

where

$$\mathcal{T}^\tau = \{\tau+1, \ldots, T\}$$

and where

$$\hat{\mathbf{c}}(t|\tau) = E[\boldsymbol{c}_t|\boldsymbol{c}_1, \ldots, \boldsymbol{c}_\tau], \quad t\in\mathcal{T}^\tau$$

is the expected value of link capacities, given the entire information available at period τ. Consequently, in problem P2, at any period τ the whole information about link capacities up to (and including) period τ is revealed. Furthermore, for the future periods we use the conditional mean values of link capacities. Since each source can declare several average delay constraints, it is conceivable that some delay constraints are expired before beginning of period τ, i.e., the cases where the active interval of the constraint occurs within the first τ periods. Another situation is when a constraint has already been active. That is, the start time is "$\leq \tau$", while the final time is "$\geq \tau$". Here, the source rate for periods $\leq \tau$

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

are already calculated. Let us denote these source rates by $x'_{st}, s \in \mathcal{S}, t \in \{1, \ldots, \tau\}$. Then, the average delay inequality is interpreted as follows: the source rates for $t < \tau$ are taken to be x'_{st}. And so, this part is fixed and the constraint should be satisfied for the remaining part at period $t \geq \tau$.

Finally, problem P2 is a particular version of problem P1 and the optimal algorithm presented in Section IV could be leveraged to find its optimal solution. However, in each period τ, we pick the optimal source rate and link margin values for just the time slot τ and then we solve the problem again for the remaining time slots. For each time slot, the optimal values of the previous periods are used as input parameters.

VI. Numerical Evaluation

This section is devoted to the illustration of our numerical experiments. First, we examine DA-DNUM on a network consisting of few links and nodes. It is followed by the description of two comparison scenarios to investigate the scalability of DA-DNUM and its superiority over existing algorithms.

A. Experiment 1: A Simple Topology

In order to facilitate the discussion of the results, we have chosen a network with time-invariant routing and topology shown in Fig. 1, and set $T = 10$. For all $t \in \{1, \ldots, 10\}$, link capacities c_{t1} and c_{t4} are drawn uniformly at random from $[4, 6]$, and c_{t2} and c_{t3} are sampled uniformly at random from $[4, 10]$. We consider logarithmic utility functions, namely $U_{st}(x_{st}) = \log x_{st}$ for all s and t, which is widely used in the literature (see, e.g., [4]). Moreover, we assume M/M/1 queuing model for all links that corresponds to delay function $D(z) = \frac{1}{z}$. We give delay indicator matrices as well as vectors $\mathbf{d}_s$ for all s below:

$$M_1 = \frac{1}{3} \times \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix},$$
$$M_2 = \frac{1}{6} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$
$$M_3 = \frac{1}{6} \times \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix},$$
$$M_4 = \frac{1}{4} \times \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$
$$\mathbf{d}_1 = [2 \quad 1]^\top, \qquad d_2 = d_3 = 2, \qquad d_4 = 2.5.$$

We stress that the above delay indicator matrices imply that for $t = 9, 10$, there is no delay constraint and hence, for these periods P1 degenerates to DNUM of [14] without delivery contracts.

Fig. 2 displays the rate allocation results obtained from DA-DNUM with step size $\gamma = 0.01$ and $\mathtt{th} = 0.01$. For the sake of comparison, Fig. 2 also shows the rate allocation result of DNUM (without delivery contracts), which is obtained by solving P1 after removal of delay constraints. Fig. 2(a) depicts final source rates of the two cases. As expected, Fig. 2(a) exhibits the same values for both DA-DNUM and DNUM for $t = 9, 10$, since there is no delay constraint associated to these periods. By contrast, for $t = 1, \ldots, 8$ source rates obtained by DA-DNUM are lower than those obtained by DNUM. This stems from the existence of at least one delay constraint in any of these periods.

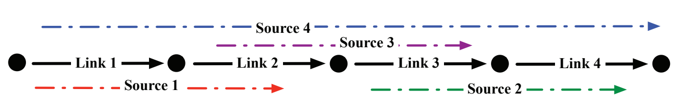


Fig. 1: Network Topology, Experiment 1

End-to-end queuing delays ϕ_{st} for all s and t are depicted in Fig. 2(b). To achieve higher system-wide aggregate utility, DA-DNUM allows some fluctuations in source delays during various periods, while guaranteing that average delays do not exceed $\mathbf{d}_s$. Such a flexibility in rate allocation, which reflects the opportunistic behavior of the algorithm, can yield a wider range of feasible rate allocation schemes in comparison with the single-period NUM, which will be discussed in the next subsection. Finally, Fig. 2(c) shows link traffics, link margins, and the amount of under-utilized link capacities. Clearly, at periods $t = 9, 10$, all links possess zero link margins. On the other hand, for $t = 1, \ldots, 8$, positive values for link margin variables (for at least one link) show that there is at least one active delay constraint imposed by the sources.

We also investigate the effectiveness of the MPC-based solution proposed in Section V. To this end, we consider a network with the same topology as in the previous simulation, and assume that link capacities are not known perfectly. More precisely, we assume that for any t and l, our knowledge of c_{tl} is no more than the support of distribution from which c_{tl} is generated (e.g., we only know that c_{t2} is drawn from $\mathrm{Unif}(4, 10)$ distribution[4]).

In the lack of knowledge of link capacities, instead of problem P1, we solve problem P2 with the following estimates: for any τ, we use $\hat{c}_1(t|\tau) = \hat{c}_4(t|\tau) = 5$ and $\hat{c}_2(t|\tau) = \hat{c}_3(t|\tau) = 7$. We compare the results of problem P2 with the knowledge of the current and the previous periods with that of problem P1, where the entire knowledge of the horizon is available. The resulting source rates and link margin values are quite similar. To illustrate the accuracy of the suboptimal MPC-based solution, we report the utility values obtained by the two solutions. The aggregated utility value for MPC-based solution is 36.16, whereas that obtained by DA-DNUM (which is the optimal value of problem P1) is 37.01, which indicates 2.2% difference.

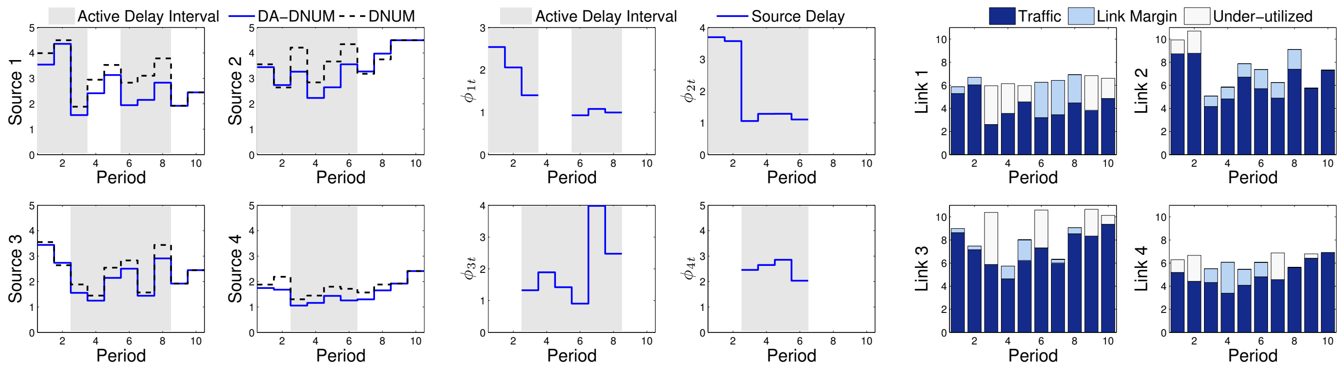
B. Experiment 2: Comparison Scenario

Next, we compare DA-DNUM with the algorithm proposed in [9] (by assuming fixed capacities) in a large-scale scenario. We remark that the algorithm proposed in [9] is based on the single-period version of NUM (the basic NUM without temporal considerations) that is tailored to delay-sensitive cases. Consequently, single-period NUM in algorithm of [9] persuades us to solve the NUM for all time instants in $\mathcal{T}$ (i.e.,

[4]We use $\mathrm{Unif}(a, b)$ to denote the uniform distribution over $[a, b]$ for all $a, b \in \mathbb{R}$ with $a < b$.

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems



(a) Optimal Source Rates (b) Source Delays (c) Link Capacities

Fig. 2: Results of Experiment 1

T separate problems). We consider a line topology shown in Fig. 3, which has 200 links and 198 sources with the following routing matrix:

$$R_t = \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & 1 & \dots & 0 & 0 \\ 1 & 1 & 1 & 1 & \dots & 0 & 0 \\ 1 & 0 & 1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

The rest of the parameters are listed in Table II. To clearly exhibit the behavior of DA-DNUM, we associate average delay constraints only for source 1 (i.e., the source traversing all links) and source 2 (namely, the one passing through the first 4 links).

To exhibit the flexibility of DA-DNUM, in this experiment we enforce $w_1 = 5$ *only* for period 2, that is source 1 has a rate requirement of 5 only at period 2. The aforementioned minimum rate demand is in conflict with the average delay requirement since the higher rate results in higher end-to-end delay according to the limited capacity of links. Nonetheless, DA-DNUM easily remedies this conflicting situation by assigning the declared minimum rate to source 1 at $t = 2$, thus enduring a larger short-term delay (around 85 instead of $\mathbf{d}_1 = 50$). Thanks to the constraints coupled over time, DA-DNUM allocates proper rates to this source in other periods so as to maintain the average delay below 50. In contrast, the single-period algorithm of [9] fails in this scenario since the underlying NUM becomes infeasible. This simple experiment indicates relatively wider set of feasible problems of DA-DNUM. One may construct several other feasible scenarios for DA-DNUM that are infeasible for the problem of [9].

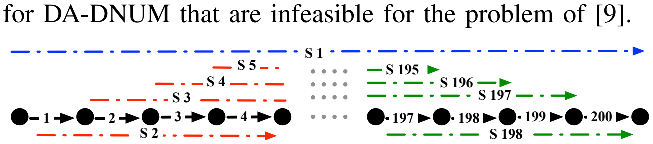


Fig. 3: Network Topology, Experiment 2

TABLE II: Parameters of Experiment 2

Parameter	Value
S	198
L	200
T	50
$c_{tl},\ \forall t, \forall l$	i.i.d. $\text{Unif}(8, 12)$ *kbps*
K_1, K_2	1
M_1, M_2	$\frac{1}{50}\mathbf{1}_{1\times 50}$
$K_s,\ \forall s \neq 1, 2$	0
$M_s,\ s \in \{3, \dots, 198\}$	$[0]_{1\times 50}$
d_1, d_2	50

TABLE III: Parameters of Experiments 3

Parameter	Value
S	20
L	20
T	20
$c_{tl},\ \forall t, \forall l$	20 *kbps*
$d_s,\ \forall s$	i.i.d. $\text{Unif}(4, 6)$

C. Experiment 3: Random Topology

Now we examine DA-DNUM for a more complex randomly generated topology. We run DA-DNUM and algorithm of [9] for a scenario with a random topology comprising 20 sources and 20 links (see Table III for the parameters). In this experiment, we associate only one delay constraint to each source. The delay constraint for each source is defined over time interval $[t_i, t_f]$, where for each source t_i and t_f are generated independently of other sources in the following way: first a time slot t_i is drawn uniformly at random from $\mathcal{T}$. If $t_i < T$, we then sample t_f uniformly at random from $\{t_i + 1, \dots, T\}$.

As it allows temporal fluctuations in source delays, DA-DNUM yields slightly better link utilization compared to the algorithm of [9]: The under-utilized link capacity averaged over all links and all periods for the algorithm of [9] is 4.06, whereas that of DA-DNUM is 3.91. Thus, DA-DNUM reduces the average under-utilized link capacity by 3.7%. The two comparison experiments therefore indicate both wider range of feasibility and better resource utilization of DA-DNUM against existing single-period approaches.

VII. Conclusion and Future Directions

We studied a multi-period NUM problem with source-driven time-coupled constraints on average end-to-end delay and ca-

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

pacity constraints. We proposed a distributed algorithm, called DA-DNUM, and provided its convergence analysis. Numerical experiments demonstrated that, compared to existing schemes, DA-DNUM admits relatively wider feasible scenarios along with higher resource utilization. This enhancement originated from multi-period problem setup that allows short-term delay fluctuations while keeping long-term delay values around the required one.

As future work, we plan to extend this work to the case of wireless scenarios, where joint delay-aware rate allocation and link scheduling is considered. Furthermore, we would like to investigate the case of non-convex delay functions, which is motivated by the case of more complicated packet arrival models. Finally, another interesting direction is to solve the problem using the state-of-the-art second-order methods to achieve faster convergence.

Acknowledgment

The authors would like to acknowledge the reviewers for their helpful comments to improve the presentation of this paper. The work of M. H. Hajiesmaili was supported by the University Grants Committee of the Hong Kong Special Administrative Region, China (Area of Excellence Grant Project No. AoE/E-02/08).

References

[1] M. H. Hajiesmaili, M. S. Talebi, and A. Khonsari, "Utility-optimal dynamic rate allocation under average end-to-end delay requirements," in *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, 2015, pp. 4842–4847.

[2] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[3] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[4] S. H. Low and D. E. Lapsley, "Optimization flow control, i: Basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, 1999.

[5] Z. Rosberg, "Rate control with end-to-end delay and rate constraints," in *Proceedings of IEEE INFOCOM Workshops*, April 2008, pp. 1–6.

[6] Y. Li, A. Papachristodoulou, M. Chiang, and A. R. Calderbank, "Congestion control and its stability in networks with delay sensitive traffic," *Computer Networks*, vol. 55, no. 1, pp. 20–32, 2011.

[7] I.-H. Hou and P. R. Kumar, "Utility maximization for delay constrained qos in wireless," in *Proceedings of the 29th IEEE International Conference on Computer Communications (INFOCOM)*, 2010, pp. 1–9.

[8] B. M. Dogahe, M. N. Murthi, X. Fan, and K. Premaratne, "A distributed congestion and power control algorithm to achieve bounded average queuing delay in wireless networks," *Telecommunication Systems*, vol. 44, no. 3-4, pp. 307–320, 2010.

[9] F. Qiu, J. Bai, and Y. Xue, "Towards optimal rate allocation in multi-hop wireless networks with delay constraints: A double-price approach," in *Proceedings of the IEEE International Conference on Communications (ICC)*, June 2012, pp. 5280–5285.

[10] H. Xiong, R. Li, A. Eryilmaz, and E. Ekici, "Delay-aware cross-layer design for network utility maximization in multi-hop networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 5, pp. 951–959, 2011.

[11] M. S. Talebi, A. Khonsari, and M. H. Hajiesmaili, "Utility-proportional bandwidth sharing for multimedia transmission supporting scalable video coding," *Computer Communications*, vol. 33, no. 13, pp. 1543–1556, 2010.

[12] T. Harks and T. Poschwatta, "Utility fair congestion control for real-time traffic," in *Proceedings of the 24th IEEE International Conference on Computer Communications (INFOCOM)*, vol. 4, 2005, pp. 2786–2791.

[13] F. Qiu, J. Bai, and Y. Xue, "Optimal rate allocation in wireless networks with delay constraints," *Ad Hoc Networks*, vol. 13, pp. 282–295, 2014.

[14] N. Trichakis, A. Zymnis, and S. Boyd, "Dynamic network utility maximization with delivery contracts," in *Proceedings of the 17th IFAC World Congress*, 2008, pp. 2907–2912.

[15] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[16] E. F. Camacho and C. B. Alba, *Model predictive control*. Springer, 2013.

[17] D. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[18] S. Shenker, "Fundamental design issues for the future internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 4, pp. 1176–1188, 1995.

[19] J. J. Jaramillo, R. Srikant, and L. Ying, "Scheduling for optimal rate allocation in ad hoc networks with heterogeneous delay constraints," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 5, pp. 979–987, 2011.

[20] M. H. Hajiesmaili, A. Khonsari, A. Sehati, and M. S. Talebi, "Content-aware rate allocation for efficient video streaming via dynamic network utility maximization," *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 2016–2027, 2012.

[21] E. Wei, A. Ozdaglar, A. Eryilmaz, and A. Jadbabaie, "A distributed newton method for dynamic network utility maximization with delivery contracts," in *Proceedings of the Annual Conference on Information Sciences and Systems (CISS)*, 2012, pp. 1–6.

[22] M. Saad, A. Leon-Garcia, and W. Yu, "Optimal network rate allocation under end-to-end quality-of-service requirements," *IEEE Transactions on Network and Service Management*, vol. 4, no. 3, pp. 40–49, 2007.

[23] S. Eswaran, A. Misra, F. Bergamaschi, and T. L. Porta, "Utility-based bandwidth adaptation in mission-oriented wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 8, no. 2, pp. 17–26, 2012.

[24] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice Hall, 1989.

[25] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[26] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

Appendix A
Proof of Theorem 1

A. Preliminaries

We briefly overview convergence of the gradient method with constant step size for solving the problem $\min_{\mathbf{z}} f(\mathbf{z})$, where f is convex. Let $\mathbf{z}^\star$ be the minimizer of the problem. If $\nabla f(\mathbf{z})$ is a Q-Lipschitz function, then the sequence $\{\mathbf{z}^{(k)}\}_{k\geq 0}$ defined by

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - \gamma \nabla f(\mathbf{z}^{(k)}),$$

converges to $\mathbf{z}^\star$ provided that $0 < \gamma < \frac{2}{Q}$ (see, e.g., [24, Proposition 3.4]). Moreover, by the nonexpansive property of the projection operator [17, Proposition B.11], convergence of the above sequence implies convergence of $\{\tilde{\mathbf{z}}^{(k)}\}_{k\geq 0}$, with

$$\tilde{\mathbf{z}}^{(k+1)} = [\tilde{\mathbf{z}}^{(k)} - \gamma \nabla f(\tilde{\mathbf{z}}^{(k)})]^+$$

to the minimizer of $\min_{\mathbf{z}\geq 0} f(\mathbf{z})$.

Let us define $\boldsymbol{\nu} = [\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_T, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_S]$. By the above result, to prove the convergence of the algorithm, it suffices to find a constant Q such that $\nabla g(\boldsymbol{\nu})$ is Q-Lipschitz. Therefore, as a consequence of [25, Theorem 9.19], we only need to show that the Hessian of $g(\boldsymbol{\nu})$ in bounded in the ℓ_2-norm.

B. Proof of the Theorem

Proof. The Hessian of $g(\boldsymbol{\nu})$, henceforth denoted by H, is a $(TL + \sum_s K_s)$-by-$(TL + \sum_s K_s)$ matrix, whose ij-element is:

$$H_{ij} = \frac{\partial^2 g(\boldsymbol{\nu})}{\partial \nu_i \partial \nu_j}.$$

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

Recall that the partial derivatives of $g(\boldsymbol{\nu})$ w.r.t. dual variables, where they exist, are given by

$$\frac{\partial g(\boldsymbol{\nu})}{\partial \lambda_{tl}} = c_{tl} - \sigma_{tl} - \sum_s (R_t)_{ls} x_{st}, \quad \forall t, \forall l,$$
$$\frac{\partial g(\boldsymbol{\nu})}{\partial \mu_{sk}} = d_{sk} - \sum_t \sum_l (M_s)_{kt} (R_t)_{ls} D(\sigma_{tl}), \quad \forall s, \forall k,$$

where

$$x_{st} = \left[U'^{-1}_{st}(\lambda^{st})\right]_{\mathcal{X}_s},$$
$$\sigma_{tl} = \left[D'^{-1}\left(-\frac{\lambda_{tl}}{\mu^{tl}}\right)\right]^+.$$

Before proceeding to calculate the elements of the Hessian, observe that we have for each source s and time t:

$$\frac{\partial x_{st}}{\partial \lambda^{st}} = \frac{\partial U'^{-1}_{st}(\lambda^{st})}{\partial \lambda^{st}} = -\beta_{st}(\lambda),$$

where

$$\beta_{st}(\lambda) = \begin{cases} \frac{-1}{U''_{st}(x_{st}(\lambda))} & \text{if } U'_{st}(W_s) \le \lambda^{st} \le U'_{st}(w_s), \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, for each link l and time t, we have that:

$$\frac{\partial \sigma_{tl}}{\partial \lambda_{tl}} = \frac{\partial}{\partial \lambda_{tl}} D'^{-1}\left(\frac{-\lambda_{tl}}{\mu^{tl}}\right) = -\frac{1}{\mu^{tl} D''(\sigma_{tl})},$$
$$\frac{\partial \sigma_{tl}}{\partial \mu^{tl}} = \frac{\partial}{\partial \mu^{tl}} D'^{-1}\left(\frac{-\lambda_{tl}}{\mu^{tl}}\right) = \frac{\lambda_{tl}}{(\mu^{tl})^2 D''(\sigma_{tl})}$$
$$= \frac{-D'(\sigma_{tl})}{\mu^{tl} D''(\sigma_{tl})},$$

where the last equality follows from (7). Hence, the elements of the Hessian are given by[5]:

$$\frac{\partial^2 g(\boldsymbol{\nu})}{\partial \lambda_{tl} \partial \mu_{sk}} = -\frac{\partial}{\partial \lambda_{tl}} \sum_{t'} \sum_{l'} (M_s)_{kt'} (R_{t'})_{l's} D(\sigma_{t'l'})$$
$$= -(M_s)_{kt} (R_t)_{ls} \cdot \frac{\partial \sigma_{tl}}{\partial \lambda_{tl}} \cdot D'(\sigma_{tl})$$
$$= (M_s)_{kt} (R_t)_{ls} \cdot \frac{D'(\sigma_{tl})}{\mu^{tl} D''(\sigma_{tl})},$$

$$\frac{\partial^2 g(\boldsymbol{\nu})}{\partial \mu_{s'k'} \partial \mu_{sk}} = -\frac{\partial}{\partial \mu_{s'k'}} \sum_t \sum_l (M_s)_{kt} (R_t)_{ls} D(\sigma_{tl})$$
$$= -\sum_t \sum_l (M_s)_{kt} (R_t)_{ls} \frac{\partial \sigma_{tl}}{\partial \mu_{s'k'}} \cdot D'(\sigma_{tl})$$
$$= -\sum_t \sum_l (M_s)_{kt} (R_t)_{ls} \frac{\partial \mu^{tl}}{\partial \mu_{s'k'}} \cdot \frac{\partial \sigma_{tl}}{\partial \mu^{tl}} \cdot D'(\sigma_{tl})$$
$$= \sum_t \sum_l (M_s)_{kt} (M_{s'})_{k't} (R_t)_{ls} (R_t)_{ls'} \cdot \frac{(D'(\sigma_{tl}))^2}{\mu^{tl} D''(\sigma_{tl})},$$

[5] We note that in points satisfying $\lambda^{st} = U'_{st}(w_s)$ or $\lambda^{st} = U'_{st}(W_s)$, the Hessian $\nabla^2 g$ may not exist. In these points, the gradient should be replaced by subgradients. Similarly to [4], to preserve the simplicity of the argument, such issues are ignored in this paper (cf. discussion on page 871 of [4]).

$$\frac{\partial^2 g(\boldsymbol{\nu})}{\partial \lambda_{t'l'} \partial \lambda_{tl}} = -\frac{\partial \sigma_{tl}}{\partial \lambda_{t'l'}} - \sum_s (R_t)_{ls} \frac{\partial x_{st}}{\partial \lambda_{t'l'}}$$
$$= \frac{\mathbb{1}\{(t', l') = (t, l)\}}{\mu^{tl} D''(\sigma_{tl})} - \sum_s (R_t)_{ls} \frac{\partial \lambda^{st}}{\partial \lambda_{t'l'}} \cdot \frac{\partial x_{st}}{\partial \lambda^{st}}$$
$$= \frac{\mathbb{1}\{(t', l') = (t, l)\}}{\mu^{tl} D''(\sigma_{tl})} + \sum_s (R_t)_{ls} (R_t)_{l's} \beta_{st}(\lambda) \mathbb{1}\{t' = t\}.$$

Hence, we derive the following bounds:

$$\left| \frac{\partial^2 g(\boldsymbol{\nu})}{\partial \lambda_{tl} \partial \mu_{sk}} \right| \le \frac{\alpha_D}{\mu_{\min} \kappa_D}, \tag{9}$$

$$\left| \frac{\partial^2 g(\boldsymbol{\nu})}{\partial \mu_{s'k'} \partial \mu_{sk}} \right| \le \frac{T L \alpha_D^2}{\mu_{\min} \kappa_D}, \tag{10}$$

$$\left| \frac{\partial^2 g(\boldsymbol{\nu})}{\partial \lambda_{t'l'} \partial \lambda_{tl}} \right| \le \frac{\mathbb{1}\{(t', l') = (t, l)\}}{\mu_{\min} \kappa_D} + \frac{S}{\kappa_U} \mathbb{1}\{t' = t\}. \tag{11}$$

Using the above inequalities, we derive an upper bound for the ℓ_2-norm of the Hessian H. Let us decompose H as follows

$$H = \left[\begin{array}{c|c} H^{(\lambda\lambda)} & H^{(\mu\lambda)} \\ \hline H^{(\lambda\mu)} & H^{(\mu\mu)} \end{array}\right]$$
$$= \left[\begin{array}{c|c} H^{(\lambda\lambda)} & 0 \\ \hline 0 & 0 \end{array}\right] + 2 \left[\begin{array}{c|c} 0 & H^{(\mu\lambda)} \\ \hline 0 & 0 \end{array}\right] + \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & H^{(\mu\mu)} \end{array}\right],$$

where $H^{(\mu\lambda)} = H^{(\lambda\mu)}$ follows from the symmetry of the Hessian. Recall that the ℓ_2-norm of any matrix A satisfies:

$$\|A\|_2 \le \sqrt{\|A\|_1 \|A\|_\infty},$$

where $\|A\|_1$ is the maximum column-sum matrix norm of A, and $\|A\|_\infty$ is the maximum row-sum matrix norm [26]. Hence, by subadditivity of norm function, we have

$$\|H\|_2 \le \sqrt{\|H^{(\lambda\lambda)}\|_1 \|H^{(\lambda\lambda)}\|_\infty} + 2\sqrt{\|H^{(\mu\lambda)}\|_1 \|H^{(\mu\lambda)}\|_\infty}$$
$$+ \sqrt{\|H^{(\mu\mu)}\|_1 \|H^{(\mu\mu)}\|_\infty}$$
$$= \|H^{(\lambda\lambda)}\|_1 + 2\sqrt{\|H^{(\mu\lambda)}\|_1 \|H^{(\mu\lambda)}\|_\infty} + \|H^{(\mu\mu)}\|_1, \tag{12}$$

where the last equality follows from the symmetry of $H^{(\lambda\lambda)}$ and $H^{(\mu\mu)}$. Using (9)-(11), we derive the following bounds:

$$\|H^{(\lambda\lambda)}\|_1 \le \frac{1}{\mu_{\min} \kappa_D} + \frac{LS}{\kappa_U},$$
$$\|H^{(\mu\mu)}\|_1 \le \frac{\alpha_D^2 T L}{\mu_{\min} \kappa_D} \Big(\sum_s K_s\Big)^2,$$
$$\|H^{(\mu\lambda)}\|_1 \le \frac{\alpha_D}{\mu_{\min} \kappa_D} \sum_s K_s,$$
$$\|H^{(\mu\lambda)}\|_\infty \le \frac{\alpha_D T L}{\mu_{\min} \kappa_D}.$$

Putting these together with (12), we obtain $\|H\|_2 \le Q$, where Q is defined in (8). Hence, the Hessian of $g(\boldsymbol{\nu})$ is upper bounded in the ℓ_2-norm by Q. By [25, Theorem 9.19], we deduce that $\nabla g(\boldsymbol{\nu})$ is Q-Lipschitz. As a consequence, if we

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2017.2677202, IEEE Transactions on Control of Network Systems

require $\gamma \in (0, \frac{2}{Q})$, the algorithm converges to the primal-dual optimal point of P1, and the proof is concluded. □

Appendix B
Proof of Proposition 2

Proof. First we derive a lower bound on μ^{tl} for all t and l so that delay constraints (4) at the optimal point, restated below, are satisfied:

$$\sum_t \sum_l (M_s)_{kt}(R_t)_{ls} D(\sigma_{tl}^\star) \le d_{sk}, \quad \forall s, \forall k \in \mathcal{K}_s. \tag{13}$$

In the case of M/M/1 packet arrival model, the delay function is $D(z) = 1/z$. It then follows from (7) that for any t and l, $D(\sigma_{tl}^\star) = \sqrt{\lambda_{tl}/\mu^{tl}}$, and therefore, the l.h.s. of (13) reads

$$\begin{aligned}\sum_t \sum_l (M_s)_{kt}(R_t)_{ls} D(\sigma_{tl}^\star) &= \sum_t \sum_l (M_s)_{kt}(R_t)_{ls} \sqrt{\frac{\lambda_{tl}}{\mu^{tl}}} \\ &\le \sum_t \sum_l \sqrt{(R_t)_{ls}\lambda_{tl}} \sqrt{\frac{(R_t)_{ls}}{\mu^{tl}}} \\ &\le \sum_t \sqrt{\sum_l (R_t)_{ls}\lambda_{tl}} \sqrt{\sum_l \frac{(R_t)_{ls}}{\mu^{tl}}} \\ &= \sum_t \sqrt{\lambda^{st}} \sqrt{\sum_l \frac{(R_t)_{ls}}{\mu^{tl}}},\end{aligned}$$

where we use the fact that $(R_t)_{ls} \in \{0, 1\}$ and $(M_s)_{kt} \le 1$ in the first inequality, and Cauchy-Schwarz inequality in the last line. Hence,

$$\mu^{tl} \ge \frac{LT^2\alpha_U}{d_{\min}^2}, \quad \forall t, \forall l, \tag{14}$$

implies:

$$\begin{aligned}\sum_t \sum_l (M_s)_{kt}(R_t)_{ls} D(\sigma_{tl}^\star) &\le \sum_t \sqrt{\lambda^{st}} \sqrt{\frac{d_{\min}^2}{LT^2\alpha_U} \sum_l (R_t)_{ls}} \\ &\le \frac{d_{\min}}{T} \sum_t \sqrt{U'_{st}(x_{st}^\star)} \sqrt{\frac{1}{L\alpha_U} \sum_l (R_t)_{ls}} \\ &\le d_{\min} \\ &\le d_{sk}.\end{aligned}$$

Hence, the lower bound on μ^{tl} in (14) guarantees that constraints (13) are satisfied and thus, we may choose $\mu_{\min} = LT^2\alpha_U d_{\min}^{-2}$.

Now we obtain upper and lower bounds on α_D and κ_D, respectively. Observe that $D'(z) = -1/z^2$ and $D''(z) = 2/z^3$ so that $|D'(z)| = (D(z))^2$. Moreover, we have that $(M_s)_{kt} \ge 1/T$ for all s and $k \in \mathcal{K}_s$. Hence, in view of constraints (3), any feasible σ_{tl} satisfies:

$$D(\sigma_{tl}) \le T \max_{s,k \in \mathcal{K}_s} d_{sk} \le T d_{\max}.$$

Hence, for all t and l,

$$|D'(\sigma_{tl})| = (D(\sigma_{tl}))^2 \le T^2 d_{\max}^2, \tag{15}$$

so that we may select $\alpha_D = T^2 d_{\max}^2$. Furthermore, constraints (1) imply that $\sigma_{tl} \le c_{tl} \le c_{\max}$ for all t and l. Hence,

$$D''(\sigma_{tl}) \ge \frac{2}{c_{\max}^3},$$

so that we may select $\kappa_D = 2c_{\max}^{-3}$, and the proof is completed. □

Mohammad H. Hajiesmaili received his B.Sc. degree from Department of Computer Engineering at Sharif University of Technology, Iran, in 2007. He received his M.Sc. and Ph.D. degrees from the Electrical and Computer Engineering Department at the University of Tehran, Iran, in 2009 and 2014, respectively. He was a researcher at the School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran, from 2008 to 2013, and a postdoctoral fellow at the Department of Information Engineering, the Chinese University of Hong Kong, from 2014 to 2016.

He is currently a postdoctoral fellow at Whiting School of Engineering, the Johns Hopkins University. His research interests include optimization, algorithm, and mechanism design in energy systems, electricity market, transportation networks, and multimedia networks.

M. Sadegh Talebi received his B.Sc. degree in electrical engineering from Iran University of Science and Technology (IUST), Iran, in 2004, and his M.Sc. degree in electrical engineering from Sharif University of Technology, Iran, in 2006.

He is currently working towards the Ph.D. degree in electrical engineering at the Department of Automatic Control, School of Electrical Engineering, KTH The Royal Institute of Technology, Sweden. Before starting his Ph.D., he served as a research engineer at the School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran. His current research interests include resource allocation in networks, sequential decision making, and learning theory.

Ahmad Khonsari received the B.Sc. degree in electrical and computer engineering from Shahid-Beheshti University, Iran, in 1991, and M.Sc. degree in computer engineering from Iran University of Science and Technology (IUST), Iran, in 1996, and Ph.D. degree in computer science from the University of Glasgow, UK, in 2003.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering, University of Tehran, Iran, and a researcher at the School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran. His research interests include performance modeling/evaluation, wired/wireless networks, distributed systems, and high performance computer architectures.

2325-5870 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.