Stackelberg Routing on Parallel Networks With Horizontal Queues

Walid Krichene, *Graduate Student Member, IEEE*, Jack D. Reilly, *Graduate Student Member, IEEE*, Saurabh Amin, *Member, IEEE*, and Alexandre M. Bayen, *Member, IEEE*

***Abstract*—This paper presents a game theoretic framework for studying Stackelberg routing games on parallel networks with horizontal queues, such as transportation networks. First, we introduce a new class of latency functions that models congestion due to the formation of physical queues. For this new class, some results from the classical congestion games literature (in which latency is assumed to be a non-decreasing function of the flow) do not hold. In particular, we find that there may exist multiple Nash equilibria that have different total costs. We provide a simple polynomial-time algorithm for computing the *best Nash equilibrium*, i.e., the one which achieves minimal total cost. Then we study the Stackelberg routing game: assuming a central authority has control over a fraction of the flow on the network (*compliant flow*), and that the remaining flow (*non-compliant*) responds selfishly, what is the best way to route the compliant flow in order to minimize the total cost? We propose a simple Stackelberg strategy, the Non-Compliant First (NCF) strategy, that can be computed in polynomial time. We show that it is optimal for this new class of latency on parallel networks. This work is applied to modeling and simulating congestion relief on transportation networks, in which a coordinator (traffic management agency) can choose to route a fraction of compliant drivers, while the rest of the drivers choose their routes selfishly.**

***Index Terms*— Game theory, Nash equilibrium, network analysis and control, routing, Stackelberg game, transportation networks.**

I. Introduction and Main Results

A. Motivation and Related Work

Routing games model the interaction between players on a network, where the cost for each player on an edge depends on the total congestion of that edge. Extensive work has been dedicated to the study of Nash equilibria (or user optimal assignments) for routing games, in which players selfishly choose the routes that minimize their individual costs (latencies) [4], [7], [8]. In general, Nash equilibria are inefficient compared to a system optimal assignment that minimizes the total cost on the network [16]. This inefficiency has been characterized for different classes of latency functions and network topologies [27], [29]. This helps understand the inefficiencies caused by congestion on numerous real-life networks, such as communication networks and traffic networks.

In order to reduce the inefficiencies due to selfish routing, many instruments have been studied, including congestion pricing [21], capacity allocation [15], and Stackelberg routing [2], [14], [25], [29]. In the Stackelberg routing game, a subset of the players, corresponding to a fraction of the total flow, hereafter called the compliant flow, is centrally assigned by a coordinator (leader), and then the remaining players (followers) choose their routes selfishly. The objective of the leader is to assign the compliant flow in a manner that minimizes a system-wide cost function, while anticipating the followers' selfish response. This setting is relevant in the planning and operation of transportation and communication networks.. In transportation networks, advances in traveler information systems have made it possible to interact with individual drivers and exchange information through GPS-enabled smartphone applications or vehicular navigation systems [31]. These devices can be used by a a traffic control center to provide routing advice that can improve the overall efficiency of the network. Naturally, the question arises on how the traffic control center should coordinate with the compliant drivers while accounting for the selfish response of other drivers: hence the importance of the Stackelberg routing framework. One might argue that the drivers who are offered routing advice are not guaranteed to follow the suggested routes, especially when these routes do not have minimal latency (in order to improve the system-wide efficiency, some drivers will be assigned routes that are sub-optimal in the Nash sense). However, in some cases, it can be reasonably assumed that a fraction of the drivers will choose the routes suggested by the coordinator, despite immediate fairness concerns. For example, some drivers may have sufficient external incentives to be compliant with the coordinator. In addition, the compliant flow may also include altruistic drivers who care about the system-wide efficiency (e.g., pollution levels).

Stackelberg routing on parallel networks has been studied extensively for the class of non-decreasing latency functions, and it is known that computing the optimal Stackelberg strategy is

Manuscript received July 25, 2012; revised May 07, 2013 and October 14, 2013; accepted October 30, 2013. Date of current version February 19, 2014. This work was supported in part by the National Science Foundation (NSF) under Award CNS-0845076 and by the Foundations Of Resilient CybEr-physical Systems (FORCES), which receives support from the NSF (under Awards CNS-1238959, CNS-1238962, CNS-1239054, and CNS-1239166). This paper was recommended by Associate Editor A. Ozdaglar.

W. Krichene is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94704 USA (e-mail: walid@eecs.berkeley.edu).

J. D. Reilly is with the Department of Civil Engineering, University of California, Berkeley, CA, 94704 USA (e-mail: jackdreilly@berkeley.edu).

S. Amin is with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: amins@mit.edu).

A. M. Bayen is with the Department of Electrical Engineering and Computer Sciences and the Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94704 USA (e-mail: bayen@berkeley.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

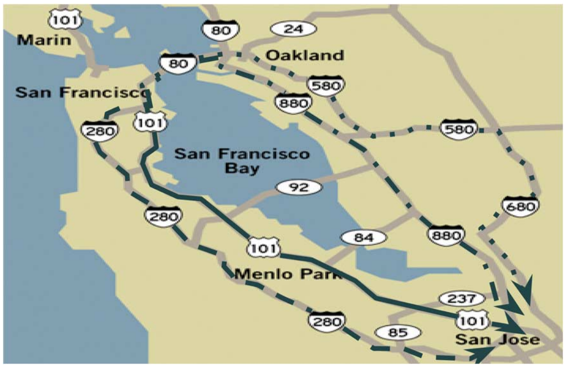


Fig. 1. Map of a parallel highway network connecting San Francisco to San Jose.

NP-hard in the size of the network [25]. This led to the design of polynomial time approximate strategies such as *Scale* and *Largest Latency First* [25], [29]. While this class of latency functions provides a good model of congestion for a broad range of networks with vertical queues, such as communication networks, it does not entirely capture congestion in networks with horizontal queues, such as transportation networks, in which queuing results in an increase in density [9], [17], [19], [24], [31], which in turn affects the latency. In order to better model the effects of density, we introduce a new class of latency functions, and we study Stackelberg routing games for this new class on parallel networks. User-equilibria for routing games with horizontal queues have been studied for example in [5], [12], [20], [30]. However, to the best of our knowledge, Stackelberg routing with horizontal queues has not been addressed so far.

We restrict our present study to parallel networks. This simple network topology is of practical importance in several situations, including traffic planning on parallel highway networks that connect two highly populated areas [6]. Fig. 1 shows one such network that connects San Francisco to San Jose. We will consider this network in Section V.

B. Congestion on Horizontal Queues

The classical model for vertical queues assumes that the latency $\ell_n(x_n)$ on a link n is a non-decreasing function of the flow x_n on that link [3], [4], [8], [26], [29]. However, for networks with horizontal queues [17], [19], [24], the latency not only depends on the flow, but also on the density. For example, on a transportation network, the latency depends on the density of cars on the road (e.g., in cars per meter), and not only on the flow (e.g., in cars per second), since for a fixed value of flow, a lower density means higher velocity, hence lower latency. These effects of changing density are not captured by models of vertical queues. In this section we describe a simplified model of congestion that takes into account both flow and density.

Let ρ_n be the density on link n, assumed to be uniform, for simplicity, and let the flow x_n be given by a continuous, concave function of the density

$$\begin{aligned} x_n^\rho : [0, \rho_n^{\max}] &\to [0, x_n^{\max}] \\ \rho_n &\mapsto x_n = x_n^\rho(\rho_n). \end{aligned}$$

Here, $x_n^{\max} > 0$ is the maximum flow or *capacity* of the link, and $\rho_n^{\max}$ is the maximum density that the link can hold. The function x_n^ρ is determined by the physical properties of the link. It is termed the flux function in conservation law theory [11], [18] and the fundamental diagram in traffic flow theory [9], [13], [23]. In general, it is a non-injective function. We make the following assumptions:

- There exists a unique density $\rho_n^{\text{crit}} \in (0, \rho_n^{\max})$ such that $x_n^\rho(\rho_n^{\text{crit}}) = x_n^{\max}$, called critical density. When $\rho_n \in [0, \rho_n^{\text{crit}}]$, the link is said to be in *free-flow*, and when $\rho_n \in (\rho_n^{\text{crit}}, \rho_n^{\max})$, it is said to be *congested*.
- In the congested regime, x_n^ρ is continuous decreasing from $(\rho_n^{\text{crit}}, \rho_n^{\max})$ onto $(0, x_n^{\max})$. In particular, $\lim_{\rho_n \to \rho_n^{\max}} x_n^\rho(\rho_n) = 0$ (the flow reduces to zero when the density approaches the maximum density).

These are standard assumptions on the flux function, following traffic flow theory [9], [13], [23]. Additionally, we assume that in the free-flow regime, x_n^ρ is linearly increasing in ρ_n, and since $x_n^\rho(\rho_n^{\text{crit}}) = x_n^{\max}$, we have in the free-flow regime $x_n^\rho(\rho_n) = x_n^{\max}\rho_n/\rho_n^{\text{crit}}$ (as a result, the flux function is non-differentiable at the critical density). The assumption of linearity in free-flow is the only restrictive assumption, and it is essential in deriving the results on optimal Stackelberg strategies. Although somewhat restrictive, this assumption is common, and the resulting flux model is widely used in modeling transportation networks, such as in [9], [22]. Fig. 2 shows examples of such flux functions.

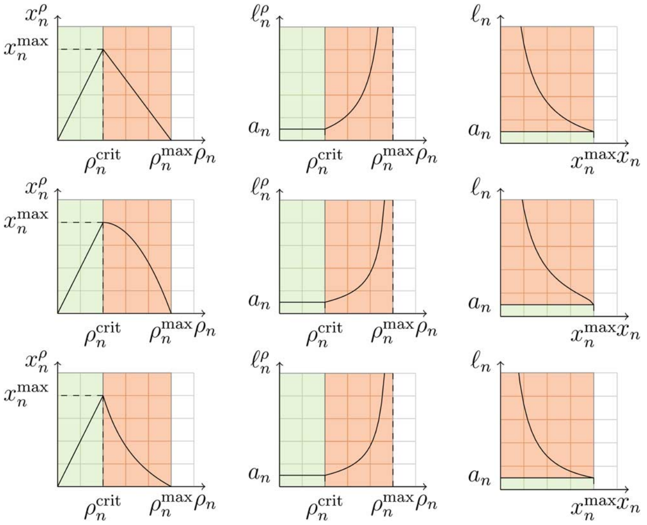


Fig. 2. Examples of flux functions for horizontal queues (left) and corresponding latency as a function of the density $\ell_n^\rho(\rho_n)$ (middle) and as a function of the flow and the congestion state $\ell_n(x_n, m_n)$ (right). The free-flow (respectively congested) regime is shaded in green (respectively red).

Since the density ρ_n and the flow x_n are assumed to be uniform on the link, the velocity v_n is given by $v_n = x_n/\rho_n$ and the latency is simply given by $L_n/v_n = L_n\rho_n/x_n$ where L_n is the length of link n. Thus to a given value of the flow, there may correspond more than one value of the latency, since the flux function is non-injective in general. To illustrate this with an example, we consider a transportation setting. A given value x_n of flow of cars on a road-segment can correspond to

- either a large concentration of cars moving slowly (high density, the road is *congested*), in which case the latency is large,

- or few cars moving fast (low density, the road is in *free-flow*), in which case the latency is small.

Therefore, the basic premise that the latency is a function of the flow does not hold for networks with horizontal queues, i.e., networks in which the density may change and impact the latency.

C. Latency Function for Horizontal Queues

Given a flux function x_n^ρ, the latency can be easily expressed as a non-decreasing function of the density

$$\begin{aligned} \ell_n^\rho : [0, \rho_n^{\max}] &\to \bar{\mathbb{R}}_+ \\ \rho_n \mapsto \ell_n^\rho(\rho_n) &= \frac{L_n \rho_n}{x_n^\rho(\rho_n)}. \end{aligned} \tag{1}$$

From the assumptions on the flux function, we have the following.

- In the free-flow regime, the flux function is linearly increasing, $x_n(\rho_n) = (x_n^{\max}/\rho_n^{\text{crit}})\rho_n$. Thus the latency is is single-valued in free-flow, $\ell_n^\rho(\rho_n) = L_n \rho_n^{\text{crit}}/x_n^{\max}$. We will denote its value by $a_n \triangleq L_n \rho_n^{\text{crit}}/x_n^{\max}$, called henceforth the *free-flow latency*.
- In the congested regime, x_n^ρ is bijective from $(\rho_n^{\text{crit}}, \rho_n^{\max})$ to $(0, x_n^{\max})$. Let

$$\begin{aligned} \rho_n^{\text{cong}} : (0, x_n^{\max}) &\to \left(\rho_n^{\text{crit}}, \rho_n^{\max}\right) \\ x_n &\mapsto \rho_n^{\text{cong}}(x_n) \end{aligned}$$

be its inverse. It maps the flow x_n to the unique congestion density that corresponds to that flow. Thus in the congested regime, latency can be expressed as a function of the flow, $x_n \mapsto \ell_n^\rho(\rho_n^{\text{cong}}(x_n))$. This function is decreasing as the composition of the decreasing function ρ_n^{cong} and the increasing function ℓ_n^ρ.

We can therefore express the latency as a function of the flow if we additionally specify the congestion state using a binary variable $m_n \in \{0, 1\}$, such that $m_n = 0$ if n is in free-flow, and $m_n = 1$ if n is congested.

Definition 1: HQSF Latency Class: A function

$$\begin{aligned} \ell_n : D_n &\to \mathbb{R}_+ \\ (x_n, m_n) &\mapsto \ell_n(x_n, m_n) \end{aligned} \tag{2}$$

defined on the domain[1]

$$D_n = [0, x_n^{\max}] \times \{0\} \cup (0, x_n^{\max}) \times \{1\}$$

is a HQSF latency function if it satisfies the following properties:

(A1) In the free-flow regime, the latency $\ell_n(\cdot, 0)$ is single-valued.

(A2) In the congested regime, the latency $x_n \mapsto \ell_n(x_n, 1)$ is decreasing on $(0, x_n^{\max})$.

(A3) $\lim_{x_n \to x_n^{\max}} \ell_n(x_n, 1) = a_n = \ell_n(x_n^{\max}, 0)$.

Property (A1) is equivalent to the assumption that the flux function is linear in free-flow, and is the only restrictive property in the sense discussed above, hence the name of the latency class. Property (A2) results from the expression of the latency as the composition $\ell_n^\rho(\rho_n^{\text{cong}}(x_n))$, where ℓ_n^ρ is increasing, and ρ_n^{cong} is decreasing. Property (A3) is equivalent to the continuity of the underlying flux function x_n^ρ.

Although it may be more natural to think of the latency as a non-decreasing function of the density, the above representation in terms of flow x_n and congestion state m_n will be useful in deriving properties of the Nash equilibria of the routing game.

Finally, we observe, as an immediate consequence of these properties, that the latency in congestion is always greater than the free-flow latency: $\forall x_n \in (0, x_n^{\max})$, $\ell_n(x_n, 1) > a_n$. Some examples of HQSF latency functions (and the underlying flux functions) are illustrated in Fig. 2. For a detailed derivation of an example latency function in a traffic setting, see Appendix A.

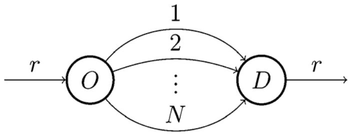


Fig. 3. Network with N parallel links under demand r.

D. Model

We consider a non-atomic routing game on a parallel network, shown in Fig. 3. Here non-atomic means that the game involves a continuum of players, where each player corresponds to an infinitesimal (non-atomic) amount of flow, [27], [28]. The network has a single source and a single sink. Connecting the source and sink are N parallel links indexed by $n \in \{1, \ldots, N\}$. We assume, without loss of generality, that the links are ordered by increasing free-flow latencies. To simplify the discussion, we further assume that free-flow latencies are distinct. Therefore we have $a_1 < a_2 < \cdots < a_N$. The network is subject to a constant positive flow demand r at the source. We will denote by (N, r) an instance of the routing game played on a network with N parallel links subject to demand r. The state of the network is given by a feasible flow assignment vector $\boldsymbol{x} \in \mathbb{R}_+^N$ such that $\sum_{n=1}^N x_n = r$ where x_n is the flow on link n, and a congestion state vector $\boldsymbol{m} \in \{0, 1\}^N$ where $m_n = 0$ if the link is in free-flow and $m_n = 1$ if the link is congested, as defined above. All physical quantities (density and flow) are assumed to be static and uniform on the link.

Every non-atomic player chooses a route in order to minimize his/her individual latency [26]. If a player chooses link n, his/her latency is given by $\ell_n(x_n, m_n)$, where ℓ_n is a HQSF latency function. We assume that players know the latency functions.

Pure Nash equilibria of the game (which we will simply refer to as Nash equilibria) are assignments $(\boldsymbol{x}, \boldsymbol{m})$ such that every player cannot improve his/her latency by switching to a different link.

Definition 2: Nash Equilibrium: A feasible assignment $(\boldsymbol{x}, \boldsymbol{m}) \in \mathbb{R}_+^N \times \{0, 1\}^N$ is a Nash equilibrium of the routing game instance (N, r) if $\forall n \in \text{supp}(\boldsymbol{x})$, $\forall k \in \{1, \ldots, N\}$, $\ell_n(x_n, m_n) \leq \ell_k(x_k, m_k)$.

[1] The latency in congestion $\ell_n(\cdot, 1)$ is defined on the open interval $(0, x_n^{\max})$. In particular, if $x_n = 0$ or $x_n = x_n^{\max}$ then the link is always considered to be in free-flow. When the link is empty ($x_n = 0$), it is naturally in free-flow. When it is at maximum capacity ($x_n = x_n^{\max}$) it is in fact on the boundary of the free-flow and congestion regions, and we say by convention that the link is in free-flow.

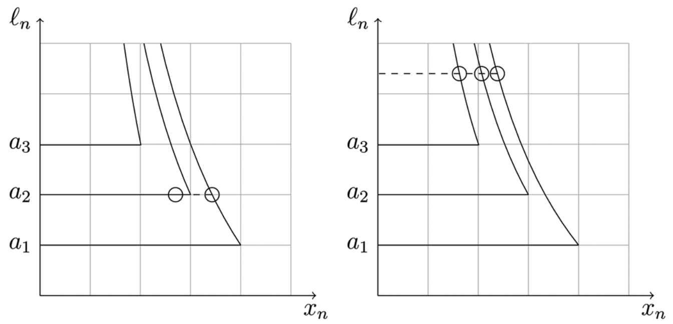


Fig. 4. Example of Nash equilibria for a three-link network. One equilibrium (left) has one link in free-flow and one congested link. A second equilibrium (right) has three congested links.

Here $\mathrm{supp}(\boldsymbol{x}) = \{n \in \{1,\ldots,N\} | x_n > 0\}$ denotes the support of $\boldsymbol{x}$. As a consequence of this definition, all links in the support of $\boldsymbol{x}$ have the same latency ℓ_0, and links that are not in the support have latency greater than or equal to ℓ_0. We will denote by $\mathrm{NE}(N,r)$ the set of Nash equilibria of the instance (N,r). We note that a Nash equilibrium for the routing game is a *static* equilibrium, we do not model dynamics of density or flow. Fig. 4 shows an example of a routing game instance and resulting Nash equilibria.

While a Nash equilibrium achieves minimal individual latencies, it does not minimize, in general, the *system cost* or *total cost* defined as follows:

Definition 3: The total cost of an assignment $(\boldsymbol{x}, \boldsymbol{m})$ is the total latency experienced by all players

$$C(\boldsymbol{x}, \boldsymbol{m}) = \sum_{n=1}^{N} x_n \ell_n(x_n, m_n). \tag{3}$$

As detailed in Section II, under the HQSF latency class, the routing game may have multiple Nash equilibria that have different total costs. We are interested, in particular, in Nash equilibria that have minimal cost, which are referred to as *best Nash equilibria* (BNE).

Definition 4: Best Nash Equilibria: The set of best Nash equilibria is the set of equilibria that minimize the total cost, i.e.,

$$\mathrm{BNE}(N,r) = \underset{(\boldsymbol{x},\boldsymbol{m}) \in \mathrm{NE}(N,r)}{\arg\min} \quad C(\boldsymbol{x}, \boldsymbol{m}). \tag{4}$$

E. Stackelberg Routing Game

In the Stackelberg routing game, a coordinator (a central authority) is assumed to have control over a positive fraction α of the total flow demand r. We call α the *compliance rate*. The coordinator wants to route the *compliant flow* αr in a way that minimizes the system cost, while anticipating the response of the rest of the players, assumed to choose their routes selfishly after the strategy of the coordinator is revealed. We will refer to the flow of selfish players $(1-\alpha)r$ as the *non-compliant flow*. More precisely, the game is played as follows:

- First, the coordinator (the leader) chooses a *Stackelberg strategy*, i.e., an assignment $\boldsymbol{s} \in \mathbb{R}_+^N$ of the compliant flow (such that $\sum_{n=1}^{N} s_n = \alpha r$).
- Then, the Stackelberg strategy $\boldsymbol{s}$ of the leader is revealed, and the non-compliant players (followers) choose their routes selfishly and form a Nash equilibrium $(\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$, *induced*[2] by strategy $\boldsymbol{s}$. By definition, the induced equilibrium $(\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$ satisfies

$$\begin{aligned} &\forall n \in \mathrm{supp}\,(\boldsymbol{t}(\boldsymbol{s})), \quad \forall k \in \{1,\ldots,N\}, \\ &\qquad \ell_n\left(s_n + t_n(\boldsymbol{s}), m_n(\boldsymbol{s})\right) \leq \ell_k\left(s_k + t_k(\boldsymbol{s}), m_k(\boldsymbol{s})\right). \end{aligned} \tag{5}$$

The total flow on the network is $\boldsymbol{s} + \boldsymbol{t}(\boldsymbol{s})$, thus the total cost is $C(\boldsymbol{s} + \boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$. Note that a Stackelberg strategy $\boldsymbol{s}$ may induce multiple Nash equilibria in general. However, we define the assignment $(\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$ to be the best such equilibrium (the one with minimal total cost, which will be shown to be unique in Section III).

We will use the following notation:

- (N, r, α) is an instance of the Stackelberg routing game played on a parallel network with N links under flow demand r with compliance rate α. Note that the routing game (N,r) is a special case of the Stackelberg routing game with $\alpha = 0$.
- $\mathrm{S}(N,r,\alpha) \subset \mathbb{R}_+^N$ is the set of Stackelberg strategies for the Stackelberg instance (N,r,α).
- $\mathrm{S}^\star(N,r,\alpha)$ is the *set of optimal Stackelberg strategies* defined as

$$\mathrm{S}^\star(N,r,\alpha) = \underset{\boldsymbol{s} \in \mathrm{S}(N,r,\alpha)}{\arg\min} \; C\left(\boldsymbol{s} + \boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s})\right). \tag{6}$$

F. Main Result

We now define a candidate Stackelberg strategy, which we call the *non-compliant first* strategy (NCF), and which we prove to be an optimal Stackelberg strategy. The NCF strategy corresponds to first computing the best Nash equilibrium $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{m}})$ of the non-compliant flow for the routing game instance $(N, (1-\alpha)r)$, then finding a particular strategy $\boldsymbol{s}$ that induces $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{m}})$.

Definition 5: The Non-Compliant First Strategy: Consider the Stackelberg instance (N,r,α). Let $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{m}})$ be the best Nash equilibrium of the non-compliant flow, $\{(\bar{\boldsymbol{t}}, \bar{\boldsymbol{m}})\} = \mathrm{BNE}(N, (1-\alpha)r)$, and $\bar{k} = \max \mathrm{supp}(\bar{\boldsymbol{t}})$ be the last link in its support. Then the non-compliant first strategy, denoted by $\mathrm{NCF}(N,r,\alpha)$, is by definition the Stackelberg strategy given by

$$\begin{aligned} \mathrm{NCF}(N,r,\alpha) = \Bigg(&0, \ldots, \overbrace{0}^{\bar{k}-1}, \overbrace{x_{\bar{k}}^{\max} - \bar{t}_{\bar{k}}}^{\bar{k}}, x_{\bar{k}+1}^{\max}, \ldots, x_{l-1}^{\max}, \\ &\alpha r - \left(\sum_{n=\bar{k}}^{l-1} x_n^{\max} - \bar{t}_{\bar{k}} \right), 0, \ldots, 0 \Bigg) \end{aligned} \tag{7}$$

where l is the maximal index in $\{\bar{k}+1, \ldots, N\}$ such that $\alpha r - (\sum_{n=\bar{k}}^{l-1} x_n^{\max} - \bar{t}_{\bar{k}}) \geq 0$.

In words, the NCF strategy saturates links one by one, by increasing index starting from link $\bar{k}$, the last link used by the non-compliant flow in the best Nash equilibrium of $(N, (1-$

[2] We note that a feasible flow assignment $\boldsymbol{s}$ of compliant flow may fail to induce a Nash equilibrium $(\boldsymbol{t}, \boldsymbol{m})$ and therefore is not considered to be a Stackelberg strategy.

TABLE I
MAIN ASSUMPTIONS AND RESULTS FOR THE STACKELBERG ROUTING GAME ON A PARALLEL NETWORK

Setting	Vertical queues	Horizontal queues, single-valued in free-flow (HQSF)
Model	$x \mapsto \ell(x)$ latency is a function of the flow $x \in [0, x^{\max}]$	$(x, m) \mapsto \ell(x, m)$ latency is a function of the flow $x \in [0, x^{\max}]$ and the congestion state $m \in \{0, 1\}$.
Assumptions	$x \mapsto \ell(x)$ is continuously non-decreasing. $x \mapsto x\ell(x)$ is convex.	$x \mapsto \ell(x, 0)$ is single-valued. $x \mapsto \ell(x, 1)$ is continuously decreasing. $\lim_{x \to x^{\max}} \ell(x, 1) = \ell(x^{\max}, 0)$.
Set of Nash equilibria	Essential uniqueness: if x, x' are Nash equilibria, then $C(x) = C(x')$ [4], [8].	No essential uniqueness in general. The number of Nash equilibria is at most $2N$ (Proposition 4) The best Nash equilibrium is a single-link-free-flow equilibrium (Lemma 2)
Optimal Stackelberg strategy	NP hard [25]	The NCF strategy is optimal and can be computed in polynomial time. (Theorem 1) The set of optimal Stackelberg strategies can be computed in polynomial time (Theorem 2)

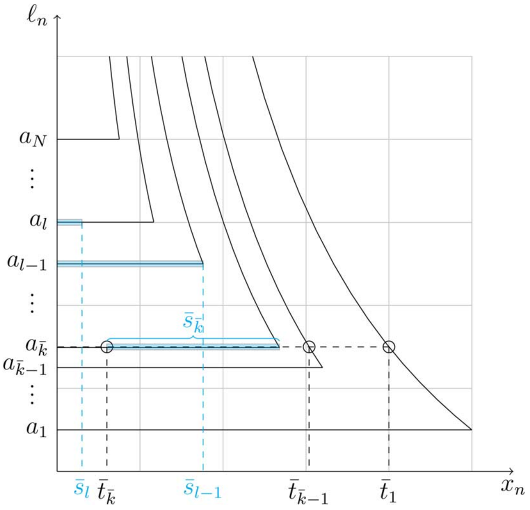


Fig. 5. Non-compliant first (NCF) strategy $\bar{\boldsymbol{s}}$ and its induced equilibrium. Circles show the best Nash equilibrium $(\bar{\boldsymbol{t}}, \boldsymbol{m})$ of the non-compliant flow $(1-\alpha)r$: link $\bar{k}$ is in free-flow, and links $\{1, \ldots, \bar{k}-1\}$ are congested. The Stackelberg strategy $\bar{\boldsymbol{s}} = \mathrm{NCF}(N, r, \alpha)$ is highlighted in blue.

$\alpha)r)$. Thus it will assign $x_{\bar{k}}^{\max} - \bar{t}_{\bar{k}}$ to link $\bar{k}$, then $x_{\bar{k}+1}^{\max}$ to link $\bar{k}+1$, $x_{\bar{k}+2}^{\max}$ to link $\bar{k}+2$ and so on, until the compliant flow is assigned entirely (see Fig. 5). The following theorem states the main result.

Theorem 1: The NCF Strategy is an Optimal Stackelberg Strategy: Under the class of HQSF latency functions, $\mathrm{NCF}(N, r, \alpha)$ is an optimal Stackelberg strategy for the Stackelberg instance (N, r, α).

We give a proof of Theorem 1 in Section III. We will also show that for the class of HQSF latency functions, best Nash equilibria can be computed in polynomial time in the size N of the network, and as a consequence, the NCF strategy can also be computed in polynomial time. This stands in contrast to previous results under the class of non-decreasing latency functions, for which computing the optimal Stackelberg strategy is NP-hard [25]. Table I summarizes the main differences between the classical setting (vertical queues) and the setting studied in this paper (horizontal queues, under the additional assumption that latency is single-valued in free-flow).

II. NASH EQUILIBRIA

In this section, we study Nash equilibria of the routing game. We show that under the class of HQSF latency functions, there may exist multiple Nash equilibria that have different costs. Then we partition the set of equilibria into congested equilibria and single-link-free-flow equilibria. Finally, we characterize the best Nash equilibrium and show that it can be computed in quadratic time in the number of links.

A. Structure and Properties of Nash Equilibria

We first give some properties of Nash equilibria. The following proposition is straightforward.

Proposition 1: Total Cost of a Nash Equilibrium: Let $(\boldsymbol{x}, \boldsymbol{m}) \in \mathrm{NE}(N, r)$ be a Nash equilibrium for the instance (N, r). Then there exists $\ell_0 > 0$ such that $\forall n \in \mathrm{supp}(\boldsymbol{x})$, $\ell_n(x_n, m_n) = \ell_0$ and $\forall n \notin \mathrm{supp}(\boldsymbol{x})$, $\ell_n(0, 0) \geq \ell_0$. The total cost of the equilibrium is then $C(\boldsymbol{x}, \boldsymbol{m}) = r\ell_0$.

Proposition 2: Let $(\boldsymbol{x}, \boldsymbol{m}) \in \mathrm{NE}(N, r)$ be a Nash equilibrium. Then $k \in \mathrm{supp}(\boldsymbol{x}) \Rightarrow \forall n < k$, link n is congested.

Proof: By contradiction, if $m_n = 0$, then $\ell_n(x_n, m_n) = a_n < a_k \leq \ell_k(x_k, m_k)$, which contradicts Definition 2 of a Nash equilibrium. ■

Corollary 1: Support of a Nash Equilibrium: Let $(\boldsymbol{x}, \boldsymbol{m}) \in \mathrm{NE}(N, r)$ be a Nash equilibrium and $k = \max \mathrm{supp}(\boldsymbol{x})$ be the last link in the support of $\boldsymbol{x}$ (i.e., the one with the largest free-flow latency). Then we have $\mathrm{supp}(\boldsymbol{x}) = \{1, \ldots, k\}$.

Proof: Since $k \in \mathrm{supp}(\boldsymbol{x})$, we have by Proposition 2 that $\forall n < k$, link n is congested, thus $n \in \mathrm{supp}(\boldsymbol{x})$ (by definition, a congested link cannot be empty). ■

No Essential Uniqueness: For the HQSF latency class, the essential uniqueness property[3] does not hold, i.e., there may exist multiple Nash equilibria that have different costs, an example is given in Fig. 4.

Single-Link-Free-Flow Equilibria and Congested Equilibria: The example shows that in general, there may exist multiple Nash equilibria that have different costs, different congestion

[3] The essential uniqueness property states that for the class of non-decreasing latency functions, all Nash equilibria have the same total cost. See for example [4], [8], [26].

state vectors, and different supports. However, not every congestion state vector $\boldsymbol{m} \in \{0,1\}^N$ can be that of a Nash equilibrium: let $(\boldsymbol{x}, \boldsymbol{m}) \in \mathrm{NE}(N, r)$ be a Nash equilibrium, and let $k = \max \operatorname{supp}(\boldsymbol{x})$ be the index of the last link in the support of $\boldsymbol{x}$. Then by Proposition 2, we have that $\forall i < k$, $m_i = 1$, and $\forall i > k$, $m_i = 0$. Thus we have

- Either $\boldsymbol{m} = (1, \ldots, 1,\ \overset{\overset{\bar{k}}{\sqcap}}{0}, 0, \ldots, 0)$ i.e., the last link in the support is in free-flow, all other links in the support are congested. In this case we call $(\boldsymbol{x}, \boldsymbol{m})$ a *single-link-free-flow equilibrium*, and denote the set of such equilibria by $\mathrm{NE}_{\mathrm{f}}(N, r)$.
- Or $\boldsymbol{m} = (1, \ldots, 1,\ \overset{\overset{\bar{k}}{\sqcap}}{1}, 0, \ldots, 0)$ i.e., all links in the support are congested. In this case we call $(\boldsymbol{x}, \boldsymbol{m})$ a *congested equilibrium*, and denote the set of such equilibria by $\mathrm{NE}_{\mathrm{c}}(N, r)$.

B. Existence of Single-Link-Free-Flow Equilibria

Let $(\boldsymbol{x}, \boldsymbol{m})$ be a single-link-free-flow equilibrium, and let $k = \max \operatorname{supp}(\boldsymbol{x})$. We have from Proposition 2 that links $\{1, \ldots, k-1\}$ are congested and link k is in free-flow. Therefore we must have $\forall n \in \{1, \ldots, k-1\}$, $\ell_n(x_n, 1) = \ell_k(x_k, 0) = a_k$. This uniquely determines the flow on the congested links:

Definition 6: Congestion Flow: Let $k \in \{2, \ldots, N\}$. Then $\forall n \in \{1, \ldots, k-1\}$, there exists a unique flow x_n such that $\ell_n(x_n, m_n) = a_k$. We denote this flow by $\hat{x}_n(k)$ and call it *k-congestion flow* on link n. It is given by

$$\hat{x}_n(k) = \ell_n(\cdot, 1)^{-1}(a_k). \tag{8}$$

We note that $\hat{x}_n(k)$ is decreasing in k, since $\ell_n(\cdot, 1)^{-1}$ is decreasing.

Proposition 3: Single-Link-Free-Flow Equilibria: $(\boldsymbol{x}, \boldsymbol{m})$ is a single-link-free-flow equilibrium if and only if $\exists k \in \{1, \ldots, N\}$ such that $0 < r - \sum_{n=1}^{k-1} \hat{x}_n(k) \leq x_k^{\max}$, and

$$\boldsymbol{x} \triangleq \left(\hat{x}_1(k), \ldots, \hat{x}_{k-1}(k), r - \sum_{n=1}^{k-1} \hat{x}_n(k), 0, \ldots, 0\right) \tag{9}$$

$$\boldsymbol{m} \triangleq (1, \ldots, 1,\ \overset{\overset{\bar{k}}{\sqcap}}{0}, \ldots, 0). \tag{10}$$

Illustrations of (10) and (9) are shown in Fig. 6.

Next, we give a necessary and sufficient condition for the existence of single-link-free-flow equilibria.

Lemma 1: Existence of Single-Link-Free-Flow Equilibria: Let

$$r^{\mathrm{NE}}(N) \triangleq \max_{k \in \{1, \ldots, N\}} \left\{ x_k^{\max} + \sum_{n=1}^{k-1} \hat{x}_n(k) \right\}. \tag{11}$$

A single-link-free-flow equilibrium exists for the instance (N, r) if and only if $r \leq r^{\mathrm{NE}}(N)$.

Proof: If a single-link-free-flow equilibrium exists, then by Proposition 3, it is of the form given by (10) and (9) for some k. The flow on link k is then given by $r - \sum_{n=1}^{k-1} \hat{x}_n(k) \leq x_k^{\max}$. Therefore $r \leq x_k^{\max} + \sum_{n=1}^{k-1} \hat{x}_n(k) \leq r^{\mathrm{NE}}(N)$.

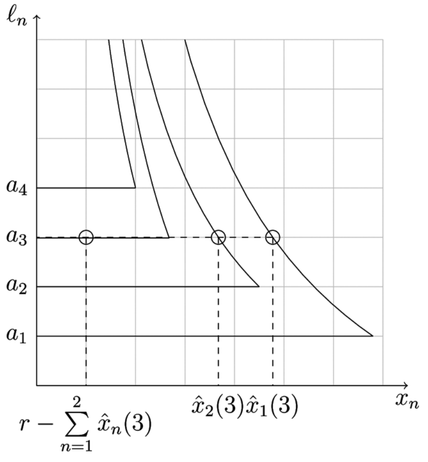


Fig. 6. Example of a single-link-free-flow equilibrium. Link 3 is in free-flow and links 1 and 2 are congested. The common latency on all links in the support is a_3.

We prove the converse by induction on the size N of the network. Let $\mathtt{P}_N$ denote the property: $\forall r \in (0, r^{\mathrm{NE}}(N)]$, there exists a single-link-free-flow equilibrium for the instance (N, r).

For $N = 1$, it is clear that if $0 < r \leq x_1^{\max}$, there is a single-link-free-flow equilibrium simply given by $(x_1, m_1) = (r, 0)$.

Now let $N \geq 1$, assume $\mathtt{P}_N$ holds and let us show $\mathtt{P}_{N+1}$. Let $0 < r \leq r^{\mathrm{NE}}(N+1)$ and consider an instance $(N+1, r)$.

Case 1: If $r \leq r^{\mathrm{NE}}(N)$, then by the induction hypothesis $\mathtt{P}_N$, there exists a single-link-free-flow equilibrium $(\boldsymbol{x}, \boldsymbol{m})$ for the instance (N, r). Then $(\boldsymbol{x}', \boldsymbol{m}')$ defined as $\boldsymbol{x}' = (x_1, \ldots, x_N, 0)$ and $\boldsymbol{m}' = (m_1, \ldots, m_N, 0)$ is clearly a single-link-free-flow equilibrium for the instance $(N+1, r)$.

Case 2: If $r^{\mathrm{NE}}(N) < r \leq r^{\mathrm{NE}}(N+1)$ then by Proposition 3, an equilibrium exists if

$$0 < r - \sum_{n=1}^{N} \hat{x}_n(N+1) \leq x_{N+1}^{\max}. \tag{12}$$

First, we note that since $r^{\mathrm{NE}}(N) < r^{\mathrm{NE}}(N+1)$, then

$$r^{\mathrm{NE}}(N+1) = x_{N+1}^{\max} + \sum_{n=1}^{N} \hat{x}_n(N+1).$$

Thus

$$r \leq r^{\mathrm{NE}}(N+1) = x_{N+1}^{\max} + \sum_{n=1}^{N} \hat{x}_n(N+1)$$

which proves the second inequality in (12). To show the first inequality, we have

$$\begin{aligned} r > r^{\mathrm{NE}}(N) &\geq x_N^{\max} + \sum_{n=1}^{N-1} \hat{x}_n(N) \\ &\geq \hat{x}_N(N+1) + \sum_{n=1}^{N-1} \hat{x}_n(N+1) \end{aligned}$$

where the last inequality results from the fact that $\hat{x}_n(N) \geq \hat{x}_n(N+1)$ and $x_N^{\max} \geq \hat{x}_N(N+1)$ by Definition 6 of congestion flow. This achieves the induction. ■

Corollary 2: The maximum demand r such that the set of Nash equilibria $\mathrm{NE}(N,r)$ is non-empty is $r^{\mathrm{NE}}(N)$.

Proof: By the previous lemma, $r^{\mathrm{NE}}(N)$ is a lower bound on the maximum demand. To show that it is also an upper bound, suppose that $\mathrm{NE}(N,r)$ is non-empty, and let $(\boldsymbol{x},\boldsymbol{m}) \in \mathrm{NE}(N,r)$ and $k = \max \operatorname{supp}(\boldsymbol{x})$. Then we have $\operatorname{supp}(\boldsymbol{x}) = \{1,\ldots,k\}$ by Corollary 1, and by Definition 2 of a Nash equilibrium, $\forall n \leq k$, $\ell_n(x_n,m_n) = \ell_k(x_k,m_k) \geq a_k$, therefore $x_n \leq \hat{x}_n(k)$. We also have $x_k \leq x_k^{\max}$. Combining the inequalities, we have

$$r = \sum_{n=1}^{k} x_n \leq x_k^{\max} + \sum_{n=1}^{k-1} \hat{x}_n(k) \leq r^{\mathrm{NE}}(N).$$

■

C. Number of Equilibria

Proposition 4: An Upper Bound on the Number of Equilibria: Consider a routing game instance (N,r). For any given $k \in \{1,\ldots,N\}$, there is at most one single-link-free-flow equilibrium and one congested equilibrium with support $\{1,\ldots,k\}$. As a consequence, by Corollary 1, the instance (N,r) has at most N single-link-free-flow equilibria and N congested equilibria.

Proof: We prove the result for single-link-free-flow equilibria, the proof for congested equilibria is similar. Let $k \in \{1,\ldots,N\}$, and assume $(\boldsymbol{x},\boldsymbol{m})$ and $(\boldsymbol{x}',\boldsymbol{m}')$ are single-link-free-flow equilibria such that $\max \operatorname{supp}(\boldsymbol{x}) = \max \operatorname{supp}(\boldsymbol{x}') = k$. We first observe that by Corollary 1, $\boldsymbol{x}$ and $\boldsymbol{x}'$ have the same support $\{1,\ldots,k\}$, and by Proposition 2, $\boldsymbol{m} = \boldsymbol{m}'$. Since link k is in free-flow under both equilibria, we have $\ell_k(x_k,m_k) = \ell_k(x'_k,m'_k) = a_k$, and by Definition 2 of a Nash equilibrium, any link in the support of both equilibria has the same latency a_k, i.e., $\forall n \leq k$, $\ell_n(x_n,1) = \ell_i(x'_n,1) = a_k$. Since the latency in congestion is injective, we have $\forall n \leq k$, $x_n = x'_n$, therefore $\boldsymbol{x} = \boldsymbol{x}'$. ■

D. Best Nash Equilibrium

In order to study the inefficiency of Nash equilibria, and the improvement of performance that we can achieve using optimal Stackelberg routing, we focus our attention on best Nash equilibria and *price of stability* [1] as a measure of their inefficiency.

Lemma 2: Best Nash Equilibrium: For a routing game instance (N,r), $r \leq r^{\mathrm{NE}}(N)$, the unique best Nash equilibrium is the single-link-free-flow equilibrium that has smallest support

$$\mathrm{BNE}(N,r) = \underset{(\boldsymbol{x},\boldsymbol{m}) \in \mathrm{NE}_{\mathrm{f}}(N,r)}{\arg\min} \left\{ \max \operatorname{supp}(\boldsymbol{x}) \right\}.$$

Proof: We first show that a congested equilibrium cannot be a best Nash equilibrium. Let $(\boldsymbol{x},\boldsymbol{m}) \in \mathrm{NE}(N,r)$ be a congested equilibrium and let $k = \max \operatorname{supp}(x)$. By Proposition 1, the cost of $(\boldsymbol{x},\boldsymbol{m})$ is $C(\boldsymbol{x},\boldsymbol{m}) = \ell_k(x_k,1)r > a_k r$. We observe that $(\boldsymbol{x},\boldsymbol{m})$ restricted to $\{1,\ldots,k\}$ is an equilibrium for the instance (k,r), thus by Corollary 2, $r \leq r^{\mathrm{NE}}k$, and by Lemma 1, there exists a single-link-free-flow equilibrium $(\boldsymbol{x}',\boldsymbol{m}')$ for (k,r), with cost $C(\boldsymbol{x}',\boldsymbol{m}') \leq a_k r$. Clearly, $(\boldsymbol{x}'',\boldsymbol{m}'')$ defined as $\boldsymbol{x}'' = (x'_1,\ldots,x'_k,0,\ldots,0)$ and $\boldsymbol{m}'' = (m'_1,\ldots,m'_k,0,\ldots,0)$, is a single-link-free-flow equilibrium for the original instance (N,r), with cost $C(\boldsymbol{x}'',\boldsymbol{m}'') = C(\boldsymbol{x}',\boldsymbol{m}') \leq a_k r < C(\boldsymbol{x},\boldsymbol{m})$, which proves that $(\boldsymbol{x},\boldsymbol{m})$ is not a best Nash equilibrium. Therefore best Nash equilibria are single-link-free-flow equilibria. And since the cost of a single-link-free-flow equilibrium $(\boldsymbol{x},\boldsymbol{m})$ is simply $C(\boldsymbol{x},\boldsymbol{m}) = a_k r$ where $k = \max \operatorname{supp}(\boldsymbol{x})$, it is clear that the smaller the support, the lower the total cost. Uniqueness follows from Proposition 4. ■

Complexity of Computing the Best Nash Equilibrium: Lemma 2 gives a simple algorithm for computing the best Nash equilibrium for any instance (N,r): simply enumerate all single-link-free-flow equilibria (there are at most N such equilibria by Proposition 4), and select the one with the smallest support. This is detailed in Algorithm 1.

Algorithm 1 Best Nash Equilibrium

```
procedure bestNE $(N, r)$
Inputs: Size of the network $N$, demand $r$
Outputs: Best Nash equilibrium $(\boldsymbol{x}, \boldsymbol{m})$
for $k \in \{1, \ldots, N\}$
    let $(\boldsymbol{x}, \boldsymbol{m}) = \texttt{freeFlowConfig}(k)$
    if $x_k \in [0, x_k^{\max}]$
        return $(\boldsymbol{x}, \boldsymbol{m})$
return No-Solution
procedure freeFlowConfig $(k)$
Inputs: Free-flow link index $k$
Outputs: Assignment $(\boldsymbol{x}, \boldsymbol{m}) = (\boldsymbol{x}^{r,k}, \boldsymbol{m}^k)$
for $n \in \{1, \ldots, N\}$
    if $n < k$
        $x_n = \hat{x}_n(k)$, $m_n = 1$
    elseif $n == k$
        $x_k = r - \sum_{n=1}^{k-1} x_n$, $m_k = 0$
    else
        $x_n = 0$, $m_n = 0$
return $(\boldsymbol{x}, \boldsymbol{m})$
```

The congestion flow values $\{\hat{x}_n(k), 1 \leq n < k \leq N\}$ can be precomputed in $O(N^2)$. There are at most N calls to `freeFlowConfig`, which runs in $O(N)$ time, thus `bestNE` runs in $O(N^2)$ time. This shows that the best Nash equilibrium can be computed in quadratic time.

III. Optimal Stackelberg Strategies

In this section, we prove our main result that NCF strategy is an optimal Stackelberg strategy (Theorem 1). Furthermore, we show that the entire set of optimal strategies $\mathbf{S}^\star(N,r,\alpha)$ can be computed in a simple way from the NCF strategy.

Let $(\bar{\boldsymbol{t}},\bar{\boldsymbol{m}})$ be the *best Nash equilibrium* for the instance $(N,(1-\alpha)r)$. It represents the best Nash equilibrium of the non-compliant flow $(1-\alpha)r$ when it is not sharing the network with the compliant flow. Let $\bar{k} = \max \operatorname{supp}(\bar{\boldsymbol{t}})$ be the last link in the support of $\bar{\boldsymbol{t}}$. Let $\bar{\boldsymbol{s}}$ be the NCF strategy defined by (7). Then the total flow $\bar{\boldsymbol{x}} = \bar{\boldsymbol{s}} + \bar{\boldsymbol{t}}$ is given by

$$\bar{\boldsymbol{x}} = \Bigg(\hat{x}_1(\bar{k}), \dots, \hat{x}_{\bar{k}-1}(\bar{k}), x_{\bar{k}}^{\max}, x_{\bar{k}+1}^{\max}, \dots, x_{l-1}^{\max}, \\ r - \sum_{n=1}^{\bar{k}-1} \hat{x}_n(\bar{k}) - \sum_{n=\bar{k}}^{l-1} x_n^{\max}, 0, \dots, 0\Bigg) \tag{13}$$

and the corresponding latencies are

$$\left(a_{\bar{k}}, \dots, \overset{\overset{\bar{k}}{\sqcap}}{a_{\bar{k}}}, a_{\bar{k}+1}, \dots, a_N \right). \tag{14}$$

Fig. 5 shows the total flow $\bar{\boldsymbol{x}}_n = \bar{\boldsymbol{s}}_n + \bar{\boldsymbol{t}}_n$ on each link. Under $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$, links $\{1, \dots, \bar{k}-1\}$ are congested and have latency $a_{\bar{k}}$, links $\{\bar{k}, \dots, l-1\}$ are in free-flow and at maximum capacity, and the remaining flow is assigned to link l.

We observe that for any Stackelberg strategy $\boldsymbol{s} \in \mathbf{S}(N, r, \alpha)$, the induced best Nash equilibrium $(\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$ is a single-link-free-flow equilibrium by Lemma 2, since $(\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$ is the best Nash equilibrium for the instance $(N, \alpha r)$ and latencies

$$\begin{aligned} \tilde{\ell}_n : \tilde{D}_n &\to \mathbb{R}_+ \\ (x_n, m_n) &\mapsto \ell_n(s_n + x_n, m_n) \end{aligned} \tag{15}$$

where $\tilde{D}_n \triangleq [0, \tilde{x}_n^{\max}] \times \{0\} \cup (0, \tilde{x}_n^{\max}) \times \{1\}$ and $\tilde{x}_n^{\max} \triangleq x_n^{\max} - s_n$.

A. Proof of Theorem 1: The NCF Strategy Is an Optimal Stackelberg Strategy

Let $\boldsymbol{s} \in S(N, r, \alpha)$ be a Stackelberg strategy and $(\boldsymbol{t}, \boldsymbol{m}) = (\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$ be the best Nash equilibrium of the non-compliant flow, induced by $\boldsymbol{s}$. Let $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{t}(\boldsymbol{s})$ and $\bar{\boldsymbol{x}} = \bar{\boldsymbol{s}} + \bar{\boldsymbol{t}}$ be the total flows. To prove Theorem 1, we seek to show that $C(\boldsymbol{x}, \boldsymbol{m}) \geq C(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$.

The proof is organized as follows: we first compare the supports of the induced equilibria (Lemma 3), then show that links $\{1, \dots, l-1\}$ are more congested under assignment $(\boldsymbol{x}, \boldsymbol{m})$ than under $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$, in the following sense: they hold less flow and have greater latency (Lemma 4). Then we conclude by showing the desired inequality.

Lemma 3: Let $k = \max \operatorname{supp}(\boldsymbol{t})$ and $\bar{k} = \max \operatorname{supp}(\bar{\boldsymbol{t}})$. Then $k \geq \bar{k}$.

In words, the last link in the support of $\boldsymbol{t}(\boldsymbol{s})$ has higher free-flow latency than the last link in the support of $\bar{\boldsymbol{t}}$.

Proof: We first note that $(\boldsymbol{s} + \boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m})$ restricted to $\operatorname{supp}(\boldsymbol{t}(\boldsymbol{s}))$ is a Nash equilibrium. Then since link k is in free-flow we have $\ell_k(s_k + t_k(\boldsymbol{s}), m_k) = a_k$, and since $k \in \operatorname{supp}(\boldsymbol{t}(\boldsymbol{s}))$, we have by definition that any other link has greater or equal latency. In particular, $\forall n \in \{1, \dots k-1\}$, $\ell_n(s_n + t_n(\boldsymbol{s}), m_n) \geq a_k$, thus $s_n + t_n(\boldsymbol{s}) \leq \hat{x}_n(k)$. Therefore we have $\sum_{n=1}^{k} s_n + t_n(\boldsymbol{s}) \leq \sum_{n=1}^{k-1} \hat{x}_n(k) + x_k^{\max}$. But $\sum_{n=1}^{k} (s_n + t_n(\boldsymbol{s})) \geq \sum_{n \in \operatorname{supp}(\boldsymbol{t})} t_n(\boldsymbol{s}) = (1-\alpha)r$ since $\operatorname{supp}(\boldsymbol{t}) \subseteq \{1, \dots, k\}$. Therefore $(1-\alpha)r \leq \sum_{n=1}^{k-1} \hat{x}_n(k) + x_k^{\max}$. By Lemma 1, there exists a single-link-free-flow equilibrium for the instance $(N, (1-\alpha)r)$ supported on the first k links. Let $(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{m}})$ be such an equilibrium. The cost of this equilibrium is $(1-\alpha)r\ell_0$ where $\ell_0 \leq a_k$ is the free-flow latency of the last link in the support of $\tilde{\boldsymbol{t}}$. Thus $C(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{m}}) \leq (1-\alpha)ra_k$. Since by definition $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{m}})$ is the *best Nash equilibrium* for the instance $(N, (1-\alpha)r)$ and has cost $(1-\alpha)ra_{\bar{k}}$, we must have $(1-\alpha)ra_{\bar{k}} \leq (1-\alpha)ra_k$, i.e., $a_{\bar{k}} \leq a_k$. ■

Lemma 4: Under assignment $(\boldsymbol{x}, \boldsymbol{m})$, the links $\{1, \dots, l-1\}$ have greater (or equal) latency and hold less (or equal) flow than under $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$, i.e., $\forall n \in \{1, \dots, l-1\}$, $\ell_n(x_n, m_n) \geq \ell_n(\bar{x}_n, \bar{m}_n)$ and $x_n \leq \bar{x}_n$.

Proof: Since $k \in \operatorname{supp}(\boldsymbol{t})$, we have by definition of a Stackelberg strategy and its induced equilibrium that $\forall n \in \{1, \dots, k-1\}$, $\ell_n(x_n, m_n) \geq \ell_k(x_k, m_k) \geq a_k$, see (5). We also have by definition of the candidate assignment $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$ and the resulting latencies given by (14), $\forall n \in \{1, \dots, \bar{k}-1\}$, n is congested and $\ell_n(x_n, m_n) = a_{\bar{k}}$. Thus using the fact that $k \geq \bar{k}$, we have $\forall n \in \{1, \dots, \bar{k}-1\}$, $\ell_n(x_n, m_n) \geq a_k \geq a_{\bar{k}} = \ell_n(\bar{x}_n, \bar{m}_n)$, and $x_n \leq \hat{x}_n(k) \leq \hat{x}_n(\bar{k}) = \bar{x}_n$.

We have from (13) that $\forall n \in \{\bar{k}, \dots, l-1\}$, n is in free-flow and at maximum capacity under $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$ (i.e., $\bar{x}_n = x_n^{\max}$ and $\ell_n(\bar{x}_n) = a_n$). Thus $\forall n \in \{\bar{k}, \dots, l-1\}$, $\ell_n(x_n, m_n) \geq a_n = \ell_n(\bar{x}_n, \bar{m}_n)$ and $x_n \leq x_n^{\max} = \bar{x}_n$. This completes the proof of the lemma. ■

We can now show the desired inequality. We have

$$\begin{aligned} C(\boldsymbol{x}, \boldsymbol{m}) &= \sum_{n=1}^{N} x_n \ell_n(x_n, m_n) \\ &= \sum_{n=1}^{l-1} x_n \ell_n(x_n, m_n) + \sum_{n=l}^{N} x_n \ell_n(x_n, m_n) \\ &\geq \sum_{n=1}^{l-1} x_n \ell_n(\bar{x}_n, \bar{m}_n) + \sum_{n=l}^{N} x_n a_l \end{aligned} \tag{16}$$

where the last inequality is obtained using Lemma 4 and the fact that $\forall n \in \{l, \dots, N\}$, $\ell_n(x_n, m_n) \geq a_n \geq a_l$. Then rearranging the terms we have

$$C(\boldsymbol{x}, \boldsymbol{m}) \geq \sum_{n=1}^{l-1} (x_n - \bar{x}_n) \ell_n(\bar{x}_n, \bar{m}_n) \\ + \sum_{n=1}^{l-1} \bar{x}_n \ell_n(\bar{x}_n, \bar{m}_n) + \sum_{n=l}^{N} x_n a_l.$$

Then we have $\forall n \in \{1, \dots, l-1\}$,

$$(x_n - \bar{x}_n)\left(\ell_n(\bar{x}_n, \bar{m}_n) - a_l\right) \geq 0$$

[by Lemma 4, $x_n - \bar{x}_n \leq 0$, and we have $\ell_n(\bar{x}_n, \bar{m}_n) \leq a_l$ by (14)]. Thus

$$\sum_{n=1}^{l-1} (x_n - \bar{x}_n) \ell_n(\bar{x}_n, \bar{m}_n) \geq \sum_{n=1}^{l-1} (x_n - \bar{x}_n) a_l \tag{17}$$

and we have

$$\begin{aligned}C(\boldsymbol{x},\boldsymbol{m}) &\geq \sum_{n=1}^{l-1}(x_n-\bar{x}_n)a_l+\sum_{n=1}^{l-1}\bar{x}_n\ell_n(\bar{x}_n,\bar{m}_n)+\sum_{n=l}^{N}x_na_l\\ &= a_l\left(\sum_{n=1}^{n}x_n-\sum_{n=1}^{l-1}\bar{x}_n\right)+\sum_{n=1}^{l-1}\bar{x}_n\ell_n(\bar{x}_n,\bar{m}_n)\\ &= a_l\left(r-\sum_{n=1}^{l-1}\bar{x}_n\right)+\sum_{n=1}^{l-1}\bar{x}_n\ell_n(\bar{x}_n,\bar{m}_n).\end{aligned}$$

But $a_l(r-\sum_{n=1}^{l-1}\bar{x}_n) = \bar{x}_l\ell_l(\bar{x}_l,\bar{m}_l)$ since $\operatorname{supp}(\bar{\boldsymbol{x}}) = \{1,\ldots,l\}$ and $\ell_l(\bar{x}_l,\bar{m}_l)=a_l$. Therefore

$$C(\boldsymbol{x},\boldsymbol{m}) \geq \bar{x}_l\ell_l(\bar{x}_l,\bar{m}_l)+\sum_{n=1}^{l-1}\bar{x}_n\ell_n(\bar{x}_n,\bar{m}_n)=C(\bar{\boldsymbol{x}},\bar{\boldsymbol{m}}).$$

This completes the proof of Theorem 1. ■

Therefore the NCF strategy is an optimal Stackelberg strategy, and it can be computed in polynomial time since it is generated in linear time after computing the best Nash equilibrium $\mathrm{BNE}(N,(1-\alpha)r)$, which was shown to be quadratic in N.

The NCF strategy is, in general, not the unique optimal Stackelberg strategy. In the next section, we show that any optimal Stackelberg strategy can in fact be easily expressed in terms of the NCF strategy.

B. Set of Optimal Stackelberg Strategies

In this section, we show that the set of optimal Stackelberg strategies $\mathbf{S}^\star(N,r,\alpha)$ can be generated from the NCF strategy. This shows in particular that the NCF strategy is robust, in a sense explained below.

Let $\bar{\boldsymbol{s}} = \mathrm{NCF}(N,r,\alpha)$ be the *non-compliant first* strategy, $\{(\bar{\boldsymbol{t}},\bar{\boldsymbol{m}})\} = \mathrm{BNE}(N,(1-\alpha)r)$ be the Nash equilibrium induced by $\bar{\boldsymbol{s}}$, and $\bar{k} = \max\operatorname{supp}(\bar{\boldsymbol{t}})$ the last link in the support of the induced equilibrium, as defined above. By definition, the NCF strategy $\bar{\boldsymbol{s}}$ assigns zero compliant flow to links $\{1,\ldots,\bar{k}-1\}$, and saturates links one by one, starting from $\bar{k}$ (see (7) and Fig. 5).

To give an example of an optimal Stackelberg strategy other than the NCF strategy, consider a strategy $\boldsymbol{s}$ defined by $\boldsymbol{s}=\bar{\boldsymbol{s}}+\boldsymbol{\epsilon}$ where

$$\boldsymbol{\epsilon}=\left(\epsilon_1,0,\ldots,0,\ \overbrace{-\epsilon_1}^{\bar{k}},0,\ldots,0\right)$$

and is such that $s_1=\epsilon_1\in[0,\hat{x}_1(\bar{k})]$, and $s_{\bar{k}}=\bar{s}_{\bar{k}}-\epsilon_1\geq 0$ (see Fig. 7). Strategy $\boldsymbol{s}$ will induce $\boldsymbol{t}(\boldsymbol{s})=\bar{\boldsymbol{t}}-\boldsymbol{\epsilon}$, and the resulting total cost is minimal since $C(\boldsymbol{s}+\boldsymbol{t}(\boldsymbol{s}))=C(\bar{\boldsymbol{s}}+\boldsymbol{\epsilon}+\bar{\boldsymbol{t}}-\boldsymbol{\epsilon})=C(\bar{\boldsymbol{s}}+\bar{\boldsymbol{t}})$. This shows that $\boldsymbol{s}$ is an optimal Stackelberg strategy. More generally, the following holds:

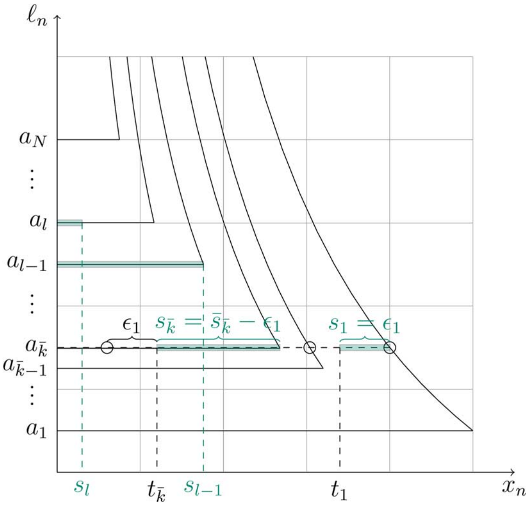


Fig. 7. Example of an optimal Stackelberg strategy $\boldsymbol{s}=\bar{\boldsymbol{s}}-\boldsymbol{\epsilon}$. The circles show the best Nash equilibrium $(\bar{\boldsymbol{t}},\boldsymbol{m})$. The strategy $\boldsymbol{s}$ is highlighted in green.

Lemma 5: Consider a Stackelberg strategy $\boldsymbol{s}$ of the form $\boldsymbol{s}=\bar{\boldsymbol{s}}+\boldsymbol{\epsilon}$ where

$$\boldsymbol{\epsilon}=\left(\epsilon_1,\epsilon_2,\ldots,\epsilon_{\bar{k}-1},-\sum_{n=1}^{\bar{k}-1}\epsilon_n,\ \overbrace{0}^{\bar{k}+1},\ldots,0\right) \tag{18}$$

and $\boldsymbol{\epsilon}$ is such that

$$\epsilon_n\in\left[0,\hat{x}_n(\bar{k})\right]\quad\forall n\in\{1,\ldots,\bar{k}-1\} \tag{19}$$

$$\bar{s}_{\bar{k}}\geq\sum_{n=1}^{\bar{k}-1}\epsilon_n. \tag{20}$$

Then $\boldsymbol{s}$ is an optimal Stackelberg strategy.

Proof: We show that $\boldsymbol{s}=\bar{\boldsymbol{s}}+\boldsymbol{\epsilon}$ is a feasible assignment of the compliant flow αr, and that the induced equilibrium of the followers is $(\boldsymbol{t}(\boldsymbol{s}),\boldsymbol{m}(\boldsymbol{s}))=(\bar{\boldsymbol{t}}-\boldsymbol{\epsilon},\bar{\boldsymbol{m}})$.

Since $\sum_{n=1}^{N}\epsilon_n = 0$ by definition (18) of $\boldsymbol{\epsilon}$, we have $\sum_{n=1}^{N}s_n=\sum_{n=1}^{N}\bar{s}_n=\alpha r$. We also have

- $\forall n\in\{1,\ldots,\bar{k}-1\}$, $s_n=\epsilon_n\in[0,\hat{x}_n(\bar{k})]$ by (19). Thus $s_n\in[0,x_n^{\max}]$.
- $s_{\bar{k}}=\bar{s}_{\bar{k}}+\epsilon_{\bar{k}}\geq 0$ by (20), and $s_{\bar{k}}\leq\bar{s}_{\bar{k}}\leq x_{\bar{k}}^{\max}$.
- $\forall n\in\{\bar{k}+1,\ldots,N\}$, $s_n=\bar{s}_n\in[0,x_n^{\max}]$.

This shows that $\boldsymbol{s}$ is a feasible assignment. To show that $\boldsymbol{s}$ induces $(\bar{\boldsymbol{t}}-\boldsymbol{\epsilon},\bar{\boldsymbol{m}})$, we need to show that $\forall n\in\operatorname{supp}(\bar{\boldsymbol{t}}-\boldsymbol{\epsilon})$, $\forall k\in\{1,\ldots,N\}$

$$\ell_n(\bar{s}_n+\epsilon_n+\bar{t}_n-\epsilon_n,\bar{m}_n)\leq\ell_k(\bar{s}_k+\epsilon_k+\bar{t}_k-\epsilon_k,\bar{m}_k).$$

This is true $\forall n\in\operatorname{supp}(\bar{\boldsymbol{t}})$, by definition of $(\bar{\boldsymbol{t}},\bar{\boldsymbol{m}})$ and (5). To conclude, we observe that $\operatorname{supp}(\bar{\boldsymbol{t}}-\boldsymbol{\epsilon})\subset\operatorname{supp}(\bar{\boldsymbol{t}})$. ■

This shows that the NCF strategy is robust to perturbations: even if the strategy $\bar{\boldsymbol{s}}$ is not realized exactly, it may still be optimal if the perturbation $\boldsymbol{\epsilon}$ satisfies the conditions given above.

The converse of the previous lemma is true. This gives a necessary and sufficient condition for optimal Stackelberg strategies, given in the following theorem.

Theorem 2: The Set of Optimal Stackelberg Strategies: The set of optimal Stackelberg strategies $\mathbf{S}^\star(N, r, \alpha)$ is the set of strategies $\boldsymbol{s}$ of form $\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{\epsilon}$ where $\bar{\boldsymbol{s}} = \mathrm{NCF}(N, r, \alpha)$ is the non-compliant first strategy, and $\boldsymbol{\epsilon}$ satisfies (18), (19), and (20).

Proof: We prove the converse of Lemma 5. Let $\boldsymbol{s} \in \mathbf{S}^\star(N, r, \alpha)$ be an optimal Stackelberg strategy, $(\boldsymbol{t}, \boldsymbol{m}) = (\boldsymbol{t}(\boldsymbol{s}), \boldsymbol{m}(\boldsymbol{s}))$ the equilibrium of non-compliant flow induced by $\boldsymbol{s}$, $k = \max \operatorname{supp}(\boldsymbol{t})$ the last link in the support of $\boldsymbol{t}$, and $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{t}$ the total flow assignment.

We first show that $\boldsymbol{x} = \bar{\boldsymbol{x}}$. By optimality of both $\boldsymbol{s}$ and $\bar{\boldsymbol{s}}$, we have $C(\boldsymbol{x}, \boldsymbol{m}) = C(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$, therefore inequalities (16) and (17) in the proof of Theorem 1 must hold with equality. In particular, to have equality in (16) we need to have

$$\sum_{n=1}^{l-1} x_n \left(\ell_n(x_n, m_n) - \ell_n(\bar{x}_n, \bar{m}_n)\right) + \sum_{n=l}^{N} x_n \left(\ell_n(x_n, m_n) - a_l\right) = 0. \quad (21)$$

The terms in both sums are non-negative. Therefore

$$x_n \left(\ell_n(x_n, m_n) - \ell_n(\bar{x}_n, \bar{m}_n)\right) = 0 \quad \forall n \in \{1, \ldots, l-1\} \quad (22)$$

$$x_n \left(\ell_n(x_n, m_n) - a_l\right) = 0 \quad \forall n \in \{l, \ldots, N\} \quad (23)$$

and to have equality in (17) we need to have

$$(x_n - \bar{x}_n)\left(\ell_n(\bar{x}_n, \bar{m}_n) - a_l\right) = 0 \quad \forall n \in \{1, \ldots, l-1\}. \quad (24)$$

Let $n \in \{1, \ldots, l-1\}$. From the expression (14) of the latencies under $\bar{\boldsymbol{x}}$, we have $\ell_n(\bar{x}_n, \bar{m}_n) < a_l$, thus from equality (24) we have $x_n - \bar{x}_n = 0$. Now let $n \in \{l+1, \ldots N\}$. We have by definition of the latency functions, $\ell_n(x_n, m_n) \geq a_n > a_l$, thus from equality (23), $x_n = 0$. We also have from the expression (13), $\bar{x}_n = 0$. Therefore $x_n = \bar{x}_n \forall n \neq l$, but since $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ are both assignments of the same total flow r, we also have $x_l = \bar{x}_l$, which proves $\boldsymbol{x} = \bar{\boldsymbol{x}}$.

Next we show that $k = \bar{k}$. We have from the proof of Theorem 1 that $k \geq \bar{k}$. Assume by contradiction that $k > \bar{k}$. Then since $k \in \operatorname{supp}(\boldsymbol{t})$, we have by definition of the induced followers' assignment in (5), $\forall n \in \{1, \ldots, N\}$, $\ell_n(x_n, m_n) \geq \ell_k(x_k, m_k)$. And since $\ell_k(x_k, m_k) \geq a_k > a_{\bar{k}}$, we have (in particular for $n = \bar{k}$) $\ell_{\bar{k}}(x_{\bar{k}}, m_{\bar{k}}) > a_{\bar{k}}$, i.e., link $\bar{k}$ is congested under $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{m}})$, thus $x_{\bar{k}} > 0$. Finally, since $\ell_{\bar{k}}(\bar{x}_{\bar{k}}, \bar{m}_{\bar{k}}) = a_{\bar{k}}$, we have $\ell_{\bar{k}}(\bar{x}_{\bar{k}}, \bar{m}_{\bar{k}}) > \ell_{\bar{k}}(\bar{x}_{\bar{k}}, \bar{m}_{\bar{k}})$. Therefore $x_{\bar{k}}(\ell_{\bar{k}}(x_{\bar{k}}, m_{\bar{k}}) - \ell_{\bar{k}}(\bar{x}_{\bar{k}}, \bar{m}_{\bar{k}})) > 0$, since $\bar{k} < k \leq l$, this contradicts (22).

Now let $\boldsymbol{\epsilon} = \boldsymbol{s} - \bar{\boldsymbol{s}}$. We want to show that $\boldsymbol{\epsilon}$ satisfies (18), (19), and (20).

First, we have $\forall n \in \{1, \ldots, \bar{k}-1\}$, $\bar{s}_n = 0$, thus $\epsilon_n = s_n - \bar{s}_n = s_n$. We also have $\forall n \in \{1, \ldots, \bar{k}-1\}, 0 \leq s_n \leq x_n$, $x_n = \bar{x}_n$ (since $\boldsymbol{x} = \bar{\boldsymbol{x}}$), and $\bar{x}_n = \hat{x}_n(\bar{k})$ [by (13)]. Therefore $0 \leq s_n \leq \hat{x}_n(\bar{k})$. This proves (19).

Second, we have $\forall n \in \{\bar{k}+1, \ldots, N\}$, $t_n = \bar{t}_n = 0$ (since $k = \bar{k}$), and $x_n = \bar{x}_n$ (since $\boldsymbol{x} = \bar{\boldsymbol{x}}$) thus $\epsilon_n = s_n - \bar{s}_n = x_n - t_n - \bar{x}_n + \bar{t}_n = 0$.

Third, we have $\sum_{n=1}^{n} \epsilon_n = 0$ since $\boldsymbol{s}$ and $\bar{\boldsymbol{s}}$ are assignments of the same compliant flow αr, thus $\epsilon_{\bar{k}} = -\sum_{n \neq \bar{k}} \epsilon_n = -\sum_{n=1}^{\bar{k}-1} \epsilon_n$. This proves (18).

Finally, we have (20) since $s_{\bar{k}} \geq 0$ by definition of $\boldsymbol{s}$. ■

IV. Price of Stability Under Optimal Stackelberg Routing

To quantify the inefficiency of Nash equilibria, and the improvement that can be achieved using Stackelberg routing, several metrics have been used including price of anarchy [26], [27] and price of stability [1]. We use price of stability as a metric, which is defined as the ratio between the cost of the best Nash equilibrium and the cost of the social optimum.[4] Let $(\boldsymbol{x}^\star, \mathbf{0})$ denote the social optimum of the instance (N, r)—the social optimum is simply the free-flow assignment that saturates links one by one by increasing index (see Appendix B). Let $\bar{\boldsymbol{s}}$ be the non-compliant first strategy $\mathrm{NCF}(N, r, \alpha)$, and $(\boldsymbol{t}(\bar{\boldsymbol{s}}), \boldsymbol{m}(\bar{\boldsymbol{s}}))$ the induced equilibrium of the followers. The price of stability of the Stackelberg instance $\mathrm{NCF}(N, r, \alpha)$ is

$$\mathrm{POS}(N, r, \alpha) = \frac{C\left(\bar{\boldsymbol{s}} + \boldsymbol{t}(\bar{\boldsymbol{s}}), \boldsymbol{m}(\bar{\boldsymbol{s}})\right)}{C(\boldsymbol{x}^\star, \mathbf{0})}$$

where $\bar{\boldsymbol{s}}$ is the NCF strategy, and $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{m}})$ its induced equilibrium. The improvement achieved by optimal Stackelberg routing with respect to the Nash equilibrium $(\alpha = 0)$ can be measured using *value of altruism* [2], defined as

$$\mathrm{VOA}(N, r, \alpha) = \frac{\mathrm{POS}(N, r, 0)}{\mathrm{POS}(N, r, \alpha)}.$$

This terminology refers to the improvement achieved by having a fraction α of altruistic (or compliant) players, compared to a situation where everyone is selfish.

We give the expressions of price of stability and value of altruism in the case of a two-link network, as a function of the compliance rate $\alpha \in [0, 1]$ and demand r.

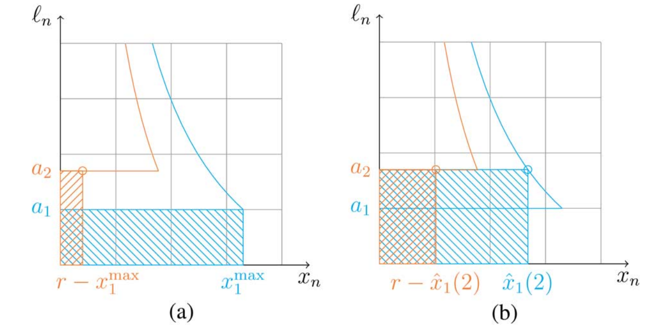
Case 1: $0 \leq (1-\alpha)r \leq x_1^{\max}$: In this case, link 1 can accommodate all the non-compliant flow, thus the induced equilibrium of the followers is $(\boldsymbol{t}(\bar{\boldsymbol{s}}), \boldsymbol{m}(\bar{\boldsymbol{s}})) = (((1-\alpha)r, 0), (0, 0))$, and by (7) the total flow induced by $\bar{\boldsymbol{s}}$ is $\bar{\boldsymbol{s}} + \boldsymbol{t}(\bar{\boldsymbol{s}}) = (x_1^{\max}, r - x_1^{\max})$ and coincides with the social optimum. Therefore, the price of stability is one.

Case 2: $x_1^{\max} < (1-\alpha)r \leq x_2^{\max} + \hat{x}_1(2)$: Observe that this case can only occur if $x_2^{\max} + \hat{x}_1(2) > x_1^{\max}$. In this case, link 1 cannot accommodate all the non-compliant flow, and the induced Nash equilibrium $(\boldsymbol{t}(\bar{\boldsymbol{s}}), \boldsymbol{m}(\bar{\boldsymbol{s}}))$ is then supported on both links. It is equal to $(\boldsymbol{x}^{2,(1-\alpha)r}, \boldsymbol{m}^2) = ((\hat{x}_1(2), (1-\alpha)r - \hat{x}_1(2)), (1, 0))$, and the total flow is $(\bar{\boldsymbol{s}} + \boldsymbol{t}(\bar{\boldsymbol{s}}) = (\hat{x}_1(2), r - \hat{x}_1(2))$, with total cost $a_2 r$ [Fig. 8(b)]. The social optimum is $(\boldsymbol{x}^\star, \boldsymbol{m}^\star) = ((x_1^{\max}, r - x_1^{\max}), (0, 0))$ (see Appendix B), with total cost $a_1 x_1^{\max} + a_2(r - x_1^{\max})$ [Fig. 8(a)]. Therefore the price of stability is

$$\mathrm{POS}(2, r, \alpha) = \frac{r a_2}{r a_2 - x_1^{\max}(a_2 - a_1)} = \frac{1}{1 - \frac{x_1^{\max}}{r}\left(1 - \frac{a_1}{a_2}\right)}.$$

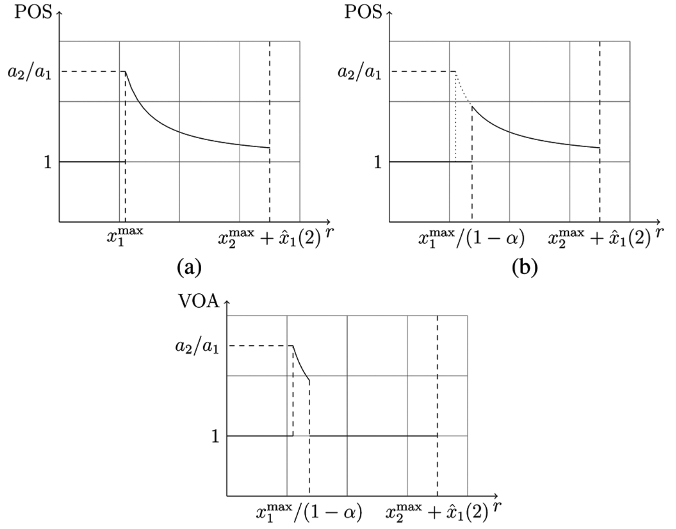
We observe that for a fixed flow demand $r > x_1^{\max}$, the price of stability is an increasing function of a_2/a_1. Intuitively, the inefficiency of Nash equilibria increases when the difference in

[4] Price of anarchy is defined as the ratio between the costs of the *worst* Nash equilibrium and the social optimum. For the case of non-decreasing latency functions, the price of anarchy and the price of stability coincide since all Nash equilibria have the same cost by the essential uniqueness property.



(a) (b)

Fig. 8. Social optimum (a) and best Nash equilibrium (b) when the demand exceeds the capacity of the first link ($r > x_1^{\max}$). The area of the shaded regions represents the total costs of each assignment.



(a) (b)

(c)

Fig. 9. Price of stability and value of altruism on a two-link network. Here we assume that $\hat{x}_1(2) + x_2^{\max} > x_1^{\max}$. (a) Price of stability, $\alpha = 0$; (b) Price of stability, $\alpha = 0.2$; (c) Value of altruism, $\alpha = 0.2$.

free-flow latency between the links increases. And as $a_2 \to a_1$, the price of stability goes to 1.

When the compliance rate is $\alpha = 0$, the price of stability attains a supremum equal to a_2/a_1, at $r = (x_1^{\max})^+$ [Fig. 9(a)]. This shows that selfish routing is most costly when the demand is slightly above critical value $r^{\mathrm{NE}}(1) = x_1^{\max}$. This also shows that for the general class of HQSF latencies on parallel networks, the price of stability is unbounded, since one can design an instance $(2, r)$ such that the maximal price of stability a_2/a_1 is arbitrarily large. Under optimal Stackelberg routing ($\alpha > 0$), the price of stability attains a supremum equal to $1/(\alpha + (1-\alpha)(a_1/a_2))$ at $r = (x_1^{\max}/(1-\alpha))^+$. We observe in particular that the supremum is decreasing in α, and that when $\alpha = 1$ (total control), the price of stability is identically one.

Therefore optimal Stackelberg routing can significantly decrease price of stability when $r \in (x_1^{\max}, x_1^{\max}/(1-\alpha))$. This can occur for small values of the compliance rate in situations where the demand slightly exceeds the capacity of the first link [Fig. 9(c)].

The same analysis can be done for a general network: given the latency functions on the links, one can compute the price of stability as a function of the flow demand r and the compliance rate α, using the form of the NCF strategy together with Algorithm 1 to compute the BNE. Computing the price of stability function reveals critical values of demand, for which optimal Stackelberg routing can lead to a significant improvement. This is discussed in further detail in the next section, using an example network with four links.

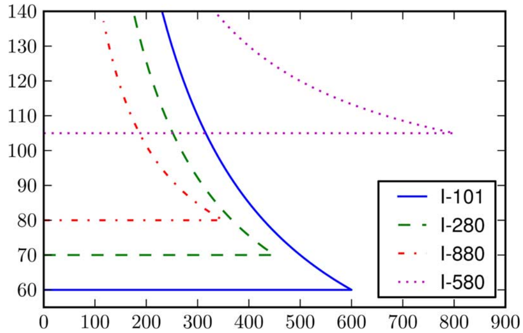


Fig. 10. Latency functions on an example highway network. Latency is in minutes, and demand is in cars/minute.

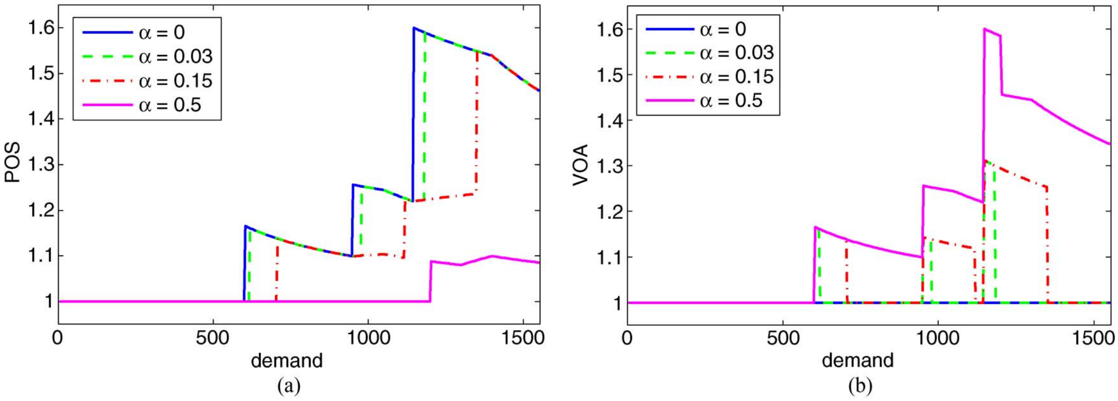
V. Numerical Results

In this section, we apply the previous results to a scenario of freeway traffic from the San Francisco Bay Area. Four parallel highways are chosen starting in San Francisco and ending in San Jose: I-101, I-280, I-880, and I-580 (Fig. 1). We analyze the inefficiency of Nash equilibria, and show how optimal Stackelberg routing (using the NCF strategy) can improve the efficiency.

Fig. 10 shows the latency functions for the highway network, assuming a triangular fundamental diagram for each highway (see Appendix A for a derivation of the latency function from a triangular fundamental diagram). Under free-flow conditions, I-101 is the fastest route available between San Francisco and San Jose. When I-101 becomes congested, other routes represent viable alternatives.

We computed price of stability and value of altruism (defined in the previous section) as a function of the demand r for different compliance rates. The results are shown in Fig. 11. We observe that for a fixed compliance rate, the price of stability is piecewise continuous in the demand [Fig. 11(a)], with discontinuities corresponding to an increase in the cardinality of the equilibrium's support (and a link transitioning from free-flow to congestion). If a transition exists for link n, it occurs at critical demand $r = r^{(\alpha)}(n)$, defined to be the infimum demand r such that n is congested under the equilibrium induced by $\mathrm{NCF}(N, r, \alpha)$.

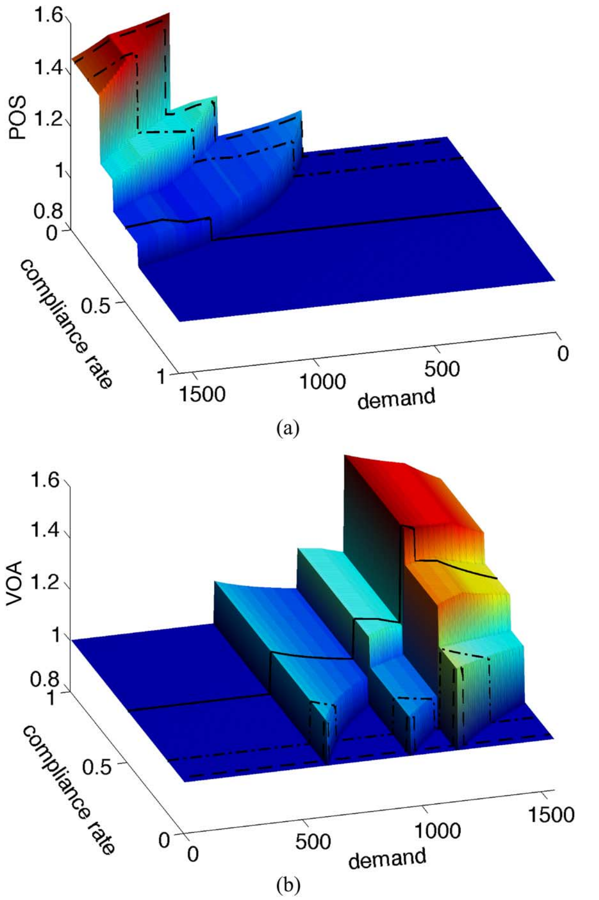
It can be shown that $r^{(\alpha)}(n) = r^{\mathrm{NE}}(n)/(1-\alpha)$, and we have in particular $r^{\mathrm{NE}}(n) = r^{(0)}(n)$. Therefore if a link n is congested under best Nash equilibrium ($r > r^{\mathrm{NE}}(n)$), optimal Stackelberg routing can decongest n if $r^{(\alpha)}(n) \geq r$. In particular, when the demand is slightly above critical demand $r^{(0)}(n)$, link n can be decongested with a small compliance rate. This is illustrated by the numerical values of price of stability on Fig. 11(a), where a small compliance rate ($\alpha = 0.05$) achieves high value of altruism when the demand is slightly above the critical values. This shows that optimal Stackelberg



(a)

(b)

Fig. 11. Price of stability (a) and value of altruism (b) as a function of the demand r for different values of compliance rate α.



(a)

(b)

Fig. 12. Price of stability (a) and value of altruism (b) as a function of the compliance rate α and demand r. Iso-α lines are plotted for $\alpha = 0.03$ (dashed), $\alpha = 0.15$ (dot-dashed), and $\alpha = 0.5$ (solid).

routing can achieve a significant improvement in efficiency, especially when the demand is near one of the critical values $r^{(\alpha)}(n)$.

Fig. 12 shows price of stability and value of altruism as a function of the demand $r \in [0, r^{\text{NE}}(N)]$ and compliance rate $\alpha \in [0, 1]$. We observe in particular that for a fixed value of demand, price of stability is a piecewise constant function of α. Computing this function can be useful for efficient planning and control, since it informs the central coordinator of the critical compliance rates that can achieve a strict improvement. For instance, if the demand on the example network is 1100 cars/min, price of stability is constant for compliance rates $\alpha \in [0.14, 0.46]$. Therefore if a compliance rate greater than 0.46 is not feasible, the controller may prefer to implement a control strategy with $\alpha = 0.14$, since further increasing the compliance rate will not improve efficiency, and may incur additional external cost (e.g., due to incentivizing more drivers).

VI. Summary and Concluding Remarks

We introduced a new class of latency functions to model congestion on networks with horizontal queues, and studied the resulting Nash equilibria for non-atomic, static routing games on parallel networks. We showed that the essential uniqueness property does not hold for the HQSF class of latency, and that the number of equilibria is at most $2N$. We also characterized the best Nash equilibrium.

In the Stackelberg routing game, we proved that the Non-compliant First (NCF) strategy is optimal, and that it can be computed in polynomial time. We illustrated these results using an example network for which we computed the decrease in inefficiency that can be achieved using optimal Stackelberg routing. This example showed that when the demand is near critical values $r^{\text{NE}}(n)$, optimal Stackelberg routing can achieve a significant improvement in efficiency, even for small values of compliance rate.

On the one hand, these results show that careful routing of a small compliant population can dramatically improve the efficiency of the network. On the other hand, they also indicate that for certain demand and compliance values, Stackelberg routing can be completely ineffective. Therefore identifying the ranges where optimal Stackelberg routing does improve the efficiency of the network is crucial for effective planning and control.

We believe this work offers several directions of future research: the work presented here only considers parallel networks under static assumptions (constant flow demand r, and static equilibria) and one question is whether these equilibria are stable in the dynamic sense, and how one may steer the system from one equilibrium to a better one: consider for example the case where the players are stuck in a congested equilibrium, and assume a coordinator has control over a fraction of the flow. Can the coordinator steer the system to a single-link-free-flow equilibrium by decongesting a link? And what is the minimal compliance rate needed to achieve this? Another question is how robust are these results? Do they hold for general network topologies? We believe that some of our results extend to general network topologies, but we foresee interesting technical challenges in formalizing these extensions.

Appendix

A. HQSF Latency Function From a Triangular Fundamental Diagram of Traffic

In this section we derive one example of a HQSF latency function ℓ_n in a traffic setting. We consider a triangular fundamental diagram, used to model traffic flow for example in [9], [10], i.e., a piecewise affine flux function x_n^ρ, given by

$$x_n^\rho(\rho_n) = \begin{cases} v_n^f \rho_n & \text{if } \rho_n \in \left[0, \rho_n^{\text{crit}}\right] \\ x_n^{\max} \frac{\rho_n - \rho_n^{\max}}{\rho_n^{\text{crit}} - \rho_n^{\max}} & \text{if } \rho_n \in \left(\rho_n^{\text{crit}}, \rho_n^{\max}\right] . \end{cases}$$

The flux function is linear in free-flow with positive slope v_n^f called free-flow speed, affine in congestion with negative slope $v_n^c \triangleq x_n^{\max}/(\rho_n^{\text{crit}} - \rho_n^{\max})$, and continuous (thus $v_n^f \rho_n^{\text{crit}} = x_n^{\max}$). By definition, it satisfies the assumptions in Section I-B. The latency is given by $L_n \rho_n / x_n^\rho(\rho_n)$ where L_n is the length of link n. It is then a simple function of the density

$$\ell_n^\rho(\rho_n) = \begin{cases} \frac{L_n}{v_n^F} & \rho_n \in \left[0, \rho_n^{\text{crit}}\right] \\ \frac{L_n \rho_n}{v_n^c(\rho_n - \rho_n^{\max})} & \rho_n \in \left(\rho_n^{\text{crit}}, \rho_n^{\max}\right] \end{cases}$$

which can be expressed as two functions of flow: a constant function $\ell_n(\cdot, 0)$ when the link is in free-flow, and a decreasing function $\ell_n(\cdot, 1)$ when the link is congested

$$\ell_n(x_n, 0) = \frac{L_n}{v_n^f}$$
$$\ell_n(x_n, 1) = L_n \left(\frac{\rho_n^{\max}}{x_n} + \frac{1}{v_n^c} \right).$$

This defines a function ℓ_n that satisfies the assumptions of Definition 1, and thus belongs to the HQSF latency class. Fig. 2 shows one example of a triangular fundamental diagram (top left) and the corresponding latency function ℓ_n (top right).

B. Social Optimal Assignments

Consider an instance (N, r) where the flow demand r does not exceed the maximum capacity of the network, i.e., $r \le \sum_n x_n^{\max}$. A social optimal assignment is an assignment that minimizes the total cost function $C(\boldsymbol{x}, \boldsymbol{m}) = \sum_n x_n \ell_n(x_n, m_n)$, i.e., it is a solution to the following Social Optimum (SO) optimization problem:

$$\underset{\substack{\boldsymbol{x} \in \prod_{n=1}^N [0, x_n^{\max}] \\ \boldsymbol{m} \in \{0,1\}^N}}{\text{minimize}} \sum_{n=1}^N x_n \ell_n(x_n, m_n) \quad (SO)$$
$$\text{subject to} \sum_{n=1}^N x_n = r.$$

Proposition 5: $(\boldsymbol{x}^\star, \boldsymbol{m}^\star)$ is optimal for (SO) only if $\forall n \in \{1, \dots, N\}, m_n^\star = 0$. ■

Proof: This follows immediately from the fact the latency on a link in congestion is always greater than the latency of the link in free-flow $\ell_n(x_n, 1) > \ell_n(x_n, 0) \forall x_n \in (0, x_n^{\max})$. ■

As a consequence of the previous proposition, and using the fact that the latency is constant in free-flow $\ell_n(x_n, 0) = a_n$, the social optimum can be computed by solving the following equivalent linear program:

$$\underset{\boldsymbol{x} \in \prod_{n=1}^N [0, x_n^{\max}]}{\text{minimize}} \sum_{n=1}^N x_n a_n$$
$$\text{subject to} \sum_{n=1}^N x_n = r.$$

Then since the links are ordered by increasing free-flow latency $a_1 < \cdots < a_N$, the social optimum is simply given by the assignment that saturates most efficient links first. Formally, if $k_0 = \max\{k | r \ge \sum_{n=1}^k x_n^{\max}\}$ then the social optimal assignment is given by $\boldsymbol{x}^\star = (x_1^{\max}, \dots, x_{k_0-1}^{\max}, r - \sum_{n=1}^{k_0-1} x_n^{\max}, 0, \dots, 0)$.

References

[1] E. Anshelevich, A. Dasgupta, J. Kleinberg, E. Tardos, T. Wexler, and T. Roughgarden, "The price of stability for network design with fair cost allocation," in *Proc. 45th Annu. IEEE Symp. Foundations of Computer Science*, 2004, pp. 295–304.

[2] A. Aswani and C. Tomlin, "Game-theoretic routing of GPS-assisted vehicles for energy efficiency," in *Proc. Amer. Control Conf. (ACC'11)*, 2011, pp. 3375–3380.

[3] M. Babaioff, R. Kleinberg, and C. H. Papadimitriou, "Congestion games with malicious players," *Games Econ. Behav.*, vol. 67, no. 1, pp. 22–35, 2009.

[4] M. Beckmann, C. B. McGuire, and C. B. Winsten, 1956.

[5] T. Boulogne, E. Altman, H. Kameda, and O. Pourtallier, "Mixed equilibrium for multiclass routing games," *IEEE Trans. Autom. Control*, vol. 47, no. 1, pp. 58–74, Jan. 2001.

[6] Caltrans, US 101 South, Corridor System Management Plan, 2010.

[7] S. Dafermos, "Traffic equilibrium and variational inequalities," *Transp. Sci.*, vol. 14, no. 1, pp. 42–54, 1980.

[8] S. C. Dafermos and F. T. Sparrow, "The traffic assignment problem for a general network," *J. Res. Nat. Bureau Standards*, vol. 73B, no. 2, pp. 91–118, 1969.

[9] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transp. Res. Part B: Methodol.*, vol. 28, no. 4, pp. 269–287, 1994.

[10] C. F. Daganzo, "The cell transmission model, part II: Network traffic," *Transp. Res. Part B: Methodol.*, vol. 29, no. 2, pp. 79–93, 1995.

[11] L. C. Evans, "Partial differential equations," in *Graduate Studies in Mathematics*. Providence, RI, USA: AMS, 1998.

[12] T. L. Friesz and R. Mookherjee, "Solving the dynamic network user equilibrium problem with state-dependent time shifts," *Transp. Res. Part B Methodol.*, vol. 40, no. 3, pp. 207–229, 2006.

[13] B. D. Greenshields, "A study of traffic capacity," *Highway Res. Board Proc.*, vol. 14, pp. 448–477, 1935.

[14] Y. A. Korilis, A. A. Lazar, and A. Orda, "Achieving network optima using stackelberg routing strategies," *IEEE/ACM Trans. Networking*, vol. 5, pp. 161–173, 1997.

[15] Y. A. Korilis, A. A. Lazar, and A. Orda, "Capacity allocation under noncooperative routing," *IEEE Trans. Autom. Control*, vol. 42, no. 3, pp. 309–325, Mar. 1997.

[16] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *Proc. 16th Annu. Conf. Theoretical Aspects of Computer Science*, 1999, pp. 404–413.

[17] J. P. Lebacque, "The Godunov scheme and what it means for first order traffic flow models," in *Proc. Int. Symp. Transportation and Traffic Theory*, 1996, pp. 647–677.

[18] R. J. LeVeque, "Finite difference methods for ordinary and partial differential equations: Steady-state and time-dependent problems," in *Classics in Applied Mathematics*. Philadelphia, PA, USA: SIAM, 2007.

[19] M. J. Lighthill and G. B. Whitham, "On kinematic waves. II. A theory of traffic flow on long crowded roads," in *Proc. R. Soc. London. Series A. Math. Phys. Sci.*, 1955, vol. 229, no. 1178, p. 317.

[20] H. K. Lo and W. Y. Szeto, "A cell-based variational inequality formulation of the dynamic user optimal assignment problem," *Transp. Res. Part B: Methodol.*, vol. 36, no. 5, pp. 421–443, 2002.

[21] A. Ozdaglar and R. Srikant, "Incentives and pricing in communication networks," in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[22] M. Papageorgiou, J. Blosseville, and H. Hadj-Salem, "Modelling and realtime control of traffic flow on the southern part of boulevard peripherique in paris: Part I: Modelling," *Transp. Res. Part A: General*, vol. 24, no. 5, pp. 345–359, 1990.

[23] M. Papageorgiou, J. M. Blosseville, and H. Hadj-Salem, "Macroscopic modelling of traffic flow on the Boulevard Périphérique in Paris," *Transp. Res. Part B: Methodol.*, vol. 23, no. 1, pp. 29–47, 1989.

[24] P. I. Richards, "Shock waves on the highway," *Oper. Res.*, vol. 4, no. 1, pp. 42–51, 1956.

[25] T. Roughgarden, "Stackelberg scheduling strategies," in *Proc. 33rd Annu. ACM Symp. Theory of Computing*, 2001, pp. 104–113, ACM.

[26] T. Roughgarden and E. Tardos, "How bad is selfish routing?," *J. ACM*, vol. 49, no. 2, pp. 236–259, 2002.

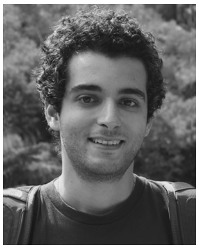
[27] T. Roughgarden and E. Tardos, "Bounding the inefficiency of equilibria in nonatomic congestion games," *Games Econ. Behav.*, vol. 47, no. 2, pp. 389–403, 2004.

[28] D. Schmeidler, "Equilibrium points of nonatomic games," *J. Stat. Phys.*, vol. 7, no. 4, pp. 295–300, 1973.

[29] C. Swamy, "The effectiveness of Stackelberg strategies and tolls for network congestion games," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1133–1142.

[30] Y. Wang, A. Messmer, and M. Papageorgiou, "Freeway network simulation and dynamic traffic assignment with METANET tools," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1776, no. -1, pp. 178–188, 2001.

[31] D. B. Work, S. Blandin, O. P. Tossavainen, B. Piccoli, and A. M. Bayen, "A traffic model for velocity data assimilation," *App. Math. Res. eXpress*, vol. 2010, no. 1, Apr. 1, 2010.



Walid Krichene (S'12) received the M.S. degree in applied mathematics from the Ecole Nationale Superieure des Mines de Paris, Paris, France, in September 2010. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA.

From September 2010 to July 2011, he worked at Criteo labs on designing techniques for predicting Click Through Rates using statistical learning theory. His research focuses on network control and optimization, applied to routing in large-scale networks. He uses game theory and online learning theory to study convergence of population dynamics to Nash equilibria; and applies online learning and optimal control to design routing and tolling schemes which improve the efficiency of the network.

Jack D. Reilly (S'12) received the B.S. degree in civil engineering from the University of California, Los Angeles, CA, USA, in 2009 and the Masters degree in civil systems engineering from the University of California (UC), Berkeley, CA, USA, in 2013. He is currently pursuing the Ph.D. degree from UC Berkeley in civil systems engineering, working on a framework for optimal ramp metering and flow reroute algorithms applied to large-scale freeway systems.

Saurabh Amin (M'09) received the B.Tech. degree in civil engineering from the Indian Institute of Technology, Roorkee, India, the M.S. degreee in transportation engineering from the University of Texas at Austin, Austin, TX, USA, and the Ph.D. degree in systems engineering from the University of California, Berkeley, CA, USA.

He is currently an Assistant Professor in the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. His research focuses on the design and implementation of high-confidence network control algorithms for infrastructure systems. He works on robust diagnostics and control problems that involve using networked systems to facilitate the monitoring and control of large-scale critical infrastructures, including transportation, water, and energy distribution systems. He also studies the effect of security attacks and random faults on the survivability of networked systems, and designs incentive-compatible control mechanisms to reduce network risks.

Alexandre M. Bayen (M'09) received the Engineering Degree in applied mathematics from the Ecole Polytechnique, Palaiseau, France, in July 1998, the M.S. degree in aeronautics and astronautics from Stanford University, Stanford, CA, USA, in June 1999, and the Ph.D. degree in aeronautics and astronautics from Stanford University in December 2003.

He was a Visiting Researcher at the NASA Ames Research Center from 2000 to 2003. Between January 2004 and December 2004, he worked as the Research Director of the Autonomous Navigation Laboratory, Laboratoire de Recherches Balistiques et Aerodynamiques, (Ministere de la Defense), Vernon, France, where he holds the rank of Major. He is an Associate Professor in the Department of Electrical Engineering and Computer Sciences, and the Department of Civil and Environmental Engineering at the University of California, Berkeley, CA, USA. He has authored one book and over 100 articles in peer-reviewed journals and conferences.

Dr. Bayen is the recipient of the Ballhaus Award from Stanford University, in 2004, of the CAREER award from the National Science Foundation, in 2009 and he is a NASA Top 10 Innovators on Water Sustainability, received in 2010. He is the recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE) Award from the White House (2010). His projects "Mobile Century" and "Mobile Millennium" received the 2008 Best of ITS Award for "Best Innovative Practic," at the ITS World Congress and a TRANNY Award from the California Transportation Foundation, in 2009. "Mobile Millennium" has been featured more than 100 times in the media, including television channels and radio stations (CBS, NBC, ABC, CNET, NPR, KGO, the BBC), and in the popular press (*Wall Street Journal*, *Washington Post*, *LA Times*).